ISSN 2007-9737

Enfoque semántico para el descubrimiento de recursos sensible al contexto sobre contenidos académicos estructurados con OAI-PMH

Arianna Becerril García[1], Rafael Lozano Espinosa[2], José Martín Molina Espinosa[2]

[1] Universidad Autónoma del Estado de México, México

[2] Tecnológico de Monterrey, Ciudad de México, México

abecerrilg@uaemex.mx, {ralozano, jose.molina}@itesm.mx

Resumen. Esencial a la noción de Web es la idea de una comunidad abierta: cualquiera puede contribuir sus ideas al todo. Esta apertura, su dimensión y dinamismo imponen retos para el desarrollo de aplicaciones de descubrimiento de recursos para el quehacer educativo o de investigación. Sin embargo, se han dado muchos esfuerzos para organizar y estructurar la masa de datos. Los repositorios académicos han adoptado el Protocolo para Cosecha de Metadatos de la Iniciativa de Archivos Abiertos (OAI-PMH, por sus siglas en inglés) y los metadatos Dublin Core para la exposición de su información. Es así, que resulta relevante el desarrollo de tecnologías que abonen en el descubrimiento de recursos de interés tomando en cuenta las necesidades de información y contexto del usuario. El presente documento describe un enfoque que considera los recursos de información estructurados con OAI-PMH, una representación ontológica y el contexto del usuario como insumos de un marco de trabajo para la construcción de aplicaciones de recuperación de información.

Palabras clave. Web semántica, recursos estructurados, sensibilidad al contexto, ontologías, OAI-PMH, redalyc.

Semantic Approach to Context-Aware Resource Discovery over Scholarly Content Structured with OAI-PMH

Abstract. Essential to the notion of the Web is the idea of an open community: anyone can contribute their ideas to the whole. This openness, the size and dynamism of the community impose challenges on the development of resource discovery applications for educational or research activities. On the other hand, there have been many efforts to organize and structure the mass of data. Scholarly repositories have adopted the Open Archives Initiative – the Protocol for Metadata Harvesting (OAI-PMH) and the Dublin Core metadata for displaying information. Thus, it is relevant to develop technologies in order to improve the discovering of resources taking into account the user information needs and the user context. This paper describes an approach which considers structured information resources with OAI-PMH, an ontological representation, and user context as inputs to a framework for building information retrieval applications.

Keywords. Semantic web, structured resources, context-awareness, ontologies, OAI-PMH, redalyc.

1. Introducción

La llamada sociedad del conocimiento considera la apropiación crítica y selectiva de la información para el desarrollo del ser humano. La World Wide Web (WWW), por su naturaleza de reunión de información vinculada, se ha convertido en la principal fuente de información y desde su introducción en 1990 ha evolucionado para enriquecer la forma en que se organiza y se expone a los usuarios; que va desde un panorama de texto e hipertexto en sitios web planos hasta estándares de estructuración de metadatos e interoperabilidad de la llamada Web 3.0.

Esta masa de información que constituye la Web, en ocasiones se siente como “de una milla de ancho pero con una pulgada de profundidad” ¿Cómo poder construir una experiencia Web más integrada, consistente y profunda? [1]. Es aquí donde se sitúa la semántica, como el proceso de comunicar la información con suficiente

significado. Así es posible construir aplicaciones inteligentes que aporten un mayor conocimiento identificando en mayor profundidad los contenidos.

El ámbito académico no ha estado exento del impacto del crecimiento de la WWW. Encontrar información relevante para el aprendizaje, la enseñanza o la investigación en el volumen de recursos y publicaciones existentes se está convirtiendo en un reto importante para los estudiantes y científicos. Aunado a ello, compartir recursos, metadatos de los recursos y datos a través de la Web es un principio central en el contexto académico y de investigación. La colaboración científica por mucho tiempo ha luchado por reusar y compartir más ampliamente el conocimiento y los datos [2].

La educación, por su parte, ha sufrido importantes cambios propiciados por el desarrollo de las tecnologías que han modificado las formas de acceso y difusión de la información y los modos de comunicación entre los individuos, entre los individuos y las máquinas y entre las propias máquinas [3].

Los portales, plataformas y bases de datos de recursos académicos disponibles en la Web conforman una gran biblioteca dinámica y creciente; con múltiples y diversos puntos de consulta que imponen retos importantes en las tareas de búsqueda y recuperación de información relevante para un estudiante, docente o investigador.

En este sentido, resulta relevante el desarrollo de tecnologías que abonen en el descubrimiento de recursos de interés tomando en cuenta las necesidades de información y características del usuario.

En el presente documento se describen los resultados de la aplicación de un modelo que considera como insumos de un marco de trabajo de recuperación de información: recursos estructurados con el Protocolo para Cosecha de Metadatos de la Iniciativa de Archivos Abiertos (OAI-PMH), una representación ontológica y el contexto del usuario.

En un trabajo previo [4] se formularon las generalidades de un modelo preliminar que no había sido implementado, es decir que se encontraba en la fase de diseño. Posteriormente, con base en él, se desarrollaron los componentes de software, y acorde a los resultados obtenidos y problemas encontrados surge el enfoque mostrado en el presente trabajo. La implementación, resultados alcanzados, validación y retos enfrentados con este nuevo acercamiento son documentados a continuación.

2. Conceptos fundamentales

2.1. OAI-PMH

La variedad de recursos de información en la Web de utilidad para un alumno, académico, profesor o científico es muy amplia, abarca libros, artículos de revistas científicas, informes, actas de congreso, tesis, pre-prints, archivos de datos, entre otros. Todos ellos disponibles a través de portales especializados, repositorios y bases de datos que usan mecanismos de descripción y exposición de sus datos.

Para que este tipo de plataformas tengan la posibilidad de intercambiar información tienen que contar con reglas de comunicación y estándares de estructuración de datos. El protocolo de interoperabilidad OAI-PMH es uno de los más utilizados para este fin.

Según el Registro de Repositorios de Acceso Abierto [5] (ROAR, por sus siglas en inglés) existen poco más de 4,000 repositorios en el mundo que implementan el protocolo OAI-PMH. Para tener una mejor idea de la cantidad de archivos de contenido intelectual disponibles en Acceso Abierto se puede acceder al proveedor de servicio OAIster [6] que cosecha menos de la mitad de los repositorios registrados en ROAR y cuenta con más de 30 millones de registros disponibles a través de OAI-PMH.

El OAI-PMH surge con la Iniciativa de Archivos Abiertos, liberada en 1999, de la necesidad de convertir los archivos en interoperables y construir servicios de recuperación de información de diversos repositorios. Su naturaleza radica en la definición de una interfaz a través de la cual un repositorio expone públicamente en la web los metadatos de los objetos digitales que almacena.

El protocolo Z39.50 [7] ya existía como un estándar que permitía la búsqueda federada a varios servidores de manera paralela. Sin

embargo, se había presentado mucha dificultad para crear servicios de búsqueda federada de alta calidad a través de un gran número de servidores autónomos, por razones como: diferentes interpretaciones de las consultas, problemas de escalabilidad, dependencia de la disponibilidad de los servidores al momento de la consulta y rendimiento sujeto a la velocidad de respuesta del servidor más lento [8].

Es así, que el OAI-PMH se consolida como un estándar de la comunidad de archivos abiertos como resultado de las ventajas que ofrece en comparación con el Z39.50.

Este protocolo es un mecanismo de baja barrera para la interoperabilidad de repositorios [9]. Define una interfaz que un servidor conectado a la red puede emplear para hacer disponible a aplicaciones externas los metadatos que describen objetos almacenados en ese servidor [8].

En el protocolo se especifican dos tipos de participantes, los proveedores de datos y los proveedores de servicio; los primeros, encargados de exponer públicamente los metadatos de su contenido y los segundos, a cargo de cosechar metadatos de los proveedores de datos para ofrecer interfaces de integración y búsqueda para el usuario final.

Hace uso de peticiones y respuestas HTTP para comunicarse entre un cosechador y un repositorio usando métodos GET o POST. Para la conformación de estas peticiones existe una URL base única que especifica el servidor y el puerto; y opcionalmente la ruta.

Dichas peticiones mejor conocidas como verbos son seis y se concatenan a la URL base. Los verbos se describen enseguida [10]:

- GetRecord: regresa los metadatos de un registro individual.
- Identify: devuelve la información acerca del repositorio.
- ListRecords: es usado para cosechar los registros de un repositorio; argumentos adicionales permiten la cosecha selectiva basada en conjuntos o temporalidad.
- ListIdentifiers: es una forma abreviada de ListRecords que trae únicamente las cabeceras de los registros.
- ListMetadataFormats: regresa los formatos de metadatos disponibles en el repositorio.
- ListSets: recupera la estructura de conjuntos de un repositorio.

Las respuestas son serializadas en XML con los metadatos de Dublin Core (descritos posteriormente). El proceso de envío – recepción de peticiones y respuestas se controla a través del denominado proceso de cosecha de metadatos. Siendo un cosechador el programa que envía peticiones a un proveedor de datos y recibe como respuesta archivos XML con metadatos Dublin Core.

2.2. Dublin Core

La Iniciativa de Metadatos Dublin Core (DC) auspicia el desarrollo de estándares de interoperabilidad a diferentes niveles, entre los que se encuentra un conjunto de metadatos para descripciones simples y genéricas popularizado por ser parte de las especificaciones del protocolo OAI-PMH.

El llamado Dublin Core no calificado es el que originalmente se utiliza para describir recursos con OAI-PMH y contempla los siguientes 15 metadatos [10]:

- dc:title,
- dc:creator,
- dc:subject,
- dc:description,
- dc:publisher,
- dc:contributor,
- dc:date,
- dc:type,
- dc:format,
- dc:identifier,
- dc:source,
- dc:language,
- dc:relation,
- dc:coverage,
- dc:rights.

2.3. Sensibilidad al contexto

La característica de sensibilidad al contexto del usuario en servicios de recuperación de información se refiere a la capacidad de percibir

información de su ambiente para otorgar resultados personalizados. Con ello es posible inferir situaciones no explicitadas y así manifestar un comportamiento inteligente.

Específicamente para el ámbito educativo muchas plataformas no toman en cuenta las diferentes necesidades del alumno y proveen la misma recuperación de información a todos los usuarios. Entonces, resulta pertinente como se menciona en [11] el uso de un modelo de estudiante para permitir la personalización efectiva de ambientes de aprendizaje.

2.4. Ontologías

Una ontología es una especificación explícita de una conceptualización [12].

Se considera conceptualización al modelado abstracto de algún fenómeno del mundo identificando sus conceptos relevantes. Es explícita dado que el tipo de conceptos usados y sus restricciones son definidos explícitamente. Es formal por el hecho de que debe ser legible para máquinas y es compartida ya que captura el conocimiento consensuado, es decir no es privativo de un individuo, sino aceptado por un grupo [12, 13].

Las ontologías habilitan a una computadora para entender la información por sí misma [14].

Para la ciencia y la educación un insumo fundamental es la bibliografía así es que la representación ontológica de referencias bibliográficas ha sido objeto de diversos desarrollos. Entre ellos se encuentra FaBiO una ontología para registrar y publicar registros bibliográficos en la Web Semántica [15]; CiTO, una ontología para citaciones bibliográficas [16]; y BIRO, la ontología de referencias bibliográficas [17].

Estas y cinco ontologías más: PRO, la ontología de roles de publicación; PSO, la ontología de estado de la publicación; PWO, la ontología del flujo de trabajo de publicación; C4O, la ontología de caracterización del contexto y conteo de citación; y DoCo, la ontología de componentes del documento conforman el conjunto SPAR (Semantic Publishing and Referencing Ontologies) compuesto de módulos para crear metadatos RDF comprehensivos para todos los aspectos de la publicación y referencia semántica [18].

Por su parte, la Ontología Bibliográfica BIBO provee los conceptos y propiedades principales para describir citaciones y referencias bibliográficas [19].

En cuanto a Dublin Core se refiere, la DCMI también ha elaborado una ontología para describir el conjunto de términos para identificar objetos digitales [20].

Por otro lado se encuentra también FOAF, un proyecto devoto de vincular personas e información usando la Web [21]. El espacio de nombres FOAF http://xmlns.com/foaf/0.1/ es usado para representar datos acerca de personas tales como, nombres, fechas de nacimiento y especialmente a la gente con la que se relacionan. Es particularmente útil para representar datos de redes sociales.

2.5. Marco de trabajo Jena

Jena, proyecto de código abierto iniciado por los Laboratorios HP en el 2000, es un marco de trabajo Java para la construcción de aplicaciones de la Web Semántica, provee bibliotecas Java para el desarrollo de código que maneje RDF, RDFS, RDFa, OWL y SPARQL alineado con las recomendaciones de la W3C [22].

Incluye un motor de inferencia basado en reglas para desempeñar razonamiento basado en ontologías OWL y RDFS y una variedad de estrategias de almacenamiento para tripletas RDF en memoria o en disco.

La API ontológica de Jena provee una interfaz de programación consistente para el desarrollo de aplicaciones ontológicas, independiente del lenguaje. Fue seleccionado para la implementación de este enfoque dada su solidez y robustez para el desarrollo de aplicaciones de la Web Semántica.

3. Descripción de la propuesta

El alcance del enfoque propuesto se restringe al uso de recursos estructurados bajo el protocolo OAI-PMH, es decir, los datos estructurados en otros formatos o bajo otros protocolos no son considerados en este modelo.

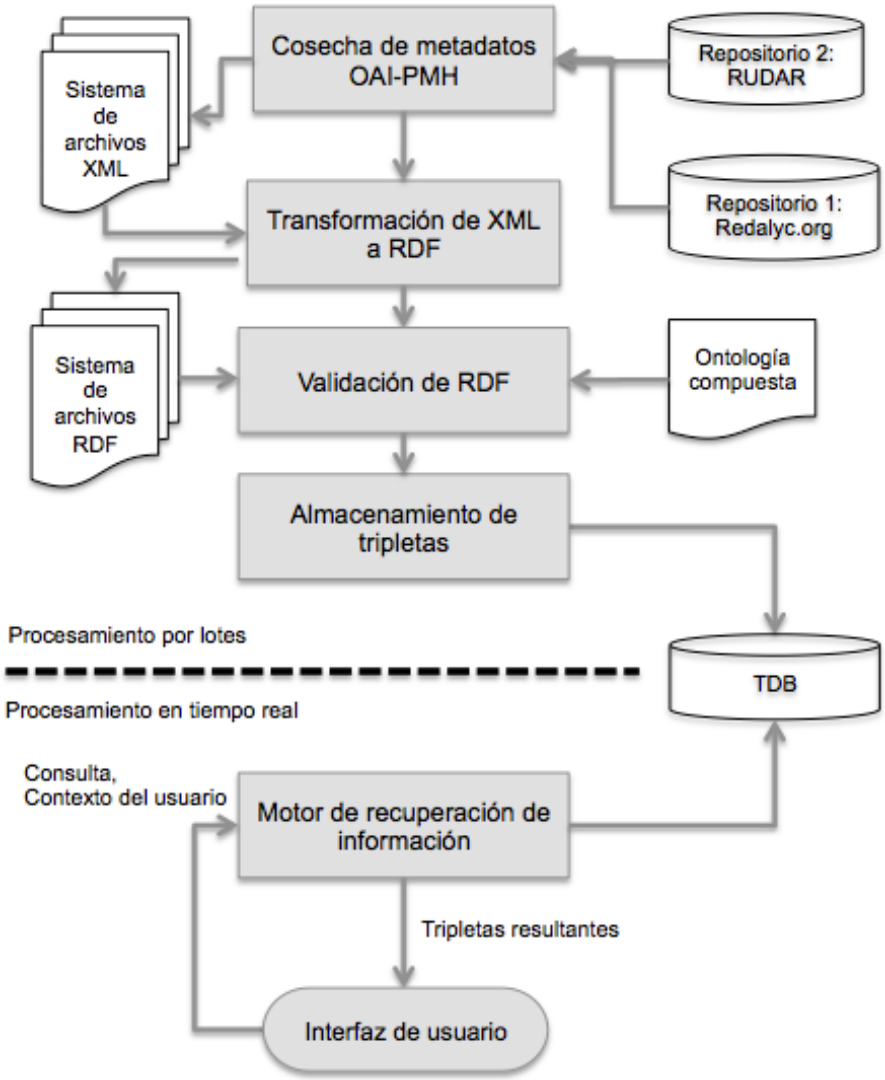

Fig. 1. Diagrama de flujo del proceso

La parte central constituye un motor de recuperación de información sobre ese conjunto de datos.

3.1. Metodología

En la figura 1 se puede observar el diagrama de flujo del proceso llevado para la implementación de la propuesta.

Se aprecia que la información requerida de entrada desde una interfaz de usuario corresponde a la consulta así como la información contextual del usuario.

Dicha información ingresa al motor de recuperación donde circula por un razonador basado en reglas que procesa la consulta sobre una base de datos de tripletas haciendo uso del API de ontologías de Jena.

El proceso de consulta es en tiempo real sobre datos colectados previamente en un procesamiento por lotes, dando como resultado

un conjunto de inferencias correspondientes a recursos de información.

Los pasos seguidos y componentes de software desarrollados se detallan en los apartados siguientes.

3.2. Cosecha de metadatos OAI-PMH

Se desarrolló una aplicación Java que realiza peticiones HTTP a los proveedores de datos haciendo uso del verbo ListRecords del OAI-PMH. Con ello, se obtienen archivos en formato XML con los datos descritos en el conjunto simple de metadatos de Dublin Core. Dentro de estos archivos XML se encuentran registros (<record>). El número de registros por archivo depende de la configuración de cada repositorio cosechado y estos pueden variar desde 1 hasta n.

3.3. Transformación a RDF

Muchas instituciones dan acceso a sus repositorios de metadatos a través de OAI-PMH pero no hacen que sus recursos sean accesibles a través de URIs desreferenciables, cosa que provoca restricciones de significado y hace que quede restringido el acceso a los metadatos [23]. Por esta razón, es necesario hacer uso de un convertidor a RDF con la intención de transformar los metadatos contenidos en un repositorio OAI-PMH a RDF.

Se han desarrollado diversos proyectos que permiten explotar datos estructurados y dotarlos de características propias de aplicaciones semánticas, como lo es OAI2LOD Server un desarrollo para exponer metadatos OAI-PMH como Linked Data [24]. Dentro de este tipo de proyectos wrapper también se puede encontrar D2R [25].

Para este trabajo se probó un desarrollo enmarcado en los llamados RDFizers, software de conversión a RDF llamado OAI2RDF como un componente en la arquitectura [26]. La herramienta realiza esta tarea con una transformación lógica que se hace a través de hojas XSLT que se invocan una vez que los datos han sido entregados.

Sin embargo, se optó por desarrollar una aplicación de software para realizar esta tarea dado que esto permitía una mejor integración del proceso de cosecha con el resto de la implementación.

Esta aplicación se encarga de leer los metadatos de los archivos XML obtenidos de la cosecha OAI-PMH utilizando el API SaxBuilder que permite recuperar los registros convirtiéndolos en objetos. Enseguida se muestra un ejemplo de registro contenido en un XML de entrada.

```
<record>
 <header>
  <identifier>oai:redalyc.org:10504408</identifier>
  <datestamp>2007-08-15</datestamp>
  <setSpec>1405-1435</setSpec>
 </header>
 <metadata>
<oai_dc:dc
xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai
_dc/"
   xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/
2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
<dc:title>La emergencia del sentido a partir…</dc:title>
<dc:creator>Edgar Morin </dc:creator>
<dc:subject>Sociología</dc:subject>
<dc:subject>El no-sentido</dc:subject>
<dc:subject>sentido del mundo</dc:subject>
<dc:subject>complejidad</dc:subject>
<dc:description>
  En este texto se reflexiona sobre el problema del …
</dc:description>
<dc:publisher>Universidad Autónoma del …
</dc:publisher>
<dc:date>2007</dc:date>
<dc:type>Artículo científico</dc:type>
<dc:format>application/pdf</dc:format>
<dc:identifier>
http://www.redalyc.org/articulo.oa?id=10504408
</dc:identifier>
<dc:relation>
http://www.redalyc.org/revista.oa?id=105
</dc:relation>
<dc:rights>Convergencia</dc:rights>
<dc:source>Convergencia. Revista de Ciencias
Sociales (México) Num.44 Vol.14</dc:source>
 <dc:language>es</dc:language>
</oai_dc:dc>
</metadata>
 </record>
```

Posteriormente, se creó un modelo en Jena a partir de listas de los objetos recuperados y se definió un esquema compatible con el estándar Dublin Core para generar un archivo RDF/XML.

Los datos del autor son enriquecidos para agregar relaciones de coautoría y más datos acerca del autor. Estos metadatos pueden extenderse en tanto los conjuntos de datos usados como fuente contengan información sobre los autores y son expresados con FOAF, de la siguiente forma:

```
<dc:creator rdf:type=[http://xmlns.com/foaf/0.1/Person]
foaf:name=[Nombre completo]
foaf:givenName=[Nombre] foaf:surname=[Apellidos]
foaf:knows=[Persona conocida o coautor]
foaf:topic_interest=[Tema de interés]
dc:description=[Ocupación, grado, etc. ]
onto:birthDate=[fecha de nacimiento] />
```

Un ejemplo de salida es el que se muestra a continuación.

```
<rdf:Description rdf:about="oai:redalyc.org:10504408">
    <dc:language>es</dc:language>
    <dc:rights>Convergencia</dc:rights>
<dc:subject>sentido del mundo</dc:subject>
<dcterms:modified rdf:resource="2007-08-15"/>
<dc:type>Artículo científico</dc:type>
<dc:source>Convergencia. Revista de Ciencias Sociales (México) Num.44 Vol.14</dc:source>
 <dc:title>La emergencia del sentido a partir del no-sentido</dc:title>
 <dcterms:isPartOf rdf:resource="set:1405-1435"/>
 <dc:format>application_pdf</dc:format>
 <dc:publisher>Universidad Autónoma del Estado de México</dc:publisher>
 <dc:subject>El no-sentido</dc:subject>
 <dc:identifier
rdf:resource="http://www.redalyc.org/articulo.oa?id=10504408"/>
 <dc:creator rdf:type="http://xmlns.com/foaf/0.1/Person"
foaf:name="Edgar Morin">Edgar Morin</dc:creator>
<dc:relation>http://www.redalyc.org/revista.oa?id=105</dc:relation>
<dc:description>En este texto se reflexiona sobre…</dc:description>
<dc:subject>Sociología</dc:subject>
<dc:date>2007</dc:date>
<dc:subject>complejidad</dc:subject>
 </rdf:Description>
```

3.4. Modelo Ontológico y validación

El componente ontológico integra dos ontologías una de ellas Dublin Core, dado que los recursos que alimentan la base de conocimiento están estructurados con metadatos DC al estar bajo el estándar OAI-PMH.

El estándar Dublin Core original incluía el nivel simple y el calificado, el primero compuesto de 15 elementos y el cual se utiliza para descripción de recursos con OAI-PMH bajo el espacio de nombres http://purl.org/dc/elements/1.1/. Sin embargo, para este trabajo se utiliza el espacio de nombres http://purl.org/dc/terms, la razón radica en que a partir del año 2012 la DCMI (Dublin Core Metadata Initiative) incorpora los dos niveles en este espacio de nombres.

La segunda ontología es FOAF. La intención del uso de FOAF en este proyecto es explorar su adaptación para tratar la información sobre los autores de los artículos, libros y otros recursos académicos. Aunado a esto, se pueden incluso expresar sus relaciones sociales basadas en coautoría para poder determinar recursos relacionados.

De este modo es posible modelar las propiedades de un recurso con su autor como en la figura 2; donde por ejemplo, un investigador es un autor de una publicación (dc:creator) pero a la vez es una persona (foaf:person) con propiedades individuales. Asimismo están representadas por un lado la relación de autoría entre una publicación y un investigador; y la de coautoría de un investigador con otro u otros con los que escribe en conjunto una publicación (figura 3).

Por otro lado, es importante recordar que las ontologías son desarrolladas a diferentes niveles de abstracción por personas distintas y para diversos propósitos. El conocimiento representado por las ontologías se dispersa debido a la existencia de muchas ontologías representando los mismos conceptos, es así, que se vuelve difícil analizar, estudiar y usar el conocimiento propagado a través de múltiples ontologías si se estudian individualmente [27].

La técnica para combinar en una sola ontología el conocimiento representado en varias ontologías es la unión o merge. Con este

ISSN 2007-9737

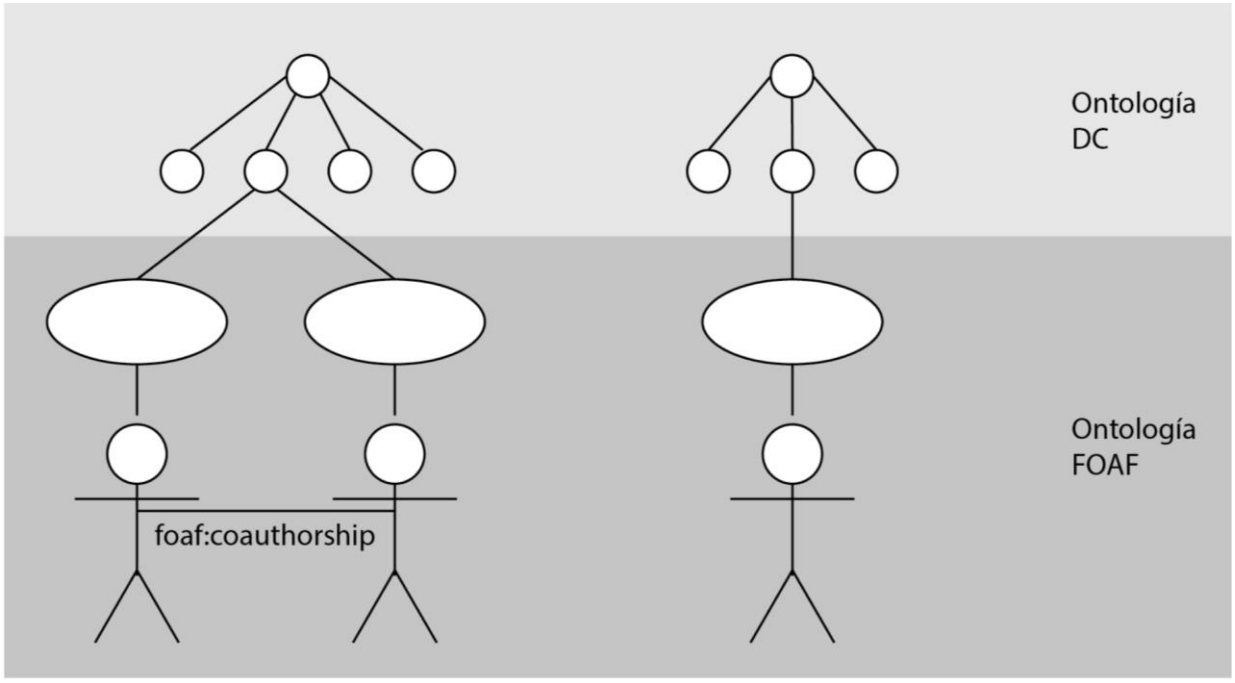

Fig. 2. Modelo ontológico con Dublin Core y Friend of a Friend

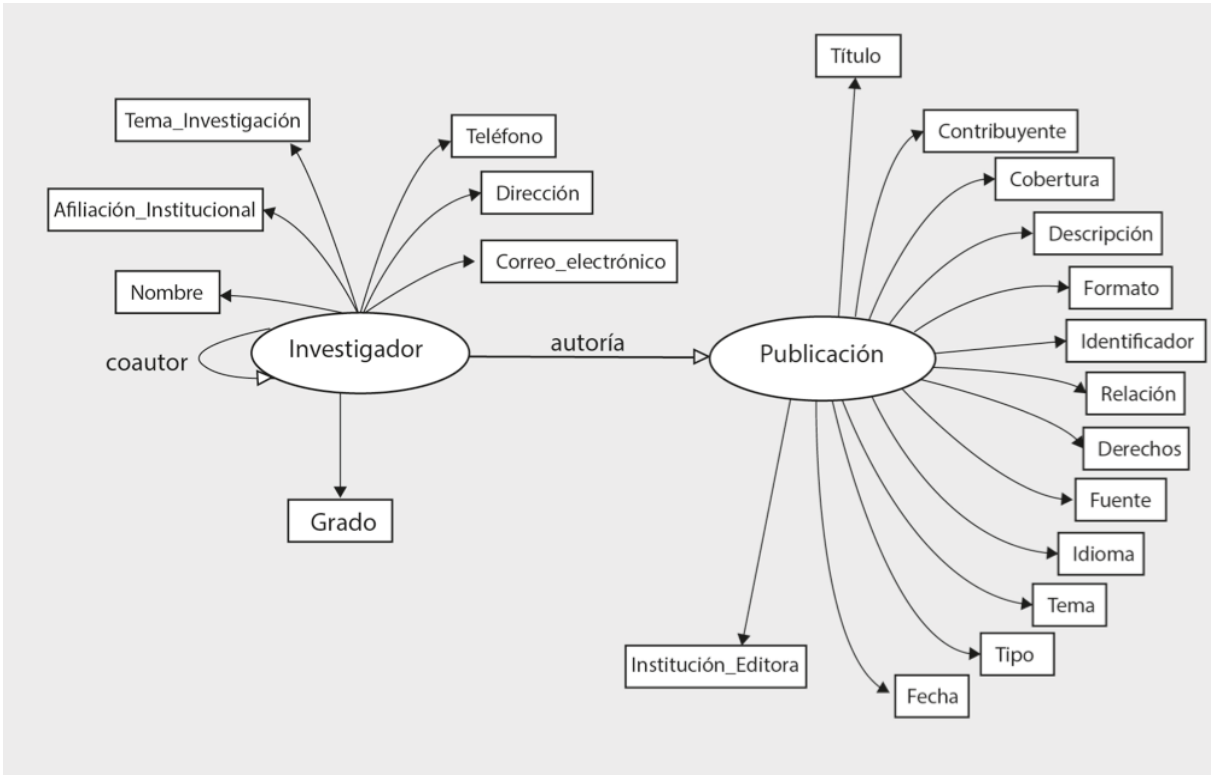

Fig. 3. Relaciones de autoría y coautoría

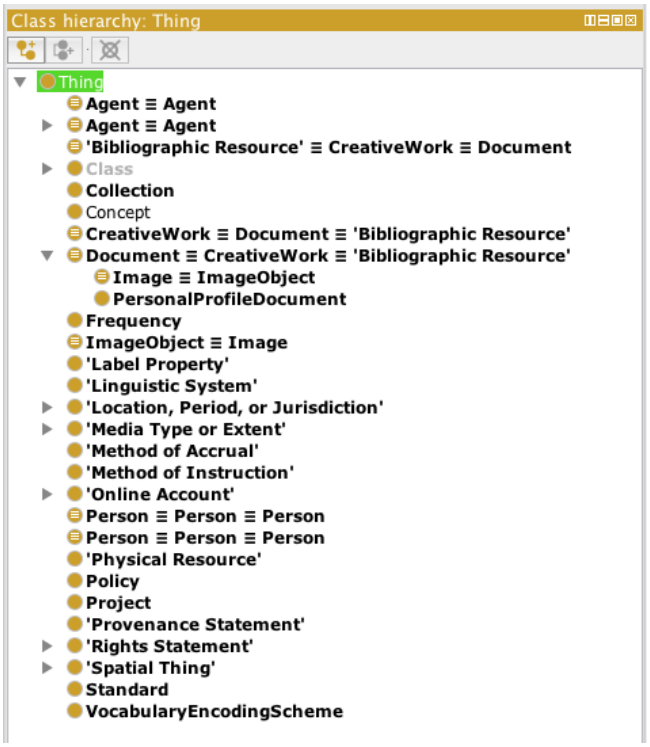

Fig. 4. Jerarquía de clases de la ontología resultante de la fusión como se muestra en Protégé

procedimiento se busca obtener una definición para expresar el conocimiento obtenido de las fuentes de información descritas, que favorezca su validación y análisis.

Con la ayuda de la herramienta Protégé [28] y su función de Refactor > Merge Ontologies se realizó la fusión de las dos ontologías de entrada.

Esta integración automática no resuelve las inconsistencias generadas después del proceso. Por ello, la ontología de salida fue sujeta a un refinamiento posterior siguiendo los pasos del algoritmo de merging propuesto en [27].

La jerarquía de clases de la ontología resultante se muestra en la figura 4 con un total de 39 propiedades.

Como parte del refinamiento se identificaron equivalencias, por ejemplo con la clase BibliographicResource de DC de la cual se hizo explícita su equivalencia a Document de FOAF que a su vez ya era equivalente a CreativeWork (Figura 5).

De la misma forma, también se establecieron otras equivalencias, entre ellas el caso de Creator (http://purl.org/dc/terms/creator) de Dublin Core con Maker (http://xmlns.com/foaf/0.1/maker) de FOAF.

3.5. Almacenamiento

El modelo contempla TDB del marco de trabajo de Jena. TDB es un componente para el almacenamiento y consulta RDF, soporta el rango completo de APIs de Jena y puede ser usado como un almacén de alto rendimiento para tripletas RDF.

3.6. Motor de recuperación de información

El motor de recuperación de información es una aplicación desarrollada en Jena haciendo uso de sus API de ontologías, razonamiento y almacenamiento. La arquitectura de sus componentes se muestra en la figura 6.

ISSN 2007-9737

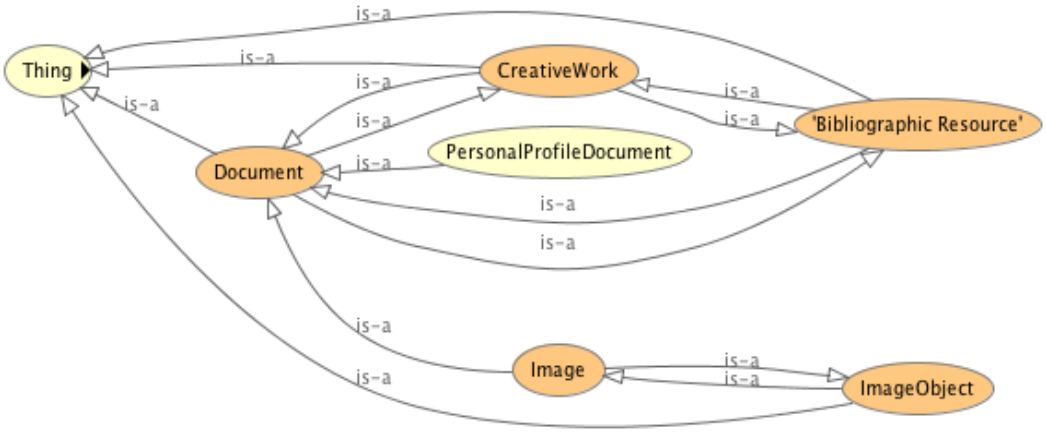

Fig. 5. Vista OWLViz de la clase Document y su equivalencia con Bibliographic Resource

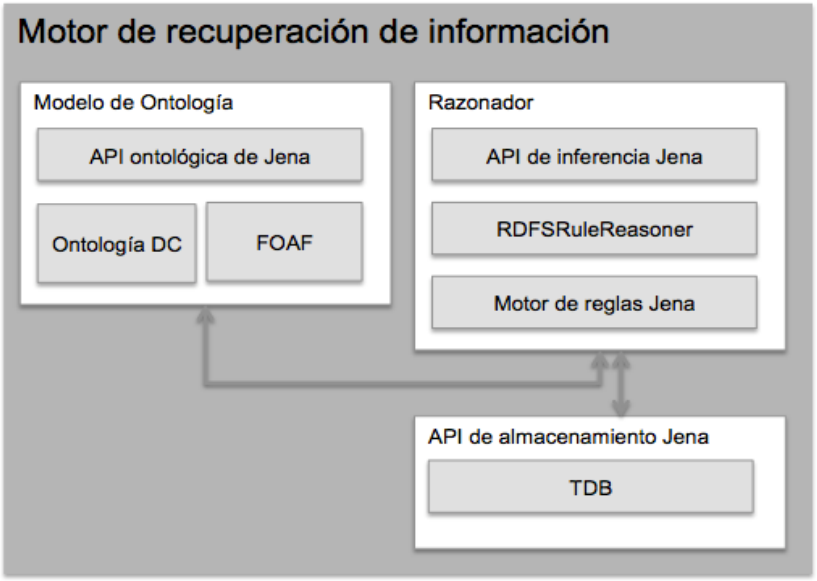

Fig. 6. Arquitectura

El subsistema de inferencia está diseñado para derivar un conjunto de enunciados a partir de la base de datos de hechos proveniente de los procesos de cosecha y transformación de recursos OAI-PMH, la información de contexto del usuario e información ontológica.

El motor de inferencia de Jena es usado para derivar enunciados RDF adicionales de la base TDB. Con fines experimentales para este desarrollo se usó el razonador OWL incluido, una implementación basada en reglas de OWL/Lite.

Cabe destacar que la inferencia se realiza sobre las coincidencias en los valores de las tripletas. Un caso, por ejemplo, es la temática que aborda un recurso de información que en Dublin Core es modelado con dc:subject, es así que los registros que tienen el mismo valor en ese atributo tienen una asociación. El modelo no contempla, en este momento, encontrar dichas coincidencias en diferentes idiomas.

La interfaz de usuario debe proveer información de contexto expresado en un perfil

del sujeto que lanza la consulta y que constituye un conjunto de características para representar circunstancias personales, profesionales, sociales o de espacio-tiempo como fecha y lugar en que se emite la consulta.

Dado que la información tratada en este trabajo es de corte académico – científico es pertinente aplicar el modelo para estudiantes.

Para este fin, se retomó el modelo de representación de estudiantes basado en ontologías para sistemas de tutoría inteligente de aprendizaje a distancia propuesto en [29]. Dicho modelo permite representar a un estudiante con cuatro clases:

- Student, representa cualquier estudiante.
- StudentCourseInformation, comprende información relevante al proceso educativo como módulos del programa que cursa, escuela, tareas, exámenes, entre otros.
- StudentCurrentActivity, se refiere al detalle de la actividad académica del año en curso.
- StudentPersonalInformation, es la información estática y permanente del estudiante.

3.7. Procesamiento de la información

3.7.1. Procesamiento en tiempo real

El proceso de consulta basada en un conjunto de parámetros que ingresan a un motor de inferencia para devolver un resultado, se realiza en tiempo real y comienza una vez que se recibe la información de entrada y concluye con el envío de resultados de salida. Es así, como este motor de recuperación de información utiliza los datos recolectados por el programa de cosecha que han sido transformados, enriquecidos y almacenados de manera centralizada con anterioridad.

3.7.2. Procesamiento por lotes

La naturaleza del funcionamiento del protocolo OAI-PMH obliga a recolectar los datos de los repositorios en procesos por lotes y en segundo plano por las siguientes razones:

Tabla 1. Resultados del proceso de cosecha OAI-PMH

	Redalyc.org	**RUDAR**
Archivos XML	17,328	121
Registros DC	346,557	12,011

a) El tiempo de cosecha de metadatos depende de los tiempos de respuesta de los repositorios individuales. Asimismo, el tiempo total de recolección de metadatos está sujeto al modo de operación del cosechador ya sea secuencial o paralelo para todos los repositorios que se deseen cosechar. Así, para el primer modo será la suma total de los tiempos de respuesta de todos los repositorios y para el modo paralelo dependerá del número de repositorios que estén siendo cosechados de manera simultánea y el tiempo de respuesta más lento de cada hilo de procesamiento.

b) La disponibilidad de los repositorios en el momento de los procesos de recolección de metadatos puede impedir que un repositorio sea localizado. Así, en un procesamiento por lotes, es posible realizar intentos de reconexión con el repositorio sin impactar el tiempo final.

4. Resultados

Con fines experimentales y de validación del modelo propuesto fueron elegidos dos repositorios que implementan OAI-PMH. Cabe resaltar que la única condición para que un repositorio sea compatible con este modelo es que implementen dicho protocolo. Los repositorios usados fueron: Redalyc.org, el portal de la Red de Revistas Científicas de América Latina, el Caribe, España y Portugal [30] y el repositorio institucional de la Universidad Roskilde llamado RUDAR (Roskilde University Digital Archive) de Dinamarca [31].

A continuación se describen los resultados obtenidos de seguir la metodología:

1. El proceso de cosecha de metadatos OAI-PMH recolectó de Redalyc.org 17,328 archivos XML conteniendo cada uno de ellos un máximo de 20 registros, que hicieron un total de 346,557 artículos científicos. La

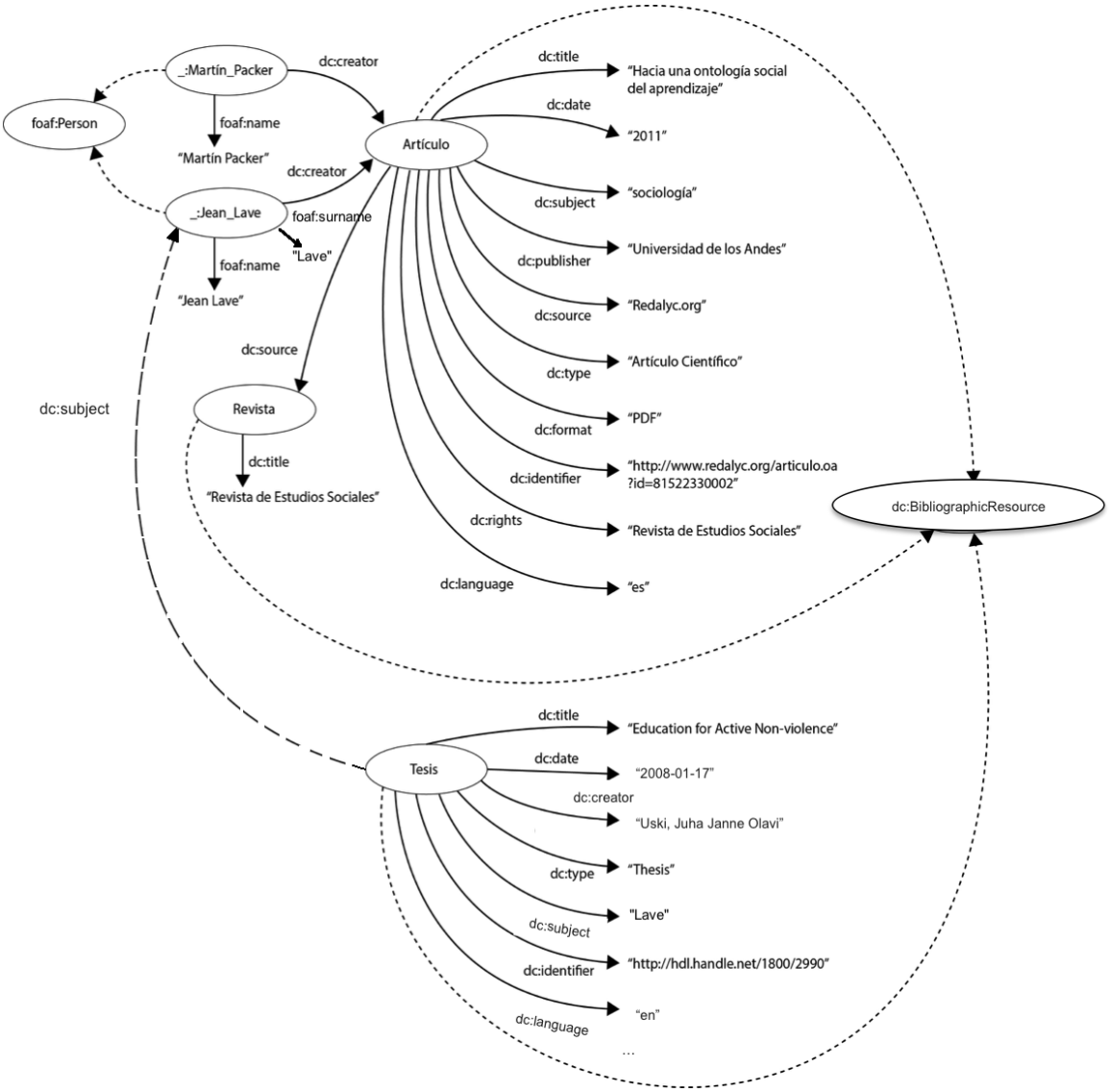

Fig. 7. Grafo de relación entre las instancias de ejemplo

cosecha con RUDAR recuperó 121 archivos con un máximo de 100 registros cada uno, haciendo un total de 12,011 recursos entre artículos, tesis y más documentos (Tabla 1).

2. Los archivos XML fueron transformados a RDF/XML resultando en total 17,449 archivos.
3. Esos archivos resultantes fueron sujetos a la validación utilizando la ontología combinada.
4. Posteriormente, la información pasó al almacén en forma de tripletas, las cuales ascendieron a un total de 7,147,338.
5. Para ejemplificar la consulta se usó como entrada el perfil del siguiente alumno:

```
<rdf:Description rdf:about="itesm:A01210238">
 <student:name
rdf:type="http://xmlns.com/foaf/0.1/Person">
Rafael R. Gómez</student:name>
```

```
<student:courseModule>Sociología</student:courseModule>
 <student:language>es</student:languaje>
<student:demographicData>br</student:demographicData>
</rdf:Description>
```

El objetivo es recuperar recursos académicos relevantes para el curso de “Sociología”.

Un recurso recuperado, entre otros, fue el correspondiente a un artículo científico titulado “Hacia una ontología social del aprendizaje”, escrito por “Jean Lave” y “Martin Packer” publicado en español en la revista “Revista de Estudios Sociales” editada en la “Universidad de Los Andes“ de Colombia en “2011” y cuya temática es “Sociología”. El resultado fue recuperado dada la coincidencia exacta con la temática (dc:subject) del artículo que es “Sociología”.

Otro recurso resultante corresponde a una tesis titulada “Education for Active Non-violence” de la autoría de “Uski, Juha Janne Olavi” publicado el “2008-01-17” y que trata de diversos temas, es decir cuenta con varios dc:subject, uno de ellos tiene como valor el texto: “Lave” que coincide con el atributo foaf:surname de la autora del artículo encontrado previamente. Es decir, esta tesis tiene como temática cuestiones relacionadas con la autora. Y aunque no contiene explicitamente la temática de “Sociología” fue recuperada dada la relación derivada.

Las relaciones entre los recursos se muestran en el grafo de la figura 7, por motivos de visualización no se incluyeron todos los datos de cada recurso en el grafo.

En resumen, los hechos obtenidos de los metadatos OAI-PMH cosechados son, entre otros:

- El artículo “Hacia una ontología social del aprendizaje” tiene una temática de “Sociología” así que es relevante para el estudiante.
- El artículo “Hacia una ontología social del aprendizaje” fue escrito en coautoría por “Jean Lave” y “Martin Packer”.
- “Lave” es un tema (dc:subject) de la tesis de “Uski, Juha Janne Olavi”.

Hecho derivado:

- La tesis de “Uski, Juha Janne Olavi” es relevante para el estudiante, ya que trata sobre una temática relacionada con una autora de un artículo que es relevante para el estudiante.

Así, es posible descubrir recursos relevantes para un usuario tomando en consideración la información de contexto a través de un perfil de usuario así como las relaciones obtenidas entre los recursos de información.

5. Trabajos relacionados

Respecto a proyectos cuyo objetivo gira en torno a la recuperación y descubrimiento de información se encuentra Sindice, el índice de la Web Semántica [32], es un proyecto patrocinado por el Digital Enterprise Research Institute (DERI) que provee un motor de búsqueda semántico de recursos marcados con RDF, microformatos, microdatos, RDFa, entre otros, indexados de la Web para exponerlos a través de una API para desarrolladores.

Freebase, una colección abierta de datos estructurados y plataforma para accederlos y manipularlos a través de una API que ha notificado su adición al proyecto Wikidata [33].

Por otro lado, hay trabajos en el campo del descubrimiento de recursos y recomendación como [34] o para ambientes de aprendizaje personales como [35] y sobre la información de Linked Data.

Todos estos proyectos si bien se enmarcan en la línea de motores semánticos no están especializados para recursos estructurados con OAI-PMH de ahí la diferencia con el modelo aquí presentado.

6. Conclusiones y trabajo futuro

El descubrimiento de recursos de información es un problema derivado del acelerado crecimiento de la Web que dificulta cada vez más la localización de información para un usuario; en lo correspondiente al ámbito educativo y de investigación es un reto importante para los estudiantes y científicos.

El enfoque presentado propone una metodología y un motor de recuperación de información basado en ontologías tomando en consideración la información de contexto a través de un perfil de usuario así como las relaciones obtenidas entre los recursos de información. Tal acercamiento permite descubrir recursos de interés personalizados a través de inferencia.

Este modelo podría extenderse para aprovechar las fuentes de información de Linked Data como insumo además de contenidos OAI-PMH, sin embargo, habría que plantear mecanismos de filtrado y selección de información académica o científica.

Adicionalmente, la propuesta puede ser enriquecida con el uso de vocabularios controlados como en [36] y/o el uso de ontologías multilingües como la desarrollada en [37] para recuperar información en diversos idiomas. Y es posible probar otros motores de inferencia como Pellet, Racer o FaCT para un completo razonamiento OWL DL.

Referencias

1. **Allemang, D. & Hendler, J. (2011).** *Semantic Web for the Working Ontologist.* 2 ed., USA, Morgan Kaufmann.
2. **Kessler, C., d'Aquin, M., & Dietze, S. (2013).** Linked Data for Science and Education. *Semantic Web Journal,* Vol. 4, No. 1, pp. 1–2.
3. **Cantillo Valero, C., Roura Redondo, M., & Sánchez Palacín, A. (2012).** Tendencias actuales en el uso de dispositivos móviles en educación. *La Educación Digital,* No. 147.
4. **Becerril García, A., Lozano Espinosa, R., & Molina Espinosa, J. (2014).** Modelo para consultas semánticas sensibles al contexto sobre recursos educativos estructurados con OAI-PMH. *Encuentro Nacional de Ciencias de la Computación (ENC),* Oaxaca, México, Nova Universitas.
5. **University of Southampton (2014).** Recuperado el 7 de julio de 2014, de Registry of Open Access Repositories, http://roar.eprints.org/
6. **OCLC. (2014).** Recuperado el 30 de 07 de 2014, de The OAIster database, http://www.oclc.org/oaister.en.html
7. **Clifford, A. L. (2001).** Metadata harvesting and the Open Archives Initiative. *ARL: A bimonthly report of Research Library Issues and Actions from ARL, CNI, and SPARC,* Association of Research Libraries, Washington, DC.
8. **ANSI/NISO (2003).** *Information Retrieval (Z39.50): Application Service Definition and Protocol Specification.* Estados Unidos, Niso Press.
9. **Lagoze, C. & Van de Sompel, H. (2001).** The Open Archives Initiative: Building a low-barrier interoperability framework. *ACM,* JCDL'01, Roanoke, VA.
10. **Lagoze, C. & Van de Sompel, H. (2015).** *The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0*.
11. **Martínez-Villaseñor, M., González-Mendoza, M., & Danvila Del Valle, I. (2014).** Enrichment of Learner Profile with Ubiquitous User Model Interoperability. *Computación y Sistemas,* Vol. 18, No. 2, pp. 359–374. DOI: 10.13053/CyS-18-2-2014-037.
12. **Gruber, T. (1995).** Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human and Computer Sciences,* Vol. 43, No. 5-6, pp. 907–928.
13. **Studer, R., Benjamins, R., & Fensel, D. (1998).** Knowledge Engineering: Principles and methods. *Data and Knowledge Engineering,* Vol. 25, No. 1-2, pp. 161–197.
14. **Abburu, S. (2012).** A Survey on Ontology Reasoners and Comparison. *International Journal of Computer Applications,* Vol. 57, No. 17, pp. 33–39.
15. **Peroni, S. & Shotton, D. (2012).** FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web,* Vol. 17, pp. 33–43. DOI: 10.1016/j.websem.2012.08.001.
16. **Shotton, D. (2010).** CiTO, the Citation Typing Ontology. *Journal of Biomedical Semantics,* Vol. 1, pp. 1–18. DOI: 10.1186/2041-1480-1-S1-S6.
17. **SPAR (2009).** Recuperado el 12 de Diciembre de 2014, de Semantic Publishing and Referencing Ontologies: http://sempublishing.sourceforge.net/
18. **Shotton, D. & Peroni, S. (2013).** Recuperado el 14 de septiembre de 2014, de BiRO, the Bibliographic Reference Ontology: http://www.essepuntato.it/lode/http://purl.org/spar/biro
19. **Giasson, F. & D'Arcus, B. (2009).** Bibliographic Ontology Specification. Recuperado el 2014, de http://bibliontology.com/
20. **DCMI. (2012).** Recuperado el 2 de abril de 2014, de DCMI Metadata Terms: http://dublincore.org/documents/2012/06/14/dcmi-terms/

21. **Brickley, D. & Miller, L. (2014).** FOAF Vocabulary Specification 0.99. Obtenido de http://xmlns.com/foaf/spec/
22. **Apache Software Foundation. (2010).** Recuperado el 2 de Febrero de 2014, de Apache Jena: http://jena.apache.org/about_jena/about.html
23. **OMediaDis. (2009).** Recuperado el 2 de agosto de 2014, de Informe Modelos de Metadatos para Contenidos Multimedia: http://omediadis.udl.cat/html/deliverables/215-Modelos_Metadatos_Contenidos_Multimedia/
24. **Haslhofer, B. & Schandl, B. (2008).** The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data. *International Workshop on Linked Data on the Web* (LDOW2008), co-located with WWW 2008. Beijing, China.
25. **Bizer, C., & Cyganiak, R. (2006).** *D2R Server – Publishing Relational Databases on the Semantic Web*. Obtenido de http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/Bizer-Cyganiak-D2R-Server-ISWC2006.pdf
26. **Mazzocchi, S. (2006).** Recuperado el 5 de enero de 2014, de OAI2RDF: http://simile.mit.edu/repository/RDFizers/oai2rdf/
27. **Stanford University (2015).** Recuperado el 15 de noviembre de 2014, de Protégé: http://protege.stanford.edu
28. **Ameen, A., Rahman Khan, K., & Rani, B. (2014).** Semi-Automatic Merging of Ontologies using Protégé. *International Journal of Computer Applications,* Vol. 85, No. 12, pp. 35–42.
29. **Panagiotopoulos, I., Kalou, A., Pierrakeas, C., & Kameas, A. (2012).** An Ontology-Based Model for Student Representation in Intelligent Tutoring Systems for Distance Learning. **I.M. Lazaros Iliadis (ed.)**, *Artificial Intelligence Applications and Innovations,* Halkidiki, Grecia, Springer.
30. **Becerril-García, A., Aguado-López, E., Rogel-Salazar, R., Garduño-Oropeza, G., & Zúñiga-Roca, M. (2012).** De un modelo centrado en la revista a un modelo centrado en entidades: la publicación y producción científica en la nueva plataforma Redalyc.org. **I.U. Oviedo (ed.),** *Aula Abierta,* Vol. 40, No. 2, pp. 53–64.
31. **Roskilde University. (2015).** Obtenido de Roskilde University Digital Archive: http://diggy.ruc.dk:8080
32. **Tummarello, G., Delbru, R., & Oren, E. (2007).** Sindice.com: Weaving the Open Linked Data. *The Semantic Web,* Springer, Berlin, Heidelberg, pp. 552–565.
33. **Vrandecic, D. & Krötzsch, M. (2014).** Wikidata: a free collaborative knowledgebase. *Communications of the ACM,* Vol. 57, No. 10, pp. 78–85.
34. **Foulonneau, M. & Grouès, V. (2012).** Common vs. Expert knowledge: making the Semantic Web an educational model. *Linked Learning.*
35. **Jeremic, Z., Jovanovic, J., & Gasevic, D. (2011).** Personal Learning Environments on the Social Semantic Web. *Semantic Web Journal,* pp. 1–30.
36. **Bakaev, M. & Avdeenko, T. (2013).** Indexing and Comparison of Multi-Dimensional Entities in a Recommender System based on Ontological Approach. *Computación y Sistemas,* Vol. 17, No. 1, pp. 5–13.
37. **Abusalah, M., Tait, J., & Oakes, M. (2009).** Cross Language Information Retrieval using Multilingual Ontology as Translation and Query Expansion Base. *Polibits,* Vol. 40, pp. 13–16.

Arianna Becerril García forma parte del equipo fundador de la Red de Revistas Científicas de América Latina y el Caribe, España y Portugal (Redalyc.org) donde se desempeña como Directora de Tecnología e Innovación. Es candidata a Doctora en Ciencias de la Computación por el Tecnológico de Monterrey en México, Maestra en Ciencias de la Computación por la misma institución e Ingeniera en Computación por la Universidad Autónoma del Estado de México (UAEM). Es profesora-investigadora de tiempo completo de la Universidad Autónoma del Estado de México. Trata temas como ingeniería del conocimiento, recuperación de información en sistemas inteligentes, web semántica, repositorios digitales, bibliometría y acceso abierto a la ciencia. Es también miembro del Consejo Asesor Internacional del Directory of Open Access Journals DOAJ y cofundadora de la Red Mexicana de Repositorios Institucionales Remeri. Cuenta con varios artículos en revistas científicas, 3 libros publicados y ha participado en más de 40 congresos nacionales e internacionales.

Rafael Lozano Espinosa es Ingeniero en Electrónica y Comunicaciones por la Universidad de las Américas. Cursó una Maestría en Computación en Sistemas de Información por la Universidad de las Américas. Cuenta además, con un Doctorado en Informática por la Universidad de Grenoble, Francia. Es Profesor

Titular en el área de Tecnologías de Información y Computación del Tecnológico de Monterrey.

José Martín Molina Espinosa es Doctor en Informática y Telecomunicaciones por el Institut National Polytechnique de Toulouse. Es profesor del Departamento de Computación en el Tecnológico de Monterrey Campus Ciudad de México. Director de la Red Nacional de Centros para Toma de Decisiones del Tecnológico de Monterrey. Ha sido líder del grupo de investigación y desarrollo en tecnologías móviles en el Tecnológico de Monterrey Campus Ciudad de México. Instructor de la Especialización de Desarrollo de Apps en iOS en la plataforma de Coursera. Fue Director de posgrados en Tecnologías de Información del ITESM, Campus Ciudad de México.

Artículo recibido el 24/09/2015; aceptado 16/01/2016.
Autor de correspondencia es Arianna Becerril García.