TECNOLÓGICO
DE MONTERREY

Biblioteca
Campus Ciudad de México

Instituto Tecnológico y de Estudios Superiores de Monterrey

Escuela de Ingeniería y Ciencias

Campus Ciudad de México

OntoOAI: Modelo Semántico para el Descubrimiento Selectivo de Conocimiento sobre Contenidos Estructurados con OAI-PMH

Tesis que presenta: Arianna Becerril García

Para obtener el grado de: Doctora en Ciencias Computacionales

Asesores de tesis: Dr. Rafael Lozano Espinosa

Dr. José Martín Molina Espinosa

México, D.F., Mayo de 2016

Declaración de autoría

Yo, Arianna Becerril García, declaro que esta disertación titulada, 'OntoOAI: Modelo Semántico para el Descubrimiento Selectivo de Conocimiento sobre Contenidos Estructurados con OAI-PMH', y el trabajo que se presenta en ella es de mi autoría.

Adicionalmente, confirmo que:

- Realicé este trabajo en su totalidad durante mi candidatura al grado de Doctor en esta universidad.
- He dado crédito a cualquier parte de esta disertación que haya sido previamente sometida para obtener un grado académico o cualquier otro tipo de titulación en ésta o cualquier otra universidad.
- He dado crédito a cualquier trabajo previamente publicado que se haya consultado en esta disertación.
- He citado el trabajo consultado de otros autores, y la fuente de donde los obtuve.
- He dado crédito a todas las fuentes de ayuda utilizadas.
- He dado crédito a las contribuciones de mis coautores, cuando los resultados corresponden a un trabajo colaborativo.
- Esta disertación es enteramente mía, con excepción de las citas indicadas.

Arianna Becerril García
México, Distrito Federal
6 de mayo de 2016

@2016 por Arianna Becerril García
Todos los derechos reservados

OntoOAI: Modelo Semántico para el Descubrimiento Selectivo de Conocimiento sobre Contenidos Estructurados con OAI-PMH

por
Arianna Becerril García

Resumen

El principio fundamental en la Web es la comunidad abierta: cualquiera puede contribuir sus ideas al todo. Esta apertura a la colaboración, su flexibilidad y tolerancia son factores que le han permitido convertirse en un recurso universal. Sin embargo, la masa creciente, dinámica y muchas veces poco estructurada de información en la Web impone retos importantes, uno de ellos: el descubrimiento de conocimiento.

El descubrimiento de conocimiento, uno de los productos finales más interesantes de la computación, resulta particularmente relevante en la tarea de un investigador, profesor o alumno, que usan la información científica para su quehacer profesional.

Entre los esfuerzos que se han dado para organizar y estructurar la masa de datos, se ubica el Protocolo para Cosecha de Metadatos de la Iniciativa de Archivos Abiertos (OAI-PMH), estándar para los repositorios de información científica y académica.

El protocolo OAI-PMH surgió con fines de interoperabilidad entre plataformas. Los recursos estructurados bajo esta especificación además de ser intercambiados entre sistemas, son usados en motores de búsqueda que habilitan la consulta a los usuarios. Sin embargo, estos datos tienen un uso potencial poco explotado en aplicaciones semánticas de recuperación de información, que permitiría construir aplicaciones *inteligentes* que aporten un mayor conocimiento identificando en mayor profundidad los contenidos.

El presente trabajo doctoral presenta OntoOAI, un modelo semántico para el descubrimiento selectivo de conocimiento sobre recursos estructurados con el protocolo OAI-PMH que permite verificar la factibilidad y dar cuenta de las limitaciones en la aplicación de tecnologías de la Web Semántica sobre dichos conjuntos de datos para descubrir conocimiento de forma selectiva, entendido como el proceso de encontrar recursos que no fueron solicitados explícitamente por un usuario pero que resultan potencialmente útiles para él con base en su contexto.

OntoOAI es probado con la combinación de tres fuentes de información: Redalyc.org, el portal de la Red de Revistas de América Latina y el Caribe, España y Portugal, el repositorio institucional de la Universidad Roskilde llamado RUDAR y DBPedia. Su aplicación permite comprobar que es factible el uso de tecnologías semánticas al obtener resultados en el descubrimiento selectivo de conocimiento y da muestra de las limitaciones en el uso de los datos OAI-PMH para este fin.

Lista de Figuras

Figura 1. Mapa de repositorios. Fuente: Repository66.org (Lewis, 2014). 7

Figura 2. Mandatos de acceso abierto. Fuente: ROARMAP (2015). 15

Figura 3. Alcance de la propuesta 17

Figura 4. Etapas del proceso KD. Fuente: Fayyad, Piatetsky-Shapiro, & Smyth (1996). 26

Figura 5. Clasificación de las medidas de interés. Fuente: Silberschatz y Tuzhilin (1995). 28

Figura 6. Modelo de pila de capas de la Web Semántica de Tim Berners Lee. Fuente: W3C (2007). 31

Figura 7. Línea del tiempo de los estándares de la W3C para la Web Semántica y XML, traducida de Bikakis, Tsinaraki, Gioldasis, Stavrakantonakis y Christodovlakis (2013). 35

Figura 8. Problema de la integración de ontologías según Guarino (1998) 43

Figura 9. Grafo RDF 44

Figura 10. Visibilidad y Protocolo OAI-PMH 53

Figura 11. Proceso de cosecha OAI-PMH 55

Figura 12. Esquema XML para metadatos Dublin Core (OpenArchives, 2002). 59

Figura 13. Metodología de OntoOAI 72

Figura 14. Extracto del *dataset PersonData* de DBpedia. 77

Figura 15. Ejemplo XML DC 78

Figura 16. Funcionamiento de XML_DC2RDF 81

Figura 17. Jerarquía de clases de la ontología DC 85

Figura 18. Jerarquía de propiedades de objeto de la ontología DC 86

Figura 19. Jerarquía de propiedades de datos de la ontología DC 86

Figura 20. Jerarquía de propiedades de anotación de la ontología DC 87
Figura 21. Jerarquía de clases de la ontología FOAF 89
Figura 22. Jerarquía de propiedades de anotación de FOAF 90
Figura 23. Jerarquía de propiedades de datos de FOAF 90
Figura 24. Jerarquía de propiedades de objeto de FOAF 91
Figura 25. Modelo ontológico con Dublin Core y Friend of a Friend 92
Figura 26. Relaciones de autoría y coautoría en la fusión de ontologías 93
Figura 27. Algoritmo de fusión de ontologías de OntoOAI basado en (Ameen, Rahman Khan, & Rani, 2014) 94
Figura 28. Declaración de equivalencia *Document* y *BibliopgraphicResource* 95
Figura 29. Jerarquía de clases de la ontología resultante de la fusión. 96
Figura 30. Equivalencia de la propiedad *Creator* con *Maker.* 98
Figura 31. Proceso de almacenamiento de tripletas 101
Figura 32. Jerarquía de clases de la ontología OntoOAIEstudiante 103
Figura 33. Grafo de clases de la ontología *OntoOAIEstudiante.owl* 104
Figura 34. Arquitectura del motor de inferencia de OntoOAI 106
Figura 35. Verificación de URL Base de Redalyc 115
Figura 36. Verificación de URL Base de RUDAR 115
Figura 37. Registro Dublin Core obtenido de Redalyc.org 117
Figura 38. Registro Dublin Core obtenido del repositorio RUDAR. 118
Figura 39. Ejemplo de salida RDF de Redalyc 119
Figura 40. Ejemplo de salida RDF de RUDAR 119
Figura 41. Ejemplo de coincidencias de autor 121

Figura 42. Ejemplo de RDF enriquecido 122

Figura 43. Composición de la base OntoOAI en TDB. 125

Figura 44. Perfil de usuario itesm:A01210238 127

Figura 45. Perfil de usuario univ:DL9510078 128

Figura 46. Resultado de búsqueda directa (ejemplo 1) 129

Figura 47. Descubrimiento selectivo de conocimiento del ejemplo 1 132

Figura 48. Visualización jerárquica de los resultados del ejemplo 1 134

Figura 49. Visualización tipo grafo de los resultados del ejemplo 1 135

Figura 50. Resultado de búsqueda directa (ejemplo 2) 136

Figura 51. Descubrimiento selectivo de conocimiento del ejemplo 2 141

Lista de Tablas

Tabla 1. Clases de RDF Schema (W3C, 2004)....46

Tabla 2. Propiedades de RDF Schema (W3C, 2004)....47

Tabla 3. Conjunto de elementos de metadatos Dublin Core, versión 1.1 (DCMI, Dublin Core Metadata Element Set, Version 1.1). 58

Tabla 4. Comparación de algunos desarrollos de software en el campo de descubrimiento y OAI-PMH. 69

Tabla 5. Comparación de *RDFizers*....80

Tabla 6. Métricas de la ontología DC (DCMI, 2016)....84

Tabla 7. Métricas de la ontología FOAF 89

Tabla 8. Análisis de equivalencia de clases entre DC y FOAF 95

Tabla 9. Análisis de equivalencias de propiedades de objeto en DC y FOAF 97

Tabla 10. Análisis de equivalencia de propiedades de datos DC y FOAF 99

Tabla 11. Métricas de la ontología resultante *DCFOAF_merged.owl*....99

Tabla 12. Resultados de la cosecha de metadatos....116

Tabla 13. Resultados del proceso de conversión a RDF....120

Tabla 14. Resultados del proceso de enriquecimiento de autores 124

Tabla 15. Resultados de la carga masiva al *triple store.* 126

Tabla 16. Extracto de resultados del ejemplo 1 de descubrimiento selectivo de conocimiento....133

Tabla 17. Extracto de resultados del ejemplo 2 de descubrimiento selectivo de conocimiento....142

Contenido

Resumen vi

Lista de Figuras vii

Lista de Tablas x

Abreviaciones y Siglas xiv

Capítulo 1 Introducción **1**

1.1 Motivación 1

1.2 Antecedentes 3

1.3 Preguntas de investigación 12

1.4 Hipótesis 12

1.5 Planteamiento del problema 13

1.6 Objetivo 13

1.7 Justificación 14

1.8 Alcance 16

1.9 Visión general de los resultados 18

1.10 Trabajos relacionados 18

1.11 Estructura del documento 21

Capítulo 2 Estado del arte **23**

2.1 Descubrimiento de conocimiento 23

2.2 Web semántica y ontologías 30

2.2.1 Tecnologías de la Web Semántica 30

2.2.2 Ontologías 36

2.2.3 Fusión de ontologías 42

2.2.4 RDF 44

2.2.5 Razonamiento 47

2.2.6 Ontologías para recursos académico – científicos 50

2.3 Protocolo para Cosecha de Metadatos de la Iniciativa de Archivos Abiertos 51
2.3.1 Dublin Core 56
2.4 Sensibilidad al contexto y descubrimiento selectivo 60
2.5 Desarrollos de sofware para descubrimiento de conocimiento y OAI-PMH 66

Capítulo 3 Modelo OntoOAI para el descubrimiento selectivo de conocimiento 71
3.1 Descripción general del modelo 71
3.2 Fuentes de información 73
3.3 Cosechador de metadatos OAI-PMH 77
3.4 Convertidor XML-DC a RDF 79
3.5 Integración XML DC con FOAF 81
3.6 Ingeniería ontológica 83
3.6.1 Ontología Dublin Core 84
3.6.2 Ontología Friend of a Friend 88
3.6.3 Fusión de ontologías 92
3.7 Almacenamiento 100
3.8 Sensibilidad al contexto 101
3.9 Motor de inferencia y descubrimiento de conocimiento 105
3.10 Procesamiento por lotes y procesamiento en tiempo real 109
3.11 Tecnologías y herramientas 110
3.11.1 Protégé 111
3.11.2 Jena 111
3.11.3 TDB 112

Capítulo 4 Caso de aplicación de OntoOAI 113
4.1 Fuentes de información seleccionadas 113
4.2 Cosecha de metadatos de Redalyc y RUDAR 116
4.3 Conversión de RDF de Redalyc y RUDAR 118
4.4 Enriquecimiento de autores y verificación del modelo de datos 120
4.5 Almacenamiento en la base OntoOAI 124
4.6 Perfil de usuario 126
4.7 Resultados de inferencia 128

4.7.1 Ejemplo 1 de descubrimiento selectivo de conocimiento 128
4.7.2 Ejemplo 2 de descubrimiento selectivo de conocimiento 135
4.8 Problemas encontrados 142

Capítulo 5 Conclusiones y trabajo futuro 144

Bibliografía 148

Apéndices 161

Artículos publicados 174

Currículum Vitae 176

Abreviaciones y Siglas

OAI-PMH	Open Archives Initiative - Protocol for Metadata Harvesting
CAS	Current Awareness Services
DSI	Diseminación Selectiva de Información
OA	Open Access
AA	Acceso Abierto

Capítulo 1 Introducción

El presente capítulo aborda los orígenes que motivaron el desarrollo del presente trabajo doctoral, antecedentes en el ámbito temático; así como los componentes del trabajo de investigación, preguntas, hipótesis, planteamiento del problema, objetivos, alcance, justificación, resultados y trabajos relacionados.

1.1 Motivación

La llamada sociedad del conocimiento considera la apropiación crítica y selectiva de la información para el desarrollo del ser humano. La World Wide Web, 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.

Las nuevas tecnologías están revolucionando la forma en que los seres humanos pueden aprender sobre el mundo y sobre sí mismos. Estas tecnologías son no sólo un medio de hacer frente a grandes datos, también son una clave para el descubrimiento de conocimiento en la era digital; y su poder se basa en la creciente disponibilidad de datos en sí. Factores como el aumento de potencia de cálculo, el crecimiento de la web, y el compromiso de gobiernos para abrir el acceso a la financiación pública de investigación están sirviendo para aumentar la disponibilidad de hechos, datos e ideas (The Hague Declaration, 2014).

Para un científico resulta no sólo indispensable sino impensable realizar su trabajo sin contar con información y mantenerse actualizado en su campo y fuera de él. Ésto ha motivado el desarrollo de formas de recuperación de información, alertas, búsquedas y diversos mecanismos en apoyo a las tareas de actualización de información para el ámbito académico o profesional.

Desde tiempos remotos, los encargados de bibliotecas, realizaron servicios de notificación de información nueva a usuarios; y a esta actividad le nombraron servicios de alertas de actualización (Kemp, 1979); misma que brindaba ventajas de accesibilidad a la información. Posteriormente, esta actividad cambió para entregar información individual y específica, y se le conoció como diseminación selectiva de información o alertas informativas (Dawra, 2004).

Actualmente, 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 y acceso a la información.

Otra característica a considerar al navegar por la red, tiene que ver con las restricciones para acceder a la información; ésta puede encontrarse de libre acceso o requerir un pago o suscripción para acceder a ella. El *Open Access* (OA) o *Acceso Abierto* (AA) es un movimiento que pugna por el acceso libre a resultados de investigación financiada con fondos públicos.

El contenido de AA se ha convertido cada vez más importante en las últimas décadas, en tanto más organizaciones estipulan que la investigación debe ser accesible

como un requerimiento para financiarla, más datos se han vuelto accesibles (Browning, Guédon, & Kaplan, 2013).

Este nuevo panorama complejo y de información masiva, presente en la cotidianidad de un alumno, un académico o un investigador, impone distintos retos en las tareas de búsqueda y recuperación de información útil para mantenerse actualizado y cubrir sus expectativas de información.

En este sentido, resulta relevante el desarrollo de tecnologías que abonen en el hallazgo de recursos de interés tomando en cuenta las necesidades de información y características del usuario, es decir, el descubrimiento de conocimiento útil para las tareas profesionales, académicas o científicas acorde a las necesidades individuales.

1.2 Antecedentes

La World Wide Web, una sociedad gigante de documentos, ha evolucionado a gran velocidad desde que Tim Berners-Lee introdujo este concepto al mundo. Él mismo señalaba desde 1994:

> "Para una computadora, la Web es un mundo aburrido y plano, carente de significado. Esto es una lástima, ya que de hecho, los documentos en la Web describen objetos reales y conceptos imaginarios, y otorgan relaciones particulares entre ellos. Por ejemplo, un documento puede describir a una persona. Agregar semántica a la Web implica dos cosas: permitir a los documentos tener información en formas legibles por máquina, y permitir la creación de vínculos con valores de relaciones. Sólo cuando tenemos este nivel adicional de la semántica, seremos

capaces de usar el poder del cómputo para ayudar a explotar la información en un grado mayor que nuestra propia lectura." (Berners Lee, 1994, p. 1)

Encontrar información relevante para el aprendizaje, la enseñanza o la investigación, dentro del 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 científico y en el educativo. La colaboración científica por mucho tiempo ha luchado por reusar y compartir más ampliamente el conocimiento y los datos (Keßler, d'Aquin, & Dietze, 2013).

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? (Allemang & Hendler, 2011, p. 2). 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.

Por otro lado, 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, comerciales y no comerciales.

Respecto a las soluciones no comerciales, desde el 2001 se viene consolidando un movimiento que promueve el acceso libre y sin restricciones al contenido científico, sobretodo cuanto éste haya sido financiado con recursos públicos. Este movimiento se llama *Acceso Abierto* y fue formalizado a través de tres declaraciones: Budapest (Chan, et al., 2002), Berlín (Max Planck Society, 2003) y Bethesda (Brown, et al., 2003).

La primera declaración establece el Acceso Abierto como promotor de la disponibilidad de un determinado contenido de forma gratuita y pública en la red, permitiendo la lectura, la descarga, copia, distribución, impresión, búsqueda o enlace a los textos completos, sin barreras económicas, legales o técnicas (Chan, et al., 2002).

En principio, cualquier tipo de contenido digital puede ser OA, ya que se puede poner online sin barreras económicas o de permisos. Por otra parte, cualquier tipo de contenido puede ser digital: textos, datos, imágenes, audio, video, multimedia, y códigos ejecutables. Se puede tener acceso abierto a música, películas, noticias, novelas y software, –y en distintos grados ya se tiene–. Pero el término "acceso abierto" fue acuñado por investigadores que trataban de eliminar las barreras de acceso a la investigación (Suber, 2015).

El inicio del Directorio de Revistas de Acceso Abierto (DOAJ) en 2003, desarrollado por la Universidad de Lund en Suecia, marcó el comienzo de la formalización y organización del Acceso Abierto para el caso de las revistas científicas. Actualmente, este directorio enlista más de 11 mil revistas de 136 países (Lund University, 2003).

Los repositorios, portales y revistas que se adhieren a este movimiento adoptan, como buena práctica, un protocolo de interoperabilidad para intercambiar información y así tener reglas de comunicación y estándares de estructuración de datos. El protocolo de interoperabilidad OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting) de la iniciativa de archivos abiertos es uno de los más utilizados para este fin.

Según el Registro de Repositorios de Acceso Abierto (Registry of Open Access Repositories), ROAR (por sus siglas en inglés) existen 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 (OCLC, 2002) que cosecha menos de la mitad de los repositorios existentes y cuenta con más de 30 millones de registros disponibles a través de OAI-PMH.

Diversas soluciones de software libre y abiertas para construir repositorios y gestionar publicaciones científicas incluyen la implementación del protocolo OAI-PMH, como DSpace (DuraSpace, 2002) y Open Journal Systems (PKP, 2001), lo que convierte a este protocolo en un estándar de facto para la publicación en Acceso Abierto.

Al 2015 se han identificado más de 32,000 instalaciones de OJS, de las cuales 8,286 contienen al menos 10 artículos publicados que hacen un total de 2.8 millones de artículos disponibles vía OAI-PMH (PKP, 2015).

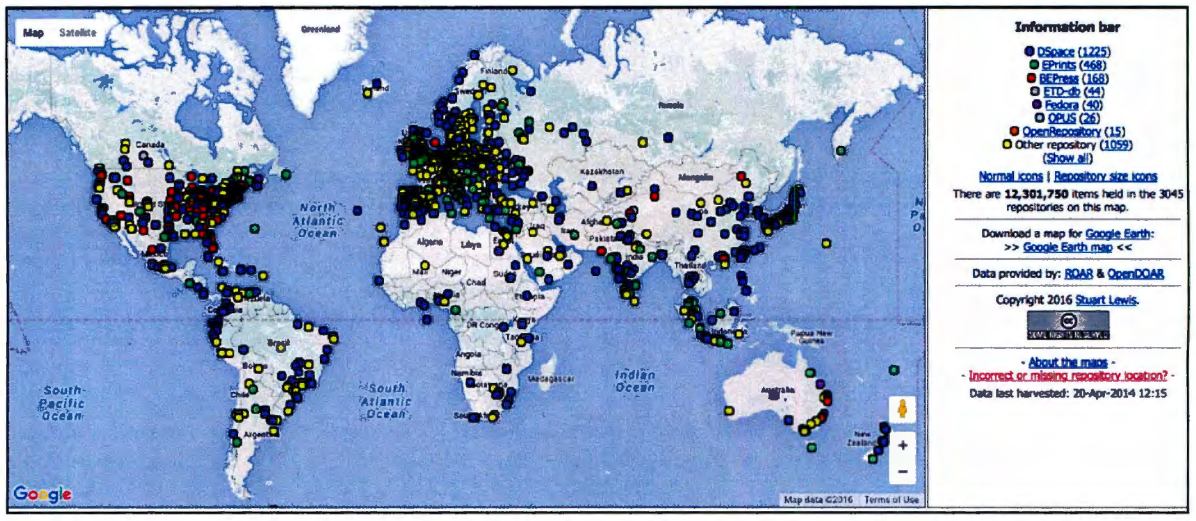

Figura 1. Mapa de repositorios. Fuente: Repository66.org (Lewis, 2014).

El mapa de repositorios Repository66.org (figura 1), con datos de ROAR y OpenDOAR, geo-localiza los repositorios que implementan el protocolo OAI-PMH, muestra 3,045 repositorios con 12,301,750 elementos cosechados para abril 2014 (Lewis, 2014).

De acuerdo con el Registro de Repositorios de Acceso Abierto ROAR, existen diversos tipos de repositorios, según lo que contienen pueden ser de: investigación institucional, multi-institucional y cross-institucional, revistas y publicaciones electrónicas, tesis, bases de datos, resúmenes, datos de investigación, *linked data*, objetos de enseñanza y aprendizaje, demostraciones, observatorios web y otros (Eprints, 2016).

También están disponibles los repositorios nacionales, que agrupan la producción científica de un país, y las redes regionales como LaReferencia, red federada de repositorios de América Latina que contiene más de un millón de elementos y agrupa 8 nodos nacionales (LaReferencia, 2012) y OpenAIRE (OpenAIRE, 2009), con más de 14

millones de publicaciones de 6,500 repositorios y revistas de acceso abierto de la región Europea.

La estructuración de datos, con la que se exponen estos recursos de información en la Web, aporta otro nivel semántico en el entendimiento de la naturaleza de ellos. Así, aplicaciones de software pueden usarlos en beneficio y mejora de los motores de recuperación de información.

El descubrimiento de la investigación se facilita a través de la navegación de contenido y los datos estructurados se prestan a una variedad de aplicaciones que pueden usar esos datos (Albert, Holmes, Börner, & Conlon, 2012).

Las bibliotecas, por su labor esencial, han tenido un interés estratégico en las herramientas y tecnologías que faciliten el descubrimiento y acceso a recursos para las comunidades que sirven.

En este ámbito, los sistemas de descubrimiento de recursos representan la siguiente generación de los catálogos OPAC (*Online Public Access Catalog*) dentro de los sistemas integrados de bibliotecas (ILS, por sus siglas en inglés, *Integrated Library Systems*). Son comúnmente conocidos como "servicios de descubrimiento de escala web", que se basan en un índice central poblado por metadatos, textos completos u otras representaciones de elementos en una colección de bibliotecas (Breeding, 2015, p. 2).

En este sentido, Schonfeld define "descubrimiento" como el proceso y la infraestructura requerida por un usuario para encontrar un elemento apropiado (Schonfeld R. , 2014).

Estos sistemas enfrentan grandes retos, el primero y de acuerdo con Breeding (2015) el ecosistema de descubrimiento actual está atado a muchos factores comerciales, por lo cual no ha sido creado un índice de descubrimiento de acceso abierto. Esta situación conlleva a la pregunta de si alternativas *Open Access* podrían conducir a costos más bajos o a incrementar la innovación.

Por otro lado, estos sistemas hacen uso del término "descubrimiento" refiriéndose a la capacidad de recuperar información de diversas fuentes dentro de un Sistema Integrado de Bibliotecas, que puede incluir catálogos OPAC, repositorios institucionales, bases de datos contratadas, publicaciones a las que se suscribe la biblioteca, adquisiciones, entre otras. Sin embargo, esa capacidad de recuperación se centra en búsquedas generadas por el usuario, y no en el "descubrimiento" basado en la recuperación semántica o basada en inferencia, que otorgue al usuario una experiencia más "inteligente" o enriquecida de obtención de información.

Cabe entonces enfatizar la diferencia entre el concepto llamado "descubrimiento de recursos" empleado en las bibliotecas y el "descubrimiento de conocimiento", objeto del presente trabajo. Usado el primero, para describir el acceso centralizado a recursos de información disponibles en una biblioteca a través de un servicio de búsqueda basado en un índice. Mientras que el descubrimiento de conocimiento, se refiere a localizar patrones e inferir conocimiento de un conjunto de datos de entrada.

El descubrimiento de conocimiento (*Knowledge Discovery o KD*) es el producto final más deseable de la computación. Encontrar nuevos fenómenos o mejorar el

conocimiento acerca de ellos está en segundo lugar en valor sólo después de preservar el mundo y el ambiente (Wiederhold, 1996).

Frawley, Piatetsky-Shapiro y Matheus (1992) definen *descubrimiento de conocimiento* como la extracción no trivial de información implícita, previamente desconocida y potencialmente útil, a partir de datos.

El reto de extraer conocimiento a partir de datos demanda investigación en estadística, bases de datos, reconocimiento de patrones, aprendizaje de máquinas, visualización de datos, optimización y cómputo de alto rendimiento, para entregar inteligencia de negocio avanzada y soluciones de descubrimiento web (IBM, 2015).

Si bien, el movimiento de *Linked Data* pretende abonar en el desarrollo de la recuperación de información a través de la estructuración y vinculación de datos y recursos en la Web, todavía hay algunas preguntas como las presentadas en el informe 2015 de la NISO sobre el futuro del descubrimiento de recursos de bibliotecas:

> "¿Hasta dónde podemos esperar que el universo de contenido de interés estará disponible como *Open Linked Data*? ¿Qué tecnologías facilitarán la mejora de la experiencia del usuario basada en *Linked Data*? ¿Qué nuevas capacidades serán posibles?" (Breeding, 2015, p. 1)

Por otro lado, la Web Semántica es una extensión de la Web actual en la cual a la información se le otorga un significado bien definido (Berners-Lee, Hendler, & Lassila, The Semantic Web, 2001). Algunas razones por las cuales es importante la Web Semántica son (Bratt, 2007):

- Para tareas que a menudo requieren la combinación de datos a través de Internet, por ejemplo .:
    - La integración de datos en una empresa
    - Hotel, transporte, información personal provienen de diferentes sitios
    - Minería de datos de bases de datos de pacientes, productos farmacéuticos o bioquímicos.
    - Referencias cruzadas de bibliotecas digitales dispares
- Para que los seres humanos entiendan cómo combinar esta información (diferentes vocabularios, idiomas, formatos)

El desarrollo de la Web Semántica, la cual busca mejorar la sensibilidad semántica de las computadoras conectadas vía Internet, requiere una representación sistemática del mundo, orientada a las computadoras; los investigadores a menudo se refieren a ese modelo del mundo como una ontología (Missikoff , Navigli, & Velardi, 2002).

Las ontologías son una forma de representar el conocimiento a través de la formalización de entidades, propiedades y relaciones dentro de un dominio; a través de ontologías es posible crear aplicaciones de la Web Semántica.

Las ventajas de las ontologías para los propósitos de la administración de la información han sido comprobadas. Cada grupo de analistas de datos necesitarían desempeñar la tarea de hacer sus métodos y conceptos compatibles con aquellos otros grupos sólo una vez si calibraran sus resultados en términos de un único lenguaje troncal canónico. Si todas las bases de datos fueran calibradas en términos de solo una ontología común (un solo conjunto estable, consistente y altamente expresivo de etiquetas de

categorías) entonces la perspectiva surgiría de aprovechar los miles de personas-año de esfuerzo que se han invertido en la creación de recursos de base de datos independientes, de tal manera que se cree, de forma más o menos automática, una única base de conocimiento integrada de una escala hasta entonces inimaginable, cumpliendo así el sueño filosófico antiguo de una "Gran Enciclopedia" comprender todo el conocimiento dentro de un único sistema (Smith, 2003).

1.3 Preguntas de investigación

¿Es factible para las tecnologías de la Web Semántica consumir los metadatos estructurados con el Protocolo para Cosecha de Metadatos OAI-PMH?

¿Cuáles son los límites o restricciones en el uso de la información estructurada con OAI-PMH para descubrir conocimiento?

¿El empleo de ontologías en la representación de conjuntos de datos estructurados con OAI-PMH resulta útil en el descubrimiento de conocimiento?

1.4 Hipótesis

Es posible aplicar tecnologías de la Web Semántica y ontologías sobre datos estructurados bajo la especificación del protocolo OAI-PMH, que permita el descubrimiento de conocimiento, entendido como la inferencia de recursos de información no solicitada, no explícita, previamente desconocida y potencialmente útil para un usuario.

1.5 Planteamiento del problema

El protocolo OAI-PMH surgió con fines de interoperabilidad entre plataformas. Los recursos estructurados bajo esta especificación además de ser intercambiados entre sistemas, son usados en motores de búsqueda que habilitan la consulta a los usuarios. Sin embargo, estos datos tienen un uso potencial poco explotado en aplicaciones semánticas que permitan el descubrimiento de conocimiento.

La Web Semántica es acerca de dos cosas. Se trata de formatos comunes para la integración y combinación de datos procedentes de diversas fuentes, cuando la web original se concentraba principalmente en el intercambio de documentos. También se trata del lenguaje para registrar cómo se relacionan los datos a los objetos del mundo real. Eso permite que una persona o una máquina, comience en una base de datos, y se mueva a través de una serie interminable de bases de datos que están conectadas pero no por medio de cables, sino por ser de la misma cosa (W3C, 2013).

1.6 Objetivo

El objetivo general del trabajo aquí presentado es diseñar un modelo semántico para el descubrimiento selectivo de conocimiento sobre recursos estructurados con el protocolo OAI-PMH que permita verificar la factibilidad y poner en evidencia las limitaciones en la aplicación de tecnologías de la Web Semántica sobre dichos conjuntos de datos para descubrir conocimiento, proceso entendido como la identificación de recursos que no fueron solicitados explícitamente por un usuario pero que resultan potencialmente útiles.

Los objetivos particulares de la investigación son:

- Proponer un modelo semántico para el descubrimiento selectivo de conocimiento.
- Habilitar los datos estructurados con OAI-PMH para su procesamiento por una aplicación semántica.
- Aplicar ontologías en la descripción y verificación del conocimiento obtenido de los recursos OAI-PMH.
- Probar mecanismos de inferencia sobre el conjunto de datos.

1.7 Justificación

Varios canales diferentes para comunicar los resultados de investigación seguirán siendo importantes en los próximos años. Sin embargo, el acceso abierto es el futuro de la publicación académica. (Finch, 2012)

El movimiento de Acceso Abierto a la ciencia toma cada vez mayor fuerza. Miles de universidades en el mundo han determinado publicar -sin costo para el usuario- su producción científica y académica. Incluso se han promulgado leyes para que un país adopte esta filosofía a nivel nacional, tal es el caso de Argentina, Perú y recientemente México (Presidencia de México, 2014).

Esas políticas también han sido implementadas a nivel institucional con los mandatos de Acceso Abierto que las universidades y centros de investigación implementan para promover la creación de repositorios y el autoarchivo de contenidos a favor de la mejora en visibilidad e impacto, la preservación de la memoria institucional y la democratización del conocimiento.

Así, se observa una curva creciente en la implementación de mandatos a nivel internacional, según lo indica el registro de mandatos y políticas de repositorios de Acceso Abierto ROARMAP (figura 1).

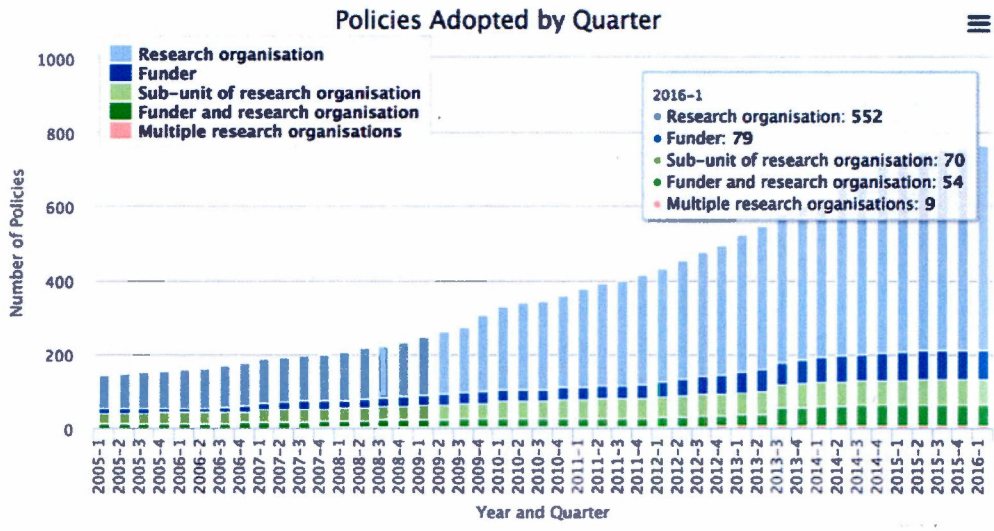

Figura 2. Mandatos de acceso abierto. Fuente: ROARMAP (2015).

La adopción de estas políticas ayudan a la población de los repositorios, consolidando estas plataformas como principales difusoras del conocimiento generado por una institución.

Las plataformas de Acceso Abierto han adoptado el protocolo OAI-PMH como estándar de facto de la comunicación e interoperabilidad, sólo basta observar el registro ROAR con miles de repositorios registrados con el protocolo validado y funcionando, (Eprints, 2016) exponiendo en la red millones de contenidos. Lo cual es muestra de la consolidación de los repositorios y la gran cantidad de material académico – científico disponible en línea bajo la filosofía de libre acceso.

Es así, que resulta relevante dedicar esfuerzos que aporten valor en el descubrimiento del conocimiento contenido en estas plataformas; de tal manera que alumnos, profesores o investigadores obtengan mayores beneficios al trabajar con ellas.

Estos desarrollos pueden abonar en los servicios que pudiera ofrecer un índice de acceso abierto ya que, según afirma Breeding (2015), no han sido creados servicios de descubrimiento de código abierto basados en un índice de acceso abierto.

Entonces, existe un área de oportunidad en el campo del descubrimiento sobre datos estructurados con OAI-PMH que incorpore técnicas de inferencia para recuperar recursos que no se hayan solicitado explícitamente a través de un buscador y que sean relevantes para el usuario.

1.8 Alcance

La propuesta se enfoca en recursos estructurados con el protocolo OAI-PMH, que se encuentran dentro de la nube de datos estructurados presentes en la Web. No contempla datos estructurados con las convenciones de la iniciativa llamada *Linked Data* o cualquier otro tipo de estructuración.

Está por demás decir que los recursos no estructurados quedan fuera del alcance del modelo presentado en este trabajo. En la figura siguiente se muestra un esquema del ámbito considerado.

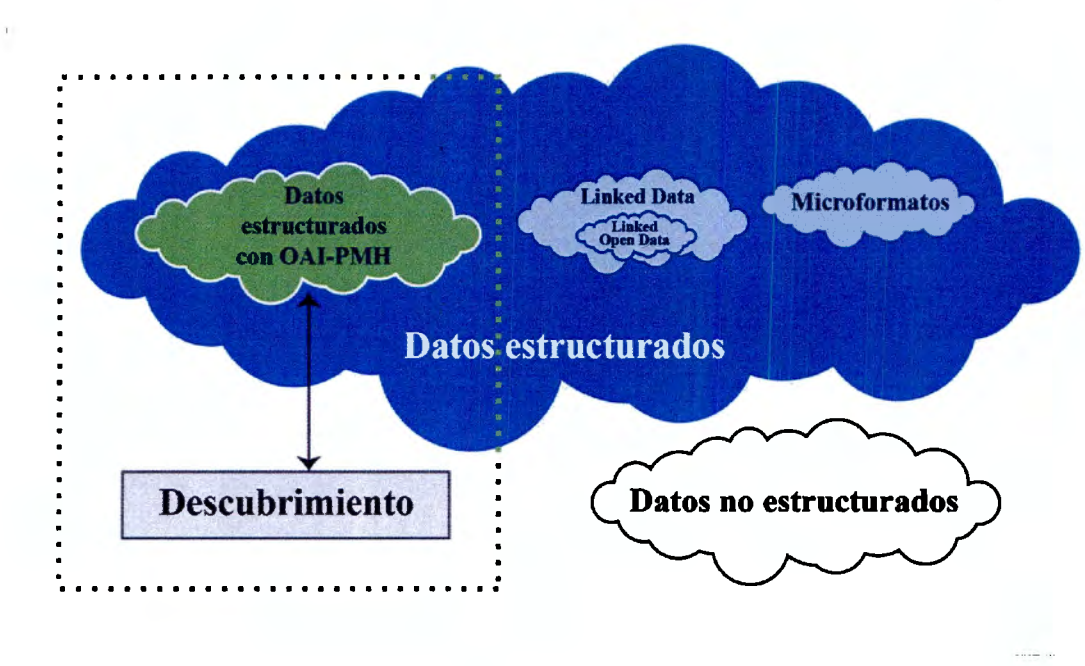

Figura 3. Alcance de la propuesta

Cabe destacar, que el manejo de grandes volúmenes de información queda fuera del alcance de la presente investigación, procesamiento propio de técnicas como *Big Data* o similares cuyo objetivo sea optimizar recursos en tiempos de respuesta, escalabilidad o disponibilidad.

El modelo OntoOAI se limita a utilizar los datos estructurados con OAI-PMH para descubrir conocimiento, entendiendo como el proceso de hallar recursos que no fueron solicitados explícitamente por un usuario pero que resultan potencialmente útiles. Se dice que es *potencialmente útil* ya que no se mide dicha utilidad desde el punto de vista del usuario, es decir, OntoOAI no contempla métricas subjetivas de *interés* únicamente obtiene los resultados arrojados por procesos de inferencia. Es así que se entiende por *potencialmente útil* lo que el proceso lógico de inferencia descubre en la base de conocimiento acorde a los parámetros de entrada.

1.9 Visión general de los resultados

Al finalizar la presente investigación se obtiene un modelo, al cual se le ha nombrado OntoOAI, con el cual fue posible desarrollar un prototipo de software que ejecuta un proceso de inferencia para el descubrimiento de conocimiento sobre datos estructurados con OAI-PMH haciendo uso de un perfil de usuario previamente determinado y que es probado en un caso de aplicación.

1.10 Trabajos relacionados

Algunos motores de la web semántica han nacido libres pero se han convertido en soluciones comerciales en pocos años o han sido adquiridos por grandes empresas. Tal es el caso de Sindice, "El índice de la Web Semántica" como se hacía llamar, que desde 2007 fue un proyecto patrocinado por el Digital Enterprise Research Institute (DERI).

Sindice ofrecía 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 (Tummarello, Delbru, & Oren, 2007).

En 2014 el equipo fundador dejó de otorgar soporte al proyecto (Franzon, 2014) ya que fundan una empresa comercial. Si bien Sindice había destacado dentro de los desarrollos de la Web Semántica, no estaba enfocado en recursos estructurados con OAI-PMH.

Por otro lado Freebase, exponía una colección abierta de datos estructurados con una plataforma para accederlos y manipularlos a través de una API. Sin embargo, en julio

de 2010 fue adquirida por Google. El desarrollo de "Google Knowledge Graph" está en parte basado en Freebase. Para diciembre de 2015 Google anunció que la API de Freebase será descontinuada (Google, 2015).

El "grafo del conocimiento" o "Knowledge Graph", la característica de la búsqueda de Google introducida en 2012, fue considerada como el primer paso en la construcción de búsquedas de nueva generación, que toca la inteligencia colectiva de la web y entiende el mundo más como la gente lo hace. A diferencia de Freebase, este desarrollo hace uso no sólo de fuentes públicas y está basado en lo que la gente busca y la información que obtiene de Google (2012).

Otro caso es la propuesta de Yahoo Inc. al crear en el 2009 Yahoo Search BOSS (Build your Own Search Engine) como una herramienta gratuita que posteriormente dejó de serlo (Schonfeld E. , 2010) y que extraía RDFa y microformatos de cada página de su índice así como archivos RSS y los hacía disponibles para el desarrollo de aplicaciones a través de un API, un servicio orientado al desarrollo de motores de búsqueda para sitios web individuales y que se ha anunciado será descontinuado pronto (Yahoo Inc., 2009).

Por su parte, Wikidata es una base de conocimiento libre que actúa como un almacenamiento central para datos estructurados de proyectos hermanos de la Fundación Wikimedia (Vrandečić & Krötzsch, 2014).

Sin embargo, debido a la generalidad de datos que manejan todos estos desarrollos ninguno contempla el tratamiento de datos de recursos estructurados con OAI-PMH.

Por otro lado, hay trabajos en el campo del descubrimiento de recursos y recomendación (Foulonneau & Grouès, 2012) o para ambientes de aprendizaje personales (Jeremić, Jovanović, & Gašević, 2011) y sobre la información de Linked Data.

Respecto a las soluciones comerciales catalogadas como "descubridores", en el campo de las bibliotecas, se encuentran por ejemplo: EBSCO Discovery Search, WorldCat Discovery Service, Summon, Ex Libris Primo, entre muchos otros. Estos servicios recuperan información de múltiples fuentes y están enfocados a la consulta de los contenidos y no a la inferencia de recursos -no solicitados- y que son de interés para un usuario, es decir no realizan descubrimiento de conocimiento como se ha definido en las secciones anteriores.

También es posible encontrar proyectos de país como la Biblioteca Nacional de Ciencia y Tecnología (National Science and Technology Library, NSTL) de China que integra nueve bibliotecas nacionales. Además de que construyeron una ontología bibliográfica reusando otras como Dublin Core, RDF Entities, Bibliographic Ontology, MarcOnt, para construir un OPAC semántico que permita visualizar las relaciones semánticas entre los objetos.

En el capítulo siguiente (apartado 2.5) se detalla una comparación que incluye 17 proyectos, de los cuales solo 4 utilizan técnicas semánticas para recuperar la información que ninguno contempla específicamente los recursos OAI-PMH, es decir, estos datos no son aprovechados para obtener una mayor semántica, son tratados de manera similar a los datos no estructurados al momento de la indexación. Asimismo, ninguno de estos 4 considera alguna información del contexto del usuario.

Del total de servicios comparados, 12 corresponden a servicios de descubrimiento para bibliotecas. Ninguno de los 12 utilizan tecnologías de la Web Semántica y no realizan procesos de descubrimiento de conocimiento con inferencia o razonamiento, de igual forma no son sensibles al contexto del usuario. Aunque 10 de ellos si indexan contenido estructurado con OAI-PMH.

1.11 Estructura del documento

El contenido de la disertación se divide en cinco capítulos. En este primero se presenta la dimensión de la investigación, los motivos que llevaron a abordar el tema, antecedentes, preguntas de investigación, objetivo, planteamiento del problema, hipótesis, justificación, alcance, trabajos relacionados e introduce los resultados obtenidos.

El segundo capítulo resume la revisión de la literatura en los temas involucrados en la investigación: descubrimiento de conocimiento, Web Semántica y ontologías, protocolo OAI-PMH, sensibilidad al contexto y descubrimiento selectivo.

La descripción del modelo OntoOAI se encuentra en el capítulo tres e incluye en diversos apartados la metodología, el detalle de las herramientas y técnicas utilizadas, la especificación de los componentes de software desarrollado y los problemas encontrados.

OntoOAI es aplicado en un caso de estudio, descrito en el capítulo cuatro, haciendo uso de la información de dos repositorios que implementan el protocolo OAI-PMH: Redalyc.org (Red de Revistas Científicas de América Latina y el Caribe, España y Portugal) y el repositorio institucional de la Universidad Roskilde de Dinamarca llamado

RUDAR (Roskilde University Digital Archive).

Finalmente, el capítulo cinco se dedica a las conclusiones de la investigación así como para mencionar las posibles líneas de investigación identificadas como trabajo futuro.

Capítulo 2 Estado del arte

Este apartado describe el panorama de las principales áreas de investigación sobre las cuales el modelo OntoOAI tiene sus bases: el descubrimiento de conocimiento, el protocolo para cosecha de metadatos OAI-PMH, las ontologías como técnica de representación del conocimiento, tecnologías de la Web Semántica, así como la sensibilidad al contexto que permite el descubrimiento selectivo.

2.1 Descubrimiento de conocimiento

Desde hace dos décadas ya se hacía especial énfasis en la urgente necesidad de una nueva generación de teorías y herramientas computacionales para asistir a los humanos en extraer información útil (conocimiento) de los volúmenes rápidamente crecientes de datos digitales (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

El descubrimiento puede ser definido como el proceso y la infraestructura requerida por un usuario para encontrar un elemento apropiado (Schonfeld R. , 2014). Entre sus diferentes fases puede incluir una búsqueda exploratoria, una búsqueda informada para localizar información específica a través del conocimiento previo; o inclusive, el conocimiento puede llegar al usuario a través de mecanismos proactivos.

En el ámbito de las bibliotecas, los sistemas de descubrimiento de recursos representan la siguiente generación de los catálogos OPAC (*Online Public Access Catalog*) dentro de los sistemas integrados de bibliotecas (ILS por sus siglas en inglés, *Integrated Library Systems*). Son comúnmente conocidos como "servicios de descubrimiento de escala web", que se basan en un índice central poblado por metadatos, textos completos u otras representaciones de elementos en una colección de bibliotecas (Breeding, 2015).

Dentro de los componentes y características de los servicios de descubrimiento en las bibliotecas, se encuentran: la interfaz de usuario final, interoperabilidad con un *link resolver*, búsqueda y recuperación, comunicación con los sistemas integrados de bibliotecas y acceso remoto a plataformas a través de un API [1].

Marshall Breeding usa el término *servicio o producto de descubrimiento* para incluir herramientas o interfaces que una biblioteca implementa para proveer a sus usuarios la capacidad de búsqueda en sus colecciones y ganar acceso a sus materiales. Estas herramientas incluyen software que puede ser instalado en un servidor local y herramientas ofrecidas como un servicio alojado, a estas últimas particularmente se les llama servicio de descubrimiento (Breeding, 2014).

Para el caso comercial, el Grupo de Trabajo en Calidad de los e-datos integrado por empresas entre las que se encuentran, entre otras, Springer, Elsevier, OCLC, JSTOR y Wiley reconocen tener un interés compartido en mejorar el descubrimiento y acceso a contenidos para usuarios de bibliotecas (Suzanne Saskia Kemperman, 2014).

Si bien a estos servicios en el campo de las bibliotecas se les ha llamado "descubridores" o "de descubrimiento"; la tarea de descubrimiento que realizan, se enfoca en la recuperación de recursos de información de las bases de datos y fuentes a las que tiene acceso una biblioteca. El descubrimiento de conocimiento tratado en el modelo OntoOAI

---

[1] Un API (*Application Programming Interface*) o interfaz de programación de aplicaciones, es el conjunto de subrutinas, funciones y procedimientos para construir software y aplicaciones, permite a dos desarrollos de software comunicarse entre sí.

está relacionado más con el concepto conocido como *Knowledge Discovery* o KD y que tiene su origen en el hallazgo de información útil en grandes volúmenes de información.

En ese sentido, el descubrimiento de conocimiento comenzó desde hace dos décadas particularmente dentro de bases de datos, como un área interdisciplinaria llamada KDD (*Knowledge Discovery in Databases*) que se enfoca en metodologías para extraer conocimiento útil a partir de datos. La frase "descubrimiento de conocimiento en bases de datos" se acuñó en el primer congreso KDD en 1989.

IBM ha jugado un papel importante dentro de KDD. Su liderazgo fue establecido en 1990 por la introducción de la minería de reglas de asociación de Rakesh Aggrawal. Otras contribuciones de IBM en KDD incluyen minería de flujos de información excesiva con técnicas de análisis de datos ligeras, minería de alto rendimiento en ambientes de ejecución paralela, y han sido pioneros en el área de minería de datos de preservación privada (IBM, 2015).

El concepto de "descubrimiento de conocimiento", para el caso de bases de datos, fue usado por Frawley, Piatetsky-Shapiro y Matheus (1992) para definir **la extracción no trivial de información implícita, previamente desconocida y potencialmente útil**, a partir de datos; presentando un marco de trabajo para el descubrimiento de conocimiento y examinando problemas al tratar con bases de datos grandes y ruidosas, el uso de conocimiento de dominio, el rol del usuario en el proceso de descubrimiento, los métodos de descubrimiento, así como la forma y uso del conocimiento descubierto. El modelo OntoOAI aborda el descubrimiento de conocimiento en el sentido definido por estos investigadores como se mostrará en el capítulo 3.

Dichos investigadores definen el proceso de descubrimiento de la siguiente forma (Frawley, Piatetsky-Shapiro, & Matheus, 1992):

Dado un conjunto de hechos (datos) F, un lenguaje L, y alguna medida de certidumbre C, se define un *patrón* como un enunciado S en L que describe relaciones entre un subconjunto F_S de F con una certidumbre c, de tal forma que S es más simple (en algún sentido) que la enumeración de todos los hechos en F_S. Un patrón que es interesante (de acuerdo a una medida de interés impuesta por el usuario) y suficientemente cierta (otra vez, de acuerdo al criterio del usuario) es llamado *conocimiento*. La salida de un programa, que monitorea el conjunto de hechos en una base de datos y produce patrones en este sentido, es *conocimiento descubierto.*

Fayyad, Piatetsky-Shapiro, & Smyth (1996) definen un proceso KDD, como el proceso no trivial de identificar patrones comprensibles, potencialmente útiles, nuevos y válidos en los datos y se compone de las etapas mostradas en la siguiente figura.

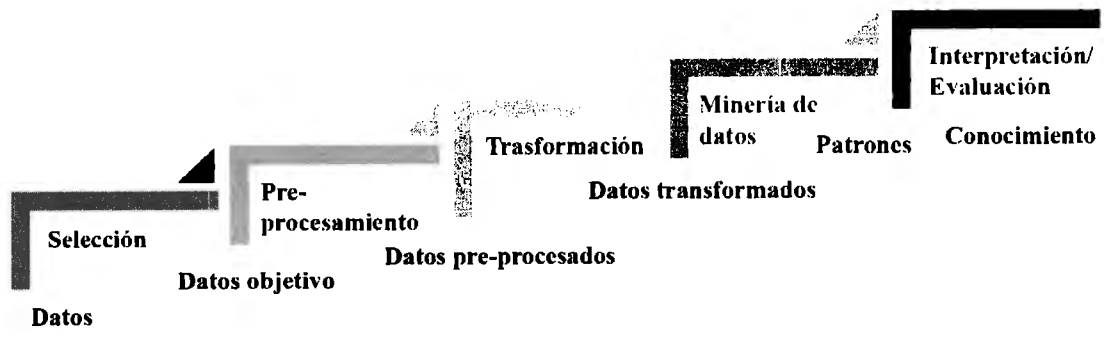

Figura 4. Etapas del proceso KD. Fuente: Fayyad, Piatetsky-Shapiro, & Smyth (1996).

Este proceso es iterativo y parte de los datos, como un conjunto de hechos. Es no trivial ya que involucra alguna búsqueda o inferencia, es decir, no es un cómputo simple de cantidades predefinidas. Los conceptos tales como novedad y comprensibilidad son subjetivos por lo que Piatetsky-Shapiro y Matheus (1994) introdujeron el concepto de *interestingness* o *interés* y retomado por Silberschatz y Tuzhilin (1995) como una medida

general del valor de un patrón, combinando la validez, utilidad, novedad y simplicidad. La función de *interés* puede ser definida explícitamente o puede ser manifestada de manera implícita a través de la ubicación ordenada de los patrones descubiertos.

Con esa métrica se puede considerar que un patrón es conocimiento si excede algún umbral de *interés*. El conocimiento, en la definición dada por estos autores, es totalmente orientado al usuario y especifico al dominio, y está determinado por cualquier función o umbral que el usuario escoja.

Un problema central del descubrimiento de conocimiento es el desarrollo de medidas para el *interés*. Tales medidas son divididas entre objetivas y subjetivas (figura 5). Las medidas de interés objetivas son aquellas que dependen únicamente en la estructura de un patrón y en los datos subyacentes usados en el proceso de descubrimiento. Las subjetivas son las que dependen en el usuario que examina el patrón y evalúan qué tan inesperado (*unexpectedness*) es un patrón encontrado y la accionabilidad (*accionability*), para verificar la sorpresa que le provoca al usuario y si puede hacer algo con el resultado, respectivamente (Silberschatz & Tuzhilin, 1995).

Cabe resaltar, que para el modelo OntoOAI la medida para calificar una sentencia como "de interés" corresponde a la clasificación de medidas objetivas, ya que, se basa en la estructura misma del patrón encontrado y los datos subyacentes al perfil del usuario y no a la enunciación explícita de interés por parte del usuario, es por ello que se menciona como "información potencialmente útil".

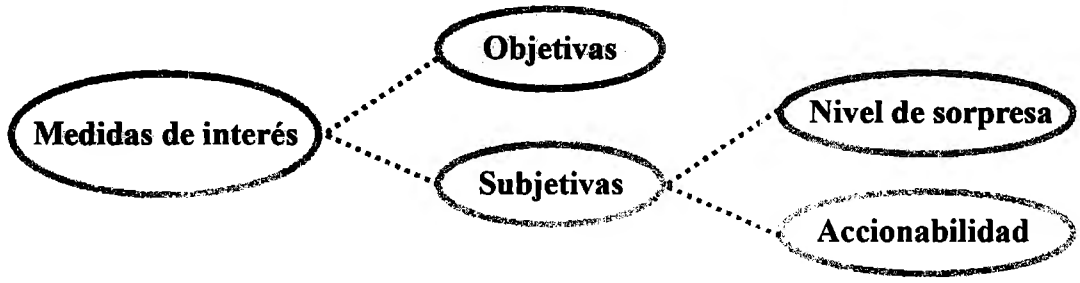

Figura 5. Clasificación de las medidas de interés. Fuente: Silberschatz y Tuzhilin (1995).

Hay una confusión común entre los conceptos de Minería de Datos (DM), Descubrimiento de Conocimiento (KD) y Descubrimiento de Conocimiento en Bases de Datos (KDD) (Kurgan & Musilek, 2006). Históricamente, la noción de encontrar patrones útiles en los datos ha recibido una variedad de nombres, incluyendo minería de datos, extracción de conocimiento, descubrimiento de información, cosecha de información, arqueología de datos, y procesamiento de patrones de datos (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Minería de Datos tiene que ver con la aplicación de métodos de bajo nivel, los cuales son algoritmos diseñados para analizar datos o extraer patrones en categorías específicas a partir de datos (Klosgen & Zytkow, 1996).

Según el proceso de KDD de la figura 4, la Minería de Datos es un paso que consiste en aplicar el análisis de datos y los algoritmos de descubrimiento que producen una enumeración particular de patrones sobre los datos.

Y finalmente, KDD se refiere a descubrir conocimiento dentro de bases de datos.

Existe otra área que trata al conocimiento, pero no en términos de descubrimiento sino de "extracción", siendo esta tarea la encargada de obtener entidades, relaciones y hechos a partir de datos no estructurados o estructurados, como bases de datos relacionales, de manera automática; es decir la creación de conocimiento.

Varios son los desarrollos en este ámbito, por ejemplo la extracción de conocimiento basada en la teoría de representación del discurso y enmarques lingüísticos Presutti, Draicchio y Gangemi (2012), la extracción de relaciones de patrones sintáctico-léxicos de Moldovan y Blanco (2012), y algunos ya representados como productos comerciales, tal es el caso de *NetOWL Extractor* que además de extraer entidades y relaciones, realiza geo-marcación y análisis de sentimiento a través de lingüística computacional y procesamiento de lenguaje natural (SRA International, Inc., 2016).

En este sentido, otros han propuesto métodos de extracción como el de Yildiz y Miksch (2007) que utiliza el contenido y la semántica predefinida de ontologías formuladas en OWL para desempeñar las tareas de extracción.

La extracción de conocimiento resulta relevante cuando la información está codificada en lenguaje natural es decir no estructurada, para llevar a cabo este tipo de procesos se han desarrollado diversos métodos y técnicas:

- Extracción de información (IE) tradicional, como una técnica de procesamiento de lenguaje natural que generalmente se encuentra en textos.
- Extracción de información basada en ontologías (OBIE), donde se usa al menos una ontología para guiar la extracción a partir de textos en lenguaje natural.
- Aprendizaje de ontologías (OL), se refiere a la creación automática de ontologías tomando como insumo texto en lenguaje natural.
- Anotación semántica, es el enriquecimiento de texto en lenguaje natural con metadatos estructurados.

El modelo OntoOAI, que se presenta en este trabajo, no trata con datos no estructurados es por ello que no se contempla la extracción del conocimiento.

Además del aprendizaje de máquina, otros campos de la inteligencia artificial se perfilaron para contribuir a los procesos de KDD como, procesamiento de lenguaje natural, problemas de planeación, manejo de incertidumbre, agentes inteligentes, representación del conocimiento incluyendo ontologías, entre otros (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Como se tratará en el apartado siguiente, la Web Semántica y las ontologías aportan técnicas y herramientas útiles para habilitar mecanismos de descubrimiento de conocimiento.

2.2 Web semántica y ontologías

La Web Semántica puede ser vista principalmente como la integración de muchas ideas y tecnologías que ya existen con el enfoque específico de mejorar la naturaleza existente de los sistemas de información basados en la web a una naturaleza orientada a una mayor semántica (Grobelnik & Mladenić, 2005).

Con el lanzamiento en 2012 de Schema.org, Google y otros han abrazado de manera efectiva la visión de la "Web Semántica". Actualmente, con el estándar RDFa y aún más con JSON-LD la marcación enriquecida está siendo más popular en los sitios web (Franzon, 2014).

2.2.1 Tecnologías de la Web Semántica

La arquitectura y tecnologías de la Web Semántica se describen en el diagrama de pila de capas (figura 6) originalmente propuesto por Tim Berners-Lee y que ha evolucionado a la par del desarrollo de las tecnologías, la mostrada en la figura corresponde a su edición de 2007.

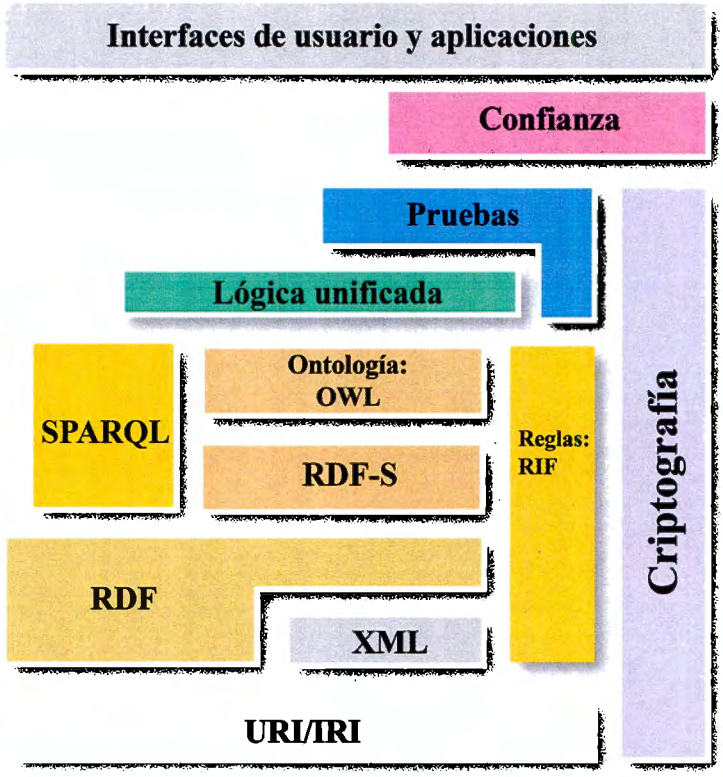

Figura 6. Modelo de pila de capas de la Web Semántica de Tim Berners Lee. Fuente: W3C (2007).

La base de la arquitectura está compuesta por el identificador uniforme de recursos (*Uniform Resource Locator* o URI), URI es una cadena de caracteres estandarizada que permite identificar de forma exclusiva los recursos en la Web, es útil para asignar nombres sin ambigüedad.

Un subconjunto de URI es *Uniform Resource Locator* (URL), que contiene un mecanismo de acceso a un documento y su localización en la red. Otro subconjunto de URI es URN que permite identificar un recurso sin que ello implique su ubicación por ejemplo un ISBN. Una variante internacional para URI es el IRI (*Internationalized Resource Identifier*), que permite el uso de caracteres Unicode en el identificador y para el que se define una asignación de URI.

La capa de XML (*Extensible Markup Language*) junto con las definiciones del espacio de nombres y del esquema XML aseguran que exista una sintaxis común. XML es

es un formato de texto simple y flexible, derivado de SGML (ISO 8879). Originalmente diseñado para cumplir con los retos de la publicación electrónica a gran escala, XML también está desempeñando un papel cada vez más importante en el intercambio de una amplia variedad de datos en la Web y en otros lugares (W3C, 2015).

El XML y el RDF conforman la capa de representación de datos en la Web Semántica. RDF (*Resource Description Framework*) es un marco de trabajo para expresar y vincular los datos incluyendo metadatos. Se basa en tripletas sujeto-predicado-objeto para formar grafos. En un apartado más adelante se describe con mayor detalle este estándar.

Si bien RDF sirve para describir los datos, cualquiera puede definir el vocabulario de términos utilizados en esa representación. Para permitir la descripción estandarizada de las taxonomías y otras construcciones ontológicas, un esquema RDF llamado RDFS fue creado y forma parte de la Web Semántica.

Una herramienta para capturar las ontologías es OWL, el Lenguaje para Ontologías Web. Es un lenguaje de la Web Semántica diseñado para representar el conocimiento rico y complejo acerca de las cosas, los grupos de cosas, y las relaciones entre las cosas.

OWL es un lenguaje computacional basado en la lógica, de tal forma que el conocimiento expresado en OWL puede ser explotado por los programas de computadoras, por ejemplo, para verificar la consistencia de ese conocimiento o para hacer explícito el conocimiento implícito. Los documentos OWL, conocidos como ontologías, pueden ser publicados en la *World Wide Web* y pueden referirse o ser referidos desde otras ontologías OWL (W3C, 2012).

El OWL es un lenguaje derivado de la lógica de descripción, y ofrece más constructores que RDFS. Está incrustado sintácticamente en RDF, y así como RDFS, proporciona vocabulario estándar adicional. OWL viene en tres niveles:

1. *OWL Lite* para taxonomías y restricciones simples
2. *OWL DL* para el pleno soporte a la lógica descripción
3. *OWL Full* para la máxima expresividad y libertad sintáctica de RDF

RDFS y OWL tienen una semántica definida y esta semántica se puede utilizar para el razonamiento dentro de ontologías y bases de conocimiento. Para proveer reglas más allá de las construcciones disponibles en estos lenguajes, se están estandarizando también los lenguajes de reglas para la web semántica: RIF y SWRL.

En la pila también se encuentra SPARQL (*Simple Protocol and RDF Query Language*), un lenguaje tipo SQL, pero que utiliza tripletas y recursos RDF, tanto para la consulta como para el retorno de los resultados. SPARQL no es solamente un lenguaje de consulta, sino que también es un protocolo para acceder a datos RDF (Obitko, 2007).

SPARQL se utiliza para expresar consultas a través de diversas fuentes de datos, cuando estos se almacenan de forma nativa como RDF o cuando pueden ser vistos como RDF a través de algún *middleware*. La versión actual de SPARQL es SPARQL 1.1, que reemplaza a la versión anterior publicada en 2008 (W3C, 2013).

Dado que ambos, RDFS y OWL se construyen sobre RDF, SPARQL se puede utilizar para consultar ontologías y bases de conocimiento directamente.

Se espera que la semántica y reglas se ejecuten en las capas bajo pruebas y los resultados sean utilizados para probar las deducciones. Las pruebas formales junto con entradas de confianza para las pruebas hacen esperar que los resultados sean de confianza, como se muestra en la capa superior de la figura de capas de la Web Semántica.

Para las entradas de confianza, se utilizan medios de criptografía, tales como firmas digitales para la verificación del origen de las fuentes. En la parte superior de estas capas, se pueden construir aplicaciones con interfaces de usuario (Obitko, 2007).

Además de las tecnologías mostradas en el diagrama de pila de capas, muchas otras han sido desarrolladas en torno a la Web Semántica y se han venido consolidando a lo largo de los últimos años, desde el desarrollo de *RDF Schema* en 1999, hasta completar una variedad de herramientas para diferentes tareas encaminadas al desarrollo de un mejor entendimiento de la información en la Web y que incluye a RDF, OWL, SPARQL y RDFa.

Paralelamente, el campo del XML ha evolucionado a la par. Desde la introducción del XML en 1996 hasta el desarrollo de XSLT, XPath, XML Schema y XQuery en sus diferentes versiones a lo largo de dos décadas.

El proceso evolutivo de todas estas tecnologías en el tiempo, se resume muy adecuadamente con el trabajo de Bikakis, Tsinaraki, Gioldasis, Stavrakantonakis y Christodovlakis (2013) y su línea del tiempo de los estándares de la W3C para la Web Semántica y XML mostrado en la siguiente figura.

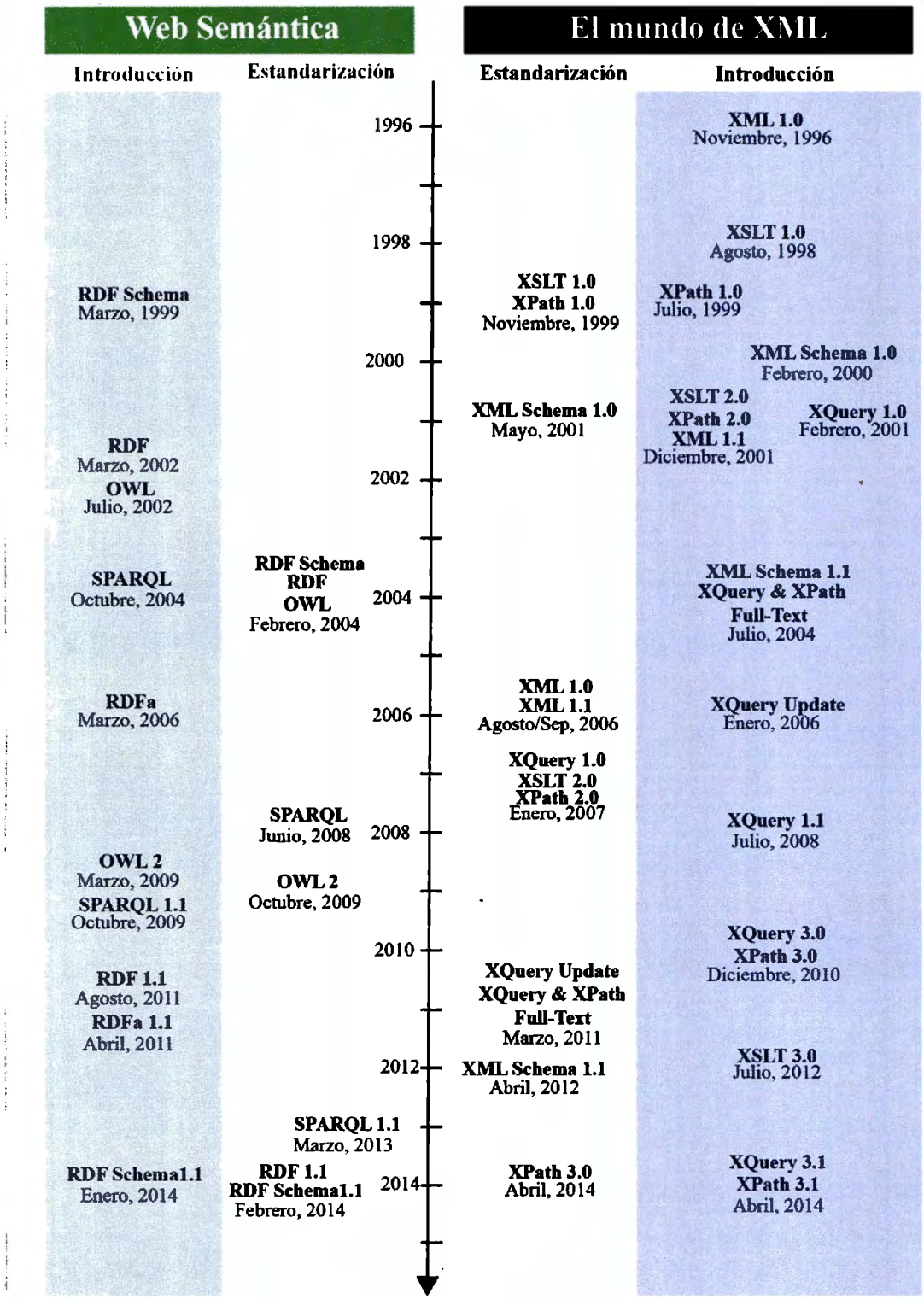

Figura 7. Línea del tiempo de los estándares de la W3C para la Web Semántica y XML, traducida de Bikakis, Tsinaraki, Gioldasis, Stavrakantonakis y Christodovlakis (2013).

2.2.2 Ontologías

La palabra ontología proviene del griego ὄντος "del ente", genitivo del participio del verbo εἰμί "ser, estar"; y λόγος "ciencia, estudio, teoría".

En la historia del término dos filósofos aparecen como los primeros en utilizarlo: Rudolf Göckelin, con su obra de 1613 *Lexicon philosophicum quo tanquam clave philosophiae fores aperiuntur, informatum opera et studio* (Göckel, 1980) y Jacob Lorhard en 1606 con *Ogdoas Scholastica, continens Diagraphen Typicam artium: Grammatices (Latinae, Graecae), Logices, Rhetorices, Astronomices, Ethices, Physices, Metaphysices, seu Ontologiae* (Lorhard, 1606).

Filosóficamente hablando, Ontología es el estudio de la naturaleza del ser y busca proveer una clasificación exhaustiva y definitiva de entidades en todas las esferas del ser.

Diferentes escuelas de filosofía ofrecen variados acercamientos a tales clasificaciones, Smith (2003) señala dos divisiones: una gran corriente compuesta por los *sustancialistas* y *flujistas*, quienes conciben la Ontología como una sustancia y los *ocurrentes* quienes favorecen la Ontología centrada en eventos o procesos. La otra división está entre los *adecuadistas* y *reduccionistas,* los primeros buscan una taxonomía de las entidades en la realidad a todos los niveles de agregación, mientras los segundos ven la realidad en términos de algunos niveles privilegiados de existentes, descomponiendo la realidad en sus constituyentes más simples .

En esta última rama se encuentra el trabajo de los *ontologistas* Aristóteles, Ingarden (1964) y Chisholm (1996) que resulta importante, ya que sus taxonomías se comparan con las de las ciencias como Biología o Química. La ontología, para los *adecuadistas*, es una empresa descriptiva. Se distingue de las ciencias específicas no sólo por su generalidad

radical sino también por su objetivo primordial o enfoque: la Ontología busca la taxonomía en lugar de la predicación o explicación (Smith, Ontology, 2003).

Durante el siglo veinte un rango de nuevas herramientas están disponibles para los ontologistas, como marcos de trabajo formales derivados del algebra, teoría de conjuntos, topología, entre otros, que junto con el lenguaje de lógica formal permiten expresar a los filósofos principios intuitivos y definiciones de una forma rigurosa y clara; y a través de la aplicación de métodos de semántica formal pueden probar sus teorías en cuanto a consistencia y completud (Smith, 2003).

El mismo término ontología ha ganado recientemente un espacio en las ciencias de la información y de la computación desde la primera vez que fue usado en esta rama por George Mealy (1967) según afirma Smith (2004).

Diferentes integradores de datos tienen sus propios términos y conceptos idiosincráticos para representar la información. Cuando se intenta reunir esta información, se tienen que encontrar métodos para resolver las incompatibilidades conceptuales y terminológicas. Inicialmente dichas incompatibilidades se resolvían caso por caso. Sin embargo, gradualmente fue evidente que proveer -una vez y para todo- una taxonomía estructural común de entidades relevantes de un dominio de aplicación proveería ventajas significativas. Esta taxonomía común es llamada ontología (Smith, 2003).

Una ontología es una especificación formal y explícita de una conceptualización compartida. Considerando *conceptualización* como el 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 (Gruber, 1995) (Studer, Benjamins, & Fensel, 1998). Las ontologías habilitan a una computadora para entender la información por sí misma (Abburu, 2012).

Para la pregunta de ¿cuáles son las ventajas de desarrollar una ontología?, se puede responder con las siguientes razones (Noy & McGuinness, 2001):

- Para compartir la comprensión común de la estructura de información entre personas o agentes de software.
- Para permitir la reutilización de conocimiento de dominio.
- Para hacer explícitas suposiciones de dominio.
- Para separar el conocimiento de dominio del conocimiento operativo.
- Para analizar el conocimiento de dominio.

Los mismos autores enuncian los componentes de una ontología:

- Clases, que corresponden a los conceptos y subclases, que representan conceptos que son más específicos que la superclase.
- Slots (también llamados roles) son las propiedades de cada concepto que describen diversas características y atributos del concepto.
- Facetas (también llamadas restricciones de roles) que describen los tipos de valores, los valores permitidos, el número de valores y cualquier otra característica que un slot puede tomar.
- Instancias u objetos de una clase.

Un bloque básico en la construcción de ontologías son los conceptos. Típicamente los conceptos están organizados jerárquicamente en una jerarquía de conceptos. Maedche y Volz (2001) definen un conjunto de conceptos y una jerarquía de conceptos, como sigue:

- Un conjunto *C* cuyos elementos son llamados conceptos

- Una **jerarquía de conceptos** $H^C : H^C$ es una relación dirigida $H^C \subseteq C \times C$ también llamada taxonomía. H^C (C_1, C_2) significa que C_1 es un subconcepto de C_2.

Los conceptos y la jerarquía de conceptos son extendidos con relaciones no-taxonómicas entre conceptos y un conjunto de axiomas, que se definen como:

- Un conjunto R cuyos elementos son llamados relaciones, los conjuntos C y R son disjuntos.
- Una función $rel : R \rightarrow C \times C$, que relaciona no-taxonómicamente los conceptos. La función *dom*: $R \rightarrow C$ con $dom (R) := \prod_1(rel (R))$ otorga el dominio de R, y el rango: *rango*: $R \rightarrow C$ con $rango (R) := \prod_2(rel (R))$ otorga el rango.
- Para $rel (R) = (C_1, C_2)$ también se escribe $R (C_1, C_2)$.
- Un conjunto de axiomas de ontología A^O, expresado en un lenguaje lógico apropiado, por ejemplo lógica de primer orden.

En algunos casos es necesario proveer un vínculo entre palabras simples y entidades ontológicas. Este mapeo es provisto por un léxico. Un léxico para la estructura ontológica $O := \{ C, R, H^C, rel, A^O \}$ es una 4-tupla $L := \{ L^C, L^R, F, G \}$ que consiste en:

- Dos conjuntos L^C y L^R, cuyos elementos son llamados entradas léxicas para conceptos y relaciones, respectivamente.
- Dos relaciones $F \subseteq L^C \times C$ y $G \subseteq L^R \times R$ llamadas referencias para conceptos y relaciones, respectivamente. Basado en F, para cada $L \in L^C$, $F (L) = \{ C \in C \mid (L, C) \in F \}$ y para $F^{-1} (C) = \{ L \in L^C \mid (L, C) \in F \}$ (G y G^{-1} son definidos análogamente).

Guarino (1998) clasifica las ontologías de acuerdo a su nivel de dependencia de una tarea o punto de vista particular, como sigue:

- Ontologías de alto nivel, que describen conceptos muy generales como espacio, tiempo, materia, objeto, evento, acción, etc., las cuales son independientes de un dominio o problema particular.
- Ontologías de dominio, que describen un vocabulario relacionado a un dominio genérico.
- Ontologías de tareas, que describen un vocabulario relacionado a una tarea o actividad genérica.
- Ontologías de aplicación, que describen los conceptos conforme a ambos, un dominio y tarea, que son comúnmente especializaciones de ambas ontologías relacionadas.

El mismo autor aclara la diferencia entre una ontología de aplicación y una base de conocimiento. La respuesta está relacionada con el propósito de una ontología, que es una base de conocimiento particular, que describe hechos que se asumen siempre ciertos por una comunidad de usuarios, en virtud del significado acordado del vocabulario utilizado.

Una base de conocimiento genérico, en cambio, puede describir hechos y afirmaciones relacionadas con un determinado estado de las cosas o un estado epistémico particular. Dentro de una base de conocimiento genérico, se pueden distinguir dos componentes: la ontología (que contiene información independiente del estado) y la base de conocimiento "central" (que contiene información de estado-dependiente).

Las ontologías se pueden usar en el tiempo de desarrollo o en el tiempo de ejecución dentro de los llamados Sistemas de Información Dirigidos por Ontologías (*Ontology-Driven Information Systems*). Para el primer escenario, la semántica de una ontología puede ayudar a realizar el análisis conceptual y modelado, o puede apoyar en la generación de alguna parte estática del sistema. En el caso del empleo de las ontologías en tiempo de ejecución, una de las razones es para habilitar la comunicación entre agentes de software, consultas en bases de datos o inclusive en interfaces de usuario.

Se han propuesto e implementado una gran variedad de proyectos para la generación de ontologías que implementan diversos métodos para la extracción de conceptos, extracción de relaciones, reutilización de ontologías, representación de ontologías, métodos para la construcción a partir de fuentes de datos etiquetados. Ding y Foo (2002) realizan una revisión de muchos de ellos.

Incluso hay trabajos que usan el descubrimiento de conocimiento -en el sentido de minería de datos- (Grobelnik & Mladenić, 2005) para la construcción de ontologías, incorporación de conocimiento de dominio, capturar datos multimodales y multilingües, entre otros (Mladenić, Grobelnik, Fortuna, & Grčar, 2009).

Es así, que la experiencia adquirida con los sistemas de información que emplean ontologías pueden apoyar los esfuerzos de los filósofos que se han ocupado no sólo con el desarrollo de las teorías ontológicas, sino también - en un campo a veces llamado ontología aplicada - con la aplicación de tales teorías en ámbitos tales como el derecho, el comercio, o la medicina (Smith, 2003).

2.2.3 Fusión de ontologías

La representación de un dominio de conocimiento puede compartir conceptualizaciones con otras representaciones de ese mismo dominio u otros, de tal forma que la unión o intersección de ontologías resulten operaciones útiles en la ingeniería de conocimiento.

Las ontologías se desarrollan a diferentes niveles de abstracción y por diferentes personas para diversos fines. El conocimiento representado por las ontologías se rompe debido a la existencia de muchas ontologías que representan los mismos conceptos, por lo tanto se hace difícil el análisis y el uso de la propagación de conocimiento a través de muchas ontologías si se estudian de forma individual. Por ello, el conocimiento representado por ontologías se pueden combinar en una sola ontología (Ameen, Rahman Khan, & Rani, 2014).

La integración de la información es un área de aplicación importante para las ontologías. Como afirma Guarino (1998) al explicar el problema de integración de ontologías, incluso si dos sistemas adoptan el mismo vocabulario, no hay garantía de que puedan ponerse de acuerdo sobre una determinada información, a menos que se comprometan a la misma conceptualización:

> Suponiendo que cada sistema tiene su propia conceptualización, una condición necesaria con el fin de llegar a un acuerdo posible es que los modelos previstos de las conceptualizaciones originales se superpongan. Dos sistemas *A* y *B*, usando el mismo lenguaje *L*, se pueden comunicar si y solo si el conjunto de modelos previstos $I_A(L)$ e $I_B(L)$ asociados a sus conceptualizaciones se traslapan (figura 8).
>
> Suponiendo ahora que estos dos conjuntos de modelos destinados se aproximan mediante dos ontologías diferentes, puede darse el caso de que las dos ontologías se solapen mientras que los modelos destinados no.

Esto significa que un enfoque de abajo hacia arriba para la integración de sistemas basados en la integración de múltiples ontologías locales podría no funcionar, especialmente si las ontologías locales sólo se centran en las relaciones conceptuales relevantes para un contexto específico, y por lo tanto sean aproximaciones *ad hoc* y débiles de los modelos previstos. Por lo tanto, parece más conveniente usar una sola ontología de nivel superior en lugar de depender de acuerdos basados en la intersección de diferentes ontologías (Guarino, 1998).

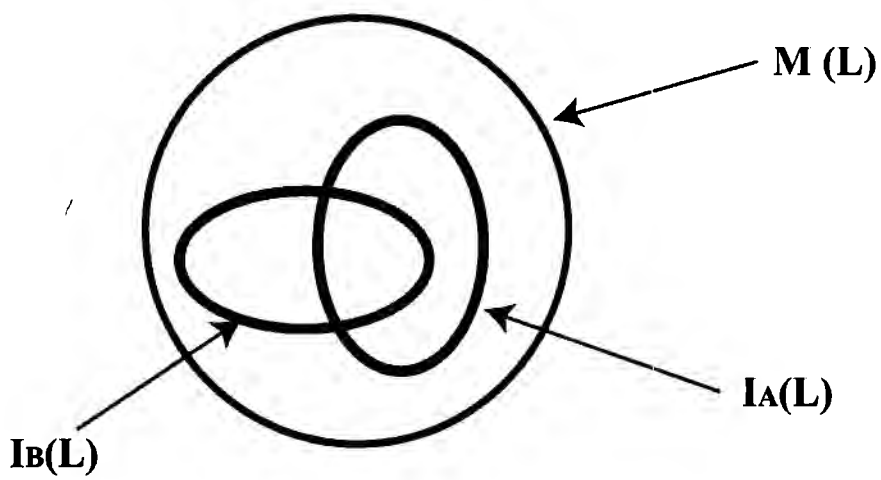

Figura 8. Problema de la integración de ontologías según Guarino (1998).

Muchas técnicas se han desarrollado para fusionar ontologías, como la expuesta por Robin & Uma (2010) donde se utiliza un algoritmo para la fusión de ontologías mediante el uso de una estrategia híbrida que realiza la concordancia, comprueba la similitud y usa funciones heurísticas. Otros proponen enfoques semánticos y sintácticos para encontrar coincidencias (McGuinness, 2000).

Con la fusión de ontologías se obtiene una nueva ontología resultante, la cual es la unión de las ontologías de entrada y contiene toda la información de ellas.

Estas consideraciones sugieren la oportunidad de implementar diferentes tipos de ontologías de acuerdo al nivel de generalidad, ya sea de alto nivel, de dominio, de tarea o de aplicación; de tal forma que permita verificar la consistencia de un conocimiento representado o el razonamiento sobre una base de conocimiento.

2.2.4 RDF

El estándar RDF (*Resource Description Framework*) es un modelo para el intercambio de datos en la Web definido por la W3C. RDF extiende la estructura vinculante de la Web para utilizar URI para nombrar la relación entre las cosas, así como los dos extremos de dicha relación (esto generalmente se conoce como una tripleta).

La estructura subyacente de cualquier expresión en RDF es una colección de tripletas, cada uno compuesto de un sujeto, un predicado y un objeto. Un conjunto de tales tripletas se llama grafo RDF.

Esto se puede ilustrar por un nodo y el diagrama de arco dirigido, en el que cada tripleta se representa como un enlace de nodo-arco-nodo (figura 9). Esta estructura de enlace forma un grafo etiquetado y dirigido, donde las aristas representan el enlace entre dos recursos, representados por los nodos del grafo (W3C, 2014). La dirección del arco debe ser apuntada hacia el objeto.

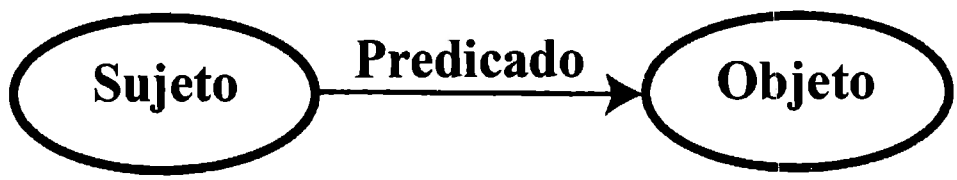

Figura 9. Grafo RDF

RDF tiene características que facilitan la fusión de datos incluso si los esquemas subyacentes son diferentes. El uso de este modelo simple permite que los datos estructurados y semi-estructurados sean mezclados, expuestos y compartidos a través de diferentes aplicaciones.

RDF proporciona una manera de expresar enunciados simples acerca de los recursos. Sin embargo, las comunidades de usuarios necesitan la capacidad de definir los vocabularios (términos) a utilizar en dichos enunciados, específicamente, para indicar que están describiendo tipos o clases específicas de recursos, y propiedades específicas en la descripción de esos recursos. Este tipo de clases y propiedades se describen como un vocabulario RDF, utilizando extensiones de RDF proporcionados por el Lenguaje de Descripción de Vocabularios RDF: RDF *Schema*.

RDF *Schema* no proporciona un vocabulario de clases y propiedades específicas de aplicaciones. En su lugar, proporciona las especificaciones necesarias para describir este tipo de clases y propiedades.

El vocabulario que comprende la especificación, consta clases y propiedades. Las clases son agrupaciones de recursos. Los miembros de una clase se conocen como instancias de la clase. Las clases son en sí mismas recursos. Las clases y las propiedades se enlistan en las tablas 1 y 2, respectivamente.

Clase	Definición
rdfs:Resource	Todas las cosas descritas en RDF se denominan recursos, y son instancias de esta clase. Esta es la clase de todo. Todas las demás clases son subclases de esta clase. *rdfs:Resource* es una instancia de *rdfs:Class*.

rdfs:Class	Esta es la clase de los recursos que son clases RDF. *rdfs:Class* es un instancia de *rdfs:Class*.
rdfs:Literal	Es la clase de los valores literales tales como cadenas y números enteros. Los valores de propiedad, tales como cadenas de texto son ejemplos de literales RDF. *rdfs:Literal* es una instancia de *rdfs:Class*. *rdfs:Literal* es una subclase de *rdfs:Resource*.
rdfs:Datatype	Es la clase de tipos de datos. Es a la vez una instancia y una subclase de *rdfs:Class*. Cada instancia de *rdfs:Datatype* es una subclase de *rdfs:Literal*.
rdf:langString	Es la clase de los valores de cadena de lenguaje de etiquetado. Es una instancia de *rdfs:Datatype* y una subclase de *rdfs:Literal*.
rdf:HTML	Es la clase de los valores literales de HTML. Es una instancia de *rdfs:Datatype* y una subclase de *rdfs:Literal*.
rdf:XMLLiteral	La clase *rdf:XMLLiteral* es la clase de los valores literales XML. Es una instancia de *rdfs:Datatype* y una subclase de *rdfs:Literal*.
rdf:Property	Es la clase de propiedades RDF. Es una instancia de *rdfs:Class*.

Tabla 1. Clases de RDF Schema (W3C, 2004).

Propiedad	Definición
rdfs:range	Es una instancia de *rdf:Property* que se utiliza para indicar que los valores de una propiedad son instancias de una o más clases.
rdfs:domain	Es una instancia de *rdf:Property* que se utiliza para indicar que cualquier recurso que tiene una determinada propiedad es una

	instancia de una o más clases.
rdf:type	Es una instancia de *rdf:Property* que se utiliza para indicar que un recurso es una instancia de una clase.
rdfs:subClassOf	Es una instancia de *rdf:Property* que se utiliza para indicar que todas las instancias de una clase son instancias de otra.
rdfs:subPropertyOf	Es una instancia de *rdf:Property* que se utiliza para indicar que todos los recursos relacionados por una propiedad también están relacionados por otra.
rdfs:label	Es una instancia de *rdf:Property* que se puede utilizar para proporcionar una versión legible del nombre de un recurso.
rdfs:comment	Es una instancia de *rdf:Property* que puede utilizarse para proporcionar una descripción legible de un recurso.

Tabla 2. Propiedades de RDF Schema (W3C, 2004).

RDF también contempla los *contenedores* que son recursos usados para representar colecciones.

Entre los formatos de serialización de RDF se encuentran *Turtle* (Beckett, Berners-Lee, Prud'hommeaux, & Carothers, 2014), *N-Triples* (Beckett, 2014), *N-Quads* (W3C, 2014), *JSON-LD* (Sporny, Longley, Kellogg, Lanthaler, & Lindström, 2014), *N3* o *Notation3* (Berners-Lee, 2005) y *RDF/XML* (W3C, 2014).

2.2.5 Razonamiento

El progreso tecnológico actual permite el almacenamiento y acceso a grandes cantidades de datos virtualmente sin costo. Aunque, el principal problema en un mundo centrado en la información sigue siendo poner apropiadamente los datos crudos recuperados en uso.

Tecnológico de Monterrey, Campus Ciudad de México
Biblioteca

El verdadero valor no está en almacenar los datos, sino en la habilidad de extraer reportes útiles y en extraer tendencias y correlaciones interesantes, a través del uso de análisis estadístico e inferencia para soportar decisiones y políticas hechas por científicos y empresas (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Las ontologías se componen de teorías formales sobre un dominio específico, y por lo tanto tienen un lenguaje lógico-formal asociado a ellas. La lógica se ha propuesto, desde los primeros días de la Inteligencia Artificial, como un marco de trabajo para la representación de conocimiento y razonamiento (*Knowledge representation and reasoning* ó KR). En este contexto, proporciona los tres componentes principales necesarios para la estructuración y el manejo de objetos del mundo real (Furletti, 2009), que son:

- **Sintaxis**: un alfabeto;
- **Semántica**: una interpretación de los símbolos del alfabeto;
- **Sistema de pruebas**: Un método para hacer pruebas

El razonamiento en ontologías y bases de conocimiento es una de las razones por las que una especificación necesita ser formal. Una base de conocimiento puede ser considerada como un conjunto de afirmaciones. El razonamiento se encarga de derivar hechos que no son expresados explícitamente ya sea en la ontología o en la base de conocimiento.

Por razonamiento lógico se entiende realizar una inferencia de una conclusión a partir de un conjunto de premisas. Existen varios métodos de razonamiento: deductivo, inductivo, analógico, analítico y sintético.

Dentro de la lógica proposicional (también llamada Lógica Booleana), el razonamiento lógico requiere la relación de la *implicación* lógica entre las sentencias, es decir, la idea de que una sentencia *se sigue lógicamente* de otra.

Las implicaciones también se conocen como *reglas* o afirmaciones *si-entonces*, y son de la forma $P \Rightarrow Q$, donde P es la premisa o antecedente y Q es su conclusión o consecuente.

Se dice que un algoritmo de inferencia que deriva sólo sentencias implicadas es *sólido* o que mantiene la verdad. La solidez es una propiedad muy deseable. Un procedimiento de inferencia no sólido tan sólo inventaría cosas poco a poco. También es muy deseable la propiedad de *completitud*: un algoritmo de inferencia es completo si puede derivar cualquier sentencia que esté implicada (Russell & Norvig, 2004).

Particularmente, un razonamiento deductivo es un argumento donde la conclusión se infiere necesariamente de las premisas. Para derivar conclusiones se aplican reglas de inferencia, entre las cuales se incluyen:

- *Modus Ponendo Ponens*, mejor conocida como *Modus Ponens* (la forma en que se afirma afirmando). Si P implica Q y P es verdadera, entonces Q es verdadera. En Inteligencia Artificial también se le denomina *encadenamiento hacia adelante.*
- *Modus Tollendo Tollens,* mejor conocida como *Modus Tollens* (la forma en que se niega negando). El argumento tiene dos premisas; si P entonces Q; la segunda premisa es *no* Q; por lo que se puede concluir lógicamente que *no* P.

Para entender el razonamiento propuesto en el presente trabajo, es importante recordar la ley de silogismo o regla de la cadena. Esta forma de razonamiento, también llamada *silogismo hipotético* o *el principio de transitividad de la implicación*, consta de dos

premisas que son proposiciones condicionales, en donde el antecedente de la segunda es el consecuente de la primera. La conclusión es también una proposición condicional cuyo antecedente es el antecedente de la primera y el consecuente es el consecuente de la segunda.

2.2.6 Ontologías para recursos académico – científicos

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 (Peroni & Shotton, 2012), CiTO, una ontología para citaciones bibliográficas (Shotton, 2010), BIRO, la ontología de referencias bibliográficas (Shotton & Peroni, 2013).

Estas y cinco ontologías más: PRO, la ontología de roles de publicación; PSO, la ontología del estado de la publicación; PWO, la ontología del flujo de trabajo de la 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 referenciación semántica (SPAR, 2013).

Por su parte, la Ontología Bibliográfica BIBO provee los conceptos y propiedades principales para describir citaciones y referencias bibliográficas (Giasson & D'Arcus, 2009).

En cuanto a Dublin Core se refiere, la DCMI también ha elaborado una ontología para describir el conjunto de términos para identificar objetos digitales en la red (DCMI, 2012) y será descrita en mayor detalle más adelante.

Por otro lado se encuentra también FOAF (*friend of a friend)*, un proyecto devoto de vincular personas e información usando la Web (Brickley & Miller, 2014). El espacio de nombres FOAF http://xmlns.com/foaf/0.1/ es usado para representar información acerca de personas, tal como nombres, fechas de nacimiento y especialmente a la gente con la que se relacionan. Es particularmente útil para representar datos de redes sociales.

FOAF, como la propia Web, es un sistema de información vinculada. Está construido utilizando tecnología descentralizada de la Web Semántica, y se ha diseñado para permitir la integración de datos a través de una variedad de aplicaciones, sitios, servicios web y sistemas de software (Brickley & Miller, 2014).

Estas dos últimas ontologías DC y FOAF son usadas en el modelo OntoOAI a través de una técnica de fusión y se detallan a mayor profundidad en el siguiente capítulo.

2.3 Protocolo para Cosecha de Metadatos de la Iniciativa de Archivos Abiertos

El éxito de la Web Semántica depende críticamente de la existencia de una suficiente cantidad de contenido semántico relevante de alta calidad (McDowell & Cafarella, 2006). Los metadatos – datos sobre los datos - y la forma en que se describe y expone la información en la web son claves para elevar la comprensión de los contenidos y la interoperabilidad entre las aplicaciones.

Los metadatos pueden reunir documentos dispares pero relacionados para crear nuevas posibilidades de investigación. Sin embargo, también pueden frustrar a un investigador cuando no son tan precisos como podrían ser (Guédon, 2013).

La educación 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 (Cantillo Valero, Roura Redondo, & Sánchez Palacín, 2012). Cuando se habla de la comunicación entre las máquinas se hace referencia a acuerdos de interoperabilidad.

Tal es el caso del protocolo OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*) que con la Iniciativa de Archivos Abiertos, liberada en 1999, surge con la intención de convertir los archivos en interoperables. El OAI-PMH es un mecanismo de baja barrera para la interoperabilidad de repositorios (Lagoze & Van de Sompel, 2001), provee un marco de trabajo para la interoperabilidad independiente de la aplicación basado en la cosecha de metadatos.

Este protocolo junto con la adhesión a la iniciativa de acceso abierto, además favorece la visibilidad de los recursos expuestos en la red bajo el OAI-PMH, ya que permite que se distribuya en múltiples plataformas con lo que se incrementan los puntos de acceso potenciales para los usuarios.

La figura siguiente esquematiza cómo un portal que expone sus contenidos en la red en un punto geográfico particular puede incrementar los puntos de acceso a nivel mundial al implementar el protocolo OAI-PMH, ya que su información será colectada por otros portales y puesta en línea en otras plataformas, lo que conlleva a la replicación de los metadatos para poblar otros buscadores en la Web incrementando el uso potencial por usuarios de diversas latitudes.

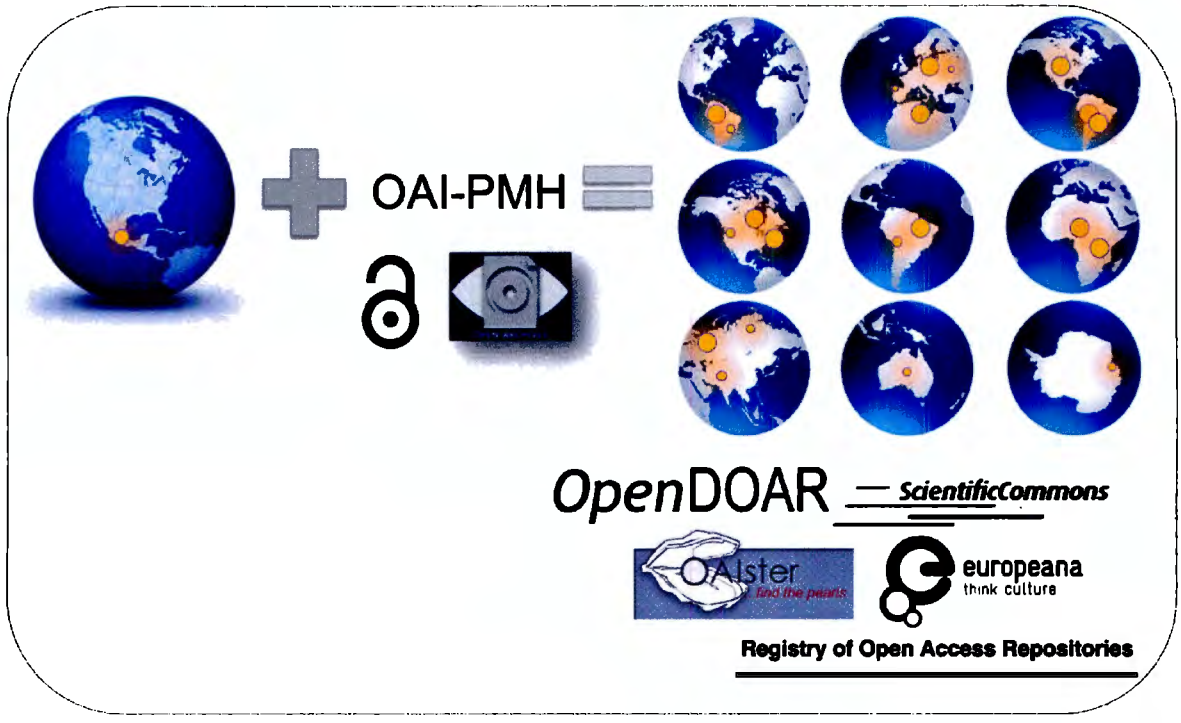

Figura 10. Visibilidad y Protocolo OAI-PMH

En este protocolo de intercambio de datos se especifican participantes y reglas para que interactúen. En el protocolo se contempran dos tipos de roles, los *proveedores de datos* y los *proveedores de servicio*:

1. Proveedores de datos: Son los encargados de poner a disposición de los proveedores de servicio los datos de identificación de cada uno de los archivos contenidos en el repositorio. Se identifican con una *BaseURL* que es una dirección URL donde se aloja el servidor y donde se concatenan los verbos. (Ejemplo: http://oai.redalyc.uaemex.mx/redalyc/oai?verb=...)
2. Proveedores de servicio: Son los encargados de realizar los procesos de cosecha o recolección de metadatos ofrecidos por los proveedores de datos para integrarlos y formar una base sobre la cual construir servicios para los usuarios.

Al proceso de solicitar y recuperar metadatos de recursos dentro del protocolo se le conoce como *cosecha de metadatos* (*metadata harvesting*), y consiste en la recuperación de registros a través de un programa que realiza peticiones *HTTP* a un proveedor de datos identificado con una *URL Base* (figura 11).

Los procesos de cosecha de metadatos generalmente son ejecutados por lotes, de tal forma que mantengan la información al día. El protocolo está concebido para recoger los metadatos y almacenarlos de manera local para que las aplicaciones de consulta de los proveedores de servicio hagan uso de ellos y los tiempos de respuesta para un usuario final no dependan de la disponibilidad de los proveedores de datos.

Si bien esta característica permite realizar aplicaciones eficientes para el usuario, demanda también altos niveles de almacenamiento y tareas exhaustivas de recolección de datos. Otros protocolos realizan la recolección de metadatos en tiempo real como es el caso del Z39.50 (ANSI/NISO, 2003) un protocolo que ya existía cuando apareció el OAI-PMH.

Z39.50 surgió como un estándar que permite la búsqueda federada a varios servidores de manera paralela. Sin embargo, ha 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 (Clifford A., 2001).

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

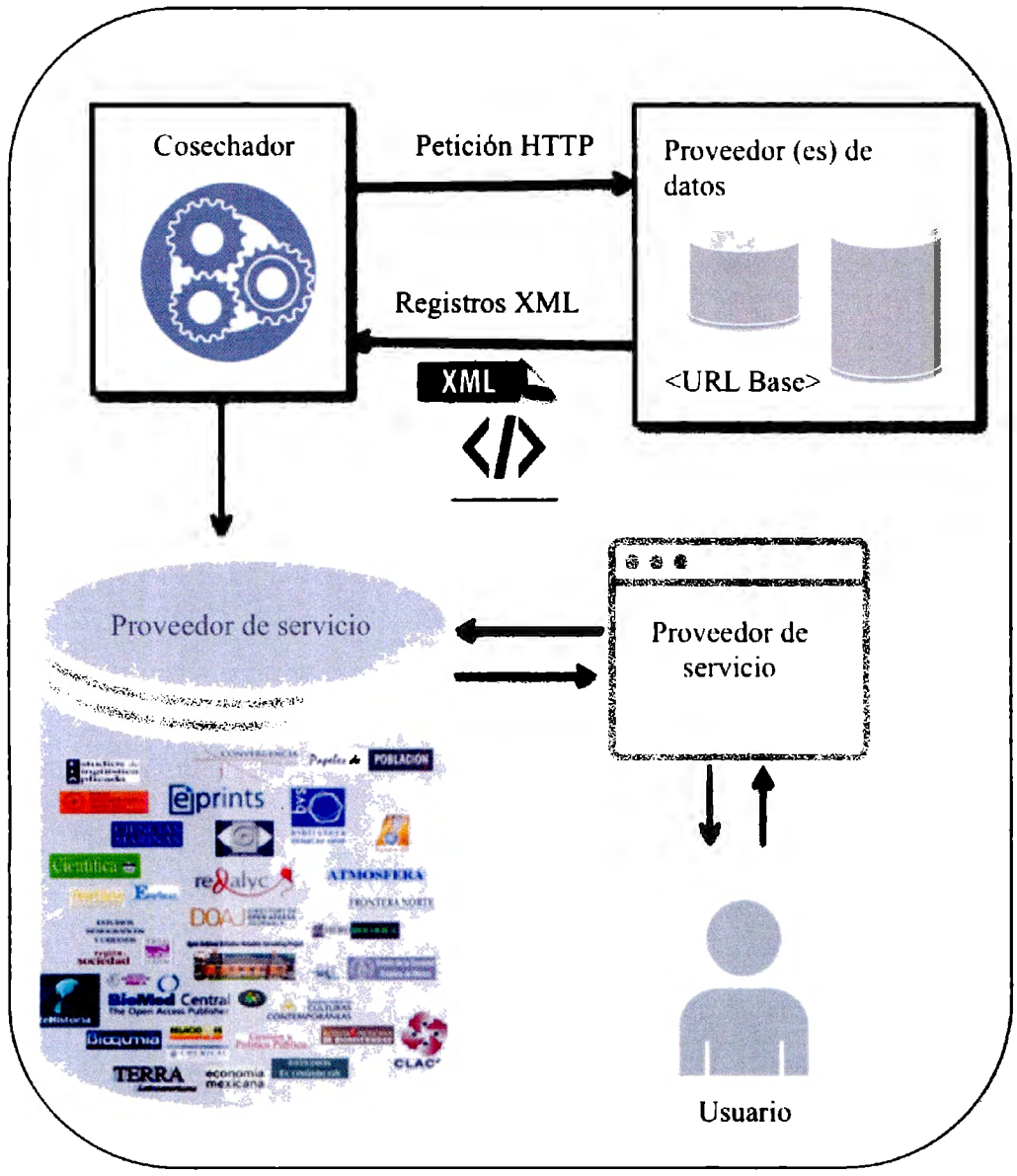

Figura 11. Proceso de cosecha OAI-PMH

Los proveedores de datos y de servicio dentro del protocolo OAI-PMH son capaces de entender y responder peticiones que incluyen alguno de los siguientes verbos:

- *Identify*: Verbo usado para conocer información acerca del repositorio como nombre, URL base, contacto técnico, entre otros.
- *GetRecord*: Permite recuperar los metadatos de un registro de manera individual a través de un identificador.

- *ListRecords:* Es el verbo principal para realizar los procesos de cosecha, a través del cual se obtienen los metadatos de los registros de un proveedor de datos.
- *ListIdentifiers*: Es una forma abreviada del verbo *ListRecords* pero que únicamente muestra los identificadores de los registros colectados.
- *ListSets:* Lista la relación de grupos en los que se agregaron los registros del repositorio, con el objetivo de realizar consultas por *Set* o grupo.
- *ListMetadataFormats*: Se usa para obtener la lista de formatos que soporta el proveedor de datos.

Los metadatos a transmitir vía OAI-PMH deberán codificarse en el formato Dublin Core dentro de un archivo XML, los cuales generalmente incluyen varios registros Dublin Core dependiendo de la configuración de cada proveedor de datos y de los cosechadores. El apartado siguiente describe este formato.

2.3.1 Dublin Core

Desde su aparición en 1995, Dublin Core (conocido como DC) estuvo orientado a promover un estándar de propósito general, sencillo y descriptivo de los recursos web sobre cualquier materia. Originalmente facilitaría la estructuración básica sin demandar una gran experiencia en la aplicación de procesos técnicos como la catalogación y clasificación de información.

La Iniciativa de Metadatos Dublin Core (DCMI, por sus siglas en inglés) 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 alcance de los metadatos está actualmente caracterizado en 4 niveles de interoperabilidad (DCMI, 2005):

Nivel 1. Definiciones compartidas de términos

- La interoperabilidad entre aplicaciones está basada en compartir definiciones en lenguaje natural dentro de un ambiente de aplicación como intranets, sistemas de bibliotecas o federación de repositorios; los participantes acuerdan los términos a usar.
- La interoperabilidad fuera del ambiente de implementación, es decir, con "el resto del mundo", no es una prioridad.

Nivel 2. Interoperabilidad semántica formal

- La interoperabilidad en aplicaciones que usan metadatos está basada en el modelo formal compartido provisto por RDF, el cual es usado para soportar *Linked Data*[2].

Nivel 3. Interoperabilidad sintáctica de conjuntos de descripciones

- Las aplicaciones son compatibles con *Linked Data* y, adicionalmente, comparten una sintaxis abstracta para registros de metadatos validables - *description set*.

Nivel 4. Interoperabilidad de perfil de conjuntos de descripciones.

---

[2] *Linked Data* es una práctica recomendada para exponer, compartir y conectar piezas de datos, información y conocimiento en la web semántica usando URIs y RDF.

- Los registros intercambiados siguen, además, un conjunto común de restricciones, usan los mismos vocabularios y reflejan un modelo compartido.

Elemento Dublin Core	Definición
dc:contributor	Una entidad responsable de hacer contribuciones al recurso.
dc:coverage	El tema temporal o espacial del recurso.
dc:creator	Una entidad responsable de la creación del recurso.
dc:date	Un punto o periodo en el tiempo asociado con un evento en el ciclo de vida del recurso.
dc:description	La explicación a detalle sobre el recurso.
dc:format	El formato de archivo, medio físico o dimensiones del recurso.
dc:identifier	Una referencia no ambigua al recurso dentro del contexto dado.
dc:language	Idioma del recurso.
dc:publisher	Una entidad responsable de poner disponible el recurso.
dc:relation	Un recurso relacionado.
dc:rights	Información acerca de los derechos del recurso.
dc:source	Un recurso relacionado desde el cual el recurso descrito es derivado.
dc:subject	El tema del recurso.
dc:title	Un nombre dado al recurso.
dc:type	La naturaleza o género del recurso.

Tabla 3. Conjunto de elementos de metadatos Dublin Core, versión 1.1 (DCMI, Dublin Core Metadata Element Set, Version 1.1).

El protocolo OAI-PMH implementa el nivel 1 de interoperabilidad con un conjunto de metadatos de 15 elementos, mostrados en la tabla 2.

Así, el OAI-PMH es capaz de describir los registros con estos metadatos y exponerlos en la Web con un formato de salida XML validado, como el mostrado a continuación.

```
<schema targetNamespace="http://www.openarchives.org/OAI/2.0/oai_dc/"
        xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
        xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns="http://www.w3.org/2001/XMLSchema"
        elementFormDefault="qualified" attributeFormDefault="unqualified">

<annotation>
  <documentation>
      XML Schema 2002-03-18 by Pete Johnston.
      Adjusted for usage in the OAI-PMH.
      Schema imports the Dublin Core elements from the DCMI schema for unqualified
Dublin Core.
      2002-12-19 updated to use simpledc20021212.xsd (instead of
simpledc20020312.xsd)
  </documentation>
</annotation>

<import namespace="http://purl.org/dc/elements/1.1/"
        schemaLocation="http://dublincore.org/schemas/xmls/simpledc20021212.xsd"/>

<element name="dc" type="oai_dc:oai_dcType"/>

<complexType name="oai_dcType">
  <choice minOccurs="0" maxOccurs="unbounded">
    <element ref="dc:title"/>
    <element ref="dc:creator"/>
    <element ref="dc:subject"/>
    <element ref="dc:description"/>
    <element ref="dc:publisher"/>
    <element ref="dc:contributor"/>
    <element ref="dc:date"/>
    <element ref="dc:type"/>
    <element ref="dc:format"/>
    <element ref="dc:identifier"/>
    <element ref="dc:source"/>
    <element ref="dc:language"/>
    <element ref="dc:relation"/>
    <element ref="dc:coverage"/>
    <element ref="dc:rights"/>
  </choice>
</complexType>

</schema>
```

Figura 12. Esquema XML para metadatos Dublin Core (OpenArchives, 2002).

La especificación del protocolo OAI-PMH indica la serialización en XML, por ello, es necesaria la adaptación de estas salidas a la codificación sintáctica de RDF/XML o algún otro formato de serialización de RDF, como factor determinante para la descripción de recursos como esquema genérico para el intercambio de metadatos entre sistemas heterogéneos y distribuidos, estándares de descripción de contenido (ontologías, mapas temáticos, tesauros, entre otros) y toda una variedad de protocolos y normas para el intercambio de información en la Web Semántica.

2.4 Sensibilidad al contexto y descubrimiento selectivo

El término Diseminación Selectiva de Información (*DSI* o *SDI* por las siglás en inglés de *Selective Dissemination of Information*) fue usado por vez primera en 1958 por Hans Peter Luhn para definir aquel servicio dedicado a la canalización de nuevos elementos de información, a partir de cualquier fuente, hasta aquellos puntos donde la probabilidad de su utilidad en el trabajo diario o interés sea elevada (Luhn, 1958).

La Diseminación Selectiva de Información es un servicio que puede ser considerado parte de los servicios de alertas de actualización o CAS (*Current Awareness Services)*, mismos que operan notificando al usuario final de manera periódica y automáticamente sobre nuevos materiales que arriben a un almacén de información. CAS es un término que se utiliza en el campo de la bibliotecología y ciencias de la información.

El desarrollo de métodos de actualización informativa se remonta al siglo XVII, con el surgimiento de las primeras sociedades académicas y sus publicaciones. En 1830, por ejemplo, fueron creadas las revistas de resúmenes para notificar a las personas sobre recientes publicaciones, seguidas de la elaboración de índices acumulativos como suministro de medios de localización de literatura existente (Kemp, 1979).

En un servicio CAS la información enviada no distingue preferencias particulares, una vez recibida, el usuario realiza un procedimiento manual de discriminación para filtrar la que cumpla con sus intereses. Sin embargo, un servicio DSI si es totalmente orientado al usuario (Dawra, 2004). Este servicio puede adecuarse a un usuario específico basándose en un perfil de recuperación de información individual, o bien, a un conjunto de personas que tienen intereses comunes, requiriéndose para tal caso la definición de un perfil colectivo (Guinchat & Menou, 1992).

La diseminación selectiva de información puede ser vista como una estrategia proactiva para la consulta de información, ya que, información de interés es conducida directamente al usuario, promoviendo un beneficio bidireccional (Becerril García, 2010):

- hacia el usuario, por hacer más fácil y eficiente el proceso de localización de información;
- hacia el acceso a la información, al aumentar la difusión y por ende la cantidad de lectores potenciales.

Muchos son los sitios y portales en la Web que ofrecen este tipo de servicio, y es posible clasificarlos básicamente bajo dos criterios:

1. Según la fuente de información: publicaciones, artículos, tablas de contenido (Table of Content, conocidas como ToC), citas, noticias, clima, movimientos en acciones de la bolsa de valores, resultados de motores de búsqueda, cambios en páginas web, entradas de blogs, actividades dentro de la industria (servicios clip), entre los más populares.
2. Según la estrategia de recuperación de información: por temáticas o áreas de conocimiento, por autores específicos, por selección de región geográfica, por

publicación, por selección de páginas o sitios web, a través de una consulta a un motor de búsqueda o con una estrategia compuesta por una combinación de los criterios anteriores.

Los servicios DSI han evolucionado a lo largo del tiempo, desde que los usuarios de una biblioteca entregaban sus preferencias por escrito para recibir materiales de interés, hasta servicios automatizados por computadora que permiten crear un perfil y utilizando *tecnología push*[3] son capaces de enviar los resultados al usuario utilizando el correo electrónico o dispositivos móviles como medios de notificación.

Sin embargo, dados los avances en el campo de la inferencia de información cabe preguntarse ¿podrían estos servicios incluir mecanismos de inferencia para "descubrir" recursos de información que resulten interesantes para un usuario?

Las computadoras se están convirtiendo en el vehículo para una creciente variedad de actividades cotidianas. Al mismo tiempo, un número creciente de usuarios no entrenados están interactuando con las computadoras, y este número seguirá aumentando a medida que las tecnologías tales como computadoras portátiles y de televisión interactiva se hacen populares. Desafortunadamente, estos avances tecnológicos no van de la mano con un cambio en la forma en que las personas interactúan con las computadoras (Maes, 1994).

Una estrategia para mejorar la interacción humano-computadora radica en considerar el ambiente del usuario, diseñar aplicaciones que sean sensibles al contexto.

---

[3] Käpylä, Niemi, y Lehtola (1998) definen la tecnología *push* como programas que reciben automáticamente información de la Web con un itinerario previamente arreglado o cuando ciertos eventos ocurran.

La computación sensible al contexto es un paradigma de la computación móvil en el cual las aplicaciones pueden descubrir y tomar ventaja de información contextual tal como localización del usuario, tiempo del día, usuarios y dispositivos vecinos, y actividad del usuario (Musumba & Nyongesa, 2013).

La definición de un perfil es una de muchas formas para capturar el contexto de un usuario. Perfilar al usuario es una parte importante de la Web Semántica ya que integra al usuario en el concepto de Web de datos con una semántica legible para las máquinas (Mladenić, Grobelnik, Fortuna, & Grčar, 2009).

Una vez definido un perfil de usuario, un agente personal[4] puede utilizar, aprender y deducir nuevos conocimientos sobre la base de este perfil para personalizar la asistencia que se ofrece a un usuario (Eyharabide & Amandi, 2012).

Los perfiles de usuario se pueden obtener de diversas maneras:

1. Preguntando al usuario, es la forma más sencilla de obtener esa información. Sin embargo, es tedioso para los usuarios dar retroalimentación explícita, ya que lleva mucho tiempo e implica esfuerzo. Aún más, los usuarios pueden no querer dar esa

---

[4] Los agentes son programas informáticos que emplean técnicas de inteligencia artificial para proporcionar asistencia activa a un usuario. Los agentes cambian radicalmente la experiencia del usuario, a través de la metáfora de que un agente puede actuar como un asistente personal. El agente adquiere su competencia aprendiendo del usuario, así como de los agentes que asisten a otros usuarios. Muchos agentes han sido construidos utilizando esta técnica, incluidos los agentes que prestan asistencia personalizada con la programación de reuniones, gestión del correo electrónico, filtrado de noticias y selección de entretenimiento (Maes, 1994).

información, no saber lo que es el contexto o cómo expresarlo con precisión (Schiaffino & Amandi, 2006).

2. A través de la inferencia del contexto relevante, mediante la observación del comportamiento del usuario, desde atributos de contexto de alto nivel hasta los de bajo nivel obtenidos a partir de la interacción del usuario con una aplicación. Este método evita preguntar al usuario, el objetivo es aprender qué contexto es relevante para el usuario durante el tiempo de ejecución sobre la base del comportamiento de observación del usuario.
3. Otro acercamiento es a través de técnicas de minería de datos enriquecidas con semántica, reunidas dentro de una ontología para producir perfiles de usuarios como la propuesta por Eyharabide y Amandi (2012), cuyo objetivo es personalizar los perfiles de usuario con información contextual con el fin de mejorar el rendimiento de un agente mediante la combinación de la minería de datos y ontologías. Se enriquece el perfil de usuario semánticamente con una representación ontológica de contexto para adaptar dinámicamente la selección de atributos de acuerdo con las preferencias del usuario.

Un perfil de usuario contiene típicamente dos tipos de información de usuario. Información independiente de la aplicación, que es principalmente información personal sobre el usuario y por otra parte, información dependiente de la aplicación, que incluye los intereses del usuario, preferencias, metas, hábitos de trabajo, patrones de comportamiento, conocimientos, necesidades, prioridades y compromisos relativos a un dominio particular (Schiaffino & Amandi, 2006).

Hay varias representaciones para perfiles de usuario, técnicas y enfoques para modelarlo. Por ejemplo, la propuesta de Grčar, Mladenič y Grobelnik (2006) se centra en la

personalización de la Web, entendida como el proceso de personalización de los sitios Web de acuerdo con el perfil del usuario específico para hacer más eficiente la navegación por Internet.

En ese mismo sentido, pero con el uso de ontologías para el modelado del usuario se encuentran propuestas como la presentada en Sieg, Mobasher y Burke (2007) que muestra una aproximación a la búsqueda personalizada que implica la construcción de modelos de contexto de usuario como perfiles ontológicos, asignando puntuaciones de interés implícitamente derivadas a los conceptos existentes en una ontología de dominio. También Golemati, Katifori, Vassilakis, Lepouras y Halatsis (2007) crean una ontología general, extensible y comprehensiva de perfil de usuario.

El modelado del usuario incluso se ha llevado a las redes sociales, Stan, Egyed-Zsigmond, Joly y Maret (2008) proponen un modelo de perfil de usuario basado en ontologías, que permite a los usuarios tener una red social sensible a las situaciones, mediante el control de la forma en que son accesibles para ciertas categorías de personas en una situación dada. Y para el área de marketing, como en Hoppe (2013) apoyándose del uso de técnicas de *Big Data.*

Novedosas técnicas, como el etiquetado social, también han sido implementadas para enriquecer las ontologías que representan los perfiles de usuario (Cantador, Szomszor, & Alani, 2008).

En lo que al ámbito académico o educativo se refiere, también se han propuesto diversos enfoques para modelar un perfil de usuario como el presentado por Panagiotopoulos, Kalou, Pierrakeas y Kameas (2012) que contiene información acerca de un estudiante tal como información demográfica, estilo de aprendizaje y desempeño académico para sistemas de tutoría inteligente para aprendizaje a distancia. Este enfoque es

usado para modelar al usuario en OntoOAI y será descrito a más detalle en los siguientes capítulos.

2.5 Desarrollos de sofware para descubrimiento de conocimiento y OAI-PMH

En la tabla siguiente se enlistan diversos desarrollos relacionados en algún punto con el modelo OntoOAI, los cuales han sido comparados de acuerdo a un conjunto de criterios clave para enfatizar sus principales características y visualizar el área de oportunidad por la cual surge la idea de desarrollar OntoOAI.

Cabe destacar que se han incluido los productos y servicios de descubrimiento de bibliotecas, que como ya se ha mencionado anteriormente difieren sustancialmente del enfoque de descubrimiento tratado en este trabajo, sin embargo, están en la comparativa con el fin de hacer evidente las diferencias.

Hay cuatro criterios importantes en esta comparación para evaluar cada desarrollo respecto a los objetivos de OntoOAI:

1. ¿Usa tecnologías de la Web Semántica? (columna WS de la tabla)
2. ¿Es especializado en la indexación de recursos estructurados con OAI-PMH o considera los metadatos de este protocolo para obtener algún valor adicional? (columna OAI-PMH)
3. ¿Permite la inferencia de recursos de información no solicitada, no explícita, previamente desconocida y potencialmente útil para un usuario, es decir, permite el descubrimiento de conocimiento? (KD)
4. ¿Considera la sensibilidad al contexto (SC) del usuario, es decir contempla un perfil de usuario?

Nombre	Descripción	Estado	WS	OAI-PMH	KD	SC
Sindice "El índice de la Web Semántica", Digital Enterprise Research Institute **(Franzon, 2014)**	Motor de búsqueda semántico de recursos marcados con RDF, microformatos, microdatos, RDFa	Inactivo, paso a ser parte de una solución comercial	✓	✗	✓	✗
SIREn (Tummarello & Delbru, 2011)	Motor de recuperación de información semántica	Activo, comercial	✓	✗	✓	✗
Freebase Metaweb, 2007 **(Google, 2015)**	Base de conocimiento con una plataforma y una API para accederla	Comprado por Google, inactivo	✓	✗	✓	✗
Search BOSS (Yahoo Inc., 2009)	Un servicio orientado al desarrollo de motores de búsqueda para sitios web sobre RDFa y microformatos	Inactivo	✓	✗	✗	✗
Knowledge Graph (Google, 2012)	Base de conocimiento usada por Google para mejorar su motor de búsqueda	Activo	✓	✗	✓	✗
Primo (ExLibris, 2006)	Permite a las bibliotecas acceder a sus colecciones	Activo, comercial	✗	✓	✗	✗
Discovery Service (OCLC)	Plataforma de búsqueda de recursos	Activo, comercial	✗	✓	✗	✗

	de bibliotecas y bases de datos externas					
BLUEcloud PAC (Sirsi Corporation)	Plataforma de servicios de bibliotecas	Activo, comercial	✗	✓	✗	✗
BiblioCore (Bibliocommons)	Nueva generación de OPAC para bibliotecas	Activo, comercial	✗	✗	✗	✗
AquaBrowser (ProQuest, 2009)	Permite a las bibliotecas acceder a sus colecciones	Activo, comercial	✗	✗	✗	✗
Summon Service (ProQuest, 2015)	Buscador para bibliotecas y otros recursos de una institución como repositorios	Activo, comercial	✗	✓	✗	✗
Encore (Innovative Interfaces)	Buscador de recursos de una biblioteca	Activo, comercial	✗	✓	✗	✗
Discovery Services (EBSCO)	Plataforma de búsqueda de recursos de bibliotecas y bases de datos externas	Activo, comercial	✗	✓	✗	✗
Blacklight (University of Virginia)	Búsqueda y recuperación de información para bibliotecas	Activo, código abierto	✗	✓	✗	✗
VuFind (Villanova University)	Búsqueda y recuperación de información para	Activo, código abierto	✗	✓	✗	✗

	bibliotecas					
eXtensible Catalog (River Campus Libraries, 2006)	Software de nueva generación para bibliotecas	Activo, código abierto	✗	✓	✗	✗
Franklin (University of Pennsylvania)	Provee acceso a múltiples colecciones	Activo, código abierto	✗	✓	✗	✗

Tabla 4. Comparación de algunos desarrollos de software en el campo de descubrimiento y OAI-PMH.

En total se revisaron 17 proyectos, de los cuales 12 corresponden a los llamados servicios de descubrimiento para bibliotecas. De estos últimos el 83% indexan contenido estructurado con OAI-PMH. Sin embargo, ninguno de los 12 utilizan tecnologías de la Web Semántica y no realizan procesos de descubrimiento de conocimiento con inferencia o razonamiento, de igual forma ninguno es sensible al contexto del usuario. Cabe mencionar que muchos de esos permiten la personalización de servicios pero para las bibliotecas no para los usuarios finales.

Respecto a los desarrollos que si hacen uso de tecnologías semánticas, que son cinco, ninguno de ellos indexa particularmente los recursos OAI-PMH, es decir, generalmente funcionan con *crawlers* para obtener la información de los sitios web, es por ello que los recursos que están estructurados con OAI-PMH no son aprovechados para obtener una mayor semántica, es decir, son tratados de manera similar a los datos no estructurados al momento de la indexación.

Asimismo, ninguno de estos cinco considera alguna información del contexto del usuario, pero cuatro de ellos realizan procesos de inferencia sobre sus bases de conocimiento.

El modelo OntoOAI surge justamente con la idea de permitir el descubrimiento de conocimiento a través de procesos de inferencia sobre recursos estructurados con OAI-PMH, considerando un perfil de usuario para descubrir información potencialmente útil para él.

En el siguiente capítulo se describe OntoOAI en detalle, la metodología para su desarrollo, arquitectura, componentes de software y las tecnologías utilizadas.

Capítulo 3 Modelo OntoOAI para el descubrimiento selectivo de conocimiento

La aplicación de tecnologías de la Web Semántica sobre datos estructurados con el protocolo OAI-PMH, para descubrir conocimiento de forma selectiva para un usuario ha dado lugar al modelo OntoOAI. La metodología, herramientas, desarrollos de software, componentes y su implementación se describen en este capítulo.

3.1 Descripción general del modelo

La secuencia de pasos (figura 13) que representa la metodología implementada para desarrollar el modelo OntoOAI muestra cada uno de los procesos llevados a cabo y que son:

1. La cosecha de metadatos de recursos que se encuentran disponibles a través de repositorios que implementan el protocolo OAI-PMH.
2. La transformación de los resultados obtenidos en XML a RDF.
3. El enriquecimiento de los enunciados (obtenidos en Dublin Core) con información adicional de autores (FOAF)
4. Validación del conocimiento obtenido.
5. Almacenamiento de tripletas
6. Recuperación de información basado en inferencia, que proveerá el servicio al usuario final. Se aprecia que la información requerida de entrada desde una interfaz de usuario corresponde a la consulta así como la información contextual del usuario. También se resaltan los componentes de software desarrollados:
   - OAI-PMH *harvester*
   - XML_DC2RDF

- DCFOAF
- DCFOAF_merged.owl

El contexto del usuario es una entrada del proceso de recuperación de información para realizar el descubrimiento selectivo.

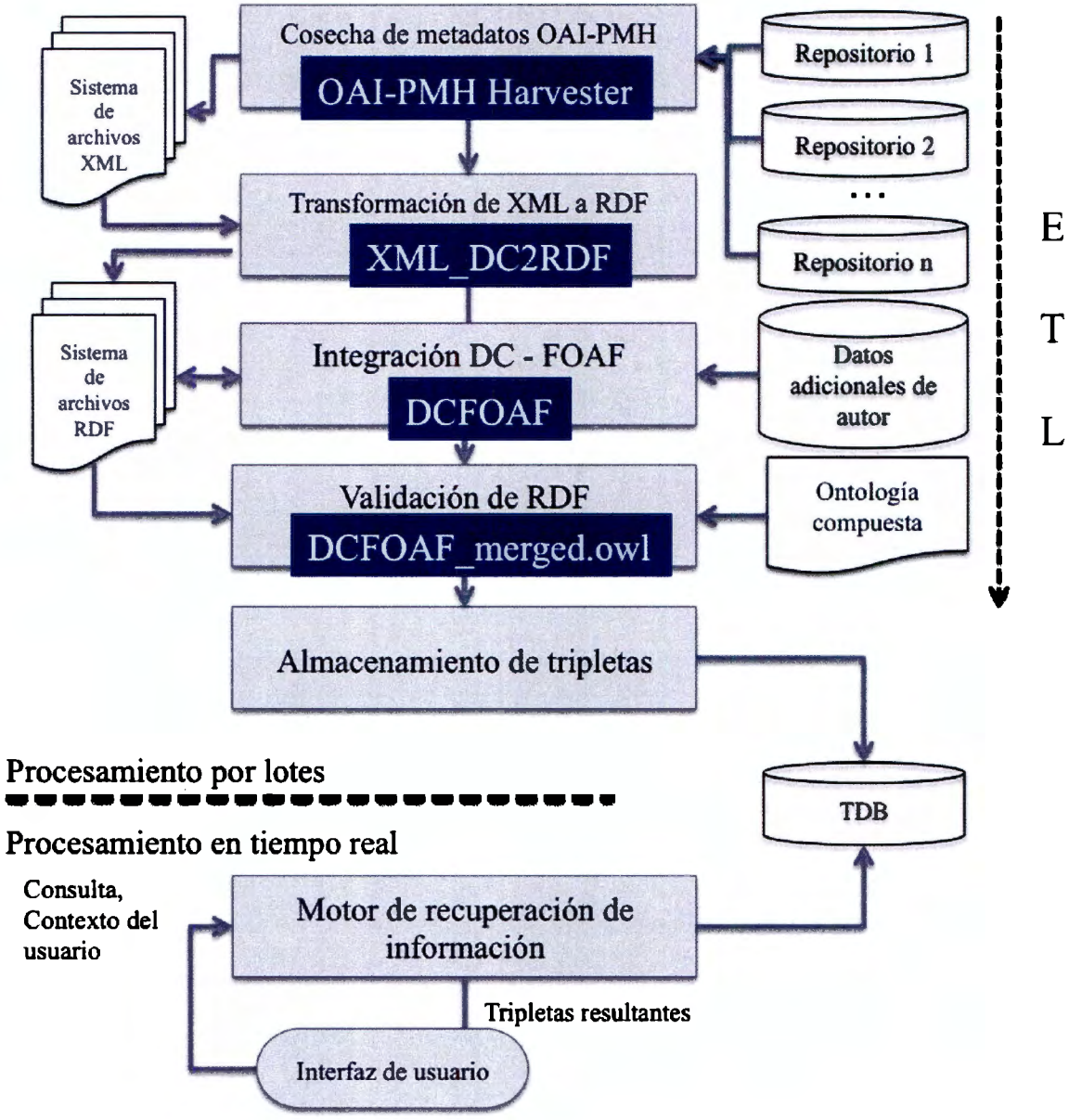

Figura 13. Metodología de OntoOAI

Este enfoque se enmarca en la lógica del proceso ETL (Extracción, Transformación y Carga), ya que la información es colectada y organizada, para posteriormente ser transformada a RDF, enriquecida y validada, con el uso de ontologías, y se carga en forma de tripletas a un almacén de datos.

Esta metodología está diseñada para agregar desde 1 a *n* repositorios, siempre y cuando sigan la especificación del protocolo OAI-PMH.

En los siguientes incisos se describe a detalle cada proceso, componente y tecnologías utilizadas.

3.2 Fuentes de información

Son tres los directorios más importantes de repositorios que implementan el protocolo OAI-PMH, en los cuales se pueden localizar sitios de la Web que cumplen esta especificación y son, por lo tanto, candidatos para aportar información en OntoOAI.

1. OpenDOAR[5], es un directorio de repositorios académicos de acceso abierto. Cada repositorio registrado en OpenDOAR ha sido revisado para controlar la calidad de los contenidos que cosecha. También funciona como proveedor de servicio (agregador de contenidos).
2. ROAR[6] - el Registro de Repositorios de Acceso Abierto - cuyo objetivo es promover el desarrollo del acceso abierto al proporcionar información sobre el

[5] http://opendoar.org/

[6] http://roar.eprints.org/

crecimiento y el estado de los repositorios en todo el mundo. De la misma forma, también lleva registro cuantitativo del crecimiento en contenidos.

3. Lista de proveedores de datos del OpenArchives.org[7], este es el sitio oficial del protocolo OAI-PMH. Para que un repositorio logre su registro en *OpenArchives* debe pasar por una validación de la implementación del estándar.

Así también, resulta relevante analizar los "proveedores de servicio" que a su vez exponen los metadatos vía el protocolo, ya que esto simplificaría de manera significativa los procesos de cosecha, dado que un proveedor de servicio ya cosecha un conjunto de repositorios.

Una vez seleccionados los sitios que cumplen con el protocolo, dentro de estos directorios, es importante localizar la URL Base de cada uno, ya que ésta será insumo del proceso de cosecha de metadatos.

Por otra parte, el modelo contempla el enriquecimiento de los metadatos con información adicional de los autores, esta información debe cumplir con la especificación FOAF, puede provenir de redes sociales de investigadores u otras fuentes de información como DBpedia[8].

DBpedia es un esfuerzo colaborativo para extraer información estructurada de Wikipedia y Wikidata y ponerla disponible en la Web. DBpedia permite consultar más sofisticadamente Wikipedia, y vincular diferentes conjuntos de datos a los obtenidos de Wikipedia. Esta iniciativa hace disponibles estos datos a través de bases de tripletas en

---

[7] http://www.openarchives.org/Register/BrowseSites
[8] http://wiki.dbpedia.org/

diferentes formatos. El conjunto de datos DBpedia utiliza una gran ontología de varios dominios que se ha derivado de Wikipedia.

La versión en inglés del 2014 del conjunto de datos de DBpedia describe 4.58 millones de "cosas" con 583 millones de "hechos". Además, proporciona versiones localizadas de DBpedia en 125 idiomas. Todas estas versiones juntas describen 38.3 millones de cosas, de las cuales 23.8 millones se traslapan con los conceptos de DBpedia en inglés. El conjunto completo de datos de DBpedia cuenta con 38 millones de resúmenes en 125 idiomas diferentes, 25.2 millones de enlaces a imágenes y 29.8 millones de enlaces a páginas web externas.

En total, la versión liberada de DBpedia 2014 consta de 3 mil millones de piezas de información (tripletas RDF) de las cuales 583 millones fueron extraídas de la edición en inglés de Wikipedia, 2.46 mil millones fueron extraídas de ediciones en otros idiomas, y alrededor de 50 millones son enlaces a bases de datos externas (DBpedia, 2016).

Todas las IRIs/URIs en los conjuntos de datos canónicos de DBpedia usan el espacio de nombres genérico http://dbpedia.org/resource/. Estos conjuntos están disponibles para su descarga en archivos de N-Triples (.nt o .nq) o en Turtle (.ttl). Dentro de estos conjuntos canónicos se encuentra *PersonData* disponible de las versiones en inglés y en alemán de Wikipedia.

Dentro de la base de *PersonData* (figura 14) se describe un recurso equivalente a una persona, por ejemplo, nombre, apellidos, definición del elemento en cuestión para describir que es una persona, descripción, fecha y lugar de nacimiento, así como fecha de fallecimiento, como sigue:

- `<http://xmlns.com/foaf/0.1/name>`
- `<http://xmlns.com/foaf/0.1/surname>`
- `<http://xmlns.com/foaf/0.1/givenName>`

- <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
  <http://xmlns.com/foaf/0.1/Person>
- <http://purl.org/dc/elements/1.1/description>
- <http://dbpedia.org/ontology/birthDate>
  <http://www.w3.org/2001/XMLSchema#date> .
- <http://dbpedia.org/ontology/birthPlace>
- <http://dbpedia.org/ontology/deathDate>
  <http://www.w3.org/2001/XMLSchema#date> .

```
# started 2015-04-29T14:11:22Z
<http://dbpedia.org/resource/Abraham_Lincoln> <http://xmlns.com/foaf/0.1/name> "Abraham Lincoln"@en .
<http://dbpedia.org/resource/Abraham_Lincoln> <http://xmlns.com/foaf/0.1/surname> "Lincoln"@en .
<http://dbpedia.org/resource/Abraham_Lincoln> <http://xmlns.com/foaf/0.1/givenName> "Abraham"@en .
<http://dbpedia.org/resource/Abraham_Lincoln> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://xmlns.com/foaf/0.1/Person> .
<http://dbpedia.org/resource/Abraham_Lincoln> <http://purl.org/dc/elements/1.1/description> "16th President of the United States"@en
.
<http://dbpedia.org/resource/Abraham_Lincoln> <http://dbpedia.org/ontology/birthDate> "1809-02-
12"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://dbpedia.org/resource/Abraham_Lincoln> <http://dbpedia.org/ontology/birthPlace>
<http://dbpedia.org/resource/Hardin_County,_Kentucky> .
<http://dbpedia.org/resource/Abraham_Lincoln> <http://dbpedia.org/ontology/birthPlace> <http://dbpedia.org/resource/Kentucky> .
<http://dbpedia.org/resource/Abraham_Lincoln> <http://dbpedia.org/ontology/deathDate> "1865-04-
15"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://dbpedia.org/resource/Aristotle> <http://xmlns.com/foaf/0.1/name> "Aristotle"@en .
<http://dbpedia.org/resource/Aristotle> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> .
<http://dbpedia.org/resource/Aristotle> <http://purl.org/dc/elements/1.1/description> "Greek philosopher"@en .
<http://dbpedia.org/resource/Aristotle> <http://dbpedia.org/ontology/birthDate> "-
0384"^^<http://www.w3.org/2001/XMLSchema#gYear> .
<http://dbpedia.org/resource/Aristotle> <http://dbpedia.org/ontology/birthPlace> <http://dbpedia.org/resource/Stagira_(ancient_city)> .
<http://dbpedia.org/resource/Aristotle> <http://dbpedia.org/ontology/deathDate> "-
0322"^^<http://www.w3.org/2001/XMLSchema#gYear> .
<http://dbpedia.org/resource/Aristotle> <http://dbpedia.org/ontology/deathPlace> <http://dbpedia.org/resource/Chalcis> .
<http://dbpedia.org/resource/Alain_Connes> <http://xmlns.com/foaf/0.1/name> "Alain Connes"@en .
<http://dbpedia.org/resource/Alain_Connes> <http://xmlns.com/foaf/0.1/surname> "Connes"@en .
<http://dbpedia.org/resource/Alain_Connes> <http://xmlns.com/foaf/0.1/givenName> "Alain"@en .
<http://dbpedia.org/resource/Alain_Connes> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> .
<http://dbpedia.org/resource/Alain_Connes> <http://purl.org/dc/elements/1.1/description> "Mathematician"@en .
<http://dbpedia.org/resource/Alain_Connes> <http://dbpedia.org/ontology/birthDate> "1947-04-
01"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://dbpedia.org/resource/Alain_Connes> <http://dbpedia.org/ontology/birthPlace> <http://dbpedia.org/resource/Draguignan> .
<http://dbpedia.org/resource/Allan_Dwan> <http://xmlns.com/foaf/0.1/name> "Allan Dwan"@en .
<http://dbpedia.org/resource/Allan_Dwan> <http://xmlns.com/foaf/0.1/surname> "Dwan"@en .
<http://dbpedia.org/resource/Allan_Dwan> <http://xmlns.com/foaf/0.1/givenName> "Allan"@en .
<http://dbpedia.org/resource/Allan_Dwan> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> .
<http://dbpedia.org/resource/Allan_Dwan> <http://purl.org/dc/elements/1.1/description> "Film director, film producer,
screenwriter"@en .
<http://dbpedia.org/resource/Allan_Dwan> <http://dbpedia.org/ontology/birthDate> "1885-04-
03"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://dbpedia.org/resource/Allan_Dwan> <http://dbpedia.org/ontology/birthPlace> <http://dbpedia.org/resource/Toronto> .
<http://dbpedia.org/resource/Allan_Dwan> <http://dbpedia.org/ontology/deathDate> "1981-12-
28"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://dbpedia.org/resource/Ayn_Rand> <http://xmlns.com/foaf/0.1/name> "Ayn Rand"@en .
<http://dbpedia.org/resource/Ayn_Rand> <http://xmlns.com/foaf/0.1/surname> "Rand"@en .
<http://dbpedia.org/resource/Ayn_Rand> <http://xmlns.com/foaf/0.1/givenName> "Ayn"@en
<http://dbpedia.org/resource/Ayn_Rand> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> .
<http://dbpedia.org/resource/Ayn_Rand> <http://purl.org/dc/elements/1.1/description> "Novelist, philosopher, playwright,
screenwriter"@en .
<http://dbpedia.org/resource/Ayn_Rand> <http://dbpedia.org/ontology/birthDate> "1905-02-
02"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://dbpedia.org/resource/Ayn_Rand> <http://dbpedia.org/ontology/birthPlace> <http://dbpedia.org/resource/Saint_Petersburg> .
<http://dbpedia.org/resource/Ayn_Rand> <http://dbpedia.org/ontology/birthPlace> <http://dbpedia.org/resource/Russian_Empire> .
<http://dbpedia.org/resource/Ayn_Rand> <http://dbpedia.org/ontology/deathDate> "1982-03-
06"^^<http://www.w3.org/2001/XMLSchema#date> .
<http://dbpedia.org/resource/Ayn_Rand> <http://dbpedia.org/ontology/deathPlace> <http://dbpedia.org/resource/New_York_City> .
<http://dbpedia.org/resource/Ayn_Rand> <http://dbpedia.org/ontology/deathPlace> <http://dbpedia.org/resource/United_States> .
<http://dbpedia.org/resource/Andre_Agassi> <http://xmlns.com/foaf/0.1/name> "Andre Agassi"@en .
<http://dbpedia.org/resource/Andre_Agassi> <http://xmlns.com/foaf/0.1/surname> "Agassi"@en .
<http://dbpedia.org/resource/Andre_Agassi> <http://xmlns.com/foaf/0.1/givenName> "Andre"@en .
```

```
<http://dbpedia.org/resource/Andre_Agassi> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> .
<http://dbpedia.org/resource/Andre_Agassi> <http://purl.org/dc/elements/1.1/description> "American tennis player"@en .
<http://dbpedia.org/resource/Andre_Agassi> <http://dbpedia.org/ontology/birthDate> "1970-04-
29"<http://www.w3.org/2001/XMLSchema#date> .
<http://dbpedia.org/resource/Andre_Agassi> <http://dbpedia.org/ontology/birthPlace> <http://dbpedia.org/resource/Las_Vegas_Valley>
.
<http://dbpedia.org/resource/Andre_Agassi> <http://dbpedia.org/ontology/birthPlace> <http://dbpedia.org/resource/Nevada> .
<http://dbpedia.org/resource/Aldous_Huxley> <http://xmlns.com/foaf/0.1/name> "Aldous Leonard Huxley"@en .
<http://dbpedia.org/resource/Aldous_Huxley> <http://xmlns.com/foaf/0.1/surname> "Huxley"@en .
<http://dbpedia.org/resource/Aldous_Huxley> <http://xmlns.com/foaf/0.1/givenName> "Aldous Leonard"@en .
<http://dbpedia.org/resource/Aldous_Huxley> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://xmlns.com/foaf/0.1/Person> .
<http://dbpedia.org/resource/Aldous_Huxley> <http://purl.org/dc/elements/1.1/description> "Writer; author"@en .
<http://dbpedia.org/resource/Aldous_Huxley> <http://dbpedia.org/ontology/birthDate> "1894-07-
<http://www.w3.org/2001/XMLSchema#date> .
<http://dbpedia.org/resource/Aldous_Huxley> <http://dbpedia.org/ontology/birthPlace> <http://dbpedia.org/resource/Surrey> .
<http://dbpedia.org/resource/Aldous_Huxley>

<http://dbpedia.org/ontology/deathDate> "1963-11-22"<http://www.w3.org/2001/XMLSchema#date> .
```

Figura 14. Extracto del *dataset PersonData* de DBpedia.

Cabe mencionar que el conjunto de datos *PersonData*, aporta en este momento solo datos de la versión en inglés de Wikipedia y no de otros idiomas.

3.3 Cosechador de metadatos OAI-PMH

El proceso de cosecha o recolección de metadatos consiste en el envío de peticiones HTTP, como se vio en el capítulo anterior, a los proveedores de datos que implementan el protocolo OAI-PMH para obtener uno o más registros correspondientes a sus recursos de información.

Se desarrolló una aplicación Java, llamada *OAI_PMH_Harvester*, para realizar éste proceso. *OAI_PMH_Havester* funciona recibiendo la respuesta de enviar al repositorio la petición:

[*URLBase*]?verb=ListRecords&metadataPrefix=oai_dc

Dicha respuesta está estructurada con el conjunto de metadatos de Dublin Core y se escribe en un archivo llamado (*ListRecords.xml*) que es manejado como una estructura de árbol utilizando la librería *Xerces Java Parser*. Se lleva una secuencia numérica de archivos del tipo: *ListRecords1.xml, ListRecords3.xml, ListRecords3.xml* ... para guardar cada lote

de metadatos que envían los proveedores de datos, hay algunos que envían conjuntos de 20 otros de 100, esa es una configuración particular de cada repositorio, por lo que el cosechador está preparado para recibir cantidades variables, utilizando el parámetro *resumptionToken.*

El cosechador también es capaz de realizar cosechas con parámetros opcionales del verbo *ListRecords* como *SET, FROM* Y *UNTIL,* para ejecutar consultas con filtros por conjunto o fechas.

Los archivos XML con DC resultantes (como el ejemplo mostrado en la figura 15) son almacenados en una directorio de donde posteriormente son tomados para su conversión a RDF.

```
<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>
```

Figura 15. Ejemplo XML DC

3.4 Convertidor XML-DC a RDF

XML_DC2RDF es el convertidor que toma la salida del cosechador representada como archivos XML con Dublin Core para convertirlos a RDF.

Esta transformación es necesaria ya que, como se confirma en el Informe Modelos de Metadatos para Contenidos Multimedia del proyecto OMediaDis (Plataforma abierta para la Gestión de Contenidos Distribución Multicanal), 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 (Ministerio para la Ciencia e Innovación de España, 2009).

Existen disponibles algunos desarrollos que permiten explotar datos estructurados dotándolos de características propias de aplicaciones semánticas, como los que se encargan de convertir a RDF, también conocidos como *RDFizers*. Herramientas que convierten datos provenientes de bases de datos relacionales, archivos CSV, XML, texto, entre otros. Un listado de ellas se encuentra en (Wikipedia, 2016).

Tres de los desarrollos más populares para convertir datos OAI-PMH fueron comparados (Tabla 5), con el objetivo de verificar la factibilidad de su implementación dentro de OntoOAI: OAI2LOD Server, un desarrollo para exponer metadatos OAI-PMH como *Linked Data* (Haslhofer & Schandl, 2008); D2R (Bizer & Cyganiak, 2006) y OAI2RDF (SIMILE, 2006).

	OAI2LOD	D2R	OAI2RDF
Descripción	Basado en D2R, expone metadatos OAI-PMH como *Linked Data.*	Publica bases de datos relacionales no-RDF como *Linked Data.*	Usa una transformación a través de hojas XSLT
OAI - PMH	Si	No	Si
Desventajas	La persistencia de los archivos RDF (los metadatos cosechados son almacenados en memoria). Probado con máximo 25,000 registros.	No trabaja con recursos del protocolo OAI-PMH.	La transformación XSLT no funciona bien para los casos de metadatos DC incompletos o en algunos tipos de implementación.

Tabla 5. Comparación de *RDFizers*

La comparación muestra que solo dos herramientas trabajan con recursos OAI-PMH. Sin embargo, para el caso de OAI2LOD se encontraron problemas importantes de demanda de memoria para cosechas de más de 25 mil registros, y dado que hay repositorios que llegan a almacenar millones de registros, este software tuvo que ser descartado.

Respecto a OAI2RDF se encontró el problema de conversiones resultantes erróneas con ausencia de metadatos Dublin Core y con algunas implementaciones de dicho estándar.

Por estas razones se optó por desarrollar una aplicación de software para efectuar esta tarea ya que, además de superar dichos inconvenientes, resultaba más adecuado para la 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 la API *SaxBuilder* (Figura 16).

Figura 16. Funcionamiento de XML_DC2RDF

SaxBuilder construye un documento JDOM a partir de la utilización de un *parser* SAX de terceros para manejar las tareas de análisis y utiliza una instancia de un *SAXHandler* para escuchar los eventos SAX[9] con el fin de construir un documento con contenido JDOM usando JDOMFactory.

JDOM es una representación Java de un documento XML. JDOM proporciona una forma de representar ese documento para la lectura fácil y eficiente, la manipulación, y la escritura.

3.5 Integración XML DC con FOAF

Los datos del autor son enriquecidos para agregar relaciones de coautoría y más datos acerca del autor. OntoOAI trabaja con fuentes de datos que describen información de personas cuando éstos son expresados con la especificación *Friend of a Friend* (FOAF).

OntoOAI utilizó para este efecto la base de DBPedia, particularmente el conjunto de datos *PersonData* como se describió en los aparados anteriores.

---

[9] http://www.saxproject.org.

Así para cada registro Dublin Core obtenido de los procesos de cosecha, se agrega información en la etiqueta *creator* que corresponde al autor o coautores del recurso.

El modelo OntoOAI espera que si un recurso de información tiene varios autores la etiqueta *creator* se repita, de tal forma que para cada autor se pueda ejercer el mismo procedimiento.

Entonces, para cada uno se especifica el tipo RDF como *http://xmlns.com/foaf/0.1/Person* y se incluyen además el nombre (*foaf:name*), apellidos (*foaf:givenName*, *foaf:surname*), personas con las que tiene una relación (*foaf:knows*) - en este caso de coautoría -, intereses *(foaf:topic_interest)*, fecha de nacimiento (*onto:birthDate),* y al final se obtiene el metadato *creator* como sigue:

<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, ya en RDF con datos de autor enriquecidos, es el que se muestra a continuación. El artículo del ejemplo es de la autoría de "Edgar Morin". El único dato que se tenía después de la cosecha de metadatos era el nombre del autor, sin embargo, después de realizar el proceso de enriquecimiento, se encontró que dicho autor tenía datos en *PersonaData* que corresponden a su página en la versión en inglés de Wikipedia.

Entonces, los datos de su artículo publicado en la revista *Convergencia* fueron adicionados con los que contiene su biografía de Wikipedia, como 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" foaf:surname="Morin" foaf:givenName="Edgar" foaf:topic_interest="Sociology" dc:description="French philosopher and sociologist" foaf:birthDate="1921-07-08">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>

3.6 Ingeniería ontológica

El modelo verifica la descripción del conocimiento obtenido después de la cosecha y transformación a través de una ontología resultado de la combinación de dos existentes: Dublin Core y FOAF.

Este proceso se realiza a través del *merging* o fusión de ontologías. A continuación se describe cada ontología y la técnica usada para la fusión, así como el proceso de validación de información para la cual se crea esta nueva ontología.

3.6.1 Ontología Dublin Core

El componente ontológico de OntoOAI 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 protocolo 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 incorpora los dos niveles en este espacio de nombres.

La ontología Dublin Core consta de 25 clases, organizadas jerárquicamente como se muestra en la figura 17.

Las métricas de la ontología DC arrojan los datos mostrados en la Tabla 6.

Axiomas	810
Clases	25
Propiedades de objeto	24
Propiedades de datos	13
Axiomas lógicos	255

Tabla 6. Métricas de la ontología DC (DCMI, 2016)

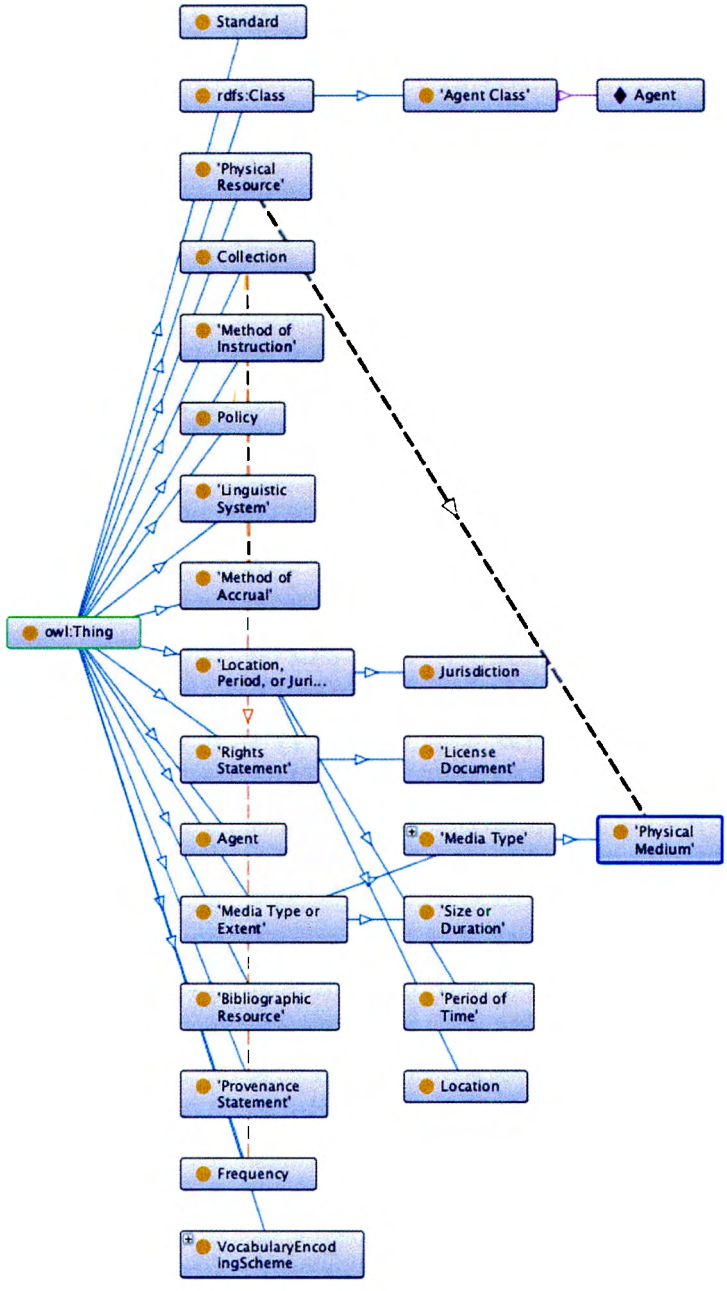

Figura 17. Jerarquía de clases de la ontología DC

Son 24 las propiedades de objeto de Dublin Core, que evidentemente son del tipo *owl:topObjectProperty*, y que son las propiedades cuyo rango corresponde a instancias de la clase *owl:Class* (Figura 18).

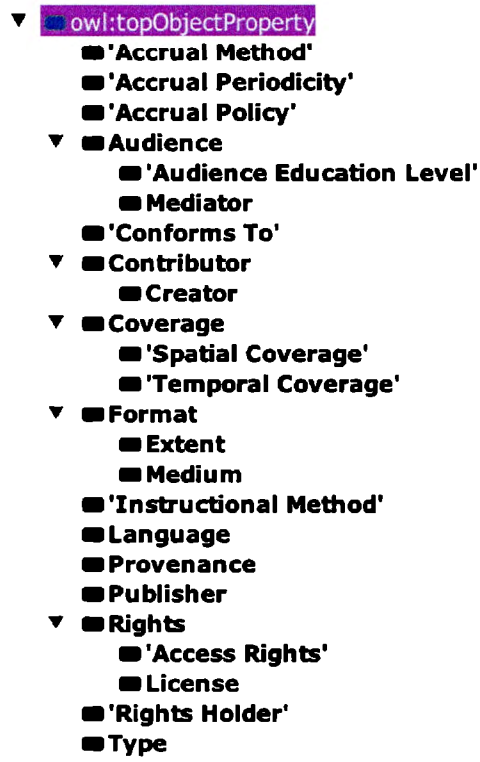

Figura 18. Jerarquía de propiedades de objeto de la ontología DC

Igualmente se pueden observar, en la figura 19, las 13 propiedades de datos, que son del tipo *owl:topDataProperty*, y que conforman el conjunto de propiedades con rango compuesto por literales e instancias de *rdfs:Datatype*.

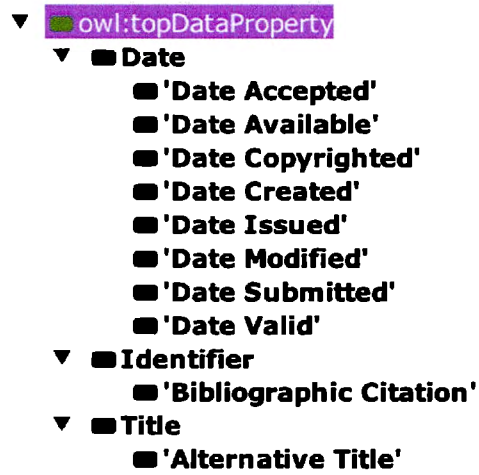

Figura 19. Jerarquía de propiedades de datos de la ontología DC

Las propiedades de anotación, que representan información adicional para clases y propiedades dentro de la ontología Dublin Core, se muestran jerárquicamente en la figura 20.

Figura 20. Jerarquía de propiedades de anotación de la ontología DC

3.6.2 Ontología Friend of a Friend

La segunda ontología que es usada en OntoOAI es *Friend of a Friend* para incorporar los datos adicionales 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.

El espacio de nombres FOAF es usado para representar información acerca de personas, tal como nombres, fechas de nacimiento, fotos, y especialmente otras personas con las que se relacionan de alguna forma. Es particularmente útil para representar datos de redes sociales.

FOAF comenzó con una observación simple (Allemang & Hendler, 2011, p. 196):

> "Si vamos a apoyar las redes sociales en la Web, los individuos deben ser capaces de tomar control de sus propios datos, almacenarlos si lo prefieren, manejarlos usando cualquier herramienta que deseen y aún así interactuar con otros usuarios, sin importar las decisiones que estos otros usuarios hayan tomado."

Aunque muchas de las más importantes redes sociales no adoptaron este principio, en la actualidad se discuten fuertemente temas de control y privacidad de los datos. FOAF es la base de muchos esfuerzos de "open social" que buscan integrar información a través de muchas redes sociales y aplicaciones disponibles en la Web.

FOAF trabaja bajo el espíritu del principio *AAA*: *Anyone can say Anything about Any topic* (Cualquiera puede decir cualquier cosa acerca de cualquier tema). En este caso, los temas de los que usualmente cualquiera habla son personas. Otras cosas que están comúnmente relacionadas con lo que se puede decir acerca de las personas y que están incluidas en FOAF son: Organizaciones (a las que pertenecen), Proyectos (donde trabajan),

Documentos (que han creado o aquellos que los describen), entre muchas otras (Allemang & Hendler, 2011).

Son 19 las clases contempladas en la ontología FOAF, su jerarquía y el resto de métricas de la ontología se muestran en la figura 21 y en la tabla 7, respectivamente.

Axiomas	559
Clases	19
Propiedades de objeto	40
Propiedades de datos	27
Axiomas lógicos	157

Tabla 7. Métricas de la ontología FOAF

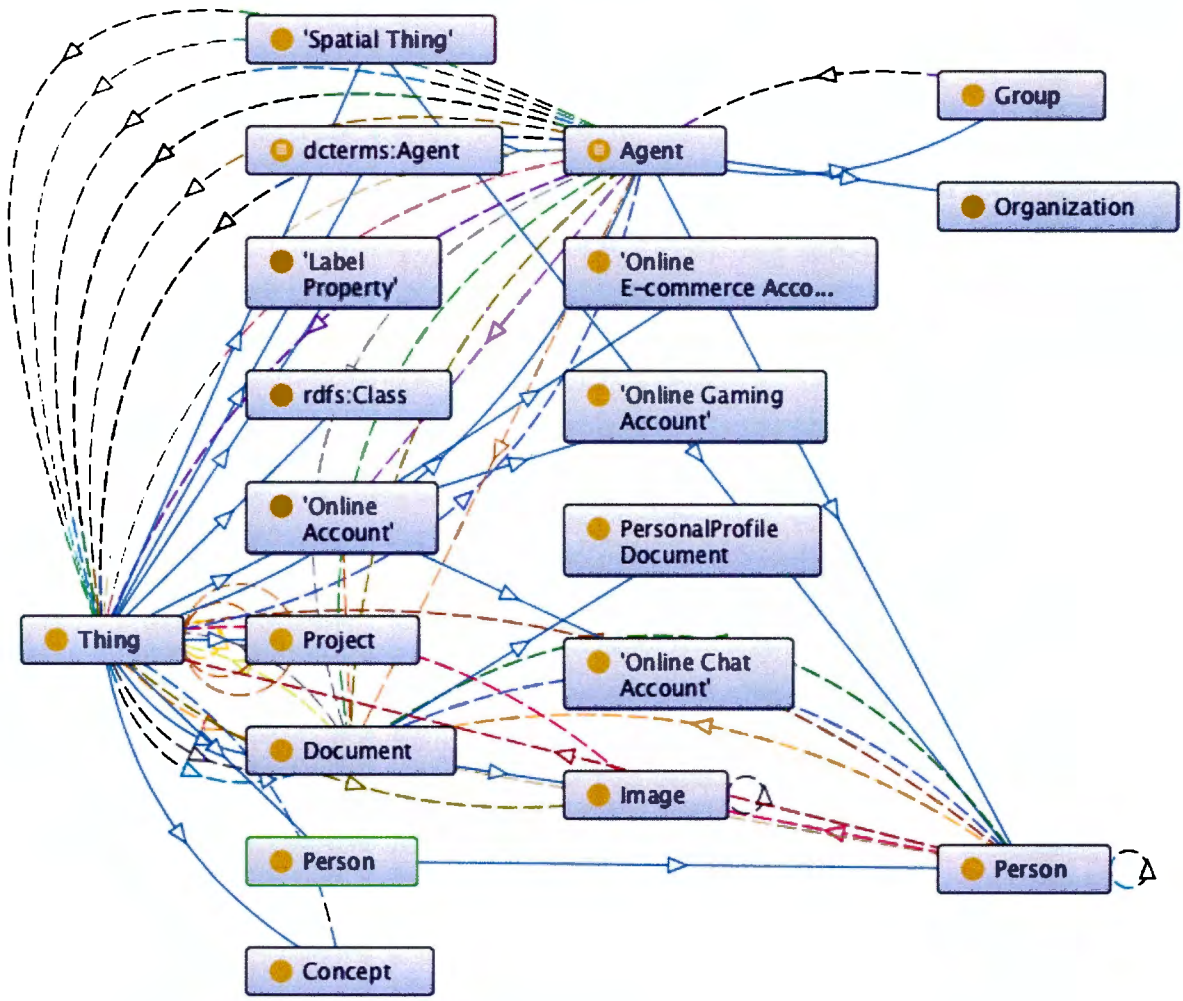

Figura 21. Jerarquía de clases de la ontología FOAF

La información adicional dada a las clases y propiedades, tanto de objeto como de datos, la definen las propiedades de anotación, cuya jerarquía es mostrada en la figura 22.

Figura 22. Jerarquía de propiedades de anotación de FOAF

Figura 23. Jerarquía de propiedades de datos de FOAF

Las 27 propiedades cuyos valores son literales o instancias del tipo *rdfs:Datatype*, es decir las propiedades de datos también se pueden observar en la figura 23.

Finalmente, FOAF define 40 propiedades de objeto (figura 24), es decir, propiedades cuyo rango corresponde a instancias de *owl:Class.*

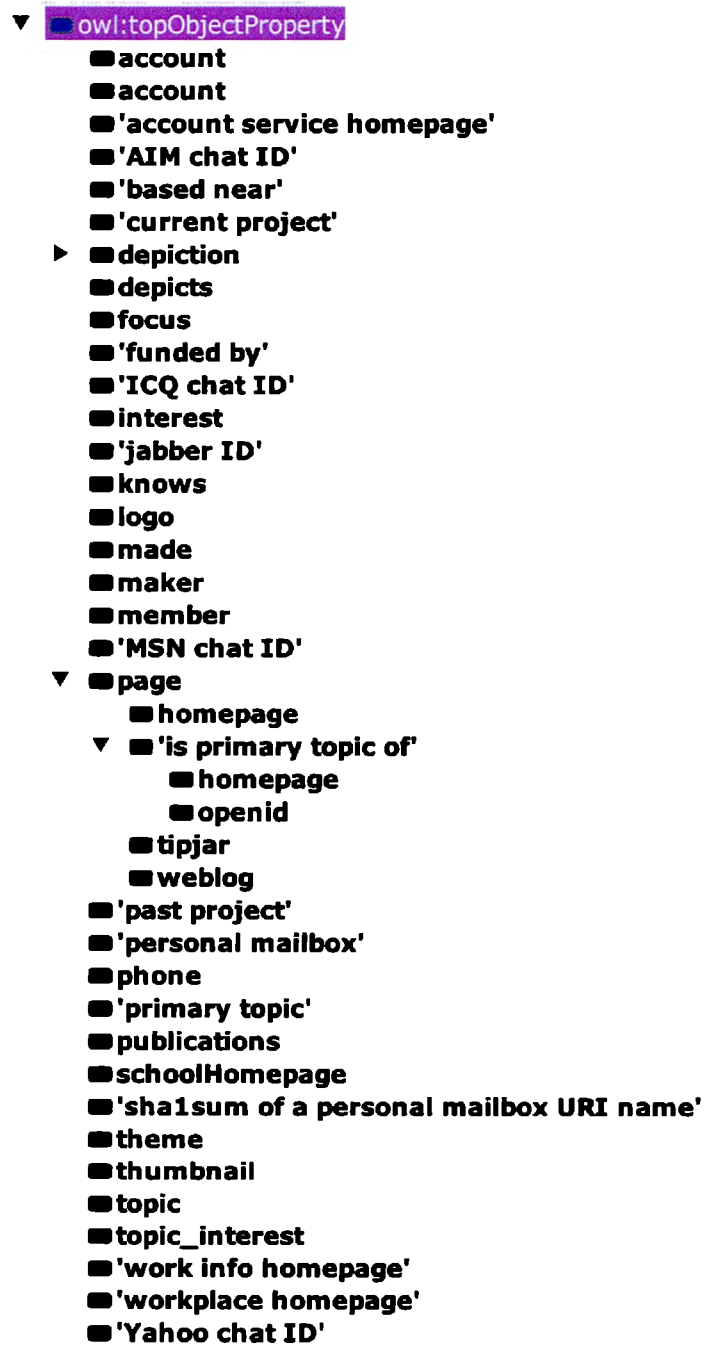

Figura 24. Jerarquía de propiedades de objeto de FOAF

3.6.3 Fusión de ontologías

La intención de la fusión de las dos ontologías, Dublin Core y FOAF, radica en generar una representación del conocimiento de la integración de datos provenientes de los repositorios OAI-PMH (Dublin Core) y otras fuentes de información que permitan enriquecer los datos del autor (FOAF), con el objetivo de construir un modelo que permita verificar la consistencia de dicha integración.

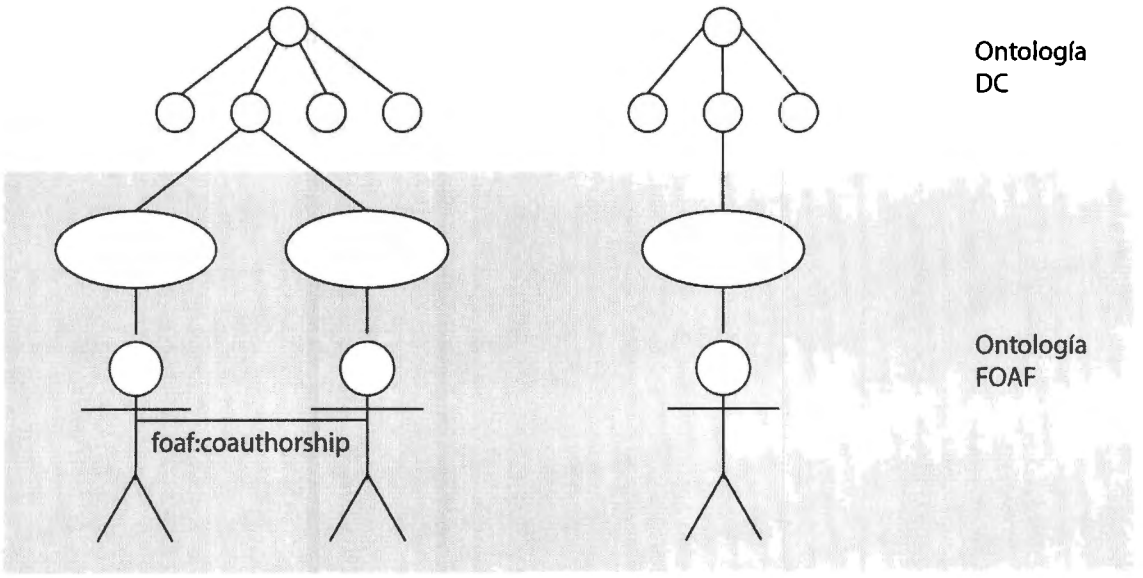

Figura 25. Modelo ontológico con Dublin Core y Friend of a Friend

El modelo ontológico que se propone se ilustra en la figura siguiente. Con él, es posible modelar las propiedades de un recurso de información (libro, artículo, entre otros) con su autor o autores; 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.

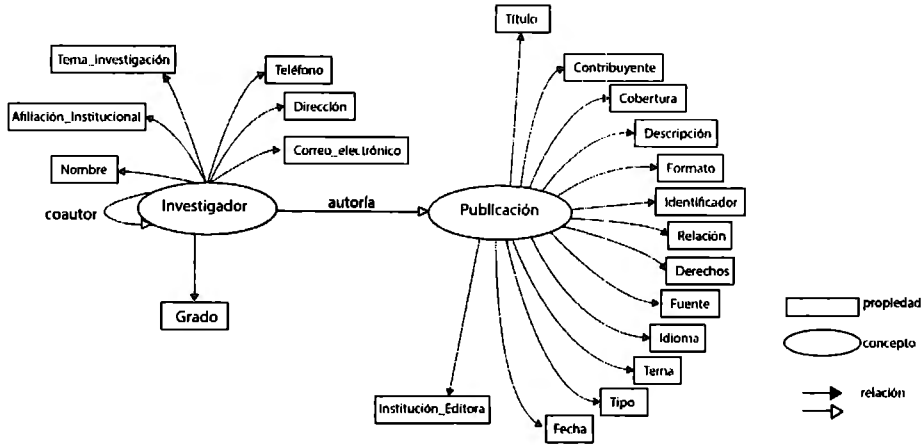

Figura 26. Relaciones de autoría y coautoría en la fusión de ontologías

En general, las operaciones con ontologías son tareas no triviales, que no se pueden resolver completamente de forma automática, debido a una diversidad de factores, como la insuficiente especificación de una ontología para encontrar similitudes con otra ontología.

Por ello, suelen hacerse de forma manual o semi-automática, donde una herramienta ayuda a encontrar posibles relaciones entre elementos de diferentes ontologías, pero la confirmación final de la relación se deja a un experto, quién tomará las decisiones, basándose en la descripción del lenguaje natural de los elementos de la ontología y el sentido común.

Con la ayuda de la herramienta Protégé y su función de *Refactor>Merge Ontologies* se realizó la fusión (*merging*) de las dos ontologías de entrada.

Como bien mencionan Ameen, Rahman Khan y Rani (2014), 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 por los mismos autores.

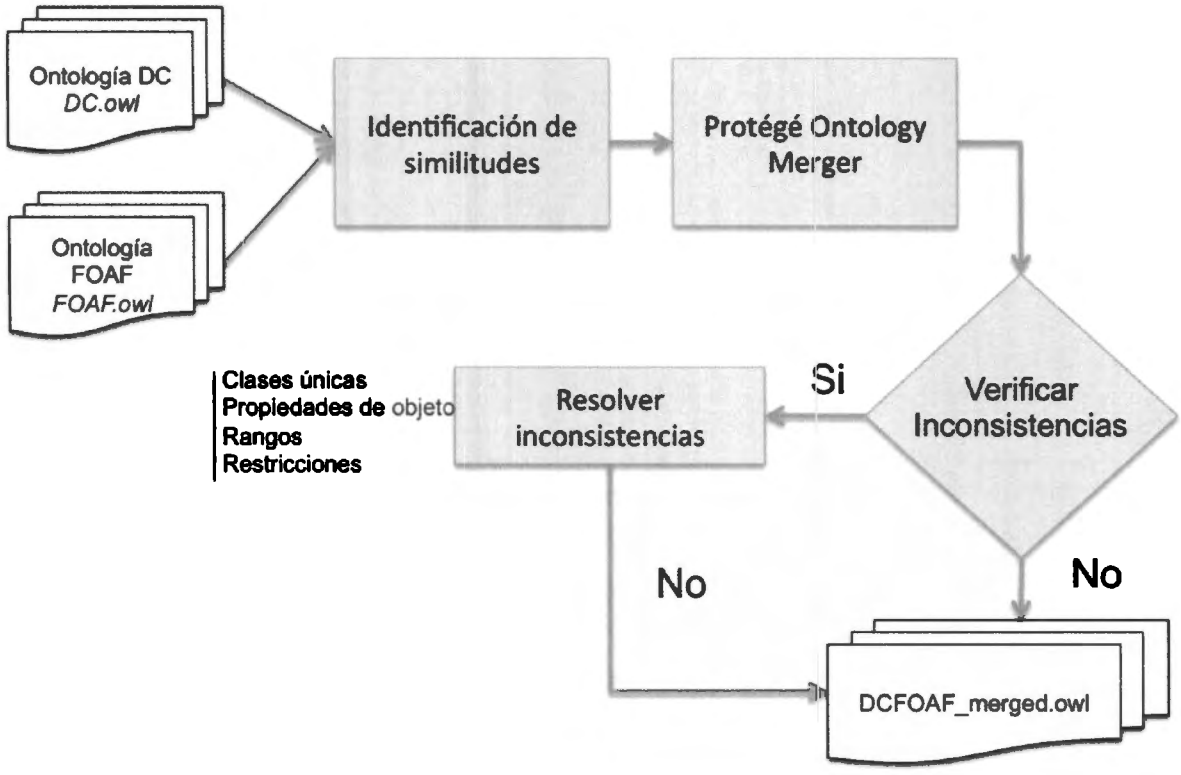

Figura 27. Algoritmo de fusión de ontologías de OntoOAI basado en (Ameen, Rahman Khan, & Rani, 2014)

El análisis de la jerarquía de clases de ambas ontologías arrojó algunas equivalencias que son marcadas en color (Tabla 8) con su correspondiente en cada ontología.

CLASES	
DC	FOAF
1. Agent	1. Agent
2. AgentClass	2. Concept
3. BibliographicResource	3. dcterms:Agent
4. Collection	4. Document
5. Frequency	5. Group
6. FileFormat	6. Image
7. LicenseDocument	7. LabelProperty
8. LinguisticSystem	8. OnlineAccount
9. LocationPeriodOrJurisdiction	9. OnlineChatAccount
10. Jurisdiction	10. OnlineEcommerceAccount
11. Location	11. OnlineGamingAccount
12. PeriodOfTime	12. Organization
13. MediaTypeOrExtent	13. Person (contactPerson)

14. MediaType	14. Person
15. MethodOfAccrual	15. PersonalProfileDocument
16. MethodOfInstruction	16. Project
17. PhysicalMedium	17. rdfs:Class
18. PhysicalResource	18. SpatialThing
19. Policy	19. Thing
20. ProvenanceStatement	
21. rdfs:Class	
22. RightsStatement	
23. SizeOrDuration	
24. Standard	
25. VocabularyEncodingScheme	

Tabla 8. Análisis de equivalencia de clases entre DC y FOAF

El proceso automático en Protégé identificó las clases idénticas en nombre (*rdfs:Class* y *Agent*) y realizó la fusión. Sin embargo para el caso de *BibliographicResource* y *Document* no lo hizo ya que el nombre no coincide, por lo que se tuvo que explicitar manualmente.

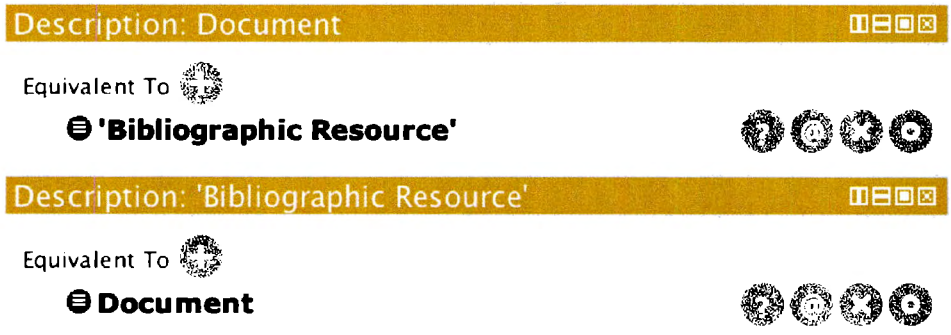

Figura 28. Declaración de equivalencia *Document* y *BibliopgraphicResource*

La jerarquía de clases de la ontología resultante se muestra en la figura 29. Además fue necesario identificar otras similitudes, para declarar equivalencias y restricciones adicionales. Así como resolver las inconsistencias generadas.

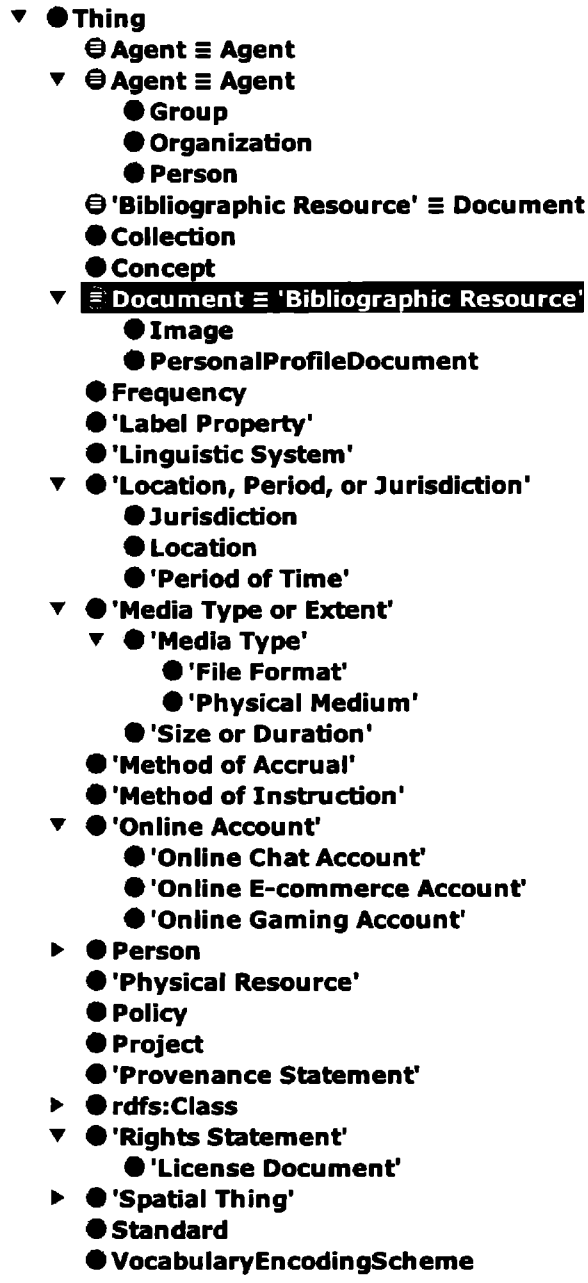

Figura 29. Jerarquía de clases de la ontología resultante de la fusión.

De forma similar fue necesario analizar las propiedades de objeto (Tabla 9) para identificar coincidencias semánticas, donde además es importante verificar y hacer coincidir rangos y dominios.

PROPIEDADES DE OBJETO	
DC	FOAF
1. accrualMethod	1. account
2. accrualPeriodicity	2. accountServiceHomepage
3. accrualPolicy	3. AIMchatID
4. audience	4. basedNear
5. audienceEducationLevel	5. currentProject
6. mediator	6. depiction
7. conformsTo	7. image
8. contributor	8. depicts
9. creator	9. focus
10. coverage	10. fundedBy
11. spatialCoverage	11. ICQchatID
12. temporalCoverage	12. interest
13. format	13. jabberID
14. extent	14. knows
15. medium	15. logo
16. instructionalMethod	16. made
17. language	17. maker
18. provenance	18. member
19. publisher	19. MSNchatID
20. rights	20. page
21. accessRights	21. homepage
22. license	22. isPrimaryTopicOf
23. rightsHolder	23. homepage
24. type	24. openId
	25. tipjar
	26. weblog
	27. pastProject
	28. personalMailbox
	29. phone
	30. primaryTopic
	31. publications
	32. schoolHomepage
	33. sha1sumOfPersonalMailbox
	34. theme
	35. thumbnail
	36. topic
	37. topic_interest
	38. workInfoHomepage
	39. workPlaceHomepage
	40. yahooChatID

Tabla 9. Análisis de equivalencias de propiedades de objeto en DC y FOAF

Respecto a estas propiedades se definió *Creator* (http://purl.org/dc/terms/creator) de Dublin Core con *Maker* (http://xmlns.com/foaf/0.1/maker) de FOAF como similares (fig. 30).

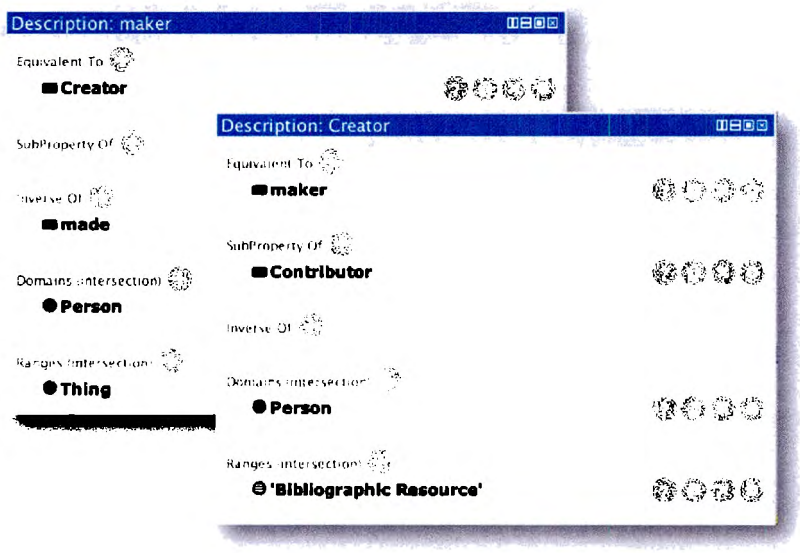

Figura 30. Equivalencia de la propiedad *Creator* con *Maker*.

Las propiedades *knows*, *topic_interest* fueron analizados ya que resultan relevantes para la descripción de los recursos en OntoOAI, como se ha indicado anteriormente.

Respecto a las propiedades de datos se ajusto la correspondiente a *title,* también de forma manual dada una variación de mayúscula en los nombres.

PROPIEDADES DE DATOS	
DC	FOAF
1. date	1. accountName
2. dateAccepted	2. age
3. dateAvailable	3. birthday
4. dateCopyrighted	4. DNAchecksum
5. dateCreated	5. family_name
6. dateIssued	6. familyName
7. dateModified	7. firstName
8. dateSubmitted	8. geekcode
9. dateValid	9. gender
10. identifier	10. givenName
11. bibliographicCitation	11. givenName
12. Title	12. jabberID

13. alternativeTitle	13. lastName 14. myersBriggs 15. name 16. nickName 17. AIMchatID 18. ICQchatID 19. MSNchatID 20. SkypeID 21. yahooChatID 22. plan 23. sha1sum(hex) 24. sha1sumPersonalMailboxURI Name 25. status 26. surname 27. title

Tabla 10. Análisis de equivalencia de propiedades de datos DC y FOAF

La ontología resultante llamada *DCFOAF_merged.owl* es verificada haciendo uso de un razonador, dado que el modelo afirmado y el inferido resultan similares se dice que la ontología es consistente. Las métricas de la ontología resultante se muestran en la siguiente tabla.

Clases	42
Propiedades de objeto	66
Propiedades de datos	40

Tabla 11. Métricas de la ontología resultante *DCFOAF_merged.owl*

Así con esta ontología se verifica la consistencia de los archivos RDF/XML obtenidos en las fases anteriores.

3.7 Almacenamiento

OntoOAI usa TDB del *framework* Jena como *triple store*. Una vez que los datos son colectados, transformados, enriquecidos y validados pasan al almacén de tripletas.

Al finalizar dichos procesos se obtienen un conjunto de grafos RDF serializados en archivos RDF/XML. En cada uno de ellos se describen una cantidad predeterminada de recursos de información, la cual depende de la configuración del repositorio fuente; como se verá en el caso de estudio, cada repositorio cosechado responde con un número diferente de recursos por archivo.

Cada uno de esos recursos es descrito por los 12 metadatos Dublin Core, que al ser repetibles pueden representar al menos 12 elementos por recurso, estos conforman las propiedades de cada uno junto con las propiedades que fueron incluidas en las fases de conversión a RDF y enriquecimiento de autores. Cada una de estas propiedades da origen a una tripleta dentro de TDB.

La carga de tripletas se hace utilizando *TDBLoader* de Jena que permite la carga masiva y la generación de un índice al mismo tiempo. Realiza operaciones de carga masiva con más eficiencia que la simple lectura de RDF en un modelo TDB.

Para realizar esta carga más eficiente se programaron *shell scripts* que se ejecutan directo sobre TDB, un script para cada repositorio cosechado con el número de sentencias de carga que corresponde al número de archivos colectados, como se verá en el siguiente capítulo para el caso de estudio. El diagrama siguiente muestra el proceso que se sigue.

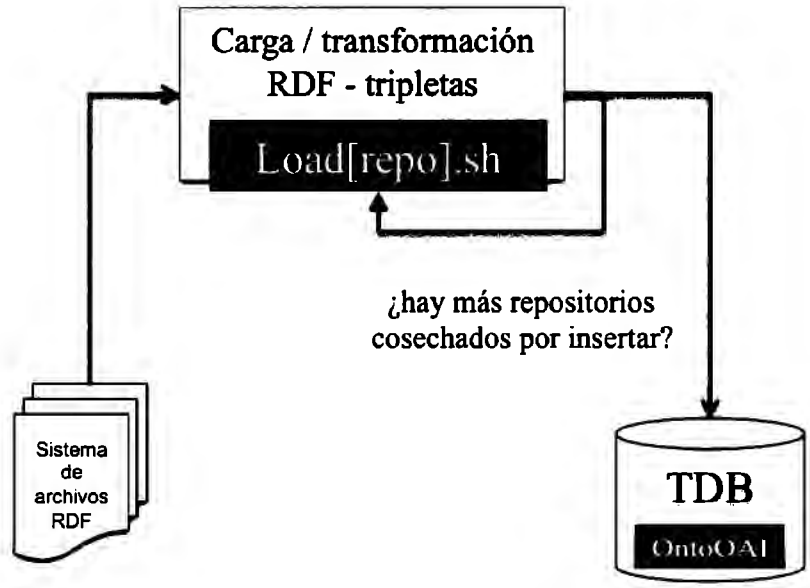

Figura 31. Proceso de almacenamiento de tripletas

3.8 Sensibilidad al contexto

La característica de sensibilidad al contexto del usuario con la que debe contar la solución propuesta refiere a la capacidad de percibir información del ambiente del usuario y otorgar resultados acorde a dicha información.

La idea fundamental con ello es inferir situaciones no explicitadas y así manifestar un comportamiento inteligente.

Por tanto, a un usuario se le asocia un perfil que constituye un conjunto de características, algunas más dinámicas que otras para representar circunstancias de espacio-tiempo escolares, personales e incluso sociales. Así un perfil de estudiante contiene sus datos de identificación, programa de estudio, periodo, cursos inscritos, entre otros.

Para el presente trabajo se adaptó el modelo de representación de estudiantes basado en ontologías para sistemas de tutoría inteligente de aprendizaje a distancia de Panagiotopoulos, Kalou, Pierrakeas y Kameas (2012), que usa las siguientes 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

La ontología propuesta por estos autores no fue localizada, además que no cumplía exactamente con las necesidades de OntoOAI; es así, que se diseñó una nueva ontología llamada *OntoOAIEstudiante*, tomando como base la ontología propuesta por dichos autores pero con una concepción nueva.

El contexto del usuario se conforma de la información conocida de antemano y se determina de una manera estática con base en el modelo *OntoOAIEstudiante*. La información puede provenir de la escuela o las oficinas de control escolar quienes serían capaces de aportar datos sobre los alumnos para recuperar información relevante para sus estudios.

Cabe resaltar que, dicha representación considera información independiente de la aplicación y no contempla información dependiente de la interacción con la aplicación.

Para crear la ontología que describe al estudiante, se retoman las cuatro clases principales de Panagiotopoulos, Kalou, Pierrakeas & Kameas (2012) pero con la siguiente jerarquía de clases.

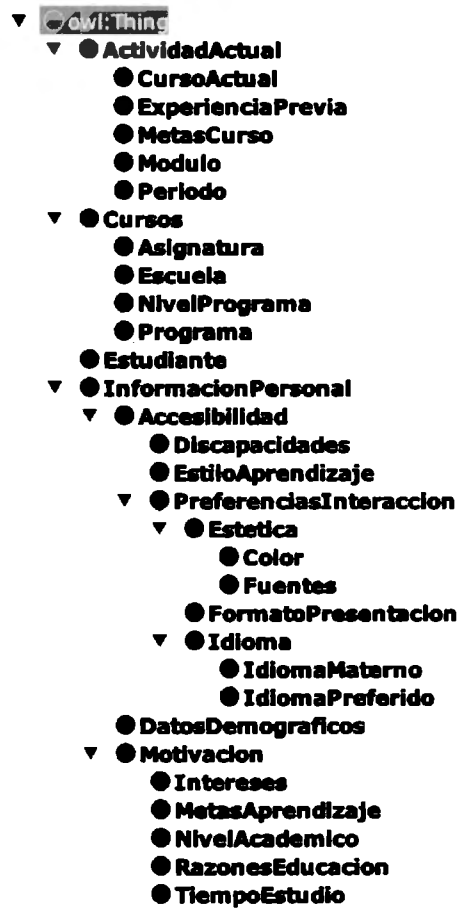

Figura 32. Jerarquía de clases de la ontología OntoOAIEstudiante

El perfil de usuario es agrupado en 4 clases principales, como sigue:

1. *Estudiante*, que representa a un alumno.
2. *Cursos*, describe las asignaturas que cursa el alumno, la escuela en la que está inscrito, programa y nivel del programa (licenciatura, maestría, doctorado).
3. *ActividadActual*, describe el periodo actual que cursa el estudiante, experiencia previa, metas del curso, módulo y periodo.
4. *InformaciónPersonal*, para representar la accesibilidad, datos demográficos y motivación del alumno.

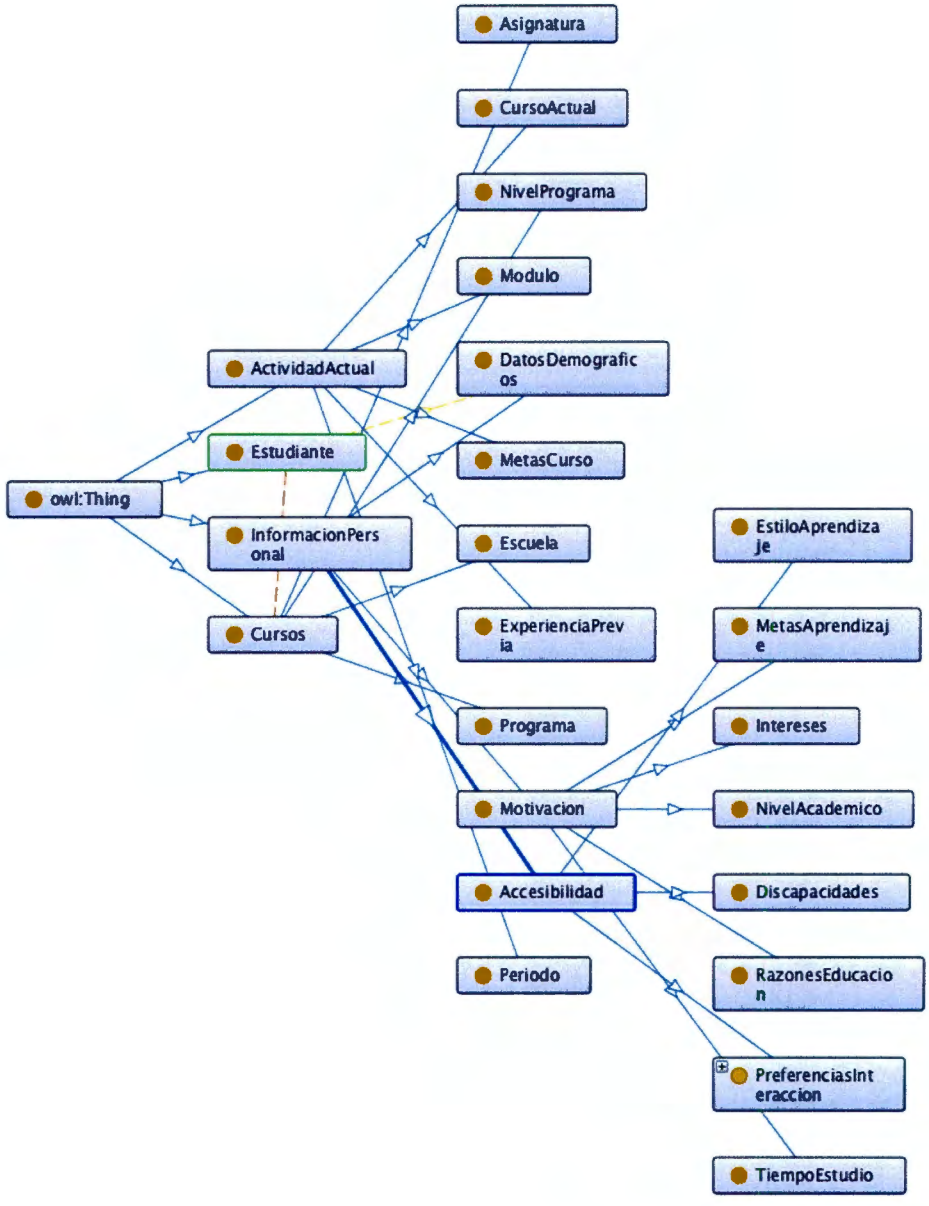

Figura 33. Grafo de clases de la ontología *OntoOAIEstudiante.owl*

Finalmente, el perfil del usuario representado en la ontología es uno de los insumos del motor de inferencia que se describe a continuación. Si bien el modelo OntoOAI puede ser útil para descubrir conocimiento para profesores, el perfil de usuario usado para probar el modelo se centra en el estudiante. Sin embargo, es fácilmente extensible a profesores, académicos o investigadores, únicamente se requeriría cambiar la representación del perfil de usuario a uno adecuado para dichos tipos de usuarios.

3.9 Motor de inferencia y descubrimiento de conocimiento

Un razonador es un programa que infiere consecuencias lógicas de un conjunto de hechos determinados explícitamente o axiomas y típicamente provee soporte automático para tareas de razonamiento tales como clasificación, *debugging* y consulta (Dentler, Cornet, Teije, & de Keizer, 2011).

OntoOAI trata el descubrimiento de conocimiento como la extracción no trivial de información implícita, previamente desconocida y potencialmente útil, concepto definido por Frawley, Piatetsky-Shapiro y Matheus (1992).

Retomando el proceso de descubrimiento de conocimiento definido por los mismos autores pero teniendo en consideración que éste se lleve a cabo a través de un razonador lógico utilizando técnicas de inferencia, se tiene:

> Dado un conjunto de hechos F (base de conocimiento compuesta por las tripletas obtenidas de los procesos de cosecha e integración), un lenguaje L (OWL), y alguna medida de certidumbre C (se considera 'cierto' al conjunto de tripletas inferidas considerando el perfil del usuario previamente definido), se define un *patrón* como un enunciado S en L que describe relaciones entre un subconjunto F_S de F con una certidumbre c, de tal forma que S es más simple (en algún sentido) que la enumeración de todos los hechos en F_S. Un patrón que es interesante (ya que se deriva de hechos) y suficientemente cierto (ya que se deriva con información considerando el perfil del usuario) es llamado *conocimiento*.

Teniendo el proceso de descubrimiento de conocimiento definido, se diseñó una aplicación que permita llevarlo a cabo. Su arquitectura así como la interacción entre los componentes se muestran en la figura siguiente.

Es importante recordar que, como se mostró en la figura 13, la información requerida de entrada corresponde a los términos de búsqueda de un usuario así como su contexto, expresado en un perfil como se explicó en el apartado anterior.

Dicha entrada circula por un razonador basado en reglas que procesa una consulta sobre una base de datos de tripletas haciendo uso del API de ontologías del *Framework Jena* de la Fundación de Software Apache (Figura 34).

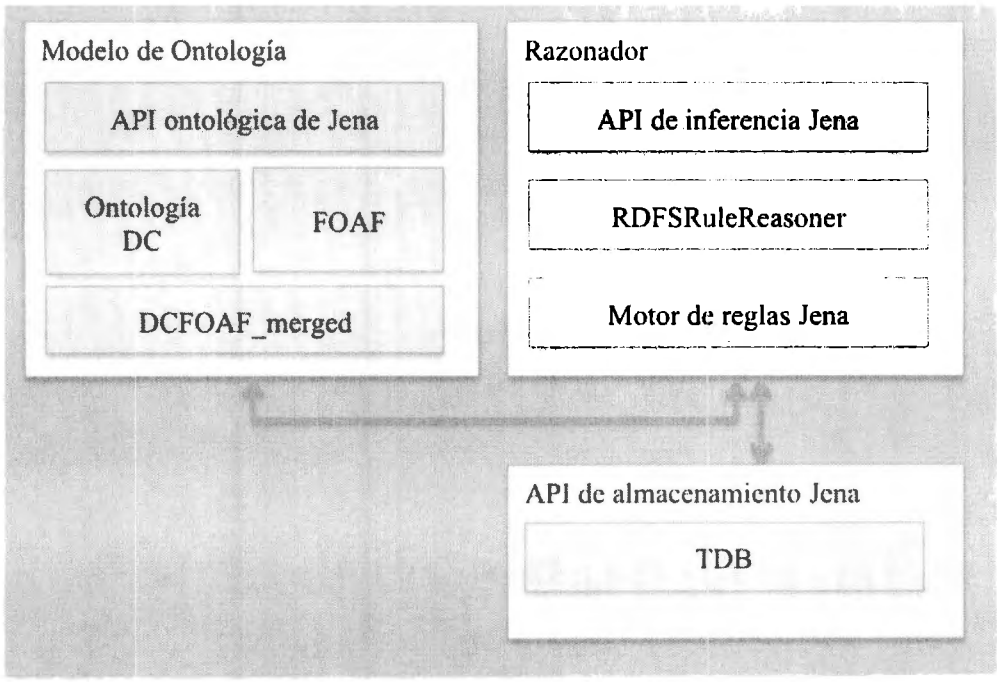

Figura 34. Arquitectura del motor de inferencia de OntoOAI

El subsistema de inferencia de Jena está diseñado para permitir un rango de motores de inferencia o razonadores para ser conectados con Jena. Tales motores son usados para derivar enunciados RDF adicionales derivados de una base RDF junto con información ontológica opcional y los axiomas y reglas asociadas con el razonador (Apache Software Foundation, 2011).

El motor de Jena infiere consecuencias lógicas de un conjunto de hechos determinados explícitamente o axiomas y típicamente provee soporte automático para

tareas de razonamiento tales como clasificación, debugging y consulta (Dentler, Cornet, Teije, & de Keizer, 2011). Es usado para derivar enunciados RDF adicionales de una base RDF junto con información ontológica opcional y los axiomas y reglas asociadas con el razonador.

Con fines experimentales para este desarrollo se usó el razonador OWL incluido, una implementación basada en reglas de OWL/Lite.

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 preparados como insumo para aplicaciones que se encarguen de la visualización para el usuario final.

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 de recursos OAI-PMH, la información de contexto del usuario e información ontológica.

El proceso de inferencia, que sigue un razonamiento deductivo *modus ponens* o con encadenamiento hacia adelante, está basado en el siguiente silogismo:

Dada las funciones

$f(a) = R_a$ que define la autoría o contribución autoral de un recurso r escrito por un autor a,

$f(a) = T_a$ que define la participación de un autor a en un recurso t, entendiendo como participación a que el autor sea el tema, título o autor del recurso,

$f(b) = T_b$ que define la participación de un autor b en un recurso t, entendiendo como participación a que el autor sea el tema, título o autor del recurso,

donde:

$R = \{ r \mid r \in R \}$, definido por los recursos de información contenidos en la base de conocimiento,

$A = \{ a \mid a \in A \}$, definido por los autores o colaboradores de los recursos de información contenidos en la base de conocimiento,

$T = \{ t \mid t \in T \}$, definido por los recursos de información contenidos en la base de conocimiento que fueron escritos por un grupo de autores determinado,

$B = \{ b \mid b \in B \}$, definido por los autores o colaboradores de un grupo de recursos de información determinado contenidos en la base de conocimiento,

Se tiene que,

Si (el recurso *r* es interesante para el usuario),

Entonces (el autor *a* del recurso *r* es interesante para el usuario).

Si (el autor *a* del recurso *r* es interesante para el usuario),

Entonces (otros trabajos T_a cuyo autor, tema o título correspondan al autor *a* son interesantes para el usuario).

Si (los trabajos T_a cuyo autor, tema o título corresponden al autor *a* son interesantes para el usuario),

Entonces (el conjunto de autores o colaboradores B de los recursos T_a es interesante para el usuario).

Si (el conjunto de autores o colaboradores B de los recursos T_a es interesante para el usuario),

Entonces (el conjunto de recursos T_b escritos por los autores B son interesantes para el usuario),

El conjunto de recursos contenidos en los conjuntos T_a y T_b representan el conocimiento descubierto acorde a su búsqueda y su contexto:

$$\text{Conocimiento descubierto} = T_a \cup T_b$$

$$|T_a \cup T_b| = \text{Total de recursos descubiertos para el usuario}$$

El conocimiento descubierto conforma la salida para el usuario, en formato de tabla con los metadatos y enlaces al texto completo de cada recurso obtenido. Esta salida también puede visualizarse de otras formas, como ejemplo se incluyeron dos en el punto 4.7.1 salida tipo red y visualización jerárquica donde se pueden observar las relaciones entre los recursos para así enriquecer la experiencia del usuario.

Una demostración de la aplicación de este razonamiento es mostrado en el capítulo 4 para dos perfiles de usuarios distintos.

3.10 Procesamiento por lotes y procesamiento en tiempo real

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:

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.

De esta forma, el procesamiento por lotes y el almacenamiento central de los resultados de la recolección permite el manejo de grandes volúmenes de información y la construcción de aplicaciones y/o servicios basados en datos integrados y disponibles.

Respecto al proceso de consulta, ésta se basa en un conjunto de parámetros que ingresan a un motor de inferencia para devolver un resultado, y se realiza en tiempo real. 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.

3.11 Tecnologías y herramientas

Las tecnologías utilizadas para el desarrollo de los componentes de OntoOAI son las que se describen a continuación.

3.11.1 Protégé

Protégé[10] es un proyecto de código abierto, auspiciado por la Universidad de Stanford, que puede ser adaptado para construir aplicaciones simples o complejas basadas en ontologías. Es una de las herramientas más estables para la edición de ontologías y que ha venido madurando con el tiempo.

La salida de Protégé puede ser integrada con sistemas basados en reglas u otros solucionadores de problemas para construir una gran variedad de sistemas inteligentes (Stanford Center for Biomedical Informatics Research, 2014).

Este desarrollo, basado en Java, soporta las especificaciones más recientes del lenguaje OWL 2. Además de ofrecer versiones para diferentes sistemas operativos. Aunque tiene un desarrollo web WebProtégé[11] útil para una rápida visualización y ediciones simples, todavía resulta más poderosa la herramienta para instalarse de modo local.

Protégé fue usado para el trabajo de ingeniería ontológica en el desarrollo de OntoOAI, incluyendo la fusión de ontologías y construcción de la correspondiente para modelar el contexto del usuario.

3.11.2 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

---

[10] http://protege.stanford.edu/
[11] http://webprotege.stanford.edu/

bibliotecas Java para el desarrollo de código que maneje RDF, RDFS, RDFa, OWL y SPARQL alineado con las recomendaciones de la W3C.

Jena 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 (Apache Software Foundation, 2011).

La *API* Ontológica de Jena provee una interfaz de programación consistente para el desarrollo de aplicaciones ontológicas, independiente del lenguaje, es decir, la *API* de Jena es neutral. Fue seleccionado para la implementación de OntoOAI dada su solidez y robustez para el desarrollo de aplicaciones de la Web Semántica.

3.11.3 TDB

TDB es un componente para el almacenamiento y consulta RDF de Jena, soporta el rango completo de *APIs* de Jena y puede ser usado como un almacén de alto rendimiento para RDF.

Un conjunto de datos de TDB debe ser accedido desde una única JVM concurrentemente para evitar corrupción de datos.

Si es necesario compartir los datos de TDB entre múltiples aplicaciones es posible usar Fuseki (Foundation, 2014) el cual provee un servidor SPARQL que puede usar TDB para el almacenamiento persistente y provee protocolos SPARQL para consulta, actualización y actualización REST sobre HTTP (Apache Software Foundation, 2011).

Capítulo 4 Caso de aplicación de OntoOAI

Este capítulo está dedicado a describir la aplicación de OntoOAI en recursos de contenido académico obtenidos de dos fuentes de información estructuradas con OAI-PMH y la base DBPedia y muestra dos ejemplos de descubrimiento selectivo de conocimiento.

A continuación se describe la implementación de cada etapa de OntoOAI, de acuerdo a la metodología descrita en el apartado 3.1 (Figura 13).

4.1 Fuentes de información seleccionadas

Haciendo uso de OpenDOAR, el directorio de repositorios que implementan OAI-PMH es posible ubicar los proveedores de datos y sus URL-Base para ser cosechados por la aplicación.

Con fines experimentales y de validación del modelo propuesto fueron localizados dos repositorios: Redalyc.org, el portal de la Red de Revistas de América Latina y el Caribe, España y Portugal (Becerril-García, Aguado-López, Rogel-Salazar, Garduño-Oropeza, & Zúñiga-Roca, 2012) y el repositorio institucional de la Universidad Roskilde llamado RUDAR (Roskilde University Digital Archive) de Dinamarca.

Redalyc, disponible en http://www.redalyc.org, es un sistema de información científica que al momento mantiene un acervo de más de 400 mil artículos a texto completo, de más de mil revistas científicas, arbitradas por pares y de acceso abierto editadas por más de 400 instituciones de Iberoamérica. Redalyc además de permitir la lectura y descarga de artículos ofrece indicadores para la medición de la producción científica para países, instituciones, revistas y autores.

Redalyc fue seleccionado en primer lugar, por ser una de las fuentes más grandes y representativas de contenido científico de Iberoamérica y en segundo lugar debido a que la autora de este trabajo ha participado en este proyecto desde sus inicios, es así que se retoma la experiencia y el conocimiento pleno de este proyecto para la aplicación de OntoOAI.

El repositorio RUDAR, que se encuentra bajo la dirección http://rudar.ruc.dk/, constituye el sistema de información que reúne, preserva, divulga y da acceso a la producción intelectual y académica de la Universidad de Roskilde, una universidad pública de Dinamarca que fue fundada en 1972.

Este repositorio concentra principalmente productos de investigación en forma digital, incluyendo *pre-prints*, libros, informes de los estudiantes, informes técnicos, artículos, tesis, actas de conferencias e imágenes.

La elección del repositorio RUDAR, para la implementación de OntoOAI, se realizó únicamente tomando en consideración que este repositorio aparece en OpenDOAR, es así que se exhibe como un sitio que implementa el protocolo OAI-PMH, y así como fue seleccionado RUDAR se pudo seleccionar cualquier otro que cumpla con esta condición mínima.

Por supuesto, fue necesario verificar que el protocolo está activo y vigente en ambos repositorios, recuperando sus respectivas URL Base, visibles en OpenDOAR, y realizando una petición HTTP del tipo *identify* (Figuras 35 y 36), cuyo resultado da cuenta de la disponibilidad del protocolo pero también arroja datos de identificación de cada uno de los proveedores de datos.

URLBaseRedalyc: **http://oai.redalyc.org/redalyc/oai**

URLBaseRudar: **http://diggy.ruc.dk:8080/dspace-oai/request**

```
▼<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
 http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
   <responseDate>2016-04-04T08:33:24Z</responseDate>
   <request verb="Identify">http://oai.redalyc.org/redalyc/oai</request>
  ▼<Identify>
     <repositoryName>redalyc</repositoryName>
     <baseURL>http://oai.redalyc.org/redalyc/oai</baseURL>
     <protocolVersion>2.0</protocolVersion>
     <adminEmail>arianna.becerril@redalyc.org</adminEmail>
     <earliestDatestamp>1995-01-01</earliestDatestamp>
     <deletedRecord>no</deletedRecord>
     <granularity>YYYY-MM-DD</granularity>
     <compression>identity</compression>
    ▼<description>
     ▼<oai-identifier xmlns="http://www.openarchives.org/OAI/2.0/oai-identifier"
       xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
       xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai-identifier
       http://www.openarchives.org/OAI/2.0/oai-identifier.xsd">
         <scheme>oai</scheme>
         <repositoryIdentifier>redalyc.org</repositoryIdentifier>
         <delimiter>:</delimiter>
         <sampleIdentifier>oai:redalyc.org:10201906</sampleIdentifier>
       </oai-identifier>
     </description>
   </Identify>
 </OAI-PMH>
```

Figura 35. Verificación de URL Base de Redalyc

```
▼<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
 http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
   <responseDate>2016-04-05T01:12:05Z</responseDate>
   <request verb="Identify">http://diggy.ruc.dk:8080/dspace-oai/request</request>
  ▼<Identify>
     <repositoryName>Roskilde University Digital Archive</repositoryName>
     <baseURL>http://diggy.ruc.dk:8080/dspace-oai/request</baseURL>
     <protocolVersion>2.0</protocolVersion>
     <adminEmail>rammer@ruc.dk</adminEmail>
     <earliestDatestamp>2001-01-01T00:00:00Z</earliestDatestamp>
     <deletedRecord>persistent</deletedRecord>
     <granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
     <compression>gzip</compression>
     <compression>deflate</compression>
    ▼<description>
     ▼<toolkit xmlns="http://oai.dlib.vt.edu/OAI/metadata/toolkit"
       xsi:schemaLocation="http://oai.dlib.vt.edu/OAI/metadata/toolkit
       http://oai.dlib.vt.edu/OAI/metadata/toolkit.xsd">
         <title>OCLC's OAICat Repository Framework</title>
        ▼<author>
           <name>Jeffrey A. Young</name>
           <email>jyoung@oclc.org</email>
           <institution>OCLC</institution>
         </author>
         <version>1.5.48</version>
         <toolkitIcon>http://alcme.oclc.org/oaicat/oaicat_icon.gif</toolkitIcon>
         <URL>http://www.oclc.org/research/software/oai/cat.shtm</URL>
       </toolkit>
     </description>
   </Identify>
 </OAI-PMH>
```

Figura 36. Verificación de URL Base de RUDAR

Adicionalmente, como se mencionó en el punto 3.2 también se usa el conjunto de datos PersonData de la base de conocimiento de DBPedia. Para este propósito se utilizó el conjunto de datos descrito con N-Triples, un archivo de aproximadamente 1GB que equivale a 8,397,081 tripletas de información de autores de la versión en inglés de Wikipedia.

4.2 Cosecha de metadatos de Redalyc y RUDAR

Con la herramienta *OAI_PMH_Harvester* desarrollada para OntoOAI fueron realizados los procesos de recolección de metadatos de cada uno de los dos repositorios, a través de la siguiente estrategia de cosecha:

[*URLBaseRedalyc*]?verb=ListRecords&metadataPrefix=oai_dc

[*URLBaseRudar*]?verb=ListRecords&metadataPrefix=oai_dc

La Tabla 12 muestra los resultados del proceso. Del repositorio Redalyc se obtuvieron 19,027 archivos XML, que corresponden a 380,540 artículos recuperados, ya que cada archivo XML contiene 20 artículos. Para el caso de Rudar, se recuperaron 153 archivos con 100 recursos cada uno, haciendo un total de 15,300 entre libros, tesis y otros tipos de contenidos.

Repositorio	Total de archivos XML obtenidos	Total recursos descritos con Dublin Core	Tiempo de cosecha	Tamaño de la cosecha
Redalyc	19,027	380,540	45.6hrs	762.9MB
RUDAR	154	15,400	1.2hrs	28.7MB
Totales	19,181	395,940	46.8hrs	791.6MB

Tabla 12. Resultados de la cosecha de metadatos

En las figuras 37 y 38 se pueden observar dos ejemplos de recursos recuperados a través del proceso de cosecha, uno de cada proveedor de datos.


```
▼<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
   <responseDate>2016-04-05T04:41:38Z</responseDate>
   <request verb="GetRecord" identifier="oai:redalyc.org:81522330002"
   metadataPrefix="oai_dc">http://oai.redalyc.org/redalyc/oai</request>
  ▼<GetRecord>
    ▼<record>
      ▼<header>
         <identifier>oai:redalyc.org:81522330002</identifier>
         <datestamp>2012-08-27</datestamp>
         <setSpec>0123-885X</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>Hacia una ontología social del aprendizaje</dc:title>
           <dc:creator>Jean Lave</dc:creator>
           <dc:creator>Martin Packer</dc:creator>
           <dc:subject>Sociología</dc:subject>
           <dc:publisher>Universidad de Los Andes</dc:publisher>
           <dc:date>2011</dc:date>
           <dc:type>Artículo científico</dc:type>
           <dc:format>application/pdf</dc:format>
           <dc:identifier>http://www.redalyc.org/articulo.oa?id=81522330002</dc:identifier>
           <dc:relation>http://www.redalyc.org/revista.oa?id=815</dc:relation>
           <dc:rights>Revista de Estudios Sociales</dc:rights>
           <dc:source>Revista de Estudios Sociales (Colombia) Num.40</dc:source>
           <dc:language>es</dc:language>
         </oai_dc:dc>
       </metadata>
     </record>
   </GetRecord>
 </OAI-PMH>
```


Figura 37. Registro Dublin Core obtenido de Redalyc.org

Si se analizan con más detalle estas muestras cosechadas, se pueden identificar las similitudes y diferencias en la implementación del protocolo, por cada repositorio. Ambos repositorios implementan adecuadamente el OAI-PMH, sin embargo, dada la flexibilidad en la especificación, cada proveedor de datos puede adoptarlo de forma distinta.

Esta característica, como se verá más adelante, ha sido factor de éxito para la interoperabilidad entre repositorios; sin embargo, también provoca ciertas limitaciones en la integración y explotación de los datos.

```
▼<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2016-04-05T21:20:10Z</responseDate>
  <request identifier="oai:rudar.ruc.dk:1800/2990" metadataPrefix="oai_dc" verb="GetRecord">http://diggy.ruc.dk:8080/dspace-oai/request</request>
 ▼<GetRecord>
  ▼<record>
   ▼<header>
     <identifier>oai:rudar.ruc.dk:1800/2990</identifier>
     <datestamp>2010-03-03T01:16:17Z</datestamp>
     <setSpec>hdl_1800_116</setSpec>
     <setSpec>hdl_1800_60</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>Education for Active Non-violence</dc:title>
      <dc:creator>Oaki, Juha Janne Olavi</dc:creator>
      <dc:subject>Silo</dc:subject>
      <dc:subject>Freire</dc:subject>
      <dc:subject>Greenleaf</dc:subject>
      <dc:subject>Wenger</dc:subject>
      <dc:subject>Lave</dc:subject>
      <dc:subject>nonviolence</dc:subject>
      <dc:subject>non-violence</dc:subject>
      <dc:subject>Mumbai</dc:subject>
      <dc:subject>Humanist Movement</dc:subject>
      <dc:subject>leadership</dc:subject>
      <dc:subject>apprenticeship</dc:subject>
      <dc:subject>humanism</dc:subject>
     ▼<dc:description>
        Education for Active Non-violence examines the ideas and the practice of the Humanist Movement, an international movement against varied forms
        of violence. The research focuses on the process of learning through participation in the Humanist Movement in Mumbai, India, and is
        especially directed towards insights about the challenges to that educational practice. The field study involves interviews of nine members
        with varied kinds of participation in the movement. Theories by Robert K. Greenleaf, Silo, Paulo Freire, Jean Lave and Etienne Wenger are
        applied for gaining perspective on the discussions of the case study; research papers about the social situation in Mumbai are also referred
        to. Case study research methodology is based on semi-structured interviews analysed through a combination of interactionist and hermeneutic
        methods.
      </dc:description>
      <dc:contributor>Carney, Stephen</dc:contributor>
      <dc:date>2008-01-17</dc:date>
      <dc:date>2008-02-01T07:06:48Z</dc:date>
      <dc:date>2008-02-01T07:06:48Z</dc:date>
      <dc:date>2008-02-01T07:06:48Z</dc:date>
      <dc:type>Thesis</dc:type>
     ▼<dc:type>
        Pædagogik og uddannelsesstudier / Pedagogy Educational Studies - not master thesis
      </dc:type>
      <dc:identifier>http://hdl.handle.net/1800/2990</dc:identifier>
      <dc:language>en</dc:language>
     </oai_dc:dc>
    </metadata>
   </record>
  </GetRecord>
</OAI-PMH>
```


Figura 38. Registro Dublin Core obtenido del repositorio RUDAR.

4.3 Conversión de RDF de Redalyc y RUDAR

Hay que recordar que para poder elevar los datos a otro nivel semántico es necesario comenzar con la transformación a RDF de los archivos XML cosechados de Redalyc y RUDAR.

Las figuras siguientes muestran los ejemplos de los registros ahora convertidos a RDF, donde ya fueron incorporados los espacios de nombres adecuados y la estructura RDF reutilizando los metadatos Dublin Core provenientes de la cosecha.

```
▼<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dcterms="http://purl.org/dc/terms/" xmlns:dc="http://purl.org/dc/elements/1.1/"
 xmlns:foaf="http://xmlns.com/foaf/0.1/" xmlns:onto="http://dbpedia.org/ontology/">
 ▼<rdf:Description rdf:about="oai:redalyc.org:81522330006">
    <dc:subject>Cambio cultural</dc:subject>
    <dc:type>Artículo científico</dc:type>
    <dc:rights>Revista de Estudios Sociales</dc:rights>
    <dc:date>2011</dc:date>
    <dc:language>es</dc:language>
   ▶<dc:description>...</dc:description>
    <dc:identifier rdf:resource="http://www.redalyc.org/articulo.oa?id=81522330006"/>
    <dc:creator>María Cristina Tenorio</dc:creator>
    <dc:relation>http://www.redalyc.org/revista.oa?id=815</dc:relation>
    <dcterms:isPartOf rdf:resource="set:0123-885X"/>
    <dc:format>application_pdf</dc:format>
    <dc:publisher>Universidad de Los Andes</dc:publisher>
    <dc:subject>modelo cultural</dc:subject>
    <dc:source>Revista de Estudios Sociales (Colombia) Num.40</dc:source>
   ▼<dc:title>
      Escolaridad generalizada: ¿inclusión social o pérdida de la identidad cultural?
    </dc:title>
    <dc:subject>identidad cultural</dc:subject>
    <dc:subject>Sociología</dc:subject>
    <dcterms:modified rdf:resource="2012-08-27"/>
  </rdf:Description>
 ▼<rdf:Description rdf:about="oai:redalyc.org:81522307011">
    <dc:type>Artículo científico</dc:type>
    <dc:subject>comprensión</dc:subject>
   ▼<dc:title>
      Problemas del entorno y de la comunidad como fuentes de aprendizaje de la estadística
    </dc:title>
    <dc:subject>Sociología</dc:subject>
```

Figura 39. Ejemplo de salida RDF de Redalyc

```
▼<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dcterms="http://purl.org/dc/terms/" xmlns:dc="http://purl.org/dc/elements/1.1/"
 xmlns:foaf="http://xmlns.com/foaf/0.1/" xmlns:onto="http://dbpedia.org/ontology/">
 ▼<rdf:Description rdf:about="oai:rudar.ruc.dk:1800/12568">
    <dc:type>Thesis</dc:type>
    <dc:subject>Danmarks Radio</dc:subject>
    <dc:creator>Pedersen, Christian Grunert</dc:creator>
   ▶<dc:description>...</dc:description>
    <dc:subject>TV-Avisen</dc:subject>
    <dc:date>2013</dc:date>
    <dc:language>da_DK</dc:language>
    <dc:type>Dansk / Danish - Master thesis</dc:type>
    <dc:type>Journalistik / Journalism - Master thesis</dc:type>
    <dcterms:modified rdf:resource="2013-09-02"/>
    <dc:subject>Dansk</dc:subject>
    <dcterms:isPartOf rdf:resource="set:hdl_1800_53"/>
    <dc:subject>Journalistik</dc:subject>
   ▼<dc:title>
      Ned fra piedestalen - en undersøgelse af TV-Avisen 2000-2013
    </dc:title>
    <dc:subject>korrespondenter</dc:subject>
    <dc:identifier rdf:resource="http://rudar.ruc.dk/handle/1800/12568"/>
    <dc:subject>tv-nyheder</dc:subject>
  </rdf:Description>
 ▼<rdf:Description rdf:about="oai:rudar.ruc.dk:1800/12581">
    <dc:date>2013</dc:date>
   ▶<dc:description>...</dc:description>
    <dc:creator>Hviland-Jørgensen, Tobe</dc:creator>
    <dc:type>Datalogi / Computer Science - Master thesis</dc:type>
    <dc:language>en</dc:language>
```

Figura 40. Ejemplo de salida RDF de RUDAR

Para la transformación a RDF de los archivos obtenidos en la cosecha de metadatos se usó el convertidor, desarrollado para OntoOAI, llamado XML_DC2RDF. De este proceso se obtuvieron los siguientes resultados, que coinciden con los recursos cosechados previamente.

Repositorio	Total de archivos RDF obtenidos	Total recursos descritos
Redalyc	19,027	380,540
RUDAR	154	15,400
Totales	19,181	395,940

Tabla 13. Resultados del proceso de conversión a RDF

4.4 Enriquecimiento de autores y verificación del modelo de datos

Como se mencionó en el capítulo anterior, los datos de los autores son enriquecidos para agregar más información que aparece en otras fuentes de datos además de los repositorios cosechados.

Este enriquecimiento se realiza siguiendo la especificación *Friend of a Friend* y obteniendo los datos de la base de DBPedia, particularmente el conjunto de datos *PersonData* como se describió en los apartados anteriores.

Un ejemplo de coincidencias encontradas se muestra a continuación. Como se observa, el caso del autor "Edgar Morin", que en Redalyc tiene 4 artículos y cuyos datos fueron enriquecidos con los obtenidos de DBpedia, que corresponden a los datos que el autor tiene en su biografía presente Wikipedia.

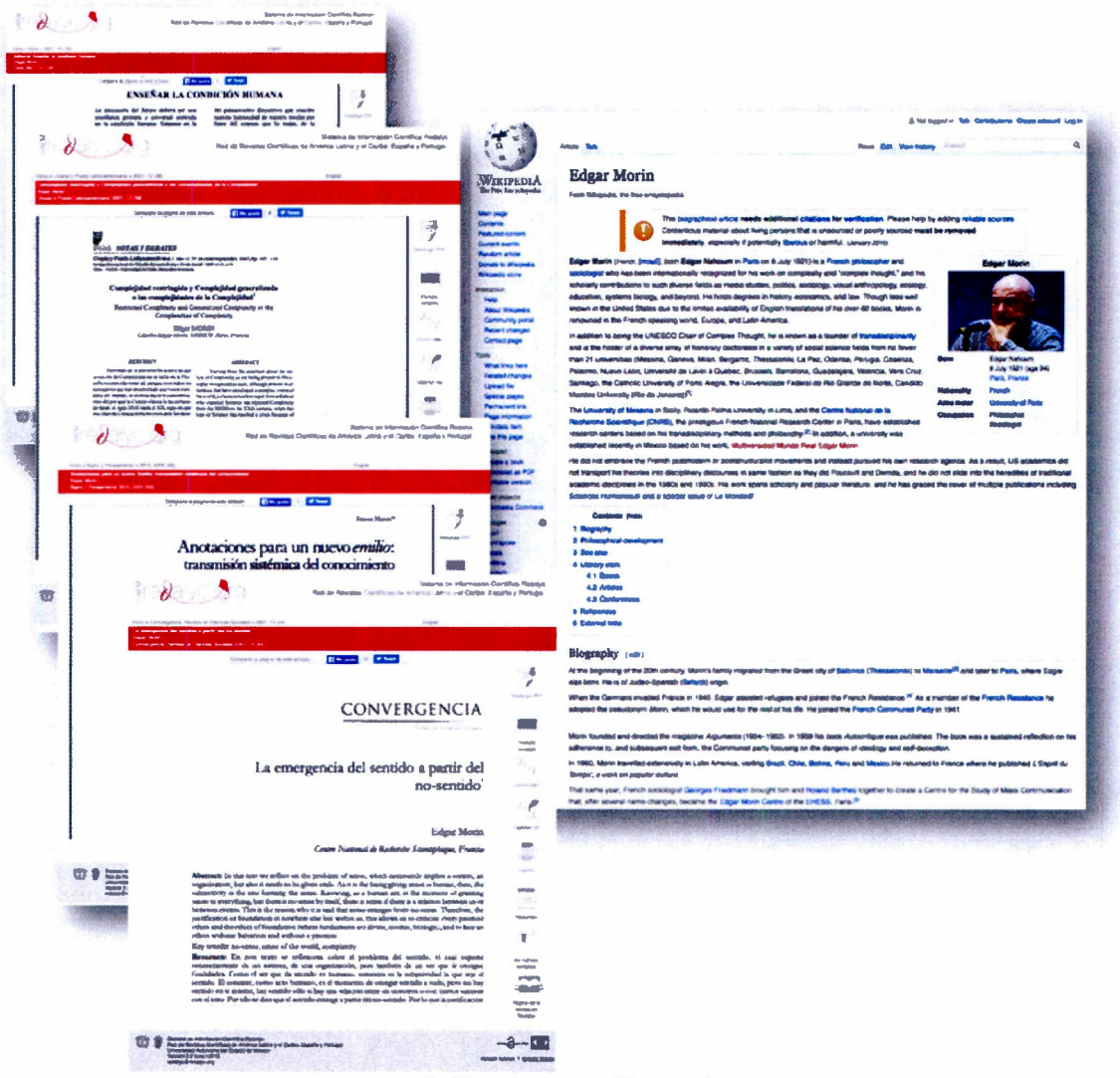

Figura 41. Ejemplo de coincidencias de autor

Entonces, para cada uno de los recursos que obtuvieron coincidencias autorales, y para cada uno de sus autores, los datos - descritos originalmente con Dublin Core - fueron enriquecidos con FOAF, como se muestra en el ejemplo de la figura 42.

```
▼<rdf:Description rdf:about="oai:redalyc.org:86019348002">
   <dc:rights>Signo y Pensamiento</dc:rights>
   <dc:subject>Comunicación</dc:subject>
   <dc:publisher>Pontificia Universidad Javeriana</dc:publisher>
   <dc:type>Artículo científico</dc:type>
   <dcterms:modified rdf:resource="2012-01-13"/>
   <dcterms:isPartOf rdf:resource="set:0120-4823"/>
   <dc:subject>Crisis cultural</dc:subject>
   <dc:identifier rdf:resource="http://www.redalyc.org/articulo.oa?id=86019348002"/>
   <dc:subject>Reforma universitaria</dc:subject>
   <dc:language>es</dc:language>
   <dc:source>Signo y Pensamiento (Colombia) Num.56 Vol.XXIX</dc:source>
  ▼<dc:title>
     Anotaciones para un nuevo Emilio: transmisión sistémica del conocimiento
   </dc:title>
   <dc:creator rdf:type="http://xmlns.com/foaf/0.1/Person" foaf:name="Edgar Morin" foaf:givenName="Edgar" foaf:surname="Morin"
   dc:description="French philosopher and sociologist" onto:birthDate="1921-07-08"/>
   <dc:relation>http://www.redalyc.org/revista.oa?id=860</dc:relation>
   <dc:format>application/pdf</dc:format>
   <dc:subject>Formación transdisciplinar</dc:subject>
   <dc:subject>Saber especializado</dc:subject>
  ▼<dc:description>
     La Universidad moderna nace de la reforma de Humboldt, quien introdujo las ciencias en plena expansión y contribuyó de manera
     decisiva a su desarrollo. Pero Humboldt era también consciente de que la Universidad debía proveer una cultura general, antes
     que una formación profesional especializada, para afrontar los problemas fundamentales y globales que se nos plantean en este
     siglo de manera cada vez más vital y urgente. Este principio cobra aún más vigencia ahora que vivimos una crisis cultural: la
     cultura de nuestro siglo necesita comunicación e intercambio entre el conocimiento científico y la cultura humanista. Sin
     embargo, en vez de irrigar como se debiera una competencia general, las disciplinas se encierran en sí mismas; la
     fragmentación del saber desintegra los problemas fundamentales y globales. Es entonces a esos problemas que se debería dedicar
     una nueva reforma de la Universidad, que no sería menos histórica que la de Humboldt. Ciertamente, cadauno de esos problemas
     requiere una competencia poli o trans-disciplinar (hay ejemplos), pero es justamente al afrontar esos problemas que se
     desarrolla esa competencia y se forman nuevos formadores.
   </dc:description>
   <dc:date>2010</dc:date>
 </rdf:Description>
```

Figura 42. Ejemplo de RDF enriquecido

El proceso de búsqueda de coincidencias se realizó con el nombre completo del autor, y con la coincidencia exacta. Está demás decir, que el problema de homonimia o polisemia (dos autores con el mismo nombre) y sinonímia (un autor publica con diferentes nombres) en los autores afecta los resultados de esta fase del proceso.

Por ello, se hace evidente que existe un área de oportunidad importante en la identificación única de personas en la web, particularmente de investigadores o autores de contenidos académico – científicos.

Sin duda este modelo puede ser mejorado, si los repositorios que exponen metadatos bajo el protocolo OAI-PMH hicieran uso de URI's desreferenciables, tanto para el caso de los recursos que describen como para los autores, sin embargo, ésta práctica no es común en los repositorios.

Aprovechar iniciativas como el identificador internacional de investigadores provisto por ORCID[12] resulta importante. Cabe mencionar que ORCID y Redalyc son interoperables ya, sin embargo, al momento ORCID no expone sus datos de autores para la Web Semántica como lo hace Wikipedia.

A este proceso de enriquecimiento de autores fueron sometidos todos los recursos cosechados por OAI-PMH de Redalyc y de RUDAR empleando un tiempo total de ejecución 36h34min.

Posterior a la fase de enriquecimiento cada uno de los RDF obtenidos, éstos fueron sometidos a un proceso de verificación con una aplicación en Jena para garantizar que cumplen con el modelo de datos descrito por la ontología DCFOAF_merged.owl descrita anteriormente.

Los resultados finales de esta etapa de enriquecimiento y verificación muestran (Tabla 14) que un 13.8% de recursos cosechados de los repositorios tuvieron al menos un autor coincidente con los individuos de *PersonData.* Haciendo un total de 60,927 autores con coincidencias. Lo cual da evidencia, de que hay recursos con más de un autor asociado.

---

[12] ORCID es un proyecto abierto, sin fines de lucro, comunitario, que ofrece un sistema para crear y mantener un registro único de investigadores y un método para vincular las actividades de investigación y los productos. Servicio disponible en http://orcid.org/

Algunos artículos no pudieron ser incluidos en los archivos RDF resultantes ya que no pasaron la verificación con respecto a la ontología, debido a que algunos recursos omitían metadatos esenciales, como título o fecha, o tenían algún problema de formación. La diferencia es clara, cuando se observa la comparación de lo obtenido en RDF (Tabla 13) y la salida de esta última fase (Tabla 14).

Repositorio	Recursos descritos	Recursos con autores en DBPedia	Autores localizados
Redalyc	380,534	52,251 (13.7%)	57,918
RUDAR	14,885	2,550 (17.13%)	3,009
Totales	395,419	54,801 (13.8%)	60,927

Tabla 14. Resultados del proceso de enriquecimiento de autores

A ese respecto se puede observar que 6 artículos de Redalyc y 515 recursos de RUDAR fueron excluidos de los archivos RDF.

4.5 Almacenamiento en la base OntoOAI

Como se explicó en el apartado 3.7, la carga de tripletas a TDB se realizó de forma masiva, y tomó 8.97hrs en total.

Los resultados arrojaron 7,917,081 tripletas almacenadas en el *triple store*. Esta base de conocimiento utiliza un espacio en disco de 522.54GB. La composición resultante de la base OntoOAI se muestra en la figura 43.

Figura 43. Composición de la base OntoOAI en TDB.

Resulta interesante observar la composición más detenidamente (Tabla 15), para conocer los datos que finalmente ingresaron al *triple store*, ya que algunos se quedaron fuera por algún error de formación, por ejemplo, de los 60,927 autores que fueron localizados con coincidencias en la fase de enriquecimiento sólo 60,354 fueron almacenados en la base de TDB.

De los 395,419 recursos resultado de la conversión a RDF, sólo 394,776 cuentan con un identificador, 379,966 con una fuente y 394,775 con fecha de publicación, por mencionar algunos.

Un dato que llama la atención es que fueron encontrados 1,600,540 *dc:subject*, es decir, un promedio de 4.04 palabras clave por recurso, que representan los temas de los que trata cada uno de ellos. Desafortunadamente, respecto a la descripción de temáticas hay una diversidad de nomenclaturas, formas de asignación e idiomas presentes en los repositorios, que impiden explotar mayormente esta información para identificar relaciones y tener una mayor semántica.

Propiedad	Tripletas
http://xmlns.com/foaf/0.1/Person>	60,354
<http://purl.org/dc/elements/1.1/identifier>	394,776
http://purl.org/dc/elements/1.1/source	379,966
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	60,354
http://purl.org/dc/elements/1.1/date	394,775
http://xmlns.com/foaf/0.1/name	60,354
http://purl.org/dc/elements/1.1/format	385,975
http://purl.org/dc/elements/1.1/description	361,577
http://purl.org/dc/terms/modified	394,772
http://dbpedia.org/ontology/birthDate	60,354
http://purl.org/dc/elements/1.1/type	405,463
http://purl.org/dc/elements/1.1/title	394,776
http://purl.org/dc/elements/1.1/publisher	382,412
http://purl.org/dc/terms/isPartOf	394,780
http://purl.org/dc/elements/1.1/subject	1,600,540
http://purl.org/dc/elements/1.1/rights	379,966
http://xmlns.com/foaf/0.1/surname	60,354
http://xmlns.com/foaf/0.1/givenName	60,354
http://purl.org/dc/elements/1.1/relation	381,980
http://purl.org/dc/elements/1.1/creator	968,903
http://purl.org/dc/elements/1.1/language	394,650

Tabla 15. Resultados de la carga masiva al *triple store*.

4.6 Perfil de usuario

Tal como se mencionó en el apartado 3.8, la sensibilidad al contexto con la que cuenta OntoOAI se basa en contar con un perfil de usuario que describa sus datos académicos acorde a la ontología OntoOAIEstudiante.

Para describir el contexto de los usuarios, usados en el caso de estudio, se definieron dos perfiles. El primero identificado con "itesm:A01210238"; identificador que se compone de las siglas de la escuela y el número de matrícula del alumno.

Este primer perfil (figura 44) describe al alumno "Rafael Gómez García", alumno del "ITESM" del programa de licenciatura en "Sociología", que cursa tres asignaturas: "Movimientos Sociales", "Epistemología", "Sociedades Modernas", cuyas preferencias de interacción apuntan a "Artículos" y "Videos", sus idiomas "es", "en", y "MX" como dato demográfico.

```
▼<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dcterms="http://purl.org/dc/terms/" xmlns:dc="http://purl.org/dc/elements/1.1/"
 xmlns:foaf="http://xmlns.com/foaf/0.1/" xmlns:onto="http://dbpedia.org/ontology/"
 xmlns:student="http://www.semanticweb.org/arianna/ontologies/2016/3/OntoOAIEstudiante">
  ▼<rdf:Description rdf:about="itesm:A01210238">
     <student:name rdf:type="http://xmlns.com/foaf/0.1/Person" foaf:name="Rafael Gómez
     García" foaf:givenName="Rafael" foaf:surname="Gómez García"/>
     <student:programa>Sociología</student:programa>
     <student:nivelPrograma>Licenciatura</student:nivelPrograma>
     <student:escuela>ITESM</student:escuela>
     <student:cursoActual>Movimientos Sociales</student:cursoActual>
     <student:cursoActual>Epistemología</student:cursoActual>
     <student:cursoActual>Sociedades modernas</student:cursoActual>
     <student:intereses>Kant</student:intereses>
     <student:intereses>Teoría del conocimiento</student:intereses>
     <student:formatoPresentacion>Artículo</student:formatoPresentacion>
     <student:formatoPresentacion>Video</student:formatoPresentacion>
     <student:idiomaMaterno>es</student:idiomaMaterno>
     <student:idiomaPreferido>en</student:idiomaPreferido>
     <student:datosDemograficos>MX</student:datosDemograficos>
   </rdf:Description>
 </rdf:RDF>
```

Figura 44. Perfil de usuario itesm:A01210238

El segundo perfil (figura 45) corresponde a una estudiante alumna de una universidad alemana identificada con "univ:DL9510078", de nombre "Margret Fintz", inscrita en un programa de "Master" en "Pedagogy Educational Studies" que cursa la asignatura "Humanism and Pedagogy" y que está interesada en los temas "Social apprenticeship" y "Johan Friedrich Herbart". Además esta alumna prefiere estudiar en recursos de tipo "article" y "thesis", sus idiomas son "de", "en" y "es" sus datos geográficos son "DE".

```
▼<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dcterms="http://purl.org/dc/terms/" xmlns:dc="http://purl.org/dc/elements/1.1/"
 xmlns:foaf="http://xmlns.com/foaf/0.1/" xmlns:onto="http://dbpedia.org/ontology/"
 xmlns:student="http://www.semanticweb.org/arianna/ontologies/2016/3/OntoOAIEstudiante">
  ▼<rdf:Description rdf:about="univ:DL9510078">
     <student:name rdf:type="http://xmlns.com/foaf/0.1/Person" foaf:name="Margret Fintz"
     foaf:givenName="Margret" foaf:surname="Fintz"/>
     <student:programa>Pedagogy Educational Studies</student:programa>
     <student:nivelPrograma>Master</student:nivelPrograma>
     <student:escuela>Deutsche Universität</student:escuela>
     <student:cursoActual>Humanism and Pedagogy</student:cursoActual>
     <student:intereses>Social apprenticeship</student:intereses>
     <student:intereses>Johann Friedrich Herbart</student:intereses>
     <student:formatoPresentacion>article</student:formatoPresentacion>
     <student:formatoPresentacion>thesis</student:formatoPresentacion>
     <student:idiomaMaterno>de</student:idiomaMaterno>
     <student:idiomaPreferido>en</student:idiomaPreferido>
     <student:idiomaPreferido>es</student:idiomaPreferido>
     <student:datosDemograficos>DE</student:datosDemograficos>
   </rdf:Description>
 </rdf:RDF>
```

Figura 45. Perfil de usuario univ:DL9510078

4.7 Resultados de inferencia

Una vez que todas las fases anteriores se llevaron a cabo y se tiene el *triple store* preparado para el razonador, se usaron los perfiles de usuario previamente definidos y se documentaron un par de ejercicios de descubrimiento selectivo de conocimiento.

4.7.1 Ejemplo 1 de descubrimiento selectivo de conocimiento

El experimento mostrado en este apartado se realizó tomando en consideración al usuario con el perfil "itesm:A01210238". Se probó una búsqueda sobre "complejidad", introduciendo como tal el término por el usuario, y es aquí donde se comienza a probar OntoOAI.

Primeramente, es importante recordar que la búsqueda es sensible al contexto, lo cual quiere decir que si un alumno de informática o biología realiza la misma búsqueda los resultados serán diferentes, dado que la "complejidad" en cómputo se aborda desde otro punto de vista diferente al sociológico.

Ahora se explicará el proceso que se efectuó para obtener dicho resultado. Primeramente, se ejecuta una búsqueda basada en coincidencias que arroja como resultado un artículo científico llamado "La emergencia del sentido a partir del no-sentido" del autor "Edgar Morin", un filósofo y sociólogo francés (esta última información obtenida de Wikipedia junto con otros datos del autor) publicado en la revista "Convergencia" en el "2007" y cuyos temas (descritos como *dc:subject*) son: "sentido del mundo", "el no-sentido", "Sociología", "complejidad". Este recurso se describe en la figura siguiente:

```
▼<rdf:Description rdf:about="oai:redalyc.org:10504408">
   <dc:language>es</dc:language>
   <dc:rights>Convergencia. Revista de Ciencias Sociales</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"
   foaf:givenName="Edgar" foaf:surname="Morin" dc:description="French philosopher and sociologist"
   onto:birthDate="1921-07-08"/>
   <dc:relation>http://www.redalyc.org/revista.oa?id=105</dc:relation>
  ▶<dc:description>...</dc:description>
   <dc:subject>Sociología</dc:subject>
   <dc:date>2007</dc:date>
   <dc:subject>complejidad</dc:subject>
 </rdf:Description>
```

Figura 46. Resultado de búsqueda directa (ejemplo 1)

Se puede observar claramente que, el artículo científico se recuperó debido a que trata un tema de "Sociología" que coincide con el nombre del programa al que el alumno está inscrito, además de que trata, entre otros, del tema de "complejidad". A partir de este recurso, que se considera de interés para el usuario ya que coincide con el término de búsqueda y su perfil, se activa el proceso de inferencia.

El diagrama mostrado en la figura 47, muestra el razonamiento detrás de la salida mostrada al usuario. Como se explicó previamente, el recurso "oai:redalyc.org:10504408", referente al artículo llamado "La emergencia del sentido a partir del no-sentido", es

localizado debido a las coincidencias en los temas que trata, que concuerdan con el tema de "complejidad" y encaja con el programa del usuario expresado en su perfil, en este caso "Sociología". El razonamiento sigue el silogismo, que se explica a continuación:

Si (el recurso "oai:redalyc.org:10504408" es interesante para el usuario),

Entonces (el autor "Edgar Morin" es interesante para el usuario).

Si (el autor "Edgar Morin" es interesante para el usuario),

Entonces (otros trabajos T_a cuyo autor, tema o título correspondan a "Edgar Morin" son interesantes para el usuario).

Si (los trabajos T_a son interesantes para el usuario),

Entonces (el conjunto de autores o colaboradores B de los recursos T_a es interesante para el usuario).

Si (el conjunto de autores B de los recursos T_a es interesante para el usuario),

Entonces (el conjunto de recursos T_b escritos por los autores B son interesantes para el usuario),

Donde:

r = "oai:redalyc.org:10504408",

a = "Edgar Morin",

Ta={"oai:redalyc.org:8601934800","oai:redalyc.org:27903809","oai:redalyc.org:49615201","oai:rudar.ruc.dk:1800/14983, "oai:rudar.ruc.dk:1800/26454"},

B = {"Michel Olsen", "Anders Sondergaard Skovbakke", "Matias Find", "Mads Christiansen", "Allan Dreyer Hansen"}

T_b={"oai:rudaruc.dk:1800/12794","oai:rudaruc.dk:1800/24877","oai:rudaruc.dk:1800/16477","oai:rudaruc.dk:1800/25805","oai:rudaruc.dk:1800/25386","oai:rudaruc.dk:1800/23930","oai:rudaruc.dk:1800/15047","oai:rudaruc.dk:1800/7734","oai:rudar

uc.dk:1800/18163","oai:rudaruc.dk:1800/18070","oai:rudaruc.dk:1800/18892","oai:rudaruc.dk:1800/13812","oai:rudaruc.dk:1800/24240","oai:rudaruc.dk:1800/25873","oai:rudaruc.dk:1800/15987","oai:rudaruc.dk:1800/4230","oai:rudaruc.dk:1800/4231","oai:rudaruc.dk:1800/24002","oai:rudaruc.dk:1800/6878","oai:rudaruc.dk:1800/24581","oai:rudaruc.dk:1800/17313","oai:rudaruc.dk:1800/3831","oai:rudaruc.dk:1800/3830","oai:rudaruc.dk:1800/3970","oai:rudaruc.dk:1800/6613","oai:rudaruc.dk:1800/23967","oai:rudaruc.dk:1800/14741","oai:rudaruc.dk:1800/23613","oai:rudaruc.dk:1800/14129","oai:rudaruc.dk:1800/23282","oai:rudaruc.dk:1800/17645","oai:rudaruc.dk:1800/19221","oai:rudaruc.dk:1800/16704","oai:rudaruc.dk:1800/23015"}

Así el conjunto de recursos contenidos en los conjuntos T_a y T_b conforman la salida para el usuario, es decir el conocimiento descubierto acorde a su búsqueda y su contexto.

$$\textit{Conocimiento descubierto} = T_b \cup T_b$$

Para el ejemplo fueron considerados sólo los recursos descubiertos de los autores de la tesis "oai:rudaruc.dk:1800/26454", que sumaron un total de 34, después de eliminar los duplicados ya que varios recursos fueron escritos en coautoría por algún autor de ese mismo grupo. Por motivos de visualización no fueron incluidos los artículos del autor Michel Olsen, autor de "oai:rudaruc:dk:1800/14983", que equivalen a 25 recursos adicionales.

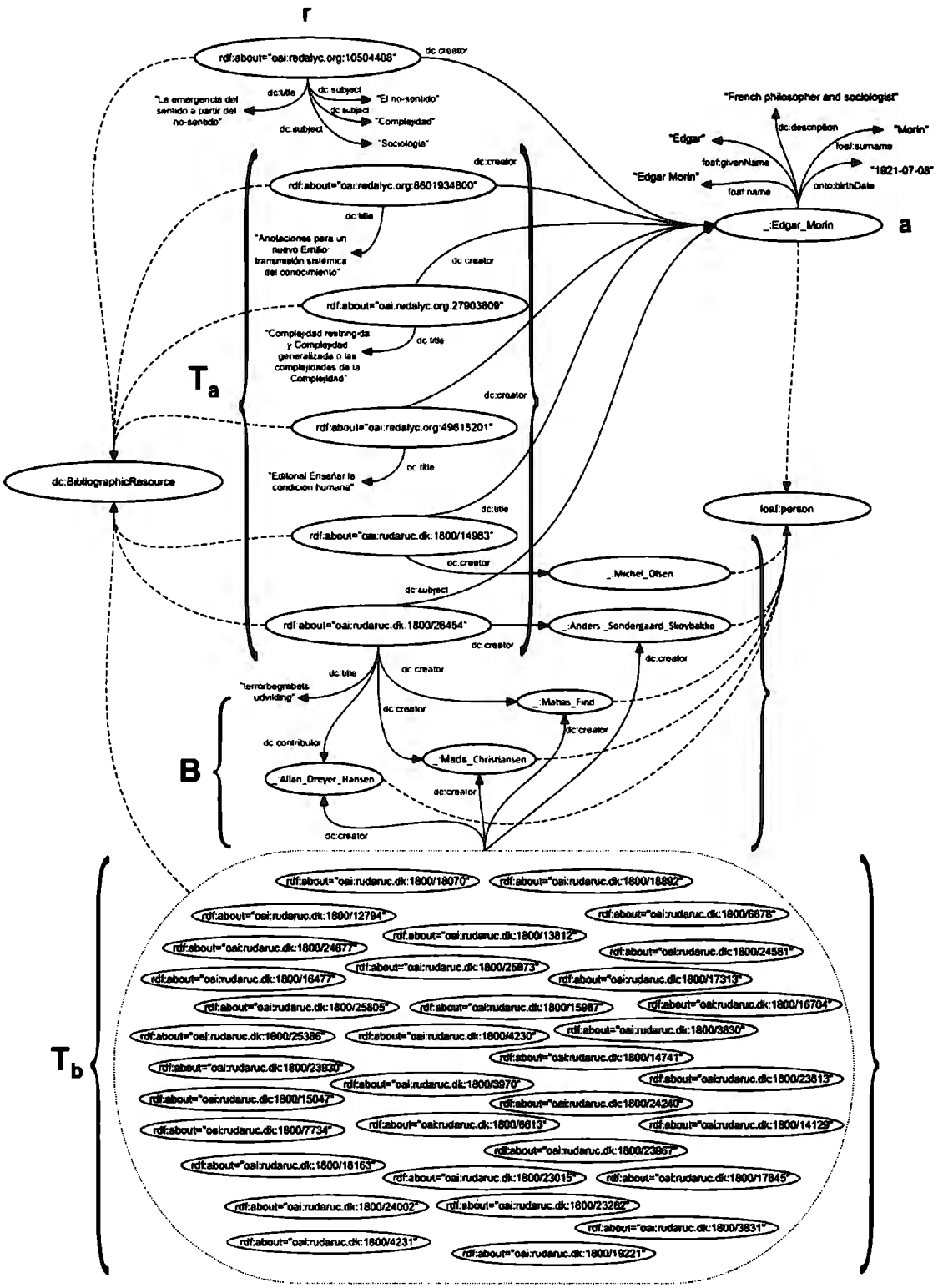

Figura 47. Descubrimiento selectivo de conocimiento del ejemplo 1

Un extracto de los resultados obtenidos para el usuario son los que se muestran en la tabla siguiente. El listado de la salida completa del ejemplo se encuentra en el Apéndice A.

1	La emergencia del sentido a partir del no-sentido	Morin, Edgar
2	Anotaciones para un nuevo Emilio: transmisión sistémica del conocimiento	Morin, Edgar
3	Complejidad restringida y Complejidad generalizada o las complejidades de la Complejidad	Morin, Edgar
4	Editorial Enseñar la condición humana	Morin, Edgar
5	Edgar Morin: La Connaissance de la Connaissance	Olsen, Michel
6	Terrorbegrebets udvikling	Skovbakke, Anders Søndergaard; Find, Matias; Christiansen, Mads; Advisor: Hansen, Allan Dreyer
7	Adgang for asylansøgere til at arbejde og bo uden for asylcentre – virkelighed eller utopi? En undersøgelse af hvad den nye ændring i det danske asylsystem indebærer, samt en vurdering af de effekter, som ændringen medfører for asylansøgere og det danske samfund. Og i forlængelse af den vil specialet give en diskussion af i hvor høj grad den nye regerings politik er et brud med den tidligere diskurs om asylansøgere.	Hansen, Allan Dreyer; Mardale, Andreea
8	PlayDay - Afprøvning af en kampagneplakat	Bjerre Ryhede, Annette; Clausen, Amalie; Nielsen, Mikkel; Rau-Andersen, René; Christiansen, Mads;Flensborg, Cecilie
9	Samspillet mellem minister og departementschef anno 2014	Dam Scheuer, John; Christiansen, Mads;Lachenmeier, Anders; Risbøl Jacobsen, Sebastian;Andreasen, Patrick
10	Etnisk diversitet i Danmark - i spændingsfeltet mellem konflikt og kontakt	Hansen, Allan Dreyer; Andersen, Sandra Louise;Jensen, Tilla Slåen; Christensen, Mathias Boberg;Hansen, Line Tauber; Würtzenfeld, Mathias

Tabla 16. Extracto de resultados del ejemplo 1 de descubrimiento selectivo de conocimiento.

Si bien en este trabajo no se plantea la exploración de formas de visualización para los resultados obtenidos, cabe resaltar, que dadas las asociaciones presentes entre los recursos es posible tomar ventaja de visualizaciones tipo grafo o jerárquicas.

Tales formas de presentar la información podrían permitir al usuario explorar a diferentes niveles los resultados de las inferencias y discriminar la información con base en las relaciones presentes. Dos ejemplos se muestran a continuación.

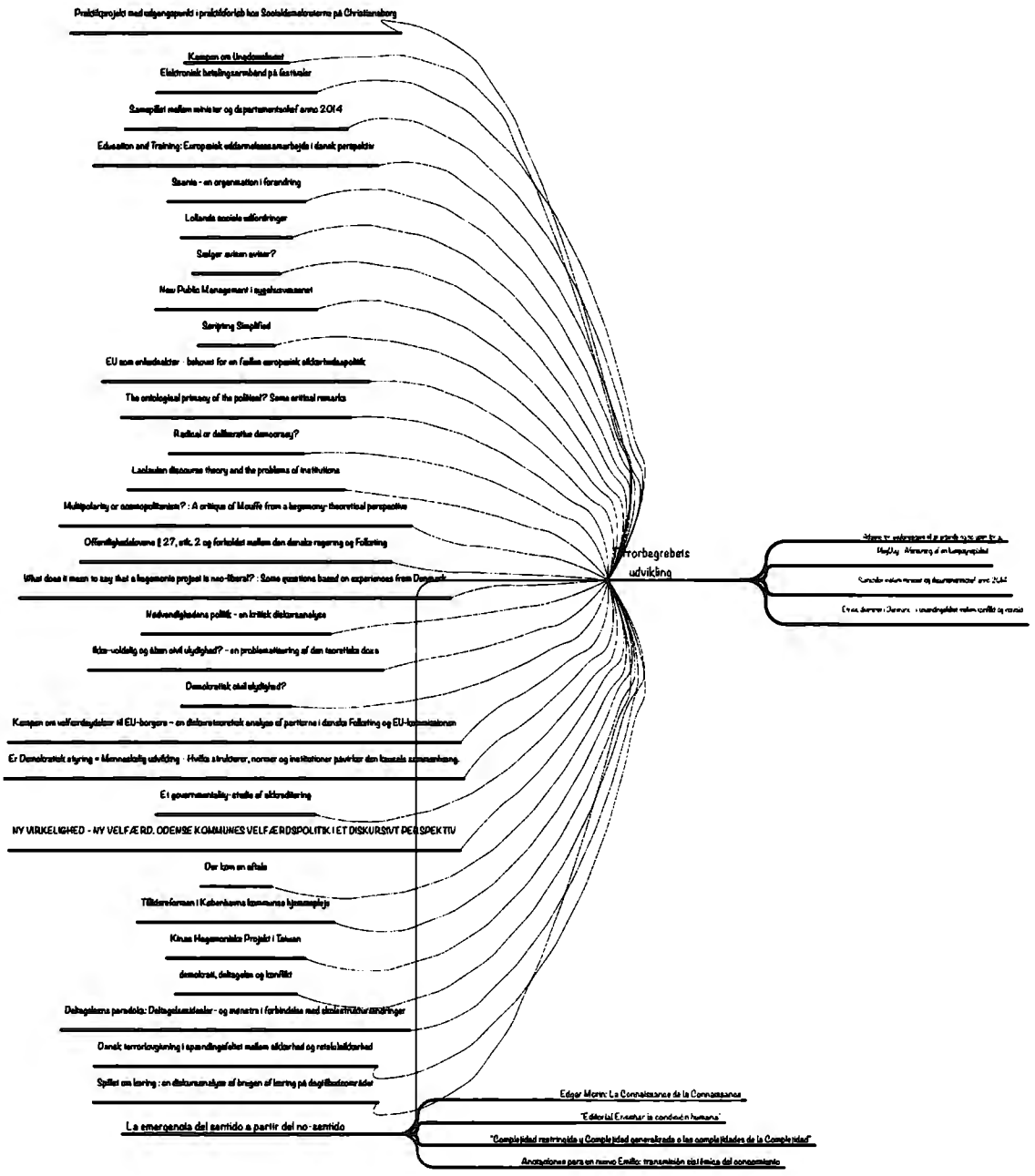

Figura 48. Visualización jerárquica de los resultados del ejemplo 1

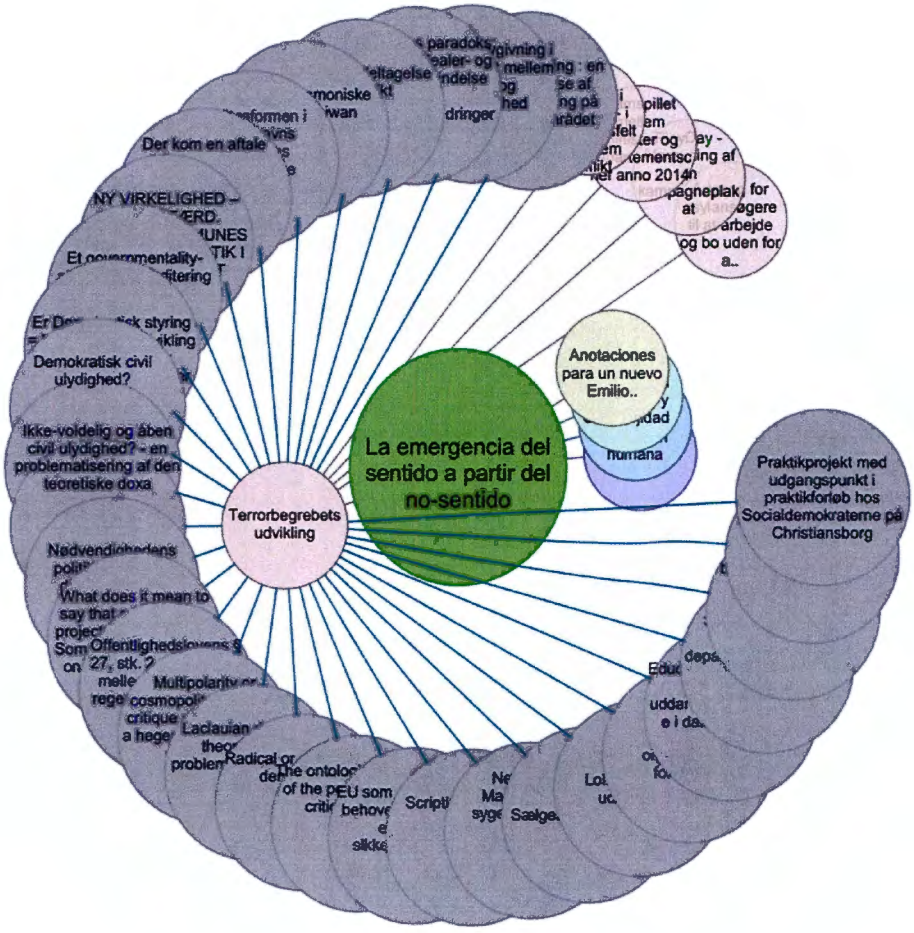

Figura 49. Visualización tipo grafo de los resultados del ejemplo 1

4.7.2 Ejemplo 2 de descubrimiento selectivo de conocimiento

Este ejercicio servirá para ejemplificar el proceso de razonamiento con otro perfil de usuario, el identificado con "univ:DL9510078", una estudiante cuyo contexto de preferencias e intereses ronda en los siguientes atributos de su perfil: "Pedagogy Educational Studies", "Humanism and Pedagogy", "Social apprenticeship", "Johan Friedrich Herbart", "article", "thesis", "German", "English", "Spanish", "DE", entre otros.

La alumna realiza una búsqueda sobre "non-violence". Considerando éste término a buscar y su información de perfil, se encuentra una tesis "oai:rudar.ruc.dk:1800/2990" de la

autoría de "Uski, Juha Janne Olavi" que corresponde justamente al programa, al cual está inscrita la estudiante "Pedagogy Educational Studies" y que trata entre otros temas el de "non-violence", publicado el "2008-01-17", los datos completos se muestran en el RDF siguiente:


```
▼<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2016-04-07T17:58:49Z</responseDate>
  <request metadataPrefix="oai_dc" verb="GetRecord"
  identifier="oai:rudar.ruc.dk:1800/2990">http://diggy.ruc.dk:8080/dspace-oai/request</request>
 ▼<GetRecord>
  ▼<record>
   ▼<header>
     <identifier>oai:rudar.ruc.dk:1800/2990</identifier>
     <datestamp>2010-03-03T01:16:17Z</datestamp>
     <setSpec>hdl_1800_118</setSpec>
     <setSpec>hdl_1800_60</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>Education for Active Non-violence</dc:title>
      <dc:creator>Uski, Juha Janne Olavi</dc:creator>
      <dc:subject>Silo</dc:subject>
      <dc:subject>Freire</dc:subject>
      <dc:subject>Greenleaf</dc:subject>
      <dc:subject>Wenger</dc:subject>
      <dc:subject>Lave</dc:subject>
      <dc:subject>nonviolence</dc:subject>
      <dc:subject>non-violence</dc:subject>
      <dc:subject>Mumbai</dc:subject>
      <dc:subject>Humanist Movement</dc:subject>
      <dc:subject>leadership</dc:subject>
      <dc:subject>apprenticeship</dc:subject>
      <dc:subject>humanism</dc:subject>
     ▼<dc:description>
        Education for Active Non-violence examines the ideas and the practice of the Humanist Movement, an
        international movement against varied forms of violence. The research focuses on the process of learning
        through participation in the Humanist Movement in Mumbai, India, and is especially directed towards
        insights about the challenges to that educational practice. The field study involves interviews of nine
        members with varied kinds of participation in the movement. Theories by Robert K. Greenleaf, Silo, Paulo
        Freire, Jean Lave and Etienne Wenger are applied for gaining perspective on the discussions of the case
        study; research papers about the social situation in Mumbai are also referred to. Case study research
        methodology is based on semi-structured interviews analysed through a combination of interactionist and
        hermeneutic methods.
      </dc:description>
      <dc:contributor>Carney, Stephen</dc:contributor>
      <dc:date>2008-01-17</dc:date>
      <dc:date>2008-02-01T07:06:48Z</dc:date>
      <dc:date>2008-02-01T07:06:48Z</dc:date>
      <dc:date>2008-02-01T07:06:48Z</dc:date>
      <dc:type>Thesis</dc:type>
     ▼<dc:type>
        Pædagogik og uddanneleesstudier / Pedagogy Educational Studies - not master thesis
      </dc:type>
      <dc:identifier>http://hdl.handle.net/1800/2990</dc:identifier>
      <dc:language>en</dc:language>
     </oai_dc:dc>
    </metadata>
   </record>
  </GetRecord>
 </OAI-PMH>
```


Figura 50. Resultado de búsqueda directa (ejemplo 2)

Dicha tesis fue localizada como resultado de la búsqueda, debido a que está clasificada con el tipo "Pedagogy Educational Studies" que coincide con el nombre del programa en el cual la alumna está inscrita, además de que trata, entre otros, del tema de "non-violence". Es así que este recurso, resultado de una estrategia de búsqueda sensible al contexto, se considera de interés para la usuaria.

Tomando ésta última, como premisa inicial de la inferencia, se da paso al proceso de razonamiento deductivo, basado en el silogismo:

Si (el recurso "oai:rudar.ruc.dk:1800/2990" es interesante para el usuario),

Entonces (los autores "Stephen Carney" y "Juha Janne Olavi Uski" son interesantes para el usuario).

Si (los autores "Stephen Carney" y "Juha Janne Olavi Uski" son interesantes para el usuario),

Entonces (otros trabajos T_a cuyo autor, tema o título correspondan a "Stephen Carney" o a "Juha Janne Olavi Uski" son interesantes para el usuario).

Si (los trabajos T_a son interesantes para el usuario),

Entonces (el conjunto de autores o colaboradores B de los recursos T_a es interesante para el usuario).

Si (el conjunto de autores B de los recursos T_a es interesante para el usuario),

Entonces (el conjunto de recursos T_b escritos por los autores B son interesantes para el usuario),

Donde:

r = "oai:rudar.ruc.dk:1800/2990",

A = {"Stephen Carney", "Juha Janne Olavi Uski"}

Ta={"oai:rudaruc.dk:1800/26144","oai:rudaruc.dk:1800/11847","oai:rudaruc.dk:18
00/4974","oai:rudaruc.dk:1800/9210","oai:rudaruc.dk:1800/1776","oai:rudaruc.dk:1
800/1781","oai:rudaruc.dk:1800/25922","oai:rudaruc.dk:1800/11460","oai:rudaruc.
dk:1800/6474","oai:rudaruc.dk:1800/4975","oai:rudaruc.dk:1800/6728","oai:rudaru
c.dk:1800/14135","oai:rudaruc.dk:1800/13337","oai:rudaruc.dk:1800/2084","oai:ru
daruc.dk:1800/5389","oai:rudaruc.dk:1800/1331","oai:rudaruc.dk:1800/4041","oai:r
udaruc.dk:1800/2738"},

|Ta| = 19,

T_b={"oai:rudaruc.dk:1800/2084","oai:rudaruc.dk:1800/2738","oai:rudaruc.dk:1800/
5389","oai:rudaruc.dk:1800/1331","oai:rudaruc.dk:1800/2618","oai:rudaruc.dk:180
0/1363","oai:rudaruc.dk:1800/4094","oai:rudaruc.dk:1800/4843","oai:rudaruc.dk:18
00/4041","oai:rudaruc.dk:1800/3815","oai:rudaruc.dk:1800/8339","oai:rudaruc.dk:1
800/8342","oai:rudaruc.dk:1800/8642","oai:rudaruc.dk:1800/2964","oai:rudaruc.dk:
1800/9257","oai:rudaruc.dk:1800/8990","oai:rudaruc.dk:1800/9833","oai:rudaruc.d
k:1800/6220","oai:rudaruc.dk:1800/6508","oai:rudaruc.dk:1800/2350","oai:rudaruc.
dk:1800/6426","oai:rudaruc.dk:1800/7025","oai:rudaruc.dk:1800/7541","oai:rudaru
c.dk:1800/8671","oai:rudaruc.dk:1800/8682","oai:rudaruc.dk:1800/8684","oai:rudar
uc.dk:1800/8687","oai:rudaruc.dk:1800/8705","oai:rudaruc.dk:1800/8589","oai:rud
aruc.dk:1800/9572","oai:rudaruc.dk:1800/6449","oai:rudaruc.dk:1800/1883","oai:ru
daruc.dk:1800/995","oai:rudaruc.dk:1800/2276","oai:rudaruc.dk:1800/7295","oai:ru
daruc.dk:1800/6544","oai:rudaruc.dk:1800/6419","oai:rudaruc.dk:1800/7142","oai:r
udaruc.dk:1800/1898","oai:rudaruc.dk:1800/4861","oai:rudaruc.dk:1800/2598","oai
:rudaruc.dk:1800/9036","oai:rudaruc.dk:1800/2854","oai:rudaruc.dk:1800/6672","o
ai:rudaruc.dk:1800/1109","oai:rudaruc.dk:1800/6441","oai:rudaruc.dk:1800/6015","

oai:rudaruc.dk:1800/7791","oai:rudaruc.dk:1800/7460","oai:rudaruc.dk:1800/10003","oai:rudaruc.dk:1800/9740","oai:rudaruc.dk:1800/4267","oai:rudaruc.dk:1800/3905","oai:rudaruc.dk:1800/501","oai:rudaruc.dk:1800/10442","oai:rudaruc.dk:1800/3307","oai:rudaruc.dk:1800/5900","oai:rudaruc.dk:1800/9154","oai:rudaruc.dk:1800/8699","oai:rudaruc.dk:1800/1779","oai:rudaruc.dk:1800/3151","oai:rudaruc.dk:1800/6540","oai:rudaruc.dk:1800/6826","oai:rudaruc.dk:1800/1822","oai:rudaruc.dk:1800/7779","oai:rudaruc.dk:1800/8597","oai:rudaruc.dk:1800/9378","oai:rudaruc.dk:1800/1487","oai:rudaruc.dk:1800/2254","oai:rudaruc.dk:1800/5941","oai:rudaruc.dk:1800/9152","oai:rudaruc.dk:1800/913","oai:rudaruc.dk:1800/1576","oai:rudaruc.dk:1800/3087","oai:rudaruc.dk:1800/6171","oai:rudaruc.dk:1800/2014","oai:rudaruc.dk:1800/7516","oai:rudaruc.dk:1800/6270","oai:rudaruc.dk:1800/3729","oai:rudaruc.dk:1800/4669","oai:rudaruc.dk:1800/7667","oai:rudaruc.dk:1800/7633","oai:rudaruc.dk:1800/8985","oai:rudaruc.dk:1800/10225","oai:rudaruc.dk:1800/3143","oai:rudaruc.dk:1800/1606","oai:rudaruc.dk:1800/3083","oai:rudaruc.dk:1800/2271","oai:rudaruc.dk:1800/6663","oai:rudaruc.dk:1800/8697","oai:rudaruc.dk:1800/8701","oai:rudaruc.dk:1800/8702","oai:rudaruc.dk:1800/8698","oai:rudaruc.dk:1800/8700","oai:rudaruc.dk:1800/9097","oai:rudaruc.dk:1800/4240","oai:rudaruc.dk:1800/4204","oai:rudaruc.dk:1800/10122","oai:rudaruc.dk:1800/7425","oai:rudaruc.dk:1800/2807","oai:rudaruc.dk:1800/1653","oai:rudaruc.dk:1800/1986","oai:rudaruc.dk:1800/2458","oai:rudaruc.dk:1800/6029","oai:rudaruc.dk:1800/8703","oai:rudaruc.dk:1800/6534","oai:rudaruc.dk:1800/6982","oai:rudaruc.dk:1800/8261","oai:rudaruc.dk:1800/9061","oai:rudaruc.dk:1800/4249"}

|Ta| = 110,

Finalmente, el conjunto de recursos contenidos en la unión de los conjuntos T_a y T_b, que suman 129, conforman la salida para la usuario, es decir el conocimiento descubierto acorde a su búsqueda y su contexto.

$$\textit{Conocimiento descubierto} = T_b \cup T_b$$

$$| T_b \cup T_b | = 129$$

El diagrama mostrado en la figura 50, muestra con grafos la relación entre los recursos encontrados y muestra los elementos r, a, T_a, T_b y B que son mencionados en el proceso deductivo.

La totalidad de los autores que componen el conjunto B (autores de los recursos T_a) no son mostrados en la figura por motivos de espacio en la visualización, son en total 65 autores contando a Jeremy Rappleye e Iveta Silova que si aparecen en el diagrama.

De igual manera los recursos de T_b sólo son indicados con nodos vacíos, sin embargo el listado completo fue mostrado anteriormente.

La tabla 16 muestra una parte de los resultados obtenidos para el usuario. El listado completo de la salida completa del ejemplo se encuentra en el Apéndice B.

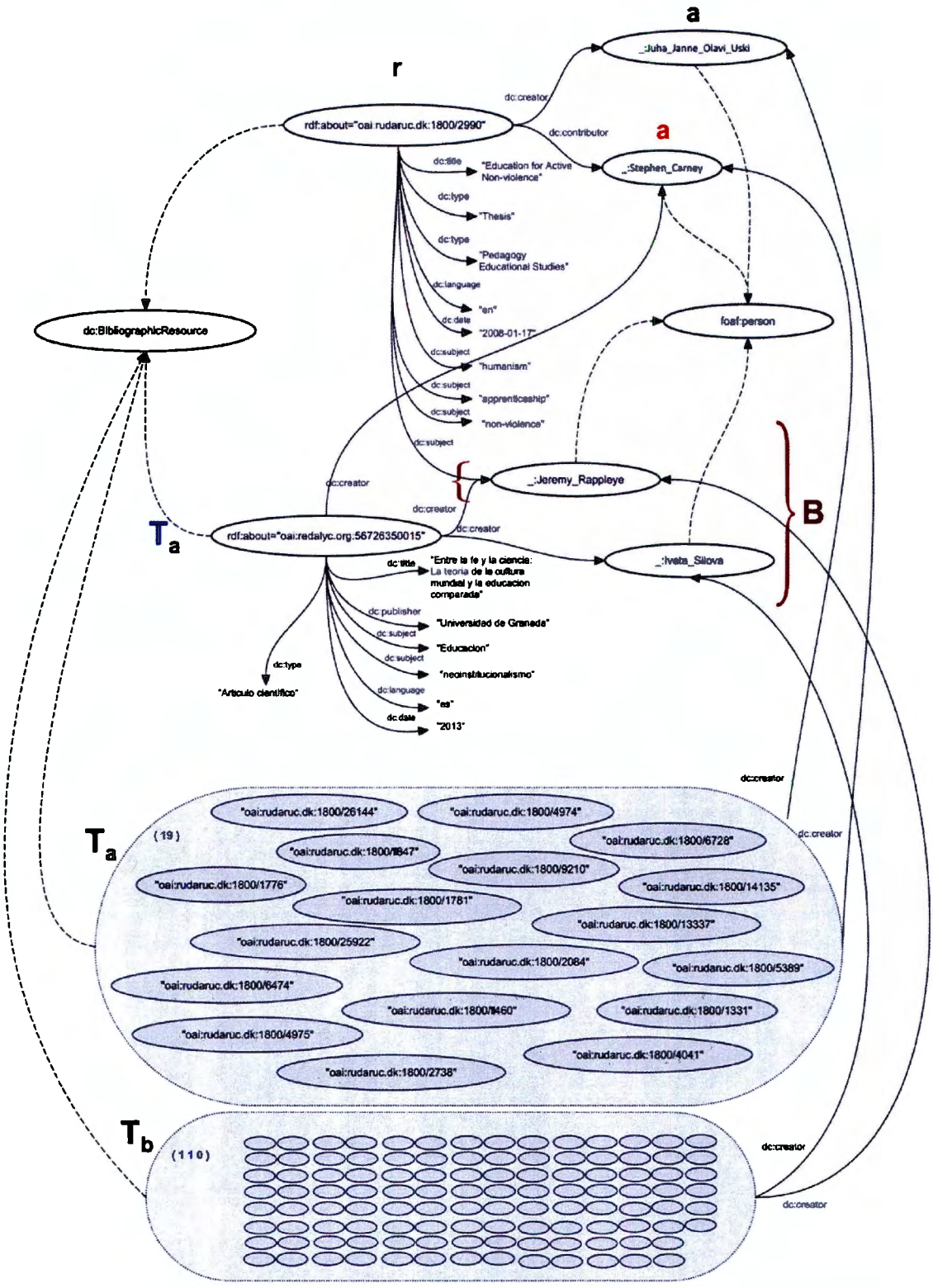

Figura 51. Descubrimiento selectivo de conocimiento del ejemplo 2

1	Entre la fe y la ciencia: La teoría de la cultura mundial y la educación comparada	Stephen Carney ; Jeremy Rappleye ; Iveta Silova ;
2	The Changing Environment of Development: From Aid to Trade	Carney, Stephen; Kehlet Hansen, Jesper Peter
3	The Decision - Youth and the Negotiation Between Choices	Carney, Stephen; Myssen, Martin; Nisted, Nina;Zmylon, Nanna Nielsen; Blomsterberg, Sofie Amalie;Birkemose, Liv; Falk, Nicklas; Pedersen, Aryono Daniel Ingemann; Christensen, Andrea Bang
4	Development through Entrepreneurship : microcredit, learning and the persistence of poverty amongst women in urban Tanzania	Sigalla, Rachel; Carney, Stephen; The Department of Psychology and Educational Studies
5	Education for Development Is Culturally Sensitive Development a Realistic Alternative to the Current Model of Development?	Carney, Stephen; Folman, Edlira; Jasinskaite, Jorune;Stræde Meiner, Mathilde; Libsekal; Wandel, Oscar;Andersen, Signe; Persson, Tina
6	Education in Flux	Carney, Stephen; Damary, Talia; Gyldenkærne, Rasmus W.; Sandvad, Nicklas; Schiøler, Maja; Yigit, Mustafa U.
7	Globalization and education	Carney, Stephen; Bak, Andreas H.
8	Micro-Finance and Poverty Alleviation in Cameroon	Carney, Stephen; Adangwi, John Atie
9	Nationalism in the New Middle East	Carney, Stephen; Heimisson, Gunnar Örn
10	Planning Process Drama - Can Drama and Learning Styles improve Education?	Carney, Stephen; Kristiansen, Anders Emil Wulff;Mottelson, Ida Nygaard; Larsen, Signe Illum Honoré

Tabla 17. Extracto de resultados del ejemplo 2 de descubrimiento selectivo de conocimiento.

4.8 Problemas encontrados

En concordancia con lo que menciona Eric Franzon, la heterogeneidad, ruido y falta de identificadores e idiosincrasias compartidas de los datos en la web es extrema (Franzon, 2014).

Sin lugar a dudas, el hecho de no contar con identificadores únicos en los sitios web resulta un problema en la integración de datos; si bien, es una buena práctica la implementación de URI's, no todos los repositorios las implementan, más aún porque resulta polémica la cuestión de qué sitio deba implementar identificadores únicos, ya que la asignación de ellas debería buscarla el sitio web "creador" del material. Para el caso de los repositorios institucionales resulta clara su función como administradores de su propia producción, sin embargo, para el caso de repositorios agregadores aún el panorama no es claro.

Todavía más interesante resulta observar que, no sólo los recursos en sí –libros, artículos, tesis, etc. – requieren una URI para vivir en la Web Semántica, sino que también cada elemento que lo compone, o al menos, los autores de estos recursos quienes demandan un ID único que permita resolver el tema de la homonimia y desambiguar la autoría.

OntoOAI se enfrentó a dichos problemas al encontrar en las asociaciones de los autores falsas aseveraciones. Por ejemplo, si el nombre de un autor resulta ser igual al de un futbolista famoso con página en Wikipedia, y si además este autor no tiene página en dicho sitio, resulta que sus artículos fueron adjudicados a futbolistas erróneamente. Un problema que sólo los autores pueden resolver haciendo uso de herramientas para identificarse e identificar sus obras.

Capítulo 5 Conclusiones y trabajo futuro

La web es uno de los desarrollos más exitosos de las últimas décadas, ésto se debe a factores clave como: arquitectura simple, conexión en red, extensible, poderosa y fácil para el usuario; universal, tolerante con la marcación imperfecta, libre, no costosa y con estándares abiertos (Bratt, 2007). Sin embargo, dichos factores también han originado sus principales debilidades: sobrecarga, heterogeneidad y falta de rigurosidad en la publicación de la información.

Mútiples esfuerzos en vía de la estructuración de la Web han surgido, formalmente desde la introducción del XML en 1996; acuerdos, estándares y protocolos se han definido con el objetivo de habilitar una mayor integración, transmisión y explotación de la información.

Sin embargo, en el camino hacia la Web Semántica los sitios avanzan a muy diferentes ritmos, se encuentran los que no incorporan ninguna forma de estructuración en los datos que exponen, hasta los que emplean los más sofisticados y minuciosos formatos para describir su información.

El protocolo para cosecha de metadatos OAI-PMH, surge como resultado de una preocupación de hacer interoperables los contenidos disponibles bajo la iniciativa de arhivos abiertos.

Son miles los repositorios, los cuales reúnen la producción intelectual, académica y científica, que participan del movimiento de Acceso Abierto, y que han adoptado esta especificación para describir, exponer e intercambiar su información. Sin embargo, el nivel de estructuración del OAI-PMH por sí mismo, no es suficiente para tomar ventaja de las bondades de la Web Semántica.

Gracias a la semántica en la Web, el software es capaz de procesar contenido, razonar con éste, combinarlo y realizar deducciones lógicas para resolver problemas automáticamente.

Como mencionan Segaran, Evans y Taylor (2009) construir modelos del mundo no es algo nuevo para los diseñadores de sistemas, pero con los sistemas semánticos, se tienen que expresar estos modelos de tal forma que los sistemas distribuidos a través de la Web puedan leer y entender los modelos con la precisión para usarlos.

OntoOAI surge con la idea de verificar la factibilidad en la aplicación de tecnologías de la Web Semántica para descubrir conocimiento, a partir de datos estructurados con OAI-PMH, a través de un modelo semántico para el descubrimiento selectivo de conocimiento que permita hallar recursos que no fueron solicitados explícitamente por un usuario pero que resultan potencialmente útiles.

La principal contribución de OntoOAI es proponer un enfoque semántico para el descubrimiento selectivo de conocimiento habilitando los datos estructurados con OAI-PMH para su procesamiento por tecnologías semánticas, de tal forma que permita ejecutar mecanismos de inferencia sobre el conjunto de datos, además de usar técnicas de fusión de ontologías para la verificación del conocimiento.

Es así que OntoOAI verifica que es factible usar estos datos con tecnologías semánticas, y que si bien se logran resultados satisfactorios en el descubrimiento de conocimiento también da cuenta de problemas y limitaciones enfrentadas.

Para construir el modelo OntoOAI, además de desarrollar varios componentes, se reutilizaron algoritmos, técnicas y herramientas propuestas por otros autores, como la aplicación del algoritmo para la fusión de ontologías de Ameen, Rahman Khan y Rani

(2014) y la adaptación del modelo de representación de estudiantes basado en ontologías de Panagiotopoulos, Kalou, Pierrakeas y Kameas (2012).

Las casi 8 millones de tripletas de la base de conocimiento resultante están disponibles para consulta o descarga, bajo la licencia *Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)*, ello con la finalidad de permitir la verificación del modelo o incluso la creación de aplicaciones nuevas sobre ellas.

Otros productos del modelo OntoOAI son los dos trabajos sometidos a dictaminación por pares y que han sido publicados:

- 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 2014.* Oaxaca, México: Nova Universitas.
- Becerril García, A., Lozano Espinosa, R., & Molina Espinosa, J. (2016). Semantic Approach to Context-Aware Resource Discovery over Scholarly Content Structured with OAI-PMH. *Computación y Sistemas, 20* (1), 127-142.

Por otro lado, este modelo deja indicadas varias líneas de trabajo futuro. OntoOAI podría extenderse para aprovechar las fuentes de información de Linked Open Data como insumo además de contenidos OAI-PMH, y aplicar los procedimientos de descubrimiento de conocimiento, desarrollando mecanismos de filtrado de información académica o científica.

Adicionalmente, la propuesta puede ser enriquecida con el uso de vocabularios controlados y/o el uso de ontologías multilingües para recuperar información en diversos idiomas.

También convendría probar otros motores de inferencia como Pellet, Racer o FaCT para un completo razonamiento OWL DL.

Respecto a la sensibilidad al contexto, el tipo de información almacenada en el perfil de usuario es estática, se determina durante el tiempo de diseño y no puede ser actualizado o modificado en tiempo de ejecución. En el sentido que apuntan Eyharabide y Amandi (2012), los perfiles de usuario de próxima generación deben ser no sólo un depósito pasivo de datos de usuario, sino también un componente activo capaz de aprender y actualizar tanto la información y el tipo de información almacenada

Finalmente, se tiene clara una línea de interés para dar seguimiento al trabajo doctoral aquí presentado, y tiene que ver con aprendizaje y extracción de ontologías con el objetivo de representar el conocimiento generado por diferentes conjuntos de contenidos, que pueden ser temáticos, geográficos o temporales. Y así generar mapas de conocimiento que permitan la comparación, por ejemplo, de la ciencia publicada en un área determinada a lo largo del tiempo, para observar su evolución y patrones así como la dinámica espacio-tiempo de las redes de colaboración de los científicos.

Bibliografía

(n.d.). (U. University of Southampton, Producer) Retrieved 7 de julio de 2014 from Registry of Open Access Repositories: http://roar.eprints.org/

Max Planck Society. (22 de octubre de 2003). From Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities: http://openaccess.mpg.de/Berlin-Declaration

Abburu, S. (2012). A Survey on Ontology Reasoners and Comparison. *International Journal of Computer Applications , 57* (17), 33-39.

Abusalah, M., Tait, J., & Oakes, M. (2009). Cross Language Information Retrieval using Multilingual Ontology as Translation and Query Expansion Base. *Polibits , 40*, 13-16.

Aguado López, E., Rogel Salazar, R., Becerril García, A., & García Flores, H. (2006). Redalyc OAI-PMH: the open archives initiative protocol for metadata harvesting (protocol version 2.0). *International Conference on Dublin Core and Metadata Applications* (pp. 1-10). Colima: Dublin Core.

Albert, P., Holmes, K., Börner, K., & Conlon, M. (2012). Research Discovery through Linked Open Data. *12th ACM/IEEE-CS joint conference on Digital Libraries* (pp. 429-430). Nueva York: ACM.

Allemang, D., & Hendler, J. (2011). *Semantic Web for the Working Ontologist* (2a edición ed.). Estados Unidos: Morgan Kaufmamm.

Ameen, A., Rahman Khan, K., & Rani, B. (2014). Semi-Automatic Merging of Ontologies using Protégé. *International Journal of Computer Applications , 85* (12), 35-42.

ANSI/NISO. (2003). *Information Retrieval (Z39.50): Application Service Definition and Protocol Specification.* Estados Unidos: Niso Press.

Apache Software Foundation. (2010). Retrieved 2 de Febrero de 2014 from Apache Jena: http://jena.apache.org/about_jena/about.html

Apache Software Foundation. (2011). Retrieved 2 de Febrero de 2014 from Apache Jena: http://jena.apache.org/about_jena/about.html

Becerril García, A. (2010). *Modelo basado en lógica difusa para la diseminación selectiva de información (FuzzyDSI).* México, D. F., México: ITESM.

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 2014.* Oaxaca, México: Nova Universitas.

Becerril García, A., Lozano Espinosa, R., & Molina Espinosa, J. (2016). Semantic Approach to Context-Aware Resource Discovery over Scholarly Content Structured with OAI-PMH. *Computación y Sistemas , 20* (1), 127-142.

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 , 40* (2), 53-64.

Beckett, D. (2014). *RDF 1.1 N-Triples*. From W3C: https://www.w3.org/TR/2014/REC-n-triples-20140225/

Beckett, D., Berners-Lee, T., Prud'hommeaux, E., & Carothers, G. (25 de febrero de 2014). *RDF 1.1 Turtle*. From W3C: https://www.w3.org/TR/2014/REC-turtle-20140225/#sec-intro Berners Lee, T. (mayo de 1994). Conferencia "W3 future directions". *1st World Wide Web Conference Geneva.*

Berners-Lee, T. (2005). *Notation 3 Logic*. From An RDF language for the Semantic Web: https://www.w3.org/DesignIssues/Notation3.html

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American* , 1-4.

Bibliocommons. (2016). From BiblioCore: http://www.bibliocommons.com/products/bibliocore

Bikakis, N., Tsinaraki, C., Gioldasis, N., Stavrakantonakis, I., & Christodovlakis, S. (2013). The XML and Semantic Web Worlds: Technologies, Interoperability and Integration. A survey of the state of the art. In I. B. Anagnostopoulos (Ed.), *Semantic Hyper/Multimedia Adaptation.* Springer.

Bizer, C., & Cyganiak, R. (2006). *D2R Server – Publishing Relational Databases on the Semantic Web*. From http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/Bizer-Cyganiak-D2R-Server-ISWC2006.pdf

Bratt, S. (2007). *Semantic Web, and Other Technologies to Watch.* Retrieved 2 de diciembre de 2015 from W3C: https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#(1)

Breeding, M. (2014). Library Resource Discovery Products: Context, Library Perspectives, and Vendor Positions. *Library Technology Reports , 50* (1), 5-58.

Breeding, M. (2015). *The future of library resource discovery. A white paper comissioned by the NISO Discovery to Delivery (D2D) Topic Comittee.* Baltimore, Maryland, USA: NISO.

Brickley, D., & Miller, L. (14 de enero de 2014). From FOAF Vocabulary Specification 0.99: http://xmlns.com/foaf/spec/

Brickley, D., & Miller, L. (14 de enero de 2014). *FOAF Vocabulary Specification 0.99.* From http://xmlns.com/foaf/spec/

Brown, P., Cabell, D., Chakravarti, A., Cohen, B., Delamothe, T., Eisen, M., et al. (20 de junio de 2003). From Declaración de Bethesda sobre Publicación de Acceso Abierto: http://ictlogy.net/articles/bethesda_es.html

Browning, S., Guédon, J.-C., & Kaplan, K. (2013). Metadata and Open Access: Reliably Finding Content and Finding Reliable Content. *Proceedings of the Charleston Library Conference*, (pp. 504-511).

Cafarella, M., Halevy, A., & Madhavan, J. (2011). Structured Data on the Web. *Communications of the ACM , 54* (2), 72-79.

Calero, C., Ruiz, F., & Piattini, M. (2012). *Ontologist for software engineering and software technology.* Ciudad Real, España: Springer.

Cantador, I., Szomszor, M., & Alani, H. (2008). Enriching Ontological User Profiles with Tagging History for Multi-Domain Recommendations. *1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008)*, (pp. 1-15). Tenerife, España.

Cantillo Valero, C., Roura Redondo, M., & Sánchez Palacín, A. (2012). Tendencias actuales en el uso de dispositivos móviles en educación. *La Educ@ción Digital , 147.*

Chan, L., Cuplinskas, D., Eisen, M., Friend, F., Genova, Y., Guédon, J.-C., et al. (2002). From Budapest Open Access Initiative: http://www.soros.org/openaccess/read.shtml

Chisholm, R. (1996). *A Realistic Theory of Categories: An Essay on Ontology.* Cambridge: Cambridge University Press.

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, Washintong, DC.

Dawra, M. (2004). *Library Science and Theories of Management.* New Delhi: Rajat.

DBpedia. (febrero de 2016). *The DBpedia Data Set (2014).* Retrieved 2014 from DBpedia: http://oldwiki.dbpedia.org/Datasets#h18-19

DCMI. (2012). Retrieved 2 de abril de 2014 from DCMI Metadata Terms: http://dublincore.org/documents/2012/06/14/dcmi-terms/

DCMI. (2016). Retrieved marzo de 2016 from Ontología Dublin Core: http://purl.org/dc/terms

DCMI. (14 de junio de 2012). *DCMI Metadata Terms*. From http://dublincore.org/documents/2012/06/14/dcmi-terms/.

DCMI. (n.d.). *Dublin Core Metadata Element Set, Version 1.1* . From Dublin Core Metadata Initiative: http://dublincore.org/documents/dces/

DCMI. (2005). *Metadata Basics*. Retrieved 5 de Mayo de 2014 from Dublin Core Metadata Initiative: http://dublincore.org/metadata-basics/

DCMI. (2012). *Metadata Basics*. Retrieved 5 de Mayo de 2014 from Dublin Core Metadata Initiative: http://dublincore.org/metadata-basics/

Dentler, K., Cornet, R., Teije, A., & de Keizer, N. (2011). Comparison of Reasoners for large Ontologies in the OWL 2 EL Profile. (I. Press, Ed.) *Semantic Web , 1*, 1-5.

Ding, Y., & Foo, S. (2002). Ontology Research and Development Part 1 – A Review of Ontology Generation. *Journal of Information Science , 2* (28), 123-136.

DuraSpace. (2002). Retrieved 2015 from DSpace: www.duraspace.org

EBSCO. (2016). From Discovery Services: https://www.ebscohost.com/discovery

Eprints. (2016). Retrieved 20 de febrero de 2016 from Registry of Open Access Repositories: http://roar.eprints.org/view/type/

ExLibris. (2006). *http://www.exlibrisgroup.com/category/PrimoOverview*. From Primo.

Eyharabide, V., & Amandi, A. (2012). agentes personales reúnen información sobre los usuarios de un perfil de usuario. Una vez definido el perfil de usuario, el agente puede utilizar, aprender y deducir nuevos conocimientos sobre la base de este perfil para personalizar la asistencia que se ofrece al usuario. *Applied Intelligence , 36* (4), 857-869.

Fayyad, U., Piatetsky-Shapir, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine , 17* (3), 37-54.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine , 17* (3), 37-54.

Foulonneau, M., & Grouès, V. (2012). Common vs. Expert knowledge: making the Semantic Web an educational model. *Linked Learning* .

Foundation, A. S. (2014). *Apache Jena.* From Fuseki: serving RDF data over HTTP : http://jena.apache.org/documentation/serving_data/

Franzon, E. (13 de mayo de 2014). End of Support for the Sindice.com search engine: history, lessons learned, and legacy. *Dataversity* .

Franzon, E. (13 de mayo de 2014). *End of Support for the Sindice.com search engine: history, lessons learned, and legacy*. Retrieved 15 de noviembre de 2015 from Dataversity: http://www.dataversity.net/end-support-sindice-com-search-engine-history-lessons-learned-legacy-guest-post/

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine , 13* (3), 57-70.

Furletti, B. (2009). *Ontology-Driven Knowledge Discovery.* Lucca, Italia: IMT Institute for Advanced Studies.

Göckel, R. (1980). *Lexicon philosophicum quo tanquam clave philosophiae fores aperiuntur, informatum opera et studio* (2a edición ed.). Re-impreso por Georg Olms.

Giasson, F., & D'Arcus, B. (Diciembre de 2009). *Bibliographic Ontology Specification.* Retrieved 2014 from http://bibliontology.com/

Golemati, M., Katifori, A., Vassilakis, C., Lepouras, G., & Halatsis, C. (2007). Creating an Ontology for the User Profile: Method and Applications. *Proceedings of the First IEEE International Conference on Research Challenges in Information Science (RCIS)* (pp. 1-7). Marruecos: IEEE.

Google. (16 de mayo de 2012). Retrieved 14 de enero de 2016 from Google Official Blog: https://googleblog.blogspot.mx/2012/05/introducing-knowledge-graph-things-not.html Google. (28 de noviembre de 2015). Retrieved 20 de enero de 2016 from Freebase: http://www.freebase.com/

Grčar, M., Mladenič, D., & Grobelnik, M. (2006). User Profiling for the Web. *Computer Science and Information Systems , 3* (2), 1-29.

Grobelnik, M., & Mladenić, D. (2005). Simple classification into large topic ontology of Web documents. *Journal of Computing and Information Technology* (4), 279–285.

Gruber, T. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human and Computer Sciences , 43* (5-6), 907-928.

Guarino, N. (1998). Formal Ontology and Information Systems. *Proceedings of FOIS'98* (pp. 3-15). Trento, Italia: Amsterdam, IOS Press.

Guédon, J.-C. (2013). Metadata for Digital Documents: A Researcher's Perspective. In S. Browning, J.-C. Guédon, L. Kaplan, & P. o. Conference (Ed.), *Metadata and Open Access: Reliably Finding Content and Finding Reliable Content.*

Guinchat, C., & Menou, M. (1992). *Introducción general a las ciencias y técnicas de la información y documentación.* Madrid, España: CSIC-UNESCO.

Haslhofer, B., & Schandl, B. (2010). Interweaving OAI-PMH Data Sources with the Linked Data Cloud. *Int. J. Metadata, Semantics and Ontologies , 1* (5), 17-31.

Haslhofer, B., & Schandl, B. (2008). The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data. *International Workshop on Linked Data on the Web (LDOW2008).* Beijing, China.

Hoppe, A. (2013). Automatic ontology-based User Profile Learning from heterogeneous Web Resources in a Big Data Context. *Proceedings of the VLDB Endowment, 6.* Riva del Garda, Trento, Italia.

IBM. (2015). From Knowledge Discovery and Data Mining: http://researcher.watson.ibm.com/researcher/view_group.php?id=144

Ingarden, R. (1964). Time and Modes of Being. In C. Thomas, *The Problem of the Existence of the World* (H. R. Michejda, Trans.). Springfield.

Innovative Interfaces . (2016). From Encore: https://www.iii.com/products/sierra/encore

Jeremić, Z., Jovanović, J., & Gašević, D. (2011). Personal Learning Environments on the Social Semantic Web. *Semantic Web Journal* , 1-30.

Jeremic, Z., Jovanovic, J., & Gasevic, D. (2011). Personal Learning Environments on the Social Semantic Web. *Semantic Web Journal* , 1-30.

Käpylä, T., Niemi, I., & Lehtola, A. (1998). Towards an accessible web by applying PUSH technology. *Fourth ERCIM Workshop, User Interfaces for All,* (pp. 19-21). ERCIM.

Keßler, C., d'Aquin, M., & Dietze, S. (2013). Linked Data for Science and Education. *Semantic Web Journal , 4* (1), 1-2.

Kemp, D. (1979). *Current awareness services.* London, U.K.: Clive Bingley.

Kessler, C., d'Aquin, M., & Dietze, S. (2013). Linked Data for Science and Education. *Semantic Web Journal , 4* (1), 1-2.

Klosgen, W., & Zytkow, J. (1996). Knowledge discovery in databases terminology. In U. Fayyad, G. Piatetsky- Shapiro, P. Smyth, & R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining* (pp. 573–592). AAAI Press.

Kurgan, L., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review , 21* (1), 1–24.

Lagoze, C., & Van de Sompel, H. (2001). The Open Archives Initiative: Building a low-barrier interoperability framework. In ACM (Ed.), *JCDL '01*. Roanoke, VA.

Lagoze, C., & Van de Sompel, H. (08 de 01 de 2015). *The Open Archives Initiative Protocol for Metadata Harvesting*. From Protocol Version 2.0 of 2002-06-14: http://www.openarchives.org/OAI/openarchivesprotocol.html

LaReferencia. (2012). Retrieved 2016 from LaReferencia: http://www.lareferencia.info/vufind/

Lewis, S. (abril de 2014). *Repository Maps*. Retrieved 20 de febrero de 2016 from Repository66.org: http://maps.repository66.org/

Lorhard, J. (1606). *Ogdoas Scholastica, continens Diagraphen Typicam artium: Grammatices (Latinae, Graecae), Logices, Rhetorices, Astronomices, Ethices, Physices, Metaphysices, seu Ontologiae.* (t. a. Uckelman, Trans.) Sangalli.

Luhn, H. P. (1958). A business intelligence system. *IBM Journal of R&D* , 314-319.

Lund University. (2003). *DOAJ*. Retrieved 3 de marzo de 2016 from Directory of Open Access Journals: www.doaj.org

Maedche, A., & Volz, R. (2001). The Ontology Extraction & Maintenance Framework Text-To-Onto. *Proceedings Workshop on Integrating Data Mining*, (pp. 1-12). Estados Unidos.

Maes, P. (1994). Agents that Reduce Work and Information Overload. *Communications of the ACM* , *37* (7), 30-40.

Martínez-Villaseñor, M., González-Mendoza, M., & Danvila Del Valle, I. (2014). Enrichment of Learner Profile with Ubiquitous User Model Interoperability. *Computación y Sistemas* , *18* (2), 359-374.

Mazzocchi, S. (2006). Retrieved 5 de enero de 2014 from OAI2RDF: http://simile.mit.edu/repository/RDFizers/oai2rdf/

Mazzocchi, S. (02 de Marzo de 2007). *SIMILE*. From OAI-PMH RDFizer: http://simile.mit.edu/wiki/OAI-PMH_RDFizer

McDowell, L., & Cafarella, M. (2006). Ontology-driven information extraction with ontosyphon. *5th international conference on The Semantic Web* (pp. 428-444). Berlín: Springer-Verlag.

McGuinness, D. (2000). The chimaera ontology environment. *AAAI/IAAI*, (pp. 1123-1124).
Mealy, G. (1967). Another look at data. *Proceedings of the Fall Joint Computer Conference* (pp. 525-534). AFIPS.

Ministerio de Ciencia e Innovación de España. (31 de 07 de 2009). *Informe Modelos de Metadatos para Contenidos Multimedia.* From http://omediadis.udl.cat/html/deliverables/215-Modelos_Metadatos_Contenidos_Multimedia/

Ministerio para la Ciencia e Innovación de España. (31 de 07 de 2009). *Informe Modelos de Metadatos para Contenidos Multimedia.* From http://omediadis.udl.cat/html/deliverables/215-Modelos_Metadatos_Contenidos_Multimedia/

Missikoff, M., Navigli, R., & Velardi, P. (2002). Integrated Approach to Web Ontology Learning and Engineering. *Computer, 35* (11), 54-57.

Mladenić, D., Grobelnik, M., Fortuna, B., & Grčar, M. (2009). Semantic Knowledge Management. In *Web, Knowledge Discovery for Semantic* (pp. 21-36). Berlin: Springer-Verlag.

Moldovan, D., & Blanco, E. (2012). Polaris: Lymba's Semantic Parser. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12.* Estambúl, Turquía: European Language Resources Association (ELRA).

Musumba, G., & Nyongesa, H. (2013). Context awareness in mobile computing: A review. *Int J Machine Learn Appl, 2* (1), 1-10.

Noy, N., & McGuinness, D. (2001). *Ontology Development 101: A Guide to Creating Your First Ontology.* Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880.

Obitko, M. (2007). *Semantic Web Architecture.* Retrieved 7 de enero de 2016 from Ontologies and Semantic Web: http://www.obitko.com/tutorials/ontologies-semantic-web/semantic-web-architecture.html

OCLC. (2002). Retrieved 30 de 07 de 2014 from The OAIster® database: http://www.oclc.org/oaister.en.html

OCLC. (2014). Retrieved 30 de 07 de 2014 from The OAIster database: http://www.oclc.org/oaister.en.html

OCLC. (2016). From WorldCat: https://www.oclc.org/worldcat-discovery.en.html OMediaDis. (31 de 07 de 2009). Retrieved 2 de agosto de 2014 from Informe Modelos de Metadatos para Contenidos Multimedia: http://omediadis.udl.cat/html/deliverables/215-Modelos_Metadatos_Contenidos_Multimedia/

OpenAIRE. (2009). Retrieved abril de 2016 from OpenAIRE: https://www.openaire.eu/project-factsheets

OpenArchives. (2002). *Esquema de validación de Dublin Core.* Retrieved 3 de Abril de 2014 from Open Archives: http://www.openarchives.org/OAI/2.0/oai_dc.xsd

OpenDOAR. (n.d.). Retrieved 30 de Julio de 2014 from Directory of Open Access Repositories: http://www.opendoar.org/

OpenDOAR. (n.d.). *Content Types in OpenDOAR Repositories - Worldwide.* Retrieved 8 de Mayo de 2014 from Directory of Open Access Repositories: http://www.opendoar.org/onechart.php?cID=&ctID=&rtID=&clID=&lID=&potID=&rSoftWareName=&search=&groupby=ct.ctDefinition&orderby=Tally%20DESC&charttype=bar&width=600&caption=Content%20Types%20in%20OpenDOAR%20Repositories%20-%20Worldwide

Panagiotopoulos, I., Kalou, A., Pierrakeas, C., & Kameas, A. (2012). An Ontology-Based Model for Student Representation in Intelligent Tutoring Systems for Distance Learning. In I. M. Lazaros Iliadis (Ed.), *Artificial Intelligence Applications and Innovations.* Halkidiki, Grecia: Springer.

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 , 17*, 33-43.

Piatetsky-Shapiro, G., & Matheus, C. (1994). The Interestingness of Deviations. *Proceedings of KDD-94* (pp. 25-36). Menlo Park, Calif: AAAI Press.
PKP. (2001). Retrieved 2016 from Open Journal Systems: https://pkp.sfu.ca/ojs/
PKP. (2015). *OJS Usage*. Retrieved 2015 from OJS: https://pkp.sfu.ca/ojs/ojs-usage/
Presidencia de México. (2014 de mayo de 2014). From Acceso abierto en la Sociedad del Conocimiento: http://www.presidencia.gob.mx/acceso-abierto-en-la-sociedad-del-conocimiento/

Presutti , V., Draicchio, F., & Gangemi, A. (2012). Knowledge Extraction based on Discourse Representation Theory and Linguistic Frames . *18th International Conference on Knowledge Engineering and Knowledge Management (EKAW2012).*
ProQuest. (2009). From AquaBrowser: http://www.proquest.com/products-services/AquaBrowser.html

ProQuest. (2015). From SummonService: http://www.proquest.com/products-services/The-Summon-Service.html

River Campus Libraries. (2006). From eXtensible Catalog: http://www.library.yorku.ca/cms/bibblog/2006/05/04/university-of-rochesters-river-campus-libraries-extensible-catalog-xc/

ROARMAP. (2015). Retrieved 18 de febrero de 2016 from Registry of Open Access Repository Mandates and Policies: http://roarmap.eprints.org/

Robin, C., & Uma, G. (2010). A novel algorithm for fully automated ontology merging using hybrid strategy. *European Journal of Scientific Research , 47* (1), 74-81.
Roskilde University. (2015). From Roskilde University Digital Archive: http://diggy.ruc.dk:8080

Russell, S., & Norvig, P. (2004). *Inteligencia artificial, un enfoque moderno* (2a ed.). Madrid, España: Pearson Prentice Hall.

Schiaffino , S., & Amandi, A. (enero-febrero de 2006). Polite personal agents. *IEEE Intelligent Systems* , 12-19.

Schonfeld, E. (17 de agosto de 2010). As Bing Takes Over Yahoo Search, SearchMonkey Dies, BOSS Is No Longer Free, But Site Explorer Still Works. *TechCrunch* .

Schonfeld, R. (2014). *Does Discovery Still Happen in the Library? Roles and Strategies for a Shifting Reality.* (ITHAKA, Ed.)

Segaran, T., Evans, C., & Taylor, J. (2009). *Programming the semantic web.* (1. ed., Ed.) Estados Unidos: O'reilly.

Shotton, D. (2010). CiTO, the Citation Typing Ontology. *Journal of Biomedical Semantics , 1*, 1-18.

Shotton, D., & Peroni, S. (2013). Retrieved 14 de septiembre de 2014 from BiRO, the Bibliographic Reference Ontology: http://www.essepuntato.it/lode/http://purl.org/spar/biro
Shotton, D., & Peroni, S. (2013). *BiRO, the Bibliographic Reference Ontology*. Retrieved 2015 de Enero from http://www.essepuntato.it/lode/http://purl.org/spar/biro

Sieg, A., Mobasher, B., & Burke, R. (2007). Ontological User Profiles for Personalized Web Search. *Proceedings of the International Conference on Artificial Intelligence and Applications* (pp. 84-91). American Association for Artificial Intelli- gence.

Silberschatz, A., & Tuzhilin, A. (1995). On Subjective Measures of Interestingness in Knowledge Discovery. *Proceedings of KDD-95* (pp. 275–281). Menlo Park, Calif: AAAI Press.

SIMILE. (2006). *OAI2RDF.* From http://simile.mit.edu/repository/RDFizers/oai2rdf/

Sirsi Corporation. (2016). From BLUEcloud PAC: http://www.sirsidynix.com/products/bluecloud-pac

Smith, B. (2004). Retrieved 15 de marzo de 2016 from Ontology and Information Systems: http://ontology.buffalo.edu/ontology(PIC).pdf

Smith, B. (2003). Ontology. In L. Floridi (Ed.), *Blackwell Guide to the Philosophy of Computing and Information* (pp. 155–166). Oxford: Blackwell.

SPAR. (2009). Retrieved 12 de Diciembre de 2014 from Semantic Publishing and Referencing Ontologies: http://sempublishing.sourceforge.net/

SPAR. (2013). *Semantic Publishing and Referencing Ontologies*. Retrieved Diciembre de 2014 from http://sempublishing.sourceforge.net/

Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., & Lindström, N. (2014). *JSON-LD 1.0*. From W3C: https://www.w3.org/TR/2014/REC-json-ld-20140116/

SRA International, Inc. (2016). From NetOwl Extractor: https://www.netowl.com/entity-extraction/

Stan, J., Egyed-Zsigmond, E., Joly, A., & Maret, P. (2008). A User Profile Ontology For Situation-Aware Social Networking. *In 3rd Workshop on Artificial Intelligence Techniques for Ambient Intelligence*, (pp. 1-5).

Stanford Center for Biomedical Informatics Research. (2014). From Protégé: http://protege.stanford.edu/

Stanford University. (2015). Retrieved 15 de noviembre de 2014 from Protégé: http://protege.stanford.edu

Studer, R., Benjamins, R., & Fensel, D. (1998). Knowledge Engineering: Principles and methods. *Data and Knowledge Engineering , 25* (1-2), 161-197.

Suber, P. (2015). *Acceso Abierto.* (R. Melero, Trans.) Toluca, Estado de México : Universidad Autónoma del Estado de México .

Suzanne Saskia Kemperman, B. B.-v. (2014). *Success Strategies for Electronic Content Discovery and Access: A Cross-Industry White Paper.* Dublin, Ohio, USA: OCLC Online Computer Library Center, Inc.

The Hague Declaration. (2014). *THE HAGUE DECLARATION ON KNOWLEDGE DISCOVERY IN THE DIGITAL AGE.* Hague, Países Bajos.

Tummarello, G., & Delbru, R. (2014). *The Sindice Project, History and Legacy*. Retrieved 3 de noviembre de 2015 from Dataversity: http://www.dataversity.net/end-support-sindice-com-search-engine-history-lessons-learned-legacy-guest-post/

Tummarello, G., Delbru, R., & Oren, E. (2007). Sindice.com: Weaving the Open Linked Data. In *The Semantic Web* (pp. 552-565). Springer Berlin Heidelberg.

University of Nottingham. (2014). Retrieved 30 de Julio de 2014 from Directory of Open Access Repositories: http://www.opendoar.org/

University of Pennsylvania. (2016). From Franklin: http://franklin.library.upenn.edu/index.html

University of Southampton. (2014). Retrieved 7 de julio de 2014 from Registry of Open Access Repositories: http://roar.eprints.org/

University of Virginia. (2016). From BlackLight: http://projectblacklight.org

Villanova University. (2016). From VuFind: http://vufind-org.github.io/vufind/

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, *57* (10), 78-85.

W3C. (2015). *Extensible Markup Language (XML)*. Retrieved 14 de abril de 2014 from https://www.w3.org/XML/

W3C. (2014). *RDF 1.1 N-Quads*. From W3C: https://www.w3.org/TR/2014/REC-n-quads-20140225/

W3C. (2014). *RDF 1.1 XML Syntax*. From W3C: http://www.w3.org/TR/2014/REC-rdf-syntax-grammar-20140225/

W3C. (10 de febrero de 2004). *RDF Vocabulary Description Language 1.0: RDF Schema*. From W3C: https://www.w3.org/TR/2004/REC-rdf-schema-20040210/#ch_classes

W3C. (n.d.). *Resource Description Framework (RDF)*. Retrieved 01 de 08 de 2014 from Semantic Web: http://www.w3.org/RDF/

W3C. (2014). *Resource Description Framework (RDF)*. Retrieved 5 de enero de 2016 from W3C: https://www.w3.org/RDF/

W3C. (2013). *Semantic Web Activity* . Retrieved 12 de enero de 2016 from W3C: https://www.w3.org/2001/sw/

W3C. (2007). *Semantic Web Stack*. Retrieved 15 de noviembre de 2015 from w3: https://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#(24)

W3C. (2013). *W3C*. (S. W. Group, Producer) Retrieved 2 de febrero de 2016 from SPARQL Query Language for RDF: https://www.w3.org/2001/sw/wiki/SPARQL

W3C. (2012). *Web Ontology Language (OWL)*. (O. W. Group, Producer) Retrieved 12 de Noviembre de 2014 from W3C: https://www.w3.org/2001/sw/wiki/OWL

Wiederhold, G. (1996). Foreword: on the barriers and future of knowledge discovery. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*. AAAI Press.

Wikipedia. (2016). *Knowledge extraction.* From Survey of Methods / Tools: https://en.wikipedia.org/wiki/Knowledge_extraction#1:1_Mapping_from_RDB_Tables.2FViews_to_RDF_Entities.2FAttributes.2FValues

Yahoo Inc. (2009). Retrieved 18 de agosto de 2015 from Yahoo Search BOSS: https://boss.yahoo.com/

Yildiz, B., & Miksch, S. (2007). ontoX - A Method for Ontology-Driven Information Extraction. *2007 International Conference on Computational Science and its Applications . III*, pp. 660-673. Berlin: Springer-Verlag.

Apéndices

Apéndice A. Resultados del ejemplo 1

1	La emergencia del sentido a partir del no-sentido	Morin, Edgar
2	Anotaciones para un nuevo Emilio: transmisión sistémica del conocimiento	Morin, Edgar
3	Complejidad restringida y Complejidad generalizada o las complejidades de la Complejidad	Morin, Edgar
4	Editorial Enseñar la condición humana	Morin, Edgar
5	Edgar Morin: La Connaissance de la Connaissance	Olsen, Michel
6	Terrorbegrebets udvikling	Skovbakke, Anders Søndergaard; Find, Matias; Christiansen, Mads; Advisor: Hansen, Allan Dreyer
7	Adgang for asylansøgere til at arbejde og bo uden for asylcentre – virkelighed eller utopi? En undersøgelse af hvad den nye ændring i det danske asylsystem indebærer, samt en vurdering af de effekter, som ændringen medfører for asylansøgere og det danske samfund. Og i forlængelse af den vil specialet give en diskussion af i hvor høj grad den nye regerings politik er et brud med den tidligere diskurs om asylansøgere.	Hansen, Allan Dreyer; Mardale, Andreea Bjerre Ryhede, Annette; Clausen, Amalie; Nielsen, Mikkel; Rau-Andersen, René; Christiansen, Mads;Flensborg, Cecilie
8	PlayDay - Afprøvning af en kampagneplakat	Dam Scheuer, John; Christiansen, Mads;Lachenmeier, Anders; Risbøl Jacobsen, Sebastian;Andreasen, Patrick
9	Samspillet mellem minister og departementschef anno 2014	Hansen, Allan Dreyer; Andersen, Sandra Louise;Jensen, Tilla Slåen; Christensen, Mathias Boberg;Hansen, Line Tauber; Würtzenfeld, Mathias
10	Etnisk diversitet i Danmark - i spændingsfeltet mellem konflikt og kontakt	Hansen, Allan Dreyer; Bech, Stinne Lyager; Plum, Maja; Institut for Samfund og Globalisering
11	Spillet om læring : en diskursanalyse af brugen af læring på dagtilbudsområdet	Hansen, Allan Dreyer; Christensen, Emma Kleist;Kauffmann, Julie Ringgaard; Würtzenfeld, Mathias
12	Dansk terrorlovgivning i spændingsfeltet mellem sikkerhed og rets(u)sikkerhed	Hansen, Allan Dreyer; Christensen, Sara
13	Deltagelsens paradoks: Deltagelsesidealer-	

	og mønstre i forbindelse med skolestrukturændringer	Ryder;Harder, Ditte Weimann
14	demokrati, deltagelse og konflikt	Hansen, Allan Dreyer; Christensen, Sara Ryder;Harder, Ditte Weimann
15	Kinas Hegemoniske Projekt i Taiwan	Hansen, Allan Dreyer; Durant, Kristin
16	Tillidsreformen i Københavns kommunes hjemmepleje	Hansen, Allan Dreyer; Henriksen, Lene Sofie
17	Der kom en aftale	Hansen, Allan Dreyer; Jensen, Inger Lisbeth Staahl
18	NY VIRKELIGHED – NY VELFÆRD. ODENSE KOMMUNES VELFÆRDSPOLITIK I ET DISKURSIVT PERSPEKTIV	Hansen, Allan Dreyer; Jensen, Mie
19	Et governmentality-studie af akkreditering	Hansen, Allan Dreyer; Jeppesen, Maja Ramsbøl
20	Er Demokratisk styring = Menneskelig udvikling - Hvilke strukturer, normer og institutioner påvirker den kausale sammenhæng.	Hansen, Allan Dreyer; Lauridsen, Laurids Sandager; Kragh, Mads Randeris
21	Kampen om velfærdsydelser til EU-borgere – en diskursteoretisk analyse af partierne i danske Folketing og EU-kommissionen	Hansen, Allan Dreyer; Nielsen, Malte Munch;Reich, Oliver Kløvedal; Christensen, Monica Hapiach
22	Demokratisk civil ulydighed?	Hansen, Allan Dreyer; Nielsen, Rasmus Svoldgaard
23	Ikke-voldelig og åben civil ulydighed? - en problematisering af den teoretiske doxa	Hansen, Allan Dreyer; Nielsen, Rasmus Svoldgaard; Andresen, Michael Kobbernagel
24	Nødvendighedens politik - en kritisk diskursanalyse	Hansen, Allan Dreyer; Refskov, Thomas;Thornelius, Morten; Vestrup, Jonas; Underbjerg, Jeppe; Otkjær, Miki; Mikkelsen, Kasper
25	What does it mean to say that a hegemonic project is neo-liberal? : Some questions based on experiences from Denmark	Hansen, Allan Dreyer; Social retfærdighed og magt; The Department of Society and Globalisation
26	Offentlighedslovens § 27, stk. 2 og forholdet mellem den danske regering og Folketing	Hansen, Allan Dreyer; Spence, Morten Madvig
27	Multipolarity or cosmopolitanism? : A critique of Mouffe from a hegemony-theoretical perspective	Hansen, Allan Dreyer; The Department of Society and Globalisation
28	Laclauian discourse theory and the problems of institutions	Hansen, Allan Dreyer; The Department of Society and Globalisation; Social retfærdighed og magt
29	Radical or deliberative democracy?	Hansen, Allan Dreyer; The Department of Society and Globalisation; Social retfærdighed og magt

30	The ontological primacy of the political? Some critical remarks	Hansen, Allan Dreyer; The Department of Society and Globalisation; Social retfærdighed og magt
31	EU som enhedsaktør - behovet for en fælles europæisk sikkerhedsapolitik	Hansen, Allan Dreyer; Thygesen, Kristoffer Bækkel; Hougaard, Jeppe Weis; Ledstrup, David;Hammer-Jakobsen, Rasmus; Jordahn, Jakob Kragh
32	Scripting Simplified	Høbye, Mads; Christiansen, Mads; Hansen, Anders Holm; Hintze, Casper Bertel Rye; Laursen, Martin Marinus Agergaard; Schou, Kristoffer Carl
33	New Public Management i sygehusvæsenet	Jacobsen, Anders; Skovbakke, Anders Søndergaard; Møller, Andreas Bang; Bækgaard, Liva Marie de Vries; Johansen, Clara Bloch;Roberts, Kirstine Kathleen
34	Sælger avisen aviser?	Jensen, Inger; Busk Jørgensen, Asbjørn;Skovbakke, Anders Søndergaard; Larsen, Louise Foged; Hougaard, Kirstine Erreboe
35	Lollands sociale udfordringer	Koch, Martin S.; Olesen, Signe Krøll; Rasmussen, Ea Pernille Feldt; Christensen, Mike Steven Lee Uslar; Christiansen, Mads; Diasso, Mark Kwating
36	Scania - en organisation i forandring	La Cour, Christian; Janet Junker, Cecilie; Just Bjerre, Christa; Find, Matias; Glipstrup-Bonde, Peter
37	Education and Training: Europæisk uddannelsessamarbejde i dansk perspektiv	Lehrmann, Natalia; Skovbakke, Anders Søndergaard; Find, Matias; Severinsen, Katrine Devantier; Hansen, Lise Duckert; Thorkildsen, Sofie Amalie Køier
38	Elektronisk betalingsarmbånd på festivaler	Madsen, Lise; Nielsen, David; Christiansen, Mads;Hansen, Mariette; Rasmussen, Mathias; Bonde, Maiken; Rasmussen, Thomas; Hansen, Johannes
39	Kampen om Ungdomshuset	Møller Christensen, Arne; Søndergaard Skovbakke, Anders; Weye Hemdorff, Frederik;Find, Matias; Dalgaard, Anna; Svane Ryby, Anne-Sofie; Køier Thorkildsen, Sofie
40	Praktikprojekt med udgangspunkt i praktikforløb hos Socialdemokraterne på Christiansborg	Zober, Cecilie Bekker, Advisor: Hansen, Allan Dreyer

Apéndice B. Resultados del ejemplo 2

1	Entre la fe y la ciencia: La teoría de la cultura mundial y la educación comparada	Stephen Carney ; Jeremy Rappleye ; Iveta Silova ;
2	The Changing Environment of Development: From Aid to Trade	Carney, Stephen; Kehlet Hansen, Jesper Peter
3	The Decision - Youth and the Negotiation Between Choices	Carney, Stephen; Myssen, Martin; Nisted, Nina;Zmylon, Nanna Nielsen; Blomsterberg, Sofie Amalie;Birkemose, Liv; Falk, Nicklas; Pedersen, Aryono Daniel Ingemann; Christensen, Andrea Bang
4	Development through Entrepreneurship : microcredit, learning and the persistence of poverty amongst women in urban Tanzania	Sigalla, Rachel; Carney, Stephen; The Department of Psychology and Educational Studies
5	Education for Development Is Culturally Sensitive Development a Realistic Alternative to the Current Model of Development?	Carney, Stephen; Folman, Edlira; Jasinskaite, Jorune;Stræde Meiner, Mathilde; Libsekal; Wandel, Oscar;Andersen, Signe; Persson, Tina
6	Education in Flux	Carney, Stephen; Damary, Talia; Gyldenkærne, Rasmus W.; Sandvad, Nicklas; Schiøler, Maja; Yigit, Mustafa U.
7	Globalization and education	Carney, Stephen; Bak, Andreas H.
8	Micro-Finance and Poverty Alleviation in Cameroon	Carney, Stephen; Adangwi, John Atie
9	Nationalism in the New Middle East	Carney, Stephen; Heimisson, Gunnar Örn
10	Planning Process Drama - Can Drama and Learning Styles improve Education?	Carney, Stephen; Kristiansen, Anders Emil Wulff;Mottelson, Ida Nygaard; Larsen, Signe Illum Honoré
11	Contemporary Maya Identity and Representations	Harbsmeier, Michael; Augoustakis, Louisa Dikea; Ayivoh, Phiby Dedei; Jensen, Signe Marie; Søgaard Kristensen, Christina; Medina, Carmen Csilla; Petersen, Freja Astrid; Saalfeldt, Rie
12	De glatte og stribede rum - fortællinger om krigen i Afghanistan	du Plessis, Erik Mygind; Rosenberg, Vibe Appel; Nossell, Jacob Yoon Egeskov; Sørensen, Kristina Louise; Johnsen, Anton Asklund; Neumann, Cæcilie
13	Un Fleurissement Fèminin - L'Affrontement du Mariage Polygame et l'Évolution du Féminisme selon le Roman «Une si Longue Lettre» de Mariama Bâ	Loddegaard, Anne; Munch, Maria Frederikke; Wallin, Sara Linnea

14	Til loven os skiller - en undersøgelse af meningstilskrivelsen af homoseksuelle i Uganda	Bull Christiansen, Lene; Fogde, Anne-Sofie; Majlund, Anne-Sofie
15	Food Choice	Singla, Rashmi; Ehlers Bloksted, Josephine; Knudsen, Anna Althea; Boie Hvid, Amanda; Bang Christensen, Andrea; Ali Ismail, Mohamed Ahmed
16	Media and democratization: a comparative study of two former communist countries (Poland and Russia)	Feldt, Jakob Egholm; Volkova, Anna
17	Utopias and dystopias - an exploration of literary genres	Mortensen, Janus; Kühn, Lena; Morales, Jedediah; Kjær Andersen, Ina; Henriksen, Kine; Reininger Ardilsø, Lena
18	Pro-poor growth in Argentina and Brazil: Combining macroeconomic policymaking and pro-poor initiatives	Wendt, Rasmus; Woodall, Marc; Engel, Denice; Kjær, Mikkel; Wingård, Rune
19	The Food Justice Movement: Discourses of Urban Gardening in New York City	Tufte, Thomas; Glasser, Anna
20	que vidimus oculis nostris - Missonæren som antropolog	Harbsmeier, Michael; Henschen, Daniel
21	Mellem diplomati og bestikkelse	Harbsmeier, Michael; Birch, Rasmus
22	I Immigranternes Fodspor	Harbsmeier, Michael; Kjærgaard, Stine
23	Suverænitet og Absolutisme	Harbsmeier, Michael; Lindhagen, Louise
24	A COMPARATIVE STUDY OF IMMIGRATION POLICIES IN DENMARK AND SWEDEN WITH A FOCUS ON PALESTINIAN MIGRANTS AS AN ETHNICAL GROUP UNDER THE RISE OF NEO NATIONALISM	Harbsmeier, Michael; Sivova, Alina
25	The Influence of the EU Integration process in the Democratization of post-communist Albania	Harbsmeier, Michael; Bruzja, Hans
26	Peacekeeping in Kosovo - a hard nut to crack	Harbsmeier, Michael; Oruci, Valentina
27	POLITICAL SOCIETIES IN TRANSITIONING HUNGARY AND POLAND FROM 1989 TO THE	Harbsmeier, Michael; Ilic, Vladimir

	RETURN OF THE REFORMED COMMUNISTS TO POWER	
28	Migrants' Engagement with Transnational Media: Iranian Diaspora living in Europe	Harbsmeier, Michael; Khayami, Ali
29	European Identity as Perceived by Others: the Portrayal of European Political, Cultural and Economic Issues in Ukrainian media	Harbsmeier, Michael; Ipatenko, Ielyzaveta
30	Historien om Dr. Faust — og om den danske folkebogsudgaven år 1588 symboliserer et brud med middelaldermentaliteten	Harbsmeier, Michael; Hansen, Helle
31	Kommunisme	Kjeldsen, Michael; Riis, Alexander Guld; Engel, Denice; Christensen, Jacob Borg; Nilsson, Ida Norvin
32	Af en anden verden - en receptionsanalyse af Danidas ulandskalendermateriale "Drømmen om Burma"	Tufte, Thomas; Rewitz, Anna Furbo
33	Empowerment Through the Media	Tufte, Thomas; Morales Navarro, Amelia
34	EU Discourse on Turkey	Fabricius, Anne; Dimitrov, Gorjan
35	Representations of Language and Multilingualism in the European Union - a Grammatical Study of Discourses and Ideologies	Fabricius, Anne; Kraft, Kamilla
36	Arbejdet med de danske trolddomsprocesser	Harbsmeier, Michael; Nielsen, Kasper Egholm
37	The Future About The Past	Harbsmeier, Michael; Bitsch, Rasmus Nielsn
38	Nordboerne - i forskning og folkesagn	Harbsmeier, Michael; Rasmussen, Christian Sejer
39	En receptionsanalyse af realitygameshowet Tab & Vind	Tufte, Thomas; Nielsen, Eva Marianne; Rysz, Liselotte
40	Communication for Sustainable Development - A case study on recycling in Malaysia	Tufte, Thomas; Gyrup, Sine; Sørensen, Carina
41	Når den offentlige sfære ikke er offentlig	Tufte, Thomas; Kaarsholm, Preben; Clement, Kristine Modvig

42	Folktales for Social Change - A Study of Dialogic Democracy, Oral Culture & Communication for Social Change in Rural Malawi	Tufte, Thomas; Lægaard, Sune; Jeppesen, Jonas Agerbæk
43	English Pronunciation	Fabricius, Anne; Hansen, Niels Agerskov
44	Desperate Housewives - A Linguistic Analysis of Sex and Interactional Roles	Fabricius, Anne; Nielsen, Camilla Bergsagel
45	Galileo Galilei og kontroversen om kometerne	Harbsmeier, Michael; Højrup, Jens; Thygesen, Jacob Meyer
46	Tidens politiske handel	Harbsmeier, Michael; Olander, Peer; Litivinova, Anna
47	I den bedste mening? Om et eksperiment med grønlandske børn og mulighederne for forsoning mellem Danmark og Grønland	Harbsmeier, Michael; Jensen, Lars; Farver, Maria Bunch
48	The Condemnation of Miquel de Molinos	Harbsmeier, Michael; Knub, Nikolaj Emil Nehammer
49	Arnold J. Toynbees historiefilosofiske koncept 'udfordring-og-svar': genese, funktion, reception, revurdering	Engberg, Charlotte; Harbsmeier, Michael; Larsen, Rune
50	JOHANN FAUSTEN OG DEN SORTE GALDE – om brydninger mellem middelalderens efterår og nyere tid belyst ved Historia Von D.	Johann Fausten (1587)
51	Fanden på væggen:Et studie af et middelalderligt karaktertræk i danske kalkmalerier.	Harbsmeier, Michael; Senbergs, Asger; Christiansen, Lasse Lund
52	Croatian trade unions - Lost in transition	Olesen, Gunnar; Tufte, Thomas; Aagaard-Poulsen, Dijana Omerbasic
53	LIMITATIONS AND POTENTIALS OF HIV/AIDS EDUTAINMENT INTERVENTIONS	Tufte, Thomas; Krofa, Ditte Rytter Peasah; Ridgewell, Guilla Joan
54	Tillid til eksperterne. I forbindelse med kommunikation om genmodificerede planter	Skjold Petersen, Lasse
55	Leaving the Classroom	Mathieu, David; Petersen, Lasse

56	THE RISE OF HOMO ISLAMICUS.AMERICAN COLLECTIVE IDENTITY MARKERS AND TERRORIST IDENTITY MARKERS AS THEY REVEAL THEMSELVES IN THE DISCURSIVE COHESION OF CONCEPTS 'TERRORISM' AND 'ISLAM' IN THE WASHINGTON POST 2001-2008	Frello, Birgitta; Fabricius, Anne; Stolovitskaia, Anastassia
57	The English language as an element of inclusion/ exclusion - The dilemmas of an 'international' university	Fabricius, Anne H.; Steinhoff, Kathrin Maria
58	The Cartoon Crisis	Blackburn, Patrick; Blomsterberg, Sofie Amalie; Birkemose, Liv; Larsen, Silas Topsøe
59	Mobility, contact and an accent norm : the case of Received Pronunciation	Fabricius, Anne H.; Preisler, Bent; Fabricius, Anne; Haberland, Hartmut; Kjærbeck, Susanne; Risager, Karen; The Department of Culture and Identity
60	History and the Perception of Suicide	Harbsmeier, Michael; Jensen, Andreas L.; Warmind, David L.; Pleimert, Troels
61	Ethnographic Fieldwork: Exploring the Origins of Methodology in Social Anthropology	Harbsmeier, Michael; Valiulyte, Ilona; Adler, Tara Maria Sellner; Stockner, Selma
62	Orientens møde med Europa	Harbsmeier, Michael; Pedersen, Nanna Blak; Buchholtz, Christine; Rasmus, Munch
63	Det Skjulte Folk	Harbsmeier, Michael; Duckert, Maja Christensen; Frederiksen, Julie Inge kure
64	The Institution of Zakat in Pakistan	Casse, Thorkil; Harbsmeier, Michael; Gilani, Syed Irfan Munawar
65	Demokratibegrebet i EF-debatten 1972	Harbsmeier, Michael; Gregersen, Andreas Melson; Nikolajsen, Morten Steinmejer
66	Europa møder Verden - om Nationalmuseets fremstilling af europæisk historie	Harbsmeier, Michael; Wiendel Rasmussen, Andreas; Meyer-Juhlin, Andreas; Ryttergaard, Emil
67	Humour and Eddie Izzard	Pecic, Zoran; Fehrn-Christensen, Alexander; W. Frausing, Anne-Sophie; Volkova, Anna; Ask Sørensen, Jeppe; Renner Alstrup, Mia; Peneva, Tatyana; Holmegaard, Thorbjørn
68	Bush Pre/Post Katrina	Fabricius, Anne; Dimitrov, Gorjan; Fruensgaard, Andreas; Kaehne, Anne
69	Language variation in Shrek	Fabricius, Anne; Pretzsch, Sabine; Snowman, Katrine; Harvey, Kathrine

70	What's your identity	Fabricius, Anne; Thorsen, Mira Skadesgård; Frederiksen, Jeppe Løvstrøm
71	Louder than Words	Fabricius, Anne; Serup, Liv; Mikkelsen, Grit; Lind, Line
72	Gunz, Girlz & Gangstaz: Danish Youth's Perceptions of America	Tufte, Thomas; Braskov, Ricky; Cremer, Justin; Ebbesen, Sidsel; Pejtersen, Jon Mark; Schmidt, Christian Grimes
73	Greenlanders versus Eskimos	Tufte, Thomas; Alachovska, Ana; Jensen, Gitte; Riolino, Martina; Ayoko, Sylvia; Veral, Selen; Pavia, Girgia
74	Voice of Mpemba - An analysis of empowerment processes in a Malawian youth club fighting	Tufte, Thomas; Kaarsholm, Preben; Jensen, Lars; Grøndahl Hansen, Mette; Grauenkær Jensen, Lise; Kromann-Larsen, Stine
75	Juventude, Comunicação e mudança social: Negociação, navegação e narração da vida de jovens em uma realidade glocal	Tufte, Thomas; Kommunikation, Journalistik og Social Forandring; Institut for Kommunikation, Virksomhed og Informationsteknologier
76	The Soul Buddyz Boardgame - tomorrow is ours - en edutainmentstrategisk kommunikationsplan	Tufte, Thomas; Kromann-Larsen, Stine; Grauenkær, Lise; Vikkelsø, Stine A.; Topperzer, Martha; Grøndahl Hansen, Mette
77	Esperanto - a hopeless project?	Haberland, Hartmut; Diggelmann, Pascale Janine; Hetroy, Emeline Mylène Solange; Hjermind, Astrid Lindgreen; Holm, Juliet Eline; Jensen, Signe Marie
78	Døende riddere i højmiddelalderen - en analyse af dødsritualer i henholdvis "The song of Roland" og "History of William Marshal"	Harbsmeier, Michael; Nielsen, Denise Lindgaard; Frimann, Sine Maria; Jensen, Anna Juel; Juul-Olsen, Sara
79	Living History	Harbsmeier, Michael; Coelli, Trent; Komonen, Maija; Petkeviciute, Laura; Hansen, Grace; Pantmann, Birgitte; Kennedy, Amanda
80	Cannibalism in Colonial Congo	Harbsmeier, Michael; Lindholm, Anders; Gehin, Laurent; Bliddal, Marie; Christensen, Josefine; Mauritzen, Alexandra
81	Hvor er Europa?	Harbsmeier, Michael; Kraft, Mathias Svane; Nielsen, Christian Lund; Stage, Trine; Thomsen, Kicki
82	Occidental and Japanese Art	Harbsmeier, Michael; Kevvai, Kadi; Bauer, Liis; Minssieux, Milene; Quist Møller, Felix; Kovalevska, Aleksandra; Harbsmeier, Jesper Tobias
83	Hvor er Europa? - En europæisk dimension i historieundervisningen i folkeskolen	Harbsmeier, Michael; Thomsen, Kicki Elisabeth; Hansen Stage, Trine Cecilie; Svane Kraft, Mathias; Lund Nielsen, Christian

84	Det normanniske kongerige på Sicilien og i Syditalien	Harbsmeier, Michael; Olsen, Jannick; Møss, Henrik; June Pejtersen, Susan; Egeborg Folsberg, Jens; Malm Nielsen, Troels
85	Development Forum - the creation of an online community	Heilesen, Simon; Larsen, Lise; Juul Bjørn, Camilla
86	Reacting on style variation - an investigation of a Smack the Pony performance	Fabricius, Anne; Nielsen, Annette; Simon, Katrine Engsager; Skjødt, Julie de Molade
87	Transnationalisme på sigt	Carøe Christiansen, Connie; Hee Pedersen, Christina; Holm, Anne; Fabricius, Anne Sophie
88	A Discourse Analysis of Beverly Hills 90210	Fabricius, Anne; Dreijer, Ida Mari; Löb, Mattias; Edlefsen, Maria Louise; Zukhritdinove, Vasila
89	A New Beginning, Obama's Speech in Cairo	Fabricius, Anne; Maganga, Eric; Cicino, Martina; Siason, Ana Lea Bernadette; Westergaard, Katrine
90	The Consequences of Mobility : Linguistic and Sociocultural Contact Zones	Preisler, Bent; Fabricius, Anne H.; Haberland, Hartmut; Kjærbeck, Susanne; Risager, Karen; The Department of Culture and Identity
91	Domains and domain loss	Haberland, Hartmut; Preisler, Bent; Fabricius, Anne; Haberland, Hartmut; Kjærbeck, Susanne; Risager, Karen; The Department of Culture and Identity
92	Deconstructing 'the domain of science' as a sociolinguistic entity in EFL societies : The relationship between English and Danish in higher education and research	Preisler, Bent; Preisler, Bent; Fabricius, Anne; Haberland, Hartmut; Kjærbeck, Susanne; Risager, Karen; The Department of Culture and Identity
93	Narratives in talk-in-interaction : Organization and construction of cultural identities	Kjærbeck, Susanne; Bent, Preisler; Haberland, Hartmut; Fabricius, Anne; Kjaerbeck, Susanne; Risager, Karen; The Department of Culture and Identity
94	Languaculture as a Key Concept in Language and Culture Teaching	Risager, Karen; Preisler, Bent; Fabricius, Anne; Haberland, Hartmut; Kjærbeck, Susanne; Risager, Karen; The Department of Culture and Identity
95	Social Identity	Fabricius, Anne H.; Christoffersen, Mette Buhl; Lystoft, Sofie; Schulz, Michael; Vogel, Klara; Boukaouit, Dounia; Brøndum, Amalie
96	Turkey -a Clash of Civilisations?	Harbsmeier, Michael; Rude, Cecilie Holdt; Ursin, Malte Grane; Tsagaraki, Maria; Brudvig, Mette Rubeck; Petersen, Sif Johanna Hvid
97	Tribalism in Africa - the Complexity of Ethnicity	Harbsmeier, Michael; Asmussen, Pernille Sørup; Brochstedt, Annesofie; Jensen, Louise Ilsø; Kjær, Anna Sofia; Krogholt, Mathilde Krabbe; Mortensen, Lærke Tibert
98	Male witchcraft in early modern Europe	Harbsmeier, Michael; Bjerre, Andreas; Stigsen, Julie; Joanna, Batistaki; Møller, Maj W.; Jensen, Charlotte

		Haagen; Mortensen, Kasper Houmøller; Pipper, Cæcilie R.
99	Exploring the Republic of Letters : German Travellers in the Dutch Underground, 1690-1720	Eskildsen, Kasper Risbjerg; Nielsen, Kristian H.; Harbsmeier, Michael; Ries, Christopher J.; Humanioras teori og historie; The Department of Culture and Identity; Videnskabsstudier
100	Communication for Development	Tufte, Thomas; Omollo, Martha Akello; Kanepeja, Zane; Kløvedal, Sille; Ilsøe, Louise; Krogholt, Mathilde Krabbe; Brochstedt, Annesofie; Mortensen, Lærke Tibert; Frandsen, Pernille Holm
101	Blurred Sex and the City - An Analysis of Language and Gender in Sex and the City	Fabricius, Anne; Skov Andersen, Marie Louise; Arildsen, Vibeke; Gade Kohl, Louise; Mastilo, Tatjana; von Hedemann, Christopher Christian Johan
102	Multietnolekt	Fabricius, Anne; Sievers, Elisa; Jessen, Katrine Scott; Rasmussen, Sidsel Margrethe; Nielsen, Signe Busk; Bæk, Signe Stubager; Ganzhorn, Tina
103	Make Poverty History	Fabricius, Anne; Andersen, Henrik Bang; Acheampong, Tilda; Hesselfeldt-Nielsen, David; Holmquist, Denice; Madsen, Smita; Dircks Nybo, Morten; Schatz, Jakob
104	Developing graphical methods in sociophonetics : outline of a research agenda	Fabricius, Anne; Watt, Dominic; The Department of Culture and Identity; Medier og Kultur; Sociolinguistics, Conversation Analysis, Language Pedagogy, Sociocultural issues
105	Promoting Linguistic Diversity : Reflections on the Language Policy of European Language Policy Conferences	Daryai-Hansen, Petra Gilliyard; Preisler, Bent; Fabricius, Anne; Haberland, Hartmut; Kjærbeck, Susanne; Risager, Karen; The Department of Culture and Identity
106	De melankolske mordere	Nielsen, Carsten Tage; Nielsen, Lone; Krarup, Camilla Søs; Solgaard, Ewelina Szyburska
107	Medicinsk Kastration - En etisk vurdering	Petersen, Thomas Søbirk; Riis, Alexander Guld; Drost, Aske Mattias Folkmann; Nielsen, Lone
108	Multikulturalisme - Bhikhu Parekhs dialogiske tilgang	Lægaard, Sune; Nørtved Sørensen, Asger Abel; Nielsen, Bjørn; Aalholt Nielsen, Lone; Mikkelsen, Tor
109	The Power of Languages	Fabricius, Anne H.; Liliegren, Ann-Sofie; Peulicke Borggreen, Babette; Slott Nielsen, Christoffer; Granados Fernandez, Esther; Bellido Ponferrada, Isabel; Løwenstein Holmen, Ida; Wakimura, Yurie
110	Viral Video Campaigning on sexual transmitted diseases (STD)	Valtysson, Bjarki; Hesselfeldt-Nielsen, David; Wong, Elaine Sik-Ling; Ngassa, Jean-Luc; Petersen, Lasse; Yousaf, Saadia; Koci, Jakub
111	Looking for Belonging - ethnic minority youth as users of television	Tufte, Thomas; Holten-Andersen, Caroline; Dunger, Frida S.; Dias dos Santos, Jovane; Uski, Juha; Larsen, Lise; Nielsen, Lone; Munch, Maria Frederikke

112	Revolutionary Discourses in Postmodernity	Fabricius, Anne; Andreasen, Christian P.; Søndergaard, Mathias; Stensen, Eydfinnur A.; Uski, Juha J. O.; Petersen, Lasse; Lyall, Gavin Shaun
113	Tyge Brahe: Afskeden med Hven	Harbsmeier, Michael; Rosenberg, Vibe Appel; Engel, Denice; Uski, Juha; Lykke Larsen, Morten; Ejlerskov Jæger, Troels
114	Culture And Adult Immigrants	Holst Spenceley, Lea; Andersen, Tamar Barbara; Fogde, Anne-Sofie; Rasmussen, Ditte Ninna; Uski, Juha Janne Olavi
115	Love Refugees	Singla, Rashmi; Nielsen, Lone; Dunger, Frida S.; Holten-Andersen, Caroline; Munch, Maria Frederikke
116	Canetti as Another	Haberland, Hartmut; Nielsen, Lone; Holten-Andersen, Caroline; Skerberga, Kristine; Larsen, Lise; Dunger, Frida S.; Fischer-Nielsen, Sara; Hesselfeldt-Nielsen, David; Savio, Alejandro M.
117	Er testoteron det stof, mænd er skabt af? - Om konstruktionen af maskulinitetsbegreber i forskningsprojekter om testoteron i Danmark fra 1910'erne og frem	Hillgaard Bülow, Morten
118	Gudrun Ensslin - valget	Jensen, Henrik; Rosenberg, Vibe Appel
119	Johan Kock and the dramatic events of 1905 and 1906 in Helsinki	Hillgaard Bülow, Morten; Uski, Juha Janne Olavi
120	Udenfor - minoritetsdrenges krænkelseserfaringer og kriminalitet	Juul, Søren; Lystlund Hansen, Sara; Rosenberg, Vibe Appel
121	Reading the Global : comparative education at the end of an era	Carney, Stephen; The Department of Psychology and Educational Studies
122	Reforming the educated person in Bolivia	Carney, Stephen; Berntson, Lennart; Uski, Juha Janne Olavi
123	Social Entrepreneurship	Carney, Stephen; Wandel, Oscar; Nisted, Nina; Musa, Zaedo; Larsen, Silas; Vestdam, Jonas
124	Social Value, Social Entrepreneurship, Danish Welfare State	Carney, Stephen; Knudsen, Anna Althea; Meiner, Mathilde Stræde; Volkova, Anna; Morales, Jedediah;Jensen, Signe Marie; Hermann, Charlotte Marie;Andersen, Signe Hvalsøe
125	Looking for Belonging - ethnic minority youth as users of television	Tufte, Thomas; Holten-Andersen, Caroline; Dunger, Frida S.; Dias dos Santos, Jovane; Uski, Juha; Larsen, Lise; Nielsen, Lone; Munch, Maria Frederikke
126	Tyge Brahe: Afskeden med Hven	Harbsmeier, Michael; Rosenberg, Vibe Appel; Engel, Denice; Uski, Juha; Lykke Larsen, Morten; Ejlerskov Jæger, Troels

127	Culture And Adult Immigrants	Holst Spenceley, Lea; Andersen, Tamar Barbara; Fogde, Anne-Sofie; Rasmussen, Ditte Ninna; Uski, Juha Janne Olavi
128	Johan Kock and the dramatic events of 1905 and 1906 in Helsinki	Hillgaard Bülow, Morten; Uski, Juha Janne Olavi
129	Revolutionary Discourses in Postmodernity	Fabricius, Anne; Andreasen, Christian P.; Søndergaard, Mathias; Stensen, Eydfinnur A.; Uski, Juha J. O.;Petersen, Lasse; Lyall, Gavin Shaun

Artículos publicados

Becerril García, A., Lozano Espinosa, R., & Molina Espinosa, J. (2016). Semantic Approach to Context-Aware Resource Discovery over Scholarly Content Structured with OAI-PMH. *Computación y Sistemas , 20* (1), 127-142.

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

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 2014.* Oaxaca, México: Nova Universitas.

Modelo para consultas semánticas sensibles al contexto sobre recursos educativos estructurados con OAI-PMH

Arianna Becerril García
Universidad Autónoma del Estado de México
Toluca, México
abecerrilg@uaemex.mx

Rafael Lozano Espinosa
José Martin Molina Espinosa
Tecnológico de Monterrey
Ciudad de México, México

***Resumen*—La gran diversidad de recursos de información disponibles en la Web en la actualidad, su heterogeneidad y desagregación imponen retos importantes para el desarrollo de aplicaciones de recuperación del conocimiento. Sin embargo, existen muchos esfuerzos para organizar y estructurar la masa de datos; que van desde microformatos, y microdatos hasta protocolos de interoperabilidad cuyo objetivo radica en favorecer el compartir los datos entre aplicaciones así como permitir su integración, manipulación y explotación.**

Por lo que corresponde a la estructuración de recursos de información pertinentes para el ámbito educativo, académico y científico, se encuentra con gran aceptación el protocolo OAI-PMH. Es así, que resulta relevante construir mecanismos de búsqueda y recuperación de información sobre dichos contenidos, que aporten mayor valor a la tarea documental en este sector.

El presente documento contiene la formulación de un modelo para resolver consultas semánticas en conjuntos de metadatos estructurados de contenidos educativos disponibles en Acceso Abierto en la Web.

***Keywords*—web semántica; recursos educativos abiertos; dispositivos móviles; ontologías; OAI-PMH.**

I. Introducción

La información y el conocimiento son los principales insumos de la actividad cotidiana hoy en día, y también se constituyen en sus principales productos. La llamada sociedad del conocimiento considera la apropiación crítica y selectiva de la información para el desarrollo del ser humano. La World Wide Web, 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.

Esencial a la noción de Web es la idea de una comunidad abierta: cualquiera puede contribuir sus ideas al todo. Es esta apertura la que ha resultado en una sorprendente integralidad de temas cubiertos por la Web. Como tal, es una mezcla de diferentes análisis y presentaciones de cualquier tema de interés de cualquier persona con energía de publicar una página Web. A pesar de ser una mezcla, la Web es muy útil, pero 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 suficiente significado a un resultado en acción. Así es posible construir aplicaciones *inteligentes* que aporten un mayor conocimiento identificando en mayor profundidad los contenidos.

Por otro lado, en las últimas décadas la educación 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 [2].

Los recursos educativos accesibles en abierto en la red son materia prima para aplicaciones que favorezcan las tareas de las actividades académicas.

Aunado a ello, la recuperación de información se ve enriquecida para un usuario final cuando ésta es sensible a su contexto. En el ámbito educativo el perfil de un alumno o profesor, las asignaturas y programas que cursa, el periodo en el que se encuentra, información de espacio y tiempo de localización de un usuario, son ejemplos de datos útiles para obtener resultados más pertinentes a lo que se busca.

En el presente documento se describe un enfoque que considera los recursos de información de corte educativo disponibles en Acceso Abierto en la Web y la información sensible al contexto del usuario como insumo de un modelo para el desarrollo de un servicio de consulta semántica que provea un marco de trabajo para la construcción de aplicaciones de recuperación de información.

II. Recursos Educativos Abiertos Estructurados

La variedad de recursos de información para un alumno, académico, profesor o científico, disponibles en la Web es muy amplia. Indudablemente, dentro de los que más valor aportan a la tarea intelectual se encuentran los libros, artículos de revistas científicas, informes, actas de congreso, e-prints, archivos de datos, entre otros.

En la nube de datos disponibles en línea se encuentran diversos mecanismos usados por sitios web y repositorios para la descripción y exposición de sus datos.