Universidad de las Ciencias Informáticas

Facultad 3

Grupo de Web Semántica

Marco de trabajo basado en los datos enlazados para la interoperabilidad semántica en el protocolo OAI-PMH

Tesis presentada en opción al título de Máster en Informática Aplicada

Autor: Ing. Yusniel Hidalgo Delgado

Tutor: Dr. Amed Abel Leiva Mederos

Ciudad de la Habana, Diciembre del 2015

Declaración de autoría

Declaramos ser autores de la presente tesis y reconocemos a la UCI los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

Ing. Yusniel Hidalgo Delgado	Dr. Amed Abel Leiva Mederos
Autor	Tutor

Agradecimientos

Primeramente quiero agradecer a la Revolución y a nuestros eternos Fidel y Raúl. Sin su guía no hubiera sido posible la construcción de este proyecto que es la Universidad en la que me he formado como profesional y como persona. Agradezco a toda mi familia, en especial a mis padres, mis hermanos, mis sobrinos, mis primos y a tía Tata. Ellos siempre han estado para mi en cada paso de mi vida. Los amo mucho.

Agradezco al Dr. Amed Abel Leiva Mederos, quien además de director de tesis se ha convertido en un gran amigo. Agradezco a mis profesores del programa de maestría en Informática Aplicada. Todos son excelentes profesionales con una profunda preparación científica. A mis colegas del grupo de investigación de Web Semántica, sobre todo a los que han apoyado en la realización de esta investigación, a Ernesto, Luis, Katerín, Rocío, Yoandri, Carlos y Claudia. Todos son talentosos profesionales.

Agradezco a todos mis amigos, los que están y los que ya no están. Ellos saben lo importante que son para mi.

Agradezco a todos aquellos que de alguna manera han contribuido en mi formación y en la preparación de esta memoria de tesis.

Dedicatoria

A mi madre que, aunque lejos, siempre la guardo muy dentro de mi corazón. Gracias por hacer de mi la persona que soy hoy.

A papi, sin él, cumplir este sueño no hubiera sido posible. Te quiero mucho.

A mis hermanos Mita y Papo por confiar en mi, por apoyarme cuando más lo he necesitado.

A mis sobrinos Manuel, Alejandra y Fernanda, los quiero mucho.

A mi tía Tata, la persona más dulce y cariñosa que he conocido. Te quiero mucho. Gracias por escucharme, prometo escucharte siempre.

A mi familia toda.

Resumen

Las revistas científicas son actualmente uno de los canales de comunicación científica que regularmente utilizan los científicos para socializar sus resultados. Con el avance del movimiento de Acceso Abierto en la última década, las revistas científicas, los repositorios digitales y los catálogos en línea están proporcionando acceso abierto a todos sus contenidos. Adicionalmente, se han diseñado protocolos para el intercambio de metadatos como una solución eficiente a la interoperabilidad entre estos sistemas, siendo OAI-PMH uno de los más utilizados. OAI-PMH es un *framework* de interoperabilidad basado en la recolección de metadatos. Está diseñado para garantizar la interoperabilidad a nivel sintáctico, con muy limitadas opciones de recuperación sobre los proveedores de datos. En esta tesis se propone un marco de trabajo basado en los datos enlazados para incrementar la interoperabilidad hasta el nivel semántico de los metadatos diseminados por el protocolo OAI-PMH. Los datos enlazados se refieren a un conjunto de principios y buenas prácticas para la publicación y enlazados de datos estructurados en la web. El marco de trabajo está compuesto por dos componentes fundamentales: (1) guías metodológicas para la publicación de metadatos bibliográficos como datos enlazados y (2) plataforma informática que implementa las guías metodológicas propuestas. Las guías metodológicas siguen un enfoque iterativo e incremental. La plataforma propuesta automatiza las actividades propuestas en las guías metodológicas. Tanto las guías metodológicas como la plataforma informática fueron aplicadas en un caso de estudio para la publicación de metadatos bibliográficos como datos enlazados. Se seleccionó una muestra de nueve revistas cubanas que soportan el protocolo OAI-PMH. Los resultados obtenidos con el caso de estudio demuestran la aplicabilidad del marco de trabajo propuesto.

Abstract

Scientific journals are currently one of the channels of communication science that scientists regularly use to socialize their results. In the last decade, scientific journals, digital repositories and online catalogs have provided open access to all its contents. In addition have been designed protocols for the exchange of metadata as an efficient solution to interoperability between these systems. OAI-PMH is an interoperability framework based on metadata harvesting, one of the most widely used today. However, this protocol is designed to ensure interoperability at a syntactic level, with very limited options for querying over data providers. This thesis presents a semantic interoperability framework based on the linked data principles for bibliographic metadata disseminated by the OAI-PMH protocol. Linked data refers to a set of principles and best practices for publishing and linking structured data on the web. The framework is composed by of two major components: (1) methodological guidelines for publishing bibliographic metadata as linked data and (2) a platform that implements the proposed methodological guidelines. The methodological guidelines follow an iterative and incremental approach. The proposed platform automates the activities proposed in the methodological guidelines. Both methodological guidelines and platform were applied in a case study for publishing bibliographic metadata as linked data. A sample of nine Cuban journals that support the OAI-PMH protocol was selected. The results obtained demonstrate the applicability of the proposed framework.

Índice General

Índice de figuras

Índice de tablas

Introducción

En sus inicios, las empresas utilizaron el papel como herramienta para gestionar los procesos de negocios. De esta forma, la información se intercambiaba empleando el correo postal, lo que hacía lento y costoso dicho proceso de intercambio. Con el desarrollo de las Tecnologías de la Información y las Comunicaciones (TIC), las empresas comenzaron a desarrollar sistemas de información distribuidos. De esta forma, la comunicación entre clientes y proveedores se agilizó, sustituyendo el anterior modelo de envío de información por otro más moderno.

En este nuevo escenario, para realizar el intercambio de datos de manera eficiente, las empresas involucradas deben ponerse de acuerdo en cómo van a procesar los datos, por lo que se hizo necesario definir un conjunto de estándares que permitieran dicho intercambio.

Existen múltiples formatos y estándares para el intercambio de datos entre los sistemas de información, asociados a diferentes dominios de aplicación. En los sistemas de información empresarial, suele utilizarse el estándar eXtensible Business Reporting Language (XBRL) para el intercambio de reportes de negocio e información. En el caso de los sistemas de gestión bibliotecaria, suele emplearse el formato MAchine-Readable Cataloging (MARC21) para el intercambio de datos bibliográficos. Por otra parte, los Sistemas de Información Geográfica (SIG) suelen utilizar estándares como Keyhole Markup Language (KML) para el intercambio de datos geográficos.

Estos formatos y estándares están basados en eXtensible Markup Language (XML), un lenguaje de marcado ampliamente utilizado para intercambiar y compartir datos entre plataformas y aplicaciones. Sin embargo, el mismo posee una semántica limitada, e incluso esta semántica suele ser ambigua, es decir, no posee suficientes restricciones que permitan expresar con éxito la semántica de los datos (Domingue et al., 2011).

En el contexto cubano se ha percibido en los últimos años un notable incremento de las versiones en línea de las revistas científicas. En un diagnóstico publicado en el 2013, se pudo constatar que de una muestra de 69 revistas científicas cubanas, 61 tienen capacidades para la programación o asimilación de un sistema para la publicación en línea de la revista (Casate and Senso, 2013). En este mismo estudio se pudo constatar que en corto plazo es posible recolectar metadatos mediante el protocolo Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) de 48 revistas científicas sobre una muestra de 68, lo que supone el 70.6 % del total de las revistas encuestadas.

OAI-PMH es un *framework* de interoperabilidad basado en la recolección de metadatos. En un escenario OAI-PMH existen dos actores fundamentales, el proveedor de datos y el proveedor de servicios[1]. El proveedor de datos es el contenedor de la información o los metadatos. Suele ser cualquier sistema que proporciona los metadatos mediante una interfaz web con soporte para el protocolo OAI-PMH. Recibe peticiones OAI-PMH válidas y devuelve un conjunto de resultados normalmente estructurados en XML. Por su parte, el proveedor de servicios es cualquier sistema que realiza peticiones OAI-PMH válidas a un proveedor de datos y utiliza los metadatos recolectados para la construcción de servicios de valor añadido. Un proveedor de servicios normalmente involucra un agente software que recolecta los metadatos y una interfaz gráfica para consultarlos. El protocolo OAI-PMH sigue un mecanismo de recolección basado en la arquitectura cliente-servidor sobre el protocolo Hyper Text Transfer Protocol (HTTP) y proporciona un mecanismo básico de interoperabilidad sintáctica entre los proveedores de datos y los proveedores de servicio.

La interoperabilidad se define como la capacidad de dos o más sistemas o componentes de intercambiar información para su posterior uso (Institute of Electrical and Electronics Engineers, 1990). En el caso particular de la interoperabilidad sintáctica se refiere a la capacidad de dos sistemas de intercambiar datos mediante formatos de intercambio de datos (ej. XML) y protocolos de comunicación, como es el caso del protocolo OAI-PMH. Aunque este protocolo ha sido ampliamente adoptado a nivel internacional, este posee algunas limitaciones importantes. A continuación se presentan algunas de ellas (Hakimjavadi et al., 2012):

1. Formato de intercambio de metadatos: como se ha referido anteriormente, la especificación del protocolo establece por defecto que los metadatos bibliográficos sean intercambiados utilizando el estándar XML proporcionando una limitada interoperabilidad sintáctica. Esto significa que los metadatos son intercambiados entre proveedores de datos y proveedores de servicios, pero estos últimos no son capaces de interpretar el significado de los metadatos intercambiados. Esto ocurre debido a que el XML utilizado en las respuestas del protocolo no posee suficientes restricciones que permitan expresar con éxito la semántica de los metadatos intercambiados. Adicionalmente, algunos proveedores de datos que implementan el protocolo OAI-PMH proporcionan una interfaz web para que los humanos puedan comprender los metadatos existentes en el mismo.

[1] http://www.openarchives.org/OAI/openarchivesprotocol.html

2. Recuperación selectiva de los metadatos: el protocolo OAI-PMH utiliza un enfoque de intercambio de metadatos basado en recolección. Los proveedores de servicios normalmente incluyen un recolector de metadatos que se encargan de obtener los metadatos existentes en el proveedor de datos. En este contexto suele ser de utilidad la formulación de consultas para obtener no todos los registros, sino un subconjunto de los mismos. Consultar un proveedor de datos para obtener todos los metadatos de los registros bibliográficos de un determinado autor suele ser frecuentemente una necesidad, sin embargo, OAI-PMH no es capaz de responder a este tipo de consultas. Esto se debe a que la especificación del protocolo establece que los únicos criterios de consulta que se pueden formular al proveedor de datos son las siguientes: (1) obtener un registro dado su identificador, (2) obtener los registros que pertenecen a una colección, (3) obtener los registros en un formato de metadato especificado (Dublin Core, MARCXML) y (4) obtener los registros dado un rango de fechas especificado.

Las limitaciones anteriores están presentes en todos los proveedores de datos que utilizan el protocolo OAI-PMH para la diseminación de sus metadatos lo que afecta la capacidad de utilización de los mismos por los proveedores de servicios.

Por otro lado, en los últimos años ha emergido, en el contexto de la web semántica, una novedosa forma de publicación de datos estructurados en el contexto de la web denominada como datos enlazados. Los datos enlazados se refieren a un conjunto de principios y buenas prácticas para la publicación y enlazado de datos estructurados en la web (Berners-Lee, 2006). La publicación de datos siguiendo los principios de los datos enlazados potencia el descubrimiento y la reutilización de los datos en el espacio de la web a la vez que resuelve los problemas de interoperabilidad semántica entre los sistemas de información mediante el uso de ontologías.

En los últimos años se han definido varios estándares que soportan la publicación de datos basándose en los datos enlazados. Algunos de estos estándares son Resource Description Framework (RDF) y SPARQL Protocol and RDF Query Language (SPARQL). El primero de ellos se refiere a un modelo de datos basado en grafos dirigidos para la descripción de recursos en el contexto de la web. Por su parte, SPARQL constituye el lenguaje de consultas utilizado para consultar el modelo de datos RDF (Heath and Bizer, 2011).

La especificación del modelo de datos RDF establece que los grafos dirigidos están formados por tripletas del tipo sujeto-predicado-objeto, donde el sujeto y el objeto constituyen los nodos del grafo y el predicado constituye los arcos que unen al sujeto y el objeto. El predicado de la

tripleta corresponde con una propiedad de un vocabulario u ontología y establece la semántica de la relación. Aunque existen diversos formatos para la serialización de los grafos RDF, estos son procesables por las computadoras de manera automática.

Teniendo en cuenta las limitaciones presentes en el protocolo OAI-PMH se establece como **problema de investigación** el siguiente: ¿Cómo incrementar la interoperabilidad de los metadatos bibliográficos diseminados mediante el protocolo OAI-PMH?

El problema de investigación se enmarca en el **objeto de estudio** de la interoperabilidad semántica y el en **campo de acción** de los marcos de trabajo para la interoperabilidad semántica en el protocolo OAI-PMH.

Esta investigación se propone como **objetivo general** desarrollar un marco de trabajo basado en los principios de los datos enlazados para incrementar la interoperabilidad de los metadatos bibliográficos diseminados por el protocolo OAI-PMH.

Como **objetivos específicos** se definen los siguientes:

1. Identificar las principales aproximaciones teóricas referentes a la interoperabilidad semántica en el contexto de los metadatos bibliográficos diseminados mediante el protocolo OAI-PMH.

2. Proponer guías metodológicas para la publicación de metadatos bibliográficos siguiendo los principios de los datos enlazados diseminados mediante el protocolo OAI-PMH.

3. Desarrollar una herramienta informática basada en las guías metodológicas propuestas que contribuya al incremento de la interoperabilidad de los metadatos.

4. Evaluar la propuesta de solución mediante un caso de estudio en un escenario real de utilización.

Se plantea como **hipótesis** de la investigación que si se desarrolla un marco de trabajo basado en los principios de los datos enlazados, entonces se incrementará la interoperabilidad de los metadatos bibliográficos diseminados mediante el protocolo OAI-PMH.

Entre las estrategias de investigación que se utilizarán están la exploratoria y la explicativa. Se pretende explorar, aplicando el método analítico sintético las principales aproximaciones existentes en la literatura para resolver el problema de la interoperabilidad semántica en los metadatos bibliográficos diseminados mediante el protocolo OAI-PMH. Se pretende además la utilización del método histórico lógico para evaluar cómo se ha comportado en los últimos cinco años el uso de las ontologías para modelar el dominio de los datos que participan en el estudio.

Para la fundamentación y elaboración del marco de trabajo se pretende el uso de los métodos modelación e inductivo deductivo. En el primer caso, se pretende la modelación de la propuesta utilizando técnicas de visualización de la información así como algunos artefactos de ingeniería que ayuden a comprender sus componentes y sus interrelaciones. Finalmente se utilizará el caso de estudio como método para evaluar la aplicabilidad del modelo propuesto en escenarios reales.

El documento de tesis está estructurado en tres capítulos. A continuación una breve descripción de cada uno de ellos.

En el **capítulo 1** se enuncian los conceptos fundamentales sobre los datos enlazados los cuales permitirán una mejor comprensión de este trabajo. Se presenta un análisis del estado del arte de las principales herramientas, ontologías y vocabularios utilizados en la publicación de datos estructurados en la web de los datos, con énfasis en el dominio de los metadatos bibliográficos. Por último se analizan algunas guías metodológicas y buenas prácticas utilizadas en la publicación y consumo de datos enlazados y que pueden ser utilizadas en la publicación de datos enlazados en el contexto de las bibliotecas digitales.

En el **capítulo 2** se presenta un marco de trabajo basado en los datos enlazados para incrementar la interoperabilidad en el protocolo OAI-PMH. Como parte de la propuesta ha sido implementada una plataforma informática utilizando tecnologías actuales del desarrollo de software.

En el **capítulo 3** se desarrolla un caso de estudio en un contexto real de utilización, empleando revistas cubanas de Acceso Abierto que soportan el protocolo OAI-PMH para el intercambio de metadatos en la web. Adicionalmente, se realiza un pre experimento para evaluar las capacidades de consulta de la propuesta en comparación con el protocolo OAI-PMH. Se describen en detalle los principales resultados obtenidos con el caso de estudio y el experimento desarrollado.

Capítulo 1

Fundamentación teórica

1.1. Introducción

En este capítulo se enuncian los conceptos fundamentales sobre los datos enlazados los cuales permitirán una mejor comprensión de este trabajo. Se presenta un análisis del estado del arte de las principales herramientas, ontologías y vocabularios utilizados en la publicación de datos estructurados en la web de los datos, con énfasis en el dominio de los metadatos bibliográficos. Por último se analizan algunas guías metodológicas y buenas prácticas utilizadas en la publicación y consumo de datos enlazados y que pueden ser utilizadas en la publicación de datos enlazados en el contexto de las bibliotecas digitales.

1.2. Conceptos fundamentales

En este epígrafe se analizan los principales conceptos que son necesarios para la comprensión del mismo. Se identifican los nexos existentes entre cada uno de ellos y con el objeto de estudio de la investigación.

1.2.1. Interoperabilidad

El concepto de interoperabilidad ha sido enunciado por varios autores en distintos contextos. Según el Institute of Electrical and Electronics Engineers (IEEE), la interoperabilidad se define como la capacidad de dos o más sistemas o componentes de intercambiar información para su posterior uso. En este sentido, tres niveles de interoperabilidad son propuestos: **interoperabilidad técnica**, **interoperabilidad sintáctica** e **interoperabilidad semántica** (Institute of Electrical and Electronics Engineers, 1990), siendo estos tres niveles los más citados en la literatura (M. A. Manso et al., 2008).

La interoperabilidad técnica es aquella que soporta la comunicación entre dos sistemas mediante el uso de señales y protocolos de comunicación (ej. TCP/IP). La interoperabilidad sintáctica se refiere a la capacidad de dos sistemas de intercambiar datos mediante formatos de intercambio de datos (ej. XML) y protocolos de comunicación. Por último, si los datos intercambiados son automáticamente interpretados de manera que se extrae el significado de los mismos, entonces los

sistemas garantizan la interoperabilidad semántica.

Según el Marco de Interoperabilidad Europeo (European Interoperability Framework (EIF)) la interoperabilidad semántica permite a las organizaciones procesar información de fuentes externas. Garantiza que el significado de la información intercambiada es comprendida y preservada a través de los intercambios entre las partes. Engloba los siguientes aspectos (Commission, 2010):

- *Interoperabilidad semántica:* se refiere al significado de los elementos de datos y la relación entre ellos. Incluye el desarrollo de de vocabularios para describir los datos intercambiados para garantizar que los datos son comprendidos de la misma manera entre las partes en comunicación.
- *Interoperabilidad sintáctica:* se refiere a la descripción exacta del formato de la información que será intercambiada en términos de gramática, formato y esquemas.

El concepto de interoperabilidad semántica definido en el EIF comprende el desarrollo de vocabularios para describir de manera explicita la semántica o el significado de los datos y las relaciones que se establecen entre ellos. Por este motivo se asume este concepto a lo largo de la presente investigación. Actualmente se desarrollan estándares y tecnologías en el contexto de la web semántica para garantizar la interoperabilidad semántica entre los sistemas de información.

1.2.2. Web Semántica

La web actual se basa en un conjunto de tecnologías y estándares definidas por la World Wide Web Consortium (W3C), dentro de los que se encuentran el protocolo HTTP, las Uniform Resource Identifier (URI)s (Masinter et al., 2005) y el lenguaje de marcado Hyper Text Marckup Language (HTML) para la presentación de los contenidos. Este último carece de un mecanismo para expresar el significado de los contenidos publicados, imposibilitando a las computadoras procesar los mismos automáticamente.

En el 2001, el creador de la web Tim Berners-Lee publica un artículo acuñando el término de web semántica. La misma se define como *una extensión de la web actual en la que la información tiene un significado bien definido posibilitando a los humanos y las computadoras trabajar en cooperación* (Berners-Lee et al., 2001). En la última década, la W3C ha venido trabajando en la definición de estándares y tecnologías que contribuyan al rápido crecimiento de la web semántica.

1.2.3. Datos enlazados

En el contexto de la web semántica juegan un rol importante los datos enlazados. Los datos enlazados se refieren a un conjunto de principios y buenas prácticas para la publicación y enlazado de datos estructurados en la Web. Estos datos provienen de diferentes fuentes que pueden ser mantenidas por organizaciones con diferentes localizaciones geográficas. La idea básica de los datos enlazados es aplicar la arquitectura general de la Web a la tarea de compartir datos estructurados a escala global (Berners-Lee, 2006).
Para publicar datos estructurados como datos enlazados, es necesario seguir cuatro principios básicos:

1. Utilizar URIs como nombres para identificar los recursos en la web.
2. Utilizar URIs basadas en el protocolo HTTP, de modo que las personas puedan ver esos nombres.
3. Proporcionar información adicional, usando los estándares RDF y SPARQL.
4. Incluir enlaces a otras URIs, de modo que se potencie el descubrimiento de información adicional sobre el recurso.

El primer principio propone el uso de URIs para identificar no solo documentos web y contenido digital, sino que sirva además para referenciar a objetos del mundo real y conceptos abstractos. El segundo principio propone el uso de URIs basadas en el protocolo HTTP para identificar objetos y conceptos abstractos, posibilitando que estas URIs estén desreferenciadas sobre dicho protocolo y en cambio proporcionen una descripción del objeto o concepto identificado. El tercer principio propone el uso del modelo de datos RDF para publicar datos estructurados en la web. El cuarto principio propone el uso de enlaces para enlazar no solo documentos web, sino cualquier tipo de recurso. Por ejemplo, un enlace puede establecerse entre un lugar y una persona, o entre un lugar y una empresa.

1.2.4. Resource Description Framework (RDF)

Con el objetivo de proporcionar un formato único procesable para la representación y descripción de los datos en la web semántica, la W3C define el estándar RDF, el cual consiste en un modelo de datos simple y totalmente compatible con la arquitectura de la web actual (Heath and Bizer, 2011).

Un documento RDF consiste en un conjunto de tripletas de la forma sujeto-predicado-objeto (figura 1.1) de manera que los datos pueden ser representados en un grafo dirigido donde la primera y la tercera componerte corresponde a los nodos del grafo y la segunda componente (predicado) actúa como enlace (arco) entre dichos nodos. Al grafo dirigido descrito anteriormente se le conoce como Grafo RDF y utiliza las ontologías para la descripción formal de los datos en términos de clases y propiedades (Klyne and Caroll, 2004).

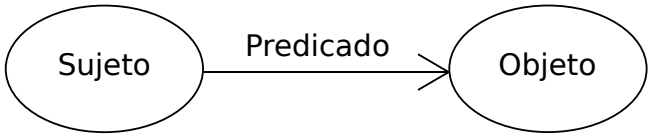

Figura 1.1: Modelo de datos basado en grafo del estándar RDF

1.2.5. Ontologías

El término ontología es utilizado con diferentes significados en diferentes comunidades. Su origen se encuentra en la filosofía y son utilizadas para estudiar la naturaleza del ser y la existencia. En computación, una definición ampliamente aceptada por la comunidad científica es la siguiente: *una ontología es una especificación formal y explícita de una conceptualización compartida* (R. Studer et al., 1998). Existen cuatro tipos diferentes de ontologías a diferentes niveles de granularidad (Steffen Staab and Rudi Studer, 2009). En un nivel superior se encuentran las ontologías fundacionales las cuales capturan conceptos generales independientes de un dominio específico. En un segundo nivel de abstracción se encuentran las ontologías de dominio las cuales modelan conceptos y relaciones que son relevantes para un dominio específico. En estas ontologías se suelen reutilizar términos de las ontologías fundacionales.

Otro tipo de ontología son las ontologías de tareas las cuales describen conceptos de una tarea en específico. A un nivel más bajo de abstracción se encuentran las ontologías de aplicaciones. Estas combinan ontologías de dominio y ontologías de tareas extendiéndolas con nuevos conceptos y relaciones más específicos.

Todas las ontologías están compuestas por cinco tipos de componentes: clases, relaciones, funciones, instancias y axiomas (Asunción Gómez-Pérez et al., 2004).

Las **clases** representan conceptos, abstractos o no, propios de un dominio específico. Las clases en una ontología se suelen organizar en taxonomías a las que se les pueden aplicar los mecanismos

de herencia.

Las **relaciones** representan un tipo de asociación entre conceptos de un dominio. Las ontologías con frecuencia poseen relaciones binarias, donde el primer argumento es conocido como dominio y el segundo como rango.

Las **funciones** son un caso particular de relaciones en las cuales el n-ésimo elemento de la relación es único para los $n-1$ elementos precedentes. Las funciones son usualmente expresadas de la forma:

$$F : C1xC2x...xCn-1 \Longrightarrow Cn \tag{1.1}$$

Las **instancias** se usan para representar elementos o individuos en una ontología.

Por último los **axiomas formales** sirven para modelar sentencias que son siempre verdaderas. Normalmente se usan para representar conocimiento que no puede ser formalmente definido por los componentes descritos anteriormente. Además, se usan para verificar la consistencia de la propia ontología.

Existe un conjunto de tecnologías que han propiciado el desarrollo computacional de las ontologías. Entre ellas se encuentran los lenguajes de definición de ontologías, siendo los más populares Ontology Web Language (OWL) y RDF Schema (RDFS). También podemos encontrar algunos razonadores tales como Pellet (Sirin et al., 2007) y FaCT++ (Tsarkov and Horrocks, 2006) los cuales permiten inferir conocimiento nuevo a partir del conocimiento modelado en las mismas.

Las ontologías constituyen elementos claves en el desarrollo de las tecnologías de la web semántica. Las mismas han sido utilizadas para modelar objetos abstractos o físicos en términos de clases y propiedades. Sin las ontologías, la web semántica carece del significado explícito de los datos y por tanto resultaría imposible que las computadoras procesen los mismos automáticamente.

1.3. Análisis de las soluciones existentes

El desarrollo de proyectos para la publicación y enlazado de datos estructurados en la web implica tomar un conjunto de decisiones técnicas y metodológicas complejas. ¿Cuáles son las herramientas existentes para la publicación de datos enlazados en el dominio de los metadatos bibliográficos?¿Cuáles son las principales limitaciones que poseen estas herramientas?¿Qué vocabularios y ontologías son los más utilizados para modelar este dominio del

conocimiento?¿Cuáles son los pasos principales que se deben tener en cuenta a la hora de acometer el desarrollo de un proyecto de este tipo? En este epígrafe se realiza un análisis crítico de las principales aproximaciones existentes para la publicación y enlazado de datos estructurados en la web de los datos, con énfasis en el dominio de los metadatos bibliográficos.

1.3.1. Herramientas para la publicación de datos enlazados

El problema de publicar metadatos bibliográficos siguiendo los principios de los datos enlazados no es nuevo. Varias instituciones alrededor del mundo han destinado ingentes esfuerzos para transformar sus metadatos a grafos RDF.

Los catálogos de las bibliotecas contienen una enorme cantidad de datos estructurados de alta calidad, sin embargo, estos datos generalmente no están disponibles en la web semántica. Una de las primeras iniciativas para transformar estos datos en datos enlazados fue la propuesta por la Unión de Catálogos de Suecia (LIBRIS) (Malmsten, 2008). Para la transformación de los registros bibliográficos en formato MARC21, se implementó un *RDF Server Wrapper*. Este servidor utiliza una aproximación basada en eXtensible Stylesheet Language Transformations (XSLT) para realizar dicha transformación.

Uno de los protocolos más extendidos para la diseminación de metadatos bibliográficos es OAI-PMH. En (Haslhofer and Schandl, 2008) y (Haslhofer and Schandl, 2010) los autores proponen *OAI2LOD Server*, una herramienta para la transformación de metadatos bibliográficos diseminados mediante OAI-PMH a datos enlazados. Esta herramienta posee algunas limitaciones, las cuales fueron resueltas por (Coppens et al., 2009).

En (Daniel Vila-Suero et al., 2013) y (Vila-Suero et al., 2012) los autores proponen la herramienta *Marimba* para la transformación de registros en MARC21 a grafos RDF. La herramienta proporciona un entorno para la alineación de clases y propiedades de ontologías con los elementos de metadatos del formato MARC21. Esta tarea es realizada manualmente por los bibliotecarios a partir de un conjunto de hojas de cálculo generadas por dicha herramienta.

En (Victor de Boer et al., 2012) los autores describen la herramienta *XMLRDF*, la cual es utilizada para transformar registros de metadatos bibliográficos en grafos RDF. La transformación es realizada en dos etapas. En una primera etapa se realiza una conversión sintáctica de los archivos XML fuentes en grafos RDF crudos. El RDF obtenido debe ser lo más cercano posible al metadato original. En una segunda etapa el RDF obtenido de esta manera es reestructurado y enriquecido. La herramienta proporciona una biblioteca dinámicamente extendida de rutinas en

Prolog para tareas comunes de conversión.

La herramienta *OAI-PMH RDFizer* (Stefano Mazzocchi, 2007) fue desarrollada por el proyecto SMILE[1] del Massachusetts Institute of Technology (MIT). Esta herramienta convierte metadatos diseminados mediante el protocolo OAI-PMH en grafos RDF. La herramienta utiliza hojas de estilo XSLT para realizar la transformación.

Todas las herramientas estudiadas poseen las siguientes limitaciones:

1. Las herramientas toman como entrada una única fuente de datos, sin que medie un proceso de integración de datos a partir de fuentes de datos heterogéneas.

2. Aplican transformaciones basadas en XSLT, lo que implica que un cambio en la fuente de datos, implica necesariamente cambios en las hojas de estilo, muchas de las cuales son editadas manualmente.

3. En la mayoría de los casos se asume que los metadatos existentes en la fuente de datos original poseen una calidad adecuada, lo que no siempre se evidencia en la práctica. Se hace necesario entonces tareas de preprocesamiento de los metadatos antes de ser publicados como datos enlazados.

1.3.2. Ontologías y vocabularios

Como se mencionó en la sección 1.2.5, las ontologías son utilizadas para modelar los datos existentes en un dominio dado. Con el objetivo de identificar aquellas ontologías que más se utilizan en el dominio de las bibliotecas e instituciones que manejan grandes volúmenes de metadatos bibliográficos, se siguió el siguiente procedimiento. En el grupo Library Linked Data del portal thedatahub.ie, se identificaron aquellas instituciones con colecciones de metadatos bibliográficos publicados siguiendo los principios de los datos enlazados, seleccionándose una muestra de siete instituciones europeas.

1. Unión de Catálogos de Suecia (Libris)

2. Biblioteca Nacional de España (BNE)

3. Biblioteca Nacional Alemana (GNB)

4. Biblioteca Nacional Británica (BNB)

[1] http://simile.mit.edu/

5. Biblioteca Nacional de Francia (BNF)

6. Europeana

7. Open University (OU)

Por cada una de estas instituciones, se identificaron las ontologías utilizadas en cada una de las colecciones publicadas como datos enlazados. En la tabla 1.1 se muestran las ontologías utilizadas por cada una de estas instituciones.

	DC	FOAF	SKOS	FRBR	ISBD	RDA	MADS	FRAD	FRSAD	BIBO	OWLT	ORE	EDM
LIBRIS	X	X	X	X									
BNE	X		X	X	X	X	X	X	X				
GNB	X	X			X					X			
BNB	X		X		X	X				X	X		
BNF	X	X	X	X		X							
Europeana	X	X										X	X
OU		X								X			

Tabla 1.1: Utilización de las ontologías por colecciones de datos

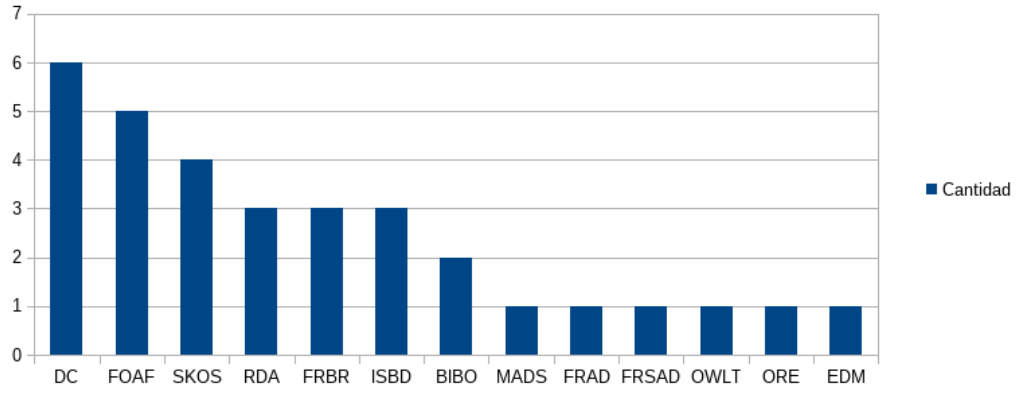

Figura 1.2: Histograma de las ontologías encontradas

Dublin Core[2]: es un esquema ampliamente utilizado para la descripción de metadatos. Esta siendo mantenido actualmente por la Dublin Core Metadata Initiative (DCMI), una organización sin ánimos de lucro establecida en Singapur. Se caracteriza por su simplicidad y generalidad y consiste en 15 campos de metadatos, entre los que se encuentran: *creator*, *subject*, *coverage*, *description* y *date*. Los mismos permiten responder a las preguntas básicas: quién, qué, dónde y

[2] http://dublincore.org/

cuándo (Coppens et al., 2009). No es de extrañar que este esquema de metadatos sea ampliamente utilizado en la descripción de recursos en el contexto de los metadatos bibliográficos, aunque ha sido utilizado en otros dominios de aplicación.

FOAF[3]: Friend of a Friend (FOAF) es una tecnología que facilita la compartición y uso de información sobre personas y sus actividades. Utiliza un formato basado en RDF para la descripción de relaciones entre personas y lugares en el contexto de la web. Este formato puede ser procesado por las computadoras de manera automática.

SKOS[4]: Simple Knowledge Organization System (SKOS) proporciona una forma estándar para la representación de Sistemas de Organización de Conocimiento tales como tesauros, taxonomías y esquemas de clasificación en el contexto de la web semántica. El mismo utiliza el estándar RDF para la codificación de la información, proporcionando un nivel de interoperabilidad entre sistemas informáticos. La especificación de SKOS está publicada actualmente como una recomendación del W3C.

RDA[5]: Resource Description and Access (RDA) es una nueva norma de catalogación de recursos digitales que pone mayor énfasis en auxiliar al usuario a encontrar, identificar, seleccionar y obtener la información deseada. Esta basado en dos modelos conceptuales desarrollados por la International Federation of Library Associations and Institutions (IFLA). Ellos son: Functional Requirements for Bibliographic Records (FRBR) y Functional Requirements for Authority Data (FRAD). FRBR y FRAD identifican las relaciones que una obra puede tener con su creador, así como las relaciones con las traducciones, interpretaciones, adaptaciones o formato físico de dicha obra (Ana Lupe Cristán and Elena Escolano Rodríguez, 2009).

FRBR[6]: FRBR es un estándar de la IFLA para modelar entidades y sus relaciones en la catalogación de registros bibliográficos. El modelo está diseñado para proporcionar las cuatro tareas siguientes: encontrar, identificar, seleccionar y obtener materiales que se correspondan con los criterios de búsqueda establecidos por los usuarios. Las entidades se han divido en tres grupos y representan los objetos claves que interesan a los usuarios de los datos bibliográficos. El primer grupo incluye los productos de creación intelectual o artística que se consignan o describen en los registros bibliográficos: **obra**, **expresión**, **manifestación** e **ítem**. El segundo grupo incluye aquellas entidades responsables del contenido intelectual o artístico, la producción y difusión

[3] http://www.foaf-project.org/
[4] http://www.w3.org/2004/02/skos/
[5] http://www.rda-jsc.org/rda.html
[6] http://vocab.org/frbr/core.html

física o la custodia de dichos productos: **personas** y **entidades corporativas**. El tercer grupo incluye un conjunto adicional de entidades que sirven como sujetos de una producción artística o intelectual: **concepto**, **objeto**, **acontecimiento** y **lugar** (Xavier Agenjo and María Luisa Martínez-Conde, 1998).

Uno de los primeros vocabularios RDF basado en el modelo fue el propuesto por Richard Newman y Ian Davis (Ian Davis and Richard Newman, 2005). Este vocabulario incluye clases RDF para los grupos de entidades 1, 2 y 3 descritos por el reporte final de FRBR.

La ontología FRBRoo (William Denton, 2008) es una aproximación diferente para la formalización de FRBR como un modelo orientado a objeto, armonizado con el modelo CIDOC-CRM[7]. Esta ontología captura y representa la semántica de la información bibliográfica facilitando su integración e intercambio.

Dublin Core se consolida como el esquema de metadatos más empleado en la actualidad, no solo en el contexto de las bibliotecas, sino en otros ámbitos de aplicación. Se distingue por su simplicidad y flexibilidad, atributos que lo sitúan como el esquema de metadatos que más ha sido reutilizado en la definición de ontologías de dominio en el contexto de las bibliotecas digitales. Adicionalmente, los estándares de la IFLA continúan ganando en aplicabilidad en distintos escenarios complejos en la gestión de grandes colecciones de metadatos.

1.3.3. Experiencias en la publicación de datos enlazados

Para la generación y publicación de colecciones de datos enlazados se requieren de un alto número de pasos, tecnologías y decisiones de diseño (Boris Villazón-Terrazas et al., 2011). Sin embargo, aún no se cuenta con una metodología de desarrollo para este tipo de proyectos que haya sido adoptada completamente por la comunidad científica. Esto ha traído consigo que los productores de datos enlazados tengan que adoptar sus propias herramientas, tecnologías y métodos para emprender este tipo de proyectos.

Existen en la literatura algunas guías metodológicas o buenas prácticas que se pueden seguir para generar y publicar datos enlazados. En (Bernadette Hyland and David Wood, 2011) los autores proponen un conjunto de seis pasos que se deben seguir para publicar datos enlazados. Estos seis pasos son: **Identificar**, **Modelar**, **Nombrar**, **Describir**, **Convertir** y **Publicar**.

En el primer paso se debe identificar y analizar la fuente de datos, ya sea una base de datos relacional, archivos Comma-Separated Values (CSV) o diccionarios de datos, con el objetivo de

[7] http://www.cidoc-crm.org/

identificar objetos de interés: lugares, personas, etc. En el segundo paso se deben modelar los datos reutilizando, tanto como sea posible, los vocabularios y ontologías existentes en el dominio de los datos. En el tercer paso se definen las URI para cada objeto dentro de la fuente de datos, teniendo en cuenta cómo los datos pueden cambiar a lo largo del tiempo. El cuarto paso consiste en describir los datos en formatos que sean entendibles tanto por los humanos (HTML, RDFa) como por las computadoras (RDF/XML). El quinto paso consiste en convertir a grafos RDF los datos previamente modelados. Existen varias serializaciones de RDF que se pueden utilizar, por ejemplo: RDF/XML[8] , Notation-3 (N3)[9], Turtle[10] , N-Triples y RDFa. Un elemento resaltado en este trabajo es la importancia que le concede al mantenimiento de los datos enlazados publicados. Es preciso contar con datos actualizados y precisos tanto como sea posible, ya que los mismos, pueden ser utilizados por aplicaciones desarrolladas por terceros.

En (Boris Villazón-Terrazas et al., 2011), se propone una guía metodológica para la publicación de datos enlazados en el dominio de las administraciones de gobierno. Esta guía asume que el proceso de publicación de datos enlazados debe tener un ciclo de vida, de la misma forma que la Ingeniería de Software.

De acuerdo a la experiencia de los autores, el proceso de publicación de datos enlazados tiene un modelo de ciclo de vida iterativo e incremental donde las mejoras continuas y extensiones hechas a los datos enlazados son resultado de realizar sucesivas iteraciones. Las guías metodológicas comprenden 5 actividades principales, las cuales se desglosan en tareas. Las actividades son: **Especificación**, **Modelado**, **Generación**, **Publicación** y **Consumo**.

En la actividad de especificación se identifican y analizan las fuentes de datos, se diseñan las URI para cada uno de los recursos y se define la licencia de los datos enlazados publicados. En la actividad de modelado se deben construir o reutilizar la o las ontologías y vocabularios necesarios para modelar los datos teniendo en cuenta el dominio de los mismos. En la actividad de generación se transforman a grafos RDF los datos existentes en las fuentes de datos. Esta actividad comprende la limpieza de los datos y el enlazados de los mismos con otras colecciones de datos enlazados previamente publicados, potenciando el descubrimiento y reutilización de los mismos. En la siguiente actividad se publican los grafos RDF en la web. Estos grafos deben estar disponibles mediante un SPARQL Endpoint para su posterior utilización. Es importante destacar que no solo se deben publicar los grafos, sino también los metadatos de estos grafos empleando el

[8] http://www.w3.org/TR/rdf-syntax-grammar/

[9] http://www.w3.org/DesignIssues/Notation3

[10] http://www.w3.org/TeamSubmission/turtle/

vocabulario VoID. La última actividad consiste en el desarrollo de herramientas que utilizan los datos enlazados publicados. Algunas de estas herramientas son: Navegadores de Datos Enlazados, Motores de Búsqueda de Datos Enlazados y aplicaciones específicas de un dominio.
Aunque existen diversos enfoques para enfrentar proyectos de datos enlazados, no existe consenso en la comunidad científica sobre cuáles son las guías metodológicas a seguir. El enfoque ha seguir depende en gran medida de la naturaleza de las fuentes de datos, la experiencia de los productores de datos enlazados y la madurez de las herramientas utilizadas para realizar este complejo proceso. Un elemento importante a tener en cuenta es la calidad de los datos estructurados que serán publicados como datos enlazados, sin embargo, este problema ha sido poco abordado en los enfoques estudiados.

1.4. Conclusiones parciales

Luego de estudiadas las principales aproximaciones existentes en la literatura especializada sobre la publicación de metadatos bibliográficos como datos enlazados, se concluye lo siguiente:

1. Las herramientas existentes para la publicación de metadatos bibliográficos como datos enlazados están diseñadas para la transformación de una única fuente de datos, sin que medie un proceso de integración de datos a partir de fuentes de datos heterogéneas. Además, ninguna realiza un preprocesamiento de los metadatos bibliográficos con el objetivo de mejorar la calidad de los mismos.

2. Dublin Core se consolida como el esquema de metadatos que mejor aceptación ha tenido en los últimos años en la definición de clases y propiedades en ontologías de dominio desarrolladas en el contexto de las bibliotecas digitales en general.

3. Aunque existen diversos enfoques para la publicación de datos como datos enlazados, aún no existe consenso en la comunidad científica de cuáles son las guías metodológicas que se deben seguir para enfrentar un proyecto de este tipo. El enfoque ha seguir dependerá en gran medida de la naturaleza de las fuentes de datos, la experiencia de los productores de datos enlazados y la madurez de las herramientas utilizadas para realizar este complejo proceso.

Capítulo 2

Propuesta de solución

2.1. Introducción

En este capítulo se presenta un marco de trabajo para incrementar la interoperabilidad semántica en el protocolo OAI-PMH. El marco de trabajo propuesto incluye la implementación de una plataforma informática utilizando tecnologías actuales del desarrollo de software.

2.2. Descripción general de la propuesta

El marco de trabajo propuesto consta de dos componentes fundamentales:

1. Guías metodológicas para la publicación de metadatos bibliográficos como datos enlazados. Las guías metodológicas consisten en cinco actividades fundamentales las cuales serán descritas en las secciones siguientes. Con la aplicación de estas guías metodológicas se pretende incrementar la interoperabilidad semántica en el protocolo OAI-PMH.

2. Plataforma informática para la extracción, publicación y consumo de metadatos bibliográficos como datos enlazados. En la concepción de la plataforma se tuvo en cuenta las actividades comprendidas en las guías metodológicas anteriormente mencionadas.

Tanto las guías metodológicas como la plataforma implementada son utilizadas en un escenario real. Los resultados obtenidos de su aplicación se describen en el capítulo 3 de esta memoria de tesis.

2.3. Guías metodológicas

En esta sección se proponen unas guías metodológicas basadas en (Boris Villazón-Terrazas et al., 2011) para la publicación de metadatos bibliográficos como datos enlazados. Estas guías metodológicas no pretenden ser exhaustivas ni aplicables a todos los dominios de aplicación. Siguen un enfoque iterativo e incremental, es decir, los resultados intermedios son mejorados luego de la aplicación de varias iteraciones. También siguen un enfoque basado en *pipeline*, donde la salida de la actividad anterior constituye la entrada a la actividad siguiente. Las guías

metodológicas propuestas consisten en 5 actividades fundamentales: (1) Extracción de datos, (2) Preprocesamiento de datos, (3) Modelado de datos, (4) Publicación de datos y (5) Consumo de datos.

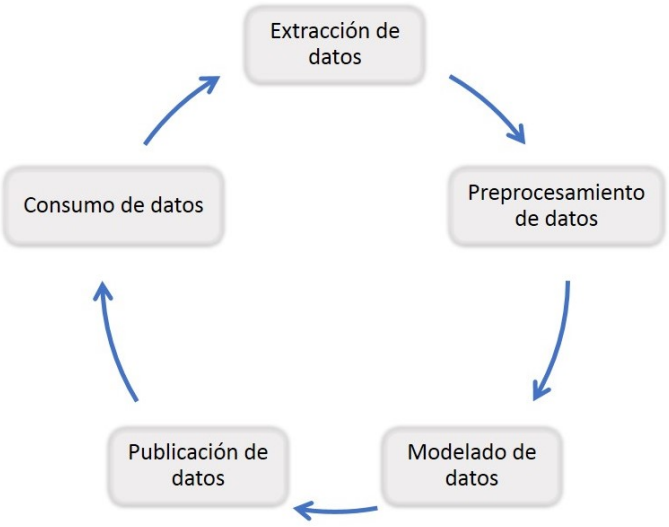

Figura 2.1: Guías metodológicas para la publicación de metadatos bibliográficos como datos enlazados

2.3.1. Extracción de datos

El objetivo de esta actividad es extraer y almacenar los metadatos bibliográficos desde los diferentes proveedores de datos que soporten el protocolo de intercambio OAI-PMH. Las entradas a esta actividad son las rutas completas a los *OAI-PMH Endpoints*. Un *OAI-PMH Endpoint* es una Uniform Resource Locator (URL) contenida en el proveedor de datos que utilizan los recolectores de metadatos para realizar las peticiones mediante el método GET del protocolo HTTP. La salida es una base de datos intermedia con los metadatos bibliográficos extraídos. A nivel tecnológico, para la realización de esta actividad es necesario contar con dos componentes esenciales: (1) recolector OAI-PMH y (2) motor de base de datos. El recolector debe implementar la especificación del protocolo OAI-PMH que aparece libremente accesible en Internet[1]. En esta especificación aparecen, entre otros elementos, los conceptos y definiciones principales asociadas al protocolo, sus características, formato de intercambio y los esquemas de petición y respuesta entre el proveedor de datos y el recolector. Por su parte, el motor de base de datos a adoptar

[1] http://www.openarchives.org/OAI/openarchivesprotocol.html

depende de los futuros usos de los datos y los requerimientos tecnológicos de la solución informática diseñada.

2.3.2. Preprocesamiento de datos

La calidad de los metadatos bibliográficos es un elemento crucial que afecta significativamente la visibilidad, el descubrimiento y la reutilización de los recursos descritos en el contexto de los datos enlazados. El objetivo de esta actividad es limpiar y normalizar algunos campos de metadatos mejorando considerablemente su calidad. Esta actividad incluye tareas de transformación de datos para normalizar campos tales como fechas, afiliaciones, nombre de autores y palabras claves. La entrada de esta actividad es la base de datos intermedia generada previamente. La salida es la misma base de datos con los metadatos limpiados y normalizados.
En esta actividad un problema común a tener en cuenta es *Entity Resolution* (ER). Se refiere al problema de identificar y agrupar diferentes manifestaciones o representaciones de un mismo objeto del mundo real (Getoor and Machanavajjhala, 2012; Gal, 2014). En el caso particular de los metadatos bibliográficos existen dos entidades que pueden ser identificadas y agrupadas tales como los nombres de los autores y sus afiliaciones. La ambigüedad en las representaciones en los nombres de los autores es un problema común en las revistas científicas en acceso abierto. Este problema surge debido a la inexistencia de sistemas de control de autoridades compartidos entre las revistas. Algo similar ocurre con las representaciones de las afiliaciones.

2.3.3. Modelado de datos

Luego del preprocesamiento de los metadatos, es necesario definir un modelo ontológico para compartir y anotar semánticamente los metadatos bibliográficos para que sean procesados tanto por los humanos como por las computadoras. El objetivo de esta actividad es determinar la o las ontologías que serán utilizadas en el modelado de los datos. La recomendación más importante en este sentido es reutilizar tanto como sea posible ontologías y vocabularios disponibles (Heath and Bizer, 2011). Se recomienda realizar las tareas y utilizar las herramientas sugeridas en (Boris Villazón-Terrazas et al., 2011) para esta actividad. La entrada a esta actividad es el esquema de la base de datos intermedia generada anteriormente. La salida es un modelo ontológico teniendo en cuenta los datos a modelar.

2.3.4. Publicación de datos

El objetivo de esta actividad es transformar en tripletas RDF los metadatos previamente extraídos, almacenados y modelados. La entrada a esta actividad es el modelo ontológico obtenido en la actividad anterior y la base de datos intermedia con los metadatos bibliográficos. La salida es uno o más grafos RDF con los metadatos bibliográficos. Esta actividad se divide en tres tareas fundamentales: transformación, enlazado y publicación.

2.3.4.1. Transformación

Según el primer principio de los datos enlazados es necesario utilizar URIs basadas en el protocolo HTTP como identificadores de los recursos en el contexto de los datos enlazados. En esta tarea es necesario tener en cuenta el diseño de las URIs que va a contener el grafo RDF a generar. Las URIs juegan un rol importante en el descubrimiento e interoperabilidad de los metadatos bibliográficos en el espacio de la web. En este sentido, se recomienda seguir las directrices establecidas en (Heath and Bizer, 2011) y (Bernadette Hyland et al., 2014).

2.3.4.2. Enlazado

El cuarto principio de los datos enlazados establece la necesidad de incluir o crear enlaces RDF hacia otras fuentes de datos en la web. Los enlaces RDF externos conectan islas de datos en un espacio de datos interconectados, permitiendo a otras aplicaciones descubrir fuentes de datos adicionales. El objetivo principal de esta tarea es generar enlaces entre el grafo generado en la tarea anterior y otros grafos similares existen en la web. Algunos de estos enlaces son *owl:sameAs* y *rdfs:seeAlso*. En el primer caso indica que dos URIs (tanto en el grafo origen como en el destino) se refieren al mismo recurso. En el segundo caso especifica que un recurso existente en el grafo destino proporciona información adicional acerca de un recurso existente en el grafo origen.

2.3.4.3. Publicación

El objetivo de esta tarea es hacer accesible en la web los grafos RDF previamente generados y enlazados. En esta tarea se debe garantizar no solo la publicación de los grafos RDF, sino además los metadatos asociados a cada uno de estos grafos. Existen al menos tres formas conocidas de publicación de los grafos RDF en la web. La primera de ellas es mediante un *SPARQL Endpoint*. En este caso, los grafos son almacenados en algún almacén de tripletas que proporciona un punto de

acceso a los datos para su consulta. La segunda forma es mediante un *Linked Data Frontend*. Este tipo de herramientas proporciona una vía para la utilización de los datos no solo para los humanos sino también para las computadoras. Finalmente, se pueden publicar los grafos RDF directamente en la web utilizando un archivo con las tripletas serializadas en algún formato conocido para este propósito. Recientemente fue formalizada una novedosa forma de publicar los grafos RDF llamada *Linked Data Fragments* (Verborgh et al., 2014). El objetivo de esta aproximación es construir servidores con clientes inteligentes, reduciendo la baja disponibilidad de los *SPARQL Endpoints* públicos. Publicar metadatos que describan a los grafos RDF resulta de utilidad tanto para los productores como para los consumidores de datos enlazados. En este sentido se puede utilizar VoID[2] para describir los grafos RDF generados, posibilitando su descubrimiento en el contexto de la web de los datos.

2.3.5. Consumo de datos

El objetivo de esta actividad es desarrollar aplicaciones del mundo real que utilicen los metadatos bibliográficos publicados como datos enlazados. En el dominio de los datos bibliográficos es posible construir bibliotecas digitales con datos enriquecidos semánticamente. Se intenta proporcionar a los usuarios acceso a bibliotecas digitales que integren diversas fuentes de datos, proporcionando servicios de valor añadido. Estas bibliotecas pueden proporcionar acceso al texto completo de las contribuciones, en adición a los metadatos de cada una de ellas. A nivel tecnológico, se utilizan los grafos RDF publicados en un *SPARQL Endpoint*, proporcionando descripciones semánticas de cada uno de los recursos existentes en los grafos RDF.

Varios casos de éxitos han sido documentados en la literatura sobre el consumo de datos bibliográficos publicados como datos enlazados. En este sentido destacan los proyectos Europeana (Haslhofer and Isaac, 2011), Biblioteca Nacional de España (Vila-Suero et al., 2012) y US Library of Congress (Summers et al., 2008).

2.4. Plataforma BM2LOD

La publicación de datos como datos enlazados involucra decisiones técnicas complejas. En los últimos años se han propuesto varias herramientas con el objetivo de transformar datos estructurados en grafos RDF siguiendo los principios de los datos enlazados. Como parte del

[2] http://www.w3.org/TR/void/

marco de trabajo propuesto se diseñó la plataforma BM2LOD (Bibliographic Metadata to Linked Open Data) (Hidalgo-Delgado et al., 2014). La plataforma reutiliza algunas de estas herramientas siguiendo un enfoque *pipeline*. Está diseñada siguiendo las guías metodológicas propuestas en la sección anterior. En la siguiente figura se muestra la arquitectura general de la plataforma.

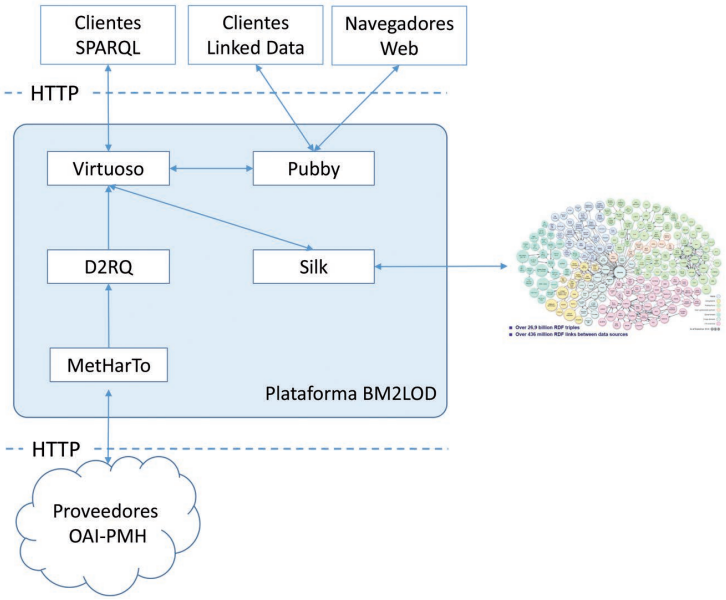

Figura 2.2: Arquitectura de la plataforma BM2LOD

MetHarTo (Hidalgo-Delgado et al., 2013): MetHarTo es un esfuerzo aún en desarrollo con la finalidad de cubrir las tareas de extracción y preprocesamiento de metadatos bibliográficos diseminados mediante el protocolo OAI-PMH. Soporta la extracción de metadatos desde el formato oai_dc (Dublin Core) aunque se pretende incorporar otros formatos y estándares tales como: oai_marc, marcxml y RFC1807. El componente es capaz de conectarse a fuentes de datos a través de un proxy. Las opciones de configuración del proxy son proporcionadas mediante el mecanismo de configuración del componente. Recolecta metadatos en lotes y de forma selectiva desde múltiples repositorios. Los metadatos recolectados son almacenados en una base de datos relacional MySQL o PosgreSQL para su uso posterior. Posee una interfaz de línea de comando para realizar algunas tareas comunes. Está desarrollada en Java siguiendo el paradigma de la Programación Orientada a Objetos. Como entrada recibe uno o varios proveedores de datos que soportan el protocolo OAI-PMH. El componente genera una base de datos intermedia que luego es utilizada por la herramienta D2RQ para la generación de tripletas RDF.

MetHarTo es considerado un *harvester*[3] que realiza peticiones OAI-PMH sobre un proveedor de datos. Este componente implementa algunas tareas de preprocesamiento de los datos, incrementando significativamente la calidad de los mismos. Específicamente se realizan dos tareas fundamentales:

1. Limpieza de datos: se intenta normalizar los metadatos bibliográficos recolectados haciendo transformaciones sobre los mismos. Algunas transformaciones comunes incluyen la eliminación del grado científico en el nombre de los autores y la homogeneización del formato de las fechas.

2. Desambiguación del nombre de los autores: la ambigüedad en el nombre de los autores se expresa mediante la representación de diferentes formas en el nombre de un mismo autor en la misma base de datos. Con esta tarea se pretende desambiguar el nombre de los autores en toda la base de datos, garantizando la unicidad en la representación de los nombres por cada uno de los autores. Para resolver el problema de la ambigüedad fue implementado un algoritmo utilizando las técnicas de conjunto de agrupamientos y la distancia de edición (Alonso-Sierra and Hidalgo-Delgado, 2014). El algoritmo implementado utiliza no solo el nombre de los autores, sino que utiliza otros metadatos tales como afiliaciones, nombre de las revistas y el título de la publicación.

D2RQ (Christian Bizer and Andy Seaborne, 2004): D2RQ es un lenguaje declarativo de alineación y plataforma para el tratamiento de bases de datos relacionales como grafos RDF virtuales. Proporciona una vía de acceso a bases de datos relacionales mediante un SPARQL Endpoint. En el proceso de alineación pueden ser utilizadas varias ontologías y vocabularios. Adicionalmente, proporciona un script programado en bash para generar bases de datos en un único archivo RDF. La herramienta fue integrada a la plataforma BM2LOD con la finalidad de generar las tripletas RDF a partir de los metadatos almacenados en la base de datos relacional obtenida con el componente MetHarTo descrito con anterioridad.

La generación de tripletas RDF utilizando D2RQ involucra la ejecución de tres scripts diferentes *generate-mapping*, *dump-rdf* y *load-graph*. Los dos primeros forman parte de la arquitectura de la herramienta D2RQ, el tercero fue diseñado y programado en la plataforma BM2LOD. A continuación se describe el funcionamiento de cada uno de los scripts mencionados:

generate-mapping: el propósito de este script es la generación de un archivo de alineación entre el

[3] http://www.openarchives.org/OAI/openarchivesprotocol.html#harvester

modelo de datos relacional y las clases y propiedades de las ontologías utilizadas para modelar el dominio de los datos. Por defecto, el script genera un archivo de alineación utilizando el lenguaje D2RQ[4]. Este archivo puede ser modificado y ajustado a las ontologías de dominio utilizadas en el modelado de los datos. La sintaxis de este script es la siguiente:

```
generate-mapping -o file.ttl -d driver.class.name -u db-user -p db-pass jdbc:url:...
```

dump-rdf: el propósito de este script es generar un archivo con el grafo RDF serializado. Este script obtiene los datos existentes en la base de datos relacional y los transforma en tripletas siguiendo la alineación obtenida con el script anterior. La sintaxis de este script es la siguiente:

```
dump-rdf file.ttl -o dump.nt
```

load-graph: el propósito de este script es cargar en el almacén de tripletas Virtuoso el grafo RDF generado con el script anterior. Utiliza la interfaz en línea de comandos ISQL[5].

Silk (Jentzsch et al., 2010; Volz et al., 2009): Silk proporciona herramientas para generar enlaces entre elementos de datos basados en la especificación de enlazado proporcionada por el usuario. Silk consiste en dos componentes principales: (1) una aplicación en consola usada para enlazar dos conjuntos de datos y (2) un servidor HTTP el cual recibe un flujo de tripletas RDF y genera enlaces entre los datos. Ambos componentes utilizan el lenguaje de configuración Silk-LSL (Silk Link Specification Language) para especificar las condiciones que deben cumplir los datos para ser enlazados. La herramienta Silk fue integrada a la plataforma BM2LOD para la generación de enlaces entre las tripletas generadas por la herramienta D2RQ y tripletas existentes en otros conjuntos de datos relacionados.

Con el propósito de automatizar el enlazado en la plataforma BM2LOD, se diseñó e implementó el script *link-graph*. Este script recibe como entrada la ruta completa al archivo de configuración de Silk y de los grafos RDF de origen y destino. Luego se genera un grafo RDF con las nuevas tripletas obtenidas en el proceso y se mezcla con el grafo de origen en el almacén de tripletas Virtuoso.

Virtuoso[6]: OpenLink Virtuoso Universal Server es una solución híbrida de almacenamiento para varios modelos de datos, incluyendo el modelo de datos relacional, basado en tripletas RDF y documentos XML. Virtuoso puede ser utilizado como un punto de integración de datos a partir de diferentes fuentes de datos heterogéneas. Virtuoso posee una versión de código abierto y varias

[4] http://d2rq.org/d2rq-language

[5] http://docs.openlinksw.com/virtuoso/isql.html

[6] http://virtuoso.openlinksw.com/

licencias comerciales. Existen versiones para la mayoría de las plataformas más utilizadas en la actualidad. En el contexto de la plataforma BM2LOD fue utilizado como almacén de tripletas para el almacenamiento y consulta de grafos RDF. Adicionalmente, proporciona un SPARQL Endpoint para acceder a las tripletas utilizando una API REST flexible.

Pubby (Cyganiak and Bizer, 2011): Pubby es un frontend de datos enlazados (Linked Data Frontend) que proporciona vistas HTML sobre los recursos existentes en un grafo RDF almacenado en un almacén de tripletas. Su funcionamiento se basa en la reescritura de URIs y el manejo de la negociación de contenidos mediante redirecciones 303 del protocolo HTTP. La herramienta es de código abierto bajo la licencia de Apache 2.0. Pubby fue integrada a la plataforma BM2LOD con la finalidad de asegurar la visualización de los grafos RDF generados por las herramientas anteriores y almacenados en Virtuoso.

2.5. Conclusiones parciales

En este capítulo se propuso un marco de trabajo basado en los datos enlazados para incrementar la interoperabilidad semántica de los metadatos bibliográficos diseminados por el protocolo OAI-PMH . El marco de trabajo propuesto está compuesto por guías metodológicas para la publicación de metadatos bibliográficos como datos enlazados y una plataforma para la extracción, transformación y consumo de dichos metadatos. Luego de presentada la propuesta se concluye lo siguiente:

1. Las guías metodológicas propuestas cubren el ciclo de vida de los proyectos de datos enlazados, con énfasis particular en el dominio de los metadatos bibliográficos. Las guías propuestas siguen un enfoque iterativo e incremental, mejorando el proceso y los resultados obtenidos en cada una de las actividades tras varias iteraciones.

2. La plataforma BM2LOD automatiza las actividades propuestas en las guías metodológicas, haciendo particular énfasis en el preprocesamiento de los metadatos bibliográficos extraídos en la etapa de extracción de metadatos. La plataforma integra un conjunto de herramientas con propósitos específicos, facilitando la labor de los productores de datos enlazados.

Capítulo 3

Validación de la propuesta

3.1. Introducción

En este capítulo se desarrolla un caso de estudio utilizando revistas cubanas de acceso abierto que soportan el protocolo OAI-PMH. El caso de estudio tiene como objetivo medir la aplicabilidad del marco de trabajo propuesto en esta memoria de tesis. Se describe en detalle las actividades que siguen el ciclo de vida de un proyecto real de datos enlazados en el contexto de los metadatos bibliográficos. Los resultados obtenidos poseen valor práctico posibilitando el desarrollo de aplicaciones informáticas con valor añadido. Adicionalmente, se describe un pre experimento desarrollado utilizando los datos aportados por el caso de estudio. Los resultados del pre experimento evidencian que la propuesta de solución incrementa la interoperabilidad semántica de los metadatos bibliográficos atendiendo al número de consultas que pueden ser formuladas utilizando ambos enfoques, mediante el uso del protocolo OAI-PMH y teniendo en cuenta el enfoque propuesto basado en los datos enlazados.

3.2. Caso de estudio: revistas cubanas

En esta sección se presenta un caso de estudio realizado con revistas cubanas de acceso abierto. El objetivo del caso de estudio es aplicar las guías metodológicas propuestas y la plataforma BM2LOD implementada para comprobar su aplicabilidad en contextos reales de utilización. El caso de estudio se realizó utilizando una computadora personal con las siguientes características técnicas:

- Procesador Intel Core i5 a 2600 MHz
- 4 GB de memoria RAM DDR3
- 500 GB de disco duro SATA-III a 5400 RPM
- Sistema operativo Ubuntu 15.04 de 64 Bits

3.2.1. Descripción de las fuentes de datos

Las fuentes de datos utilizadas en el caso de estudio son una muestra de 9 revistas científicas cubanas, dos de ellas publican artículos originales sobre informática y computación y el resto publica contribuciones sobre ciencias médicas. Para la selección de las revistas se tuvo en cuenta algunos criterios tales como: estabilidad de la revista en línea y soporte para el protocolo OAI-PMH. En la tabla 3.1 se listan las revistas participantes y sus respectivas URL OAI-PMH.

Nombre de la revista	**URL OAI-PMH**
Serie Científica	https://publicaciones.uci.cu/index.php/SC/oai
Revista Cubana de Ciencias Informáticas	http://rcci.uci.cu/index.php/rcci/oai
Revista Cub de Inform en Ciencias de la Salud	http://www.acimed.sld.cu/index.php/acimed/oai
Medisur	http://www.medisur.sld.cu/index.php/medisur/oai
Revista Finlay	http://www.revfinlay.sld.cu/index.php/finlay/oai
Rev. Cub. de Cardiolog. y Cirugía Cardiovasc.	http://www.revcardiologia.sld.cu/index.php/revcardiologia/oai
Revista Cubana de Oftalmología	http://www.revoftalmologia.sld.cu/index.php/oftalmologia/oai
Revista de Ciencias Médicas de P. del Río	http://publicaciones.pri.sld.cu/index.php/publicaciones/oai
Revista Médico Científica	http://www.revistamedicocientifica.org/index.php/rmc/oai

Tabla 3.1: Revistas cubanas utilizadas en el caso de estudio

3.2.2. Extracción de datos

La extracción de metadatos a partir de las revistas científicas que participan en el estudio se realizó con la herramienta MetHarTo previamente descrita. Luego de configurada y ejecutada la herramienta se obtuvo una base de datos en PostgreSQL con metadatos sobre cuatro entidades fundamentales: revistas, artículos, autores y colecciones. La base de datos cuenta con 5203 autores, 2711 artículos y 134 colecciones. En la tabla 3.2 se presentan las entidades y campos de metadatos extraídos de las revistas.

Entidad	Metadato	Ejemplo
Revista	título	"Medisur"
	URL	"http://www.medisur.sld.cu/index.php/medisur/oai"
	ISSN	"1727-897X"
	correo	"medisur@infoced.sld.mu"
	idioma	"Español"
Artículo	título	"Enfermedad pulmonar obstructiva crónica desde la óptica de la medicina basada en pruebas"
	resumen	"Los exacerbaciones agudas de la enfermedad pulmonar obstructiva crónica que requieren hospitalización, se asocian a un aumento de la mortalidad, que en al caso de los pacientes que precisan tratamiento en una unidad de cuidados intensivos, ascienden de un 4 % y hasta un 11-24 %."
	fecha	"2007-10-23"
	fuente	"Medisur; Vol 5, No 1 (2007) Suplemento 1; 22-28"
	identificador	"http://www.medisur.sld.cu/index.php/medisur/article/view/262"
Autor	nombre	"Rolando Delgado Figueredo"
	afiliación	"Hospital General Universitario Dr. Gustavo Aldereguía Lima. Cienfuegos"
Colección	nombre	"Revisiones Bibliográficas"

Tabla 3.2: Entidades y campos de metadatos obtenidos de las revistas

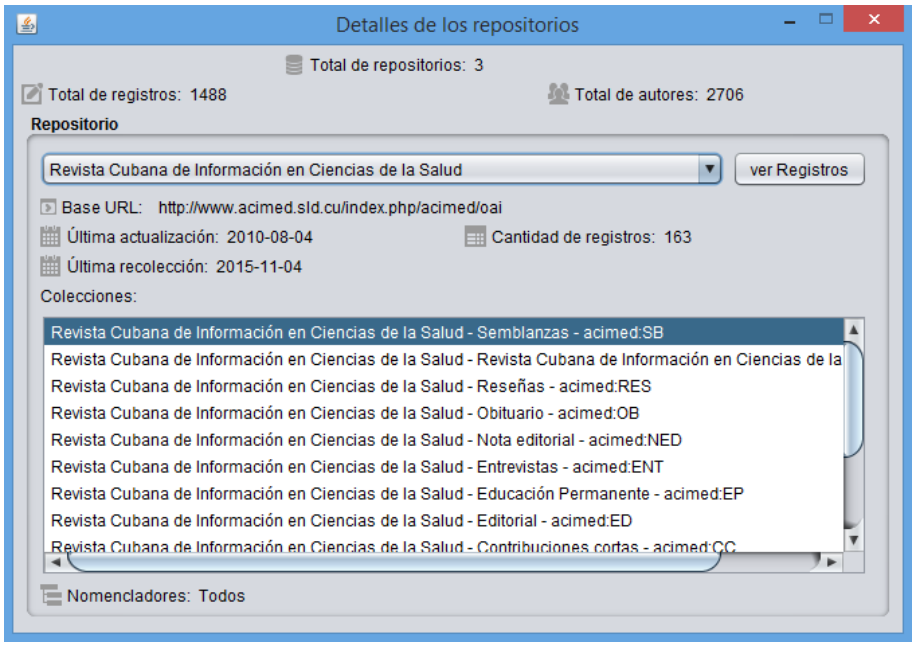

Figura 3.1: Vista de la herramienta MetHarTo

3.2.3. Preprocesamiento de datos

El preprocesamiento de los datos se realiza con la finalidad de limpiar y normalizar los metadatos obtenidos en la actividad anterior. La entrada al preprocesamiento es la base de datos intermedia obtenida en la extracción de datos. La salida de esta actividad es la misma base de datos con los metadatos limpiados y normalizados. En esta actividad se realizaron dos actividades fundamentales: (1) limpieza del nombre de los autores y (2) desambiguación del nombre de los autores. El algoritmo 1 elimina cadenas de caracteres que regularmente aparecen en el nombre de los autores. Algunas de estas cadenas son *MSc, MsC, Dr, Dra, P.Titular*. Estas cadenas se encuentran en un archivo de texto plano modificable por el administrador de la plataforma.

Algoritmo 1 Limpieza del nombre de los autores

Entrada: Lista S con cadenas de caracteres a eliminar de los nombres de los autores

Entrada: Lista N con los nombres de los autores

Salida: Lista R con los nombres de los autores limpiados

1: **Inicializar** la lista R
2: **para todo** $n_i \in N$ **hacer**
3: **si** $n_i \; contiene$ ";" **entonces**
4: $k \leftarrow split \; n_i \; for$ ";"
5: **para todo** $s_j \in S$ **hacer**
6: **si** $k \; contain \; s_j$ **entonces**
7: $r1 \leftarrow k.replace(s_j, "")$
8: $r1.trim()$
9: $R.add(r1)$
10: **fin si**
11: **fin para**
12: **fin si**
13: **fin para**
14: **devolver** $lista \; R$

Para resolver el problema de la ambigüedad en el nombre de los autores, se implementó un algoritmo que utiliza un enfoque basado en la combinación de agrupamiento y la distancia de edición. El algoritmo utiliza información del contexto de los autores para generar los grupos de autores ambiguos. Sigue un enfoque semiautomático, el algoritmo sugiere autores ambiguos y es

decisión del administrador de la plataforma unificar o no los nombre de los autores que se refieren a una misma persona. Este algoritmo fue descrito en (Alonso-Sierra and Hidalgo-Delgado, 2014).

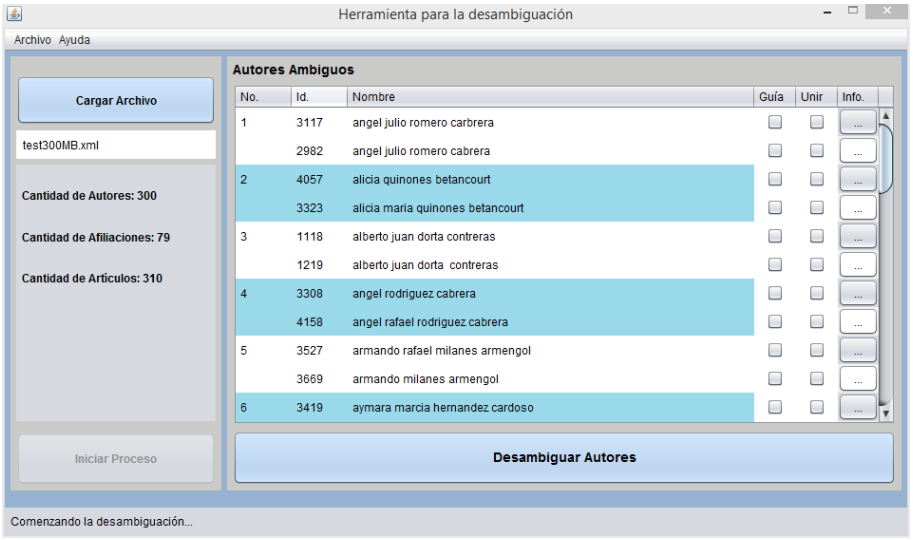

Figura 3.2: Componente para la desambiguación del nombre de los autores

3.2.4. Modelado de datos

En esta actividad se modelan los metadatos bibliográficos extraídos y preprocesados con clases y propiedades propias de una o varias ontologías y vocabularios. A partir de los resultados obtenidos en el estudio descrito en la sección 1.3.2 se analizaron las ontologías existentes en la actualidad para modelar metadatos bibliográficos. Adicionalmente, se utilizaron herramientas para la búsqueda de ontologías y vocabularios existentes en la web, tales como: Falcons[1], *Linked Open Vocabularies*[2] y Swoogle[3]. Teniendo en cuenta el esquema de datos de la base de datos obtenida en la actividad anterior, se alinea el esquema con las ontologías y vocabularios identificados previamente. En la tabla 3.3 se muestran las clases y propiedades de las ontologías utilizadas para modelar los datos. En el caso particular de las *ObjectProperty* se utilizaron las que se muestran en la tabla 3.4. La alineación anterior se materializa en la plataforma BM2LOD utilizando la herramienta D2RQ. Esta herramienta genera una alineación automática utilizando como entrada el esquema de la base de datos relacional. Este esquema se modifica y se personaliza atendiendo al modelo ontológico anterior.

[1] http://ws.nju.edu.cn/falcons/ontologysearch/
[2] http://lov.okfn.org/dataset/lov
[3] http://swoogle.umbc.edu/

Entidad	Clase	Metadato	Propiedad
Revista	fabio:Journal	título	dc: title, rdfs:label
		URL	bibo: uri
		ISSN	bibo: issn
		correo	foaf:mbox
		idioma	dc: language
Artículo	fabio: JournalArticle	título	dc: title, rdfs:label
		resumen	fabio: abstract
		fecha	dc: date
		fuente	dc: source
		identificador	bibo: uri
Autor	foaf: Person	nombre	foaf: name
		afiliación	swrc: affiliation
Colección	fabio: ItemCollection	nombre	foaf: name

Tabla 3.3: Alineación entre campos de metadatos y las ontologías

Entidades	ObjectProperty
Artículo –>Revista	bibo:produced_in
Artículo –>Colección	bibo:isPartOf
Artículo –>Autor	bibo:list_of_authors

Tabla 3.4: ObjectProperty que relacionan las clases de las ontologías utilizadas

3.2.5. Publicación de datos

La publicación de los metadatos bibliográficos como datos enlazados involucra tres tareas fundamentales: (1) generación del grafo RDF teniendo en cuenta la alineación obtenida en la actividad anterior, (2) generación de enlaces entre el grafo RDF obtenido y grafos similares existentes en la web y finalmente (3) publicación del grafo RDF en la web. A continuación se describen el procedimiento y las herramientas utilizadas en cada una de las tareas anteriores.

3.2.5.1. Generación

En esta tarea se utiliza la herramienta D2RQ. Esta herramienta proporciona un script programado en bash para la generación automática del grafo RDF a partir del esquema y los datos

almacenados en una base de datos relacional. El grafo generado es almacenado entonces en el almacén de tripletas Virtuoso. Para ello se diseñó y programó un script en bash que automáticamente ejecuta el script *"dump-rdf"* y envía el grafo RDF generado al Virtuoso mediante la interfaz en línea de comandos ISQL.

3.2.5.2. Enlazado

La generación de enlaces entre el grafo RDF generado en la tarea anterior y grafos similares publicados en la web es una de las tareas fundamentales para el enriquecimiento de recursos publicados como datos enlazados. En esta tarea se generan enlaces del tipo *owl:sameAs* entre el grafo RDF obtenido y los grafos RDF de DBLP[4]. DBLP es una base de datos bibliográfica especializada en ciencias de la computación y áreas afines. Proporciona metadatos de revistas, monografías y actas de congresos de impacto en la comunidad científica.

Para generar los enlaces se utiliza la herramienta Silk, la cual proporciona un lenguaje declarativo de alineación Silk-LSL[5]. Aunque Silk proporciona acceso a SPARQL Endpoints remotos para realizar el enlazado, en este caso de estudio se utilizó una versión offline de DBLP evitándose los problemas derivados de la velocidad de conexión y latencia de la red.

A continuación se muestra un fragmento del archivo donde se especifican las opciones utilizadas por Silk para generar los enlaces. En el caso de estudio se generaron enlaces entre los autores teniendo en cuenta el nombre de los mismos. Para establecer la comparación entre los nombres de los autores entre ambos grafos, origen y destino, se utilizó la distancia de Levenshtein con un umbral de 1, lo que significa que solo se enlazaran aquellos autores que contengan exactamente el mismo nombre en ambos grafos RDF.

3.2.5.3. Publicación

La publicación de los datos enlazados consiste en proporcionar acceso a los grafos RDF en la web. Existen al menos tres variantes diferentes para realizar esta tarea: (1) mediante un SPARQL Endpoint, (2) mediante un *Linked Data Frontend* y (3) subir los grafos RDF materializados a un servidor web. En el caso de estudio se emplearon las dos primeras variantes. En el primer caso se publicó el grafo RDF mediante el SPARQL Endpoint proporcionado por el almacén de tripletas Virtuoso. Adicionalmente, se instaló y configuró sobre el SPARQL Endpoint el *Linked*

[4] http://dblp.l3s.de/d2r/

[5] http://silk-framework.com/

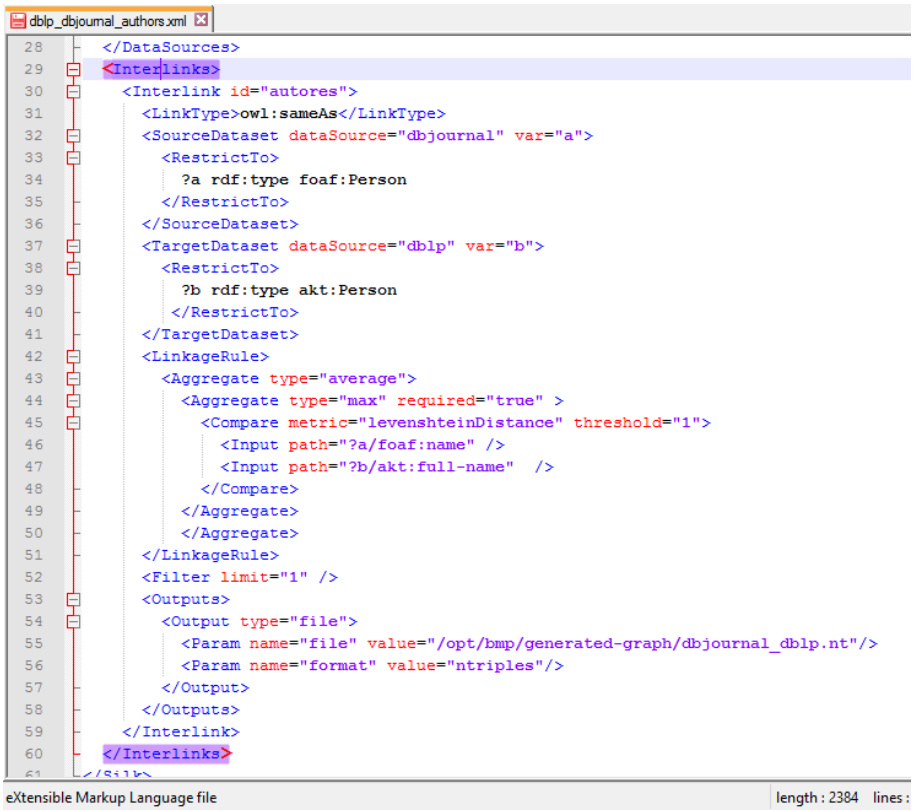

Figura 3.3: Fragmento del archivo de configuración de Silk

Data Frontend Pubby, posibilitando que las tripletas RDF fueran accesibles tanto para los humanos (navegadores web) como para las computadoras (obtención de tripletas RDF). La siguiente figura muestra un fragmento del grafo RDF en la herramienta Pubby.

Figura 3.4: Vista de la herramienta Pubby

3.2.6. Consumo de datos

En un proyecto de datos enlazados, no basta con tener los grafos RDF publicados y accesibles en la web. Se hace necesario entonces el desarrollo de herramientas informáticas que consuman los datos publicados. En el contexto particular del caso de estudio, se diseñó e implementó una biblioteca digital que combina patrones de interacción y componentes de la arquitectura de la información para proporcionar mayor interactividad en su uso por parte de los usuarios. La biblioteca digital implementa tanto la búsqueda textual como la búsqueda facetada. Ambos enfoques han sido ampliamente utilizados en sistemas de recuperación de información siguiendo el enfoque tradicional, sin embargo, no se han encontrado en la literatura abundantes enfoques que apliquen estos paradigmas de búsqueda en tareas de consumo de datos enlazados (Cordoví-García et al., 2014).

Tanto la búsqueda textual como la búsqueda facetada fueron implementadas utilizando consultas parametrizadas escritas en el lenguaje SPARQL. En la búsqueda textual se utilizó el operador

bif:contain proporcionado por el procesador SPARQL del servidor Virtuoso. El uso de este operador es característico del Virtuoso, el cual genera un índice optimizado para la realización de consultas sobre los literales existentes en los grafos RDF almacenados.

La biblioteca digital fue implementada en el lenguaje de programación PHP, haciendo uso de la biblioteca de clases EasyRDF[6] para el manejo de las consultas SPARQL en el servidor Virtuoso. Adicionalmente, se utilizo Jquery[7] como *framework* JavaScript para el manejo de interacciones AJAX complejas. En la siguiente figura se muestra una vista de la página principal de la biblioteca digital implementada.

Figura 3.5: Vista de la biblioteca digital implementada

3.3. Diseño experimental

El marco de trabajo propuesto está basado en los principios de los datos enlazados por lo que los metadatos bibliográficos resultantes se encuentran publicados utilizando el modelo de datos RDF. La hipótesis inicial de la investigación quedó expresada de la siguiente manera:

Si se desarrolla un ***marco de trabajo basado en los principios de los datos enlazados****, entonces se* ***incrementará la interoperabilidad de los metadatos bibliográficos*** *diseminados mediante el protocolo OAI-PMH.*

[6] http://www.easyrdf.org/
[7] https://jquery.com/

Partiendo de la hipótesis planteada se deduce que la variable dependiente esta dada por el incremento de la interoperabilidad de los metadatos bibliográficos diseminados mediante el protocolo OAI-PMH. Para probar la hipótesis planteada, se diseñó un pre experimento con pre y post prueba sobre un único grupo. El objetivo del pre experimento es comprobar si se incrementa o no el nivel de interoperabilidad de los metadatos bibliográficos diseminados teniendo en cuenta el número de consultas que soportan cada uno de los enfoques.

De manera general, el pre experimento consiste en realizar un conjunto de consultas previamente diseñadas sobre la muestra seleccionada, primero mediante el protocolo OAI-PMH (O_1) y luego utilizando el marco de trabajo propuesto (O_2). Luego se compara el número de consultas obtenidas en cada una de las observaciones y se arriba a conclusiones.

El diseño del pre experimento quedaría de la siguiente manera:

$$\text{G} \quad O_1 \quad \text{X} \quad O_2$$

Donde:

G : es el grupo sobre el que se realizarán las observaciones. En el caso particular de este estudio G está compuesto por las nueve revistas científicas utilizadas en el caso de estudio desarrollado en la sección anterior.

O_1 : se corresponde con la primera observación sobre el grupo objeto de estudio. Se calcula el número de consultas que es capaz de responder cualquier revista científica de la muestra y que soporte el protocolo OAI-PMH. Teniendo en cuenta que el protocolo OAI-PMH es un estándar y por tanto es implementado de igual manera por todas las revistas científicas que participan en el estudio, basta con elegir una y realizar las consultas.

X : se corresponde con la aplicación sobre el grupo en estudio del marco de trabajo propuesto. Para ello se tendrán en cuenta los resultados obtenidos con el caso de estudio desarrollado en la sección anterior.

O_2 : se corresponde con la segunda observación. Se tomará en cuenta los resultados obtenidos en el caso de estudio desarrollado en la sección anterior. Se calculará, al igual que en O_1, el número de consultas que es capaz de responder la propuesta de solución. Para ello se utilizará el lenguaje de consultas SPARQL.

A continuación se listan las 12 consultas diseñadas y que se realizarán sobre ambos enfoques, OAI-PMH y datos enlazados.

Q1: *Obtener el nombre de la revista X*

Q2: *Obtener la URL de la revista X*

Q3: *Obtener el correo electrónico del administrador de la revista X*

Q4: *Obtener las colecciones de la revista X*

Q5: *Obtener los artículos de la revista X desde la fecha Y hasta la fecha Z*

Q6: *Obtener los artículos de la colección X de la revista Y*

Q7: *Obtener el artículo X dado su identificador*

Q8: *Obtener los autores del artículo X de la revista Y*

Q9: *Obtener los artículos del autor X publicados en la revista Y*

Q10: *Obtener los artículos que en el título contengan el término X*

Q11: *Obtener los autores con el nombre X*

Q12: *Obtener los cinco autores con más artículos publicados*

Las 12 consultas fueron formuladas tanto en el formato de URL del protocolo OAI-PMH como en el lenguaje de consultas SPARQL siempre que fuese posible. A continuación se ilustran cada una de las formulaciones para la consulta **Q1**:
OAI-PMH:

```
http://rcics.sld.cu/index.php/acimed/oai?verb=Identify
```

SPARQL:

```
SELECT ?nombre WHERE {
?s a fabio:Journal.
?s dc:title ?nombre.
}
LIMIT 10
```

Luego de formuladas y ejecutadas las consultas se obtuvieron los resultados que se muestran en la tabla 3.5.

3.4. Análisis de los resultados

Luego de obtenidos los resultados de las observaciones del pre experimento realizado, se puede afirmar que el enfoque basado en datos enlazados para la publicación de metadatos bibliográficos diseminados por el protocolo OAI-PMH incrementa la cantidad de consultas que se pueden realizar sobre dicho protocolo. Teniendo en cuenta lo anterior se acepta la hipótesis planteada inicialmente

Consulta	O_1	O_2	**Consulta**	O_1	O_2
Q1	✓	✓	Q7	✓	✓
Q2	✓	✓	Q8	–	✓
Q3	✓	✓	Q9	–	✓
Q4	✓	✓	Q10	–	✓
Q5	✓	✓	Q11	–	✓
Q6	✓	✓	Q12	–	✓

Tabla 3.5: Resultados de las observaciones obtenidas en el pre experimento

como verdadera. A continuación se detallan cada uno de los resultados obtenidos en la etapa de evaluación de la propuesta de solución:

1. El formato de intercambio de metadatos que proporcionan las respuestas del protocolo OAI-PMH se sustituye por el modelo de datos RDF. Este modelo de datos es procesable automáticamente por las computadoras.

2. El enfoque propuesto utiliza clases y propiedades de las ontologías para la anotación semántica de los metadatos intercambiados quedando explicita la semántica o significado de los mismos.

3. El enfoque propuesto es capaz de describir las relaciones existentes entre las entidades (revista, artículo, autor, colección) que son intercambiadas. Esto último no es soportado por el protocolo OAI-PMH.

4. El enfoque propuesto incrementa el número de consultas que se pueden realizar sobre los metadatos intercambiados utilizando para ello el estándar SPARQL.

3.5. Conclusiones parciales

Teniendo en cuenta los resultados obtenidos en el caso de estudio y el pre experimento desarrollados en este capítulo, se concluye que:

1. El caso de estudio desarrollado demostró que las guías metodológicas propuestas cubren el ciclo de vida de los proyectos de datos enlazados en el contexto de los metadatos bibliográficos.

2. La plataforma propuesta es capaz de integrar metadatos desde diferentes proveedores de datos que soporten el protocolo OAI-PMH. Con el caso de estudio desarrollado se obtuvo un resultado tangible con utilidad práctica a corto plazo.

3. El enfoque basado en datos enlazados incrementa el número de consultas que se pueden realizar sobre los metadatos bibliográficos y por ende incrementa la interoperabilidad semántica en el contexto del protocolo OAI-PMH.

Conclusiones

Atendiendo a los objetivos propuestos con esta investigación, se concluye que:

1. La publicación de metadatos bibliográficos siguiendo los principios de los datos enlazados aún enfrenta diversos desafíos los cuales están estrechamente relacionados con la madurez de las herramientas informáticas existentes para realizar este complejo proceso, la completitud de las guías metodológicas existentes y la diversidad de ontologías y vocabularios para modelar metadatos bibliográficos.

2. El marco de trabajo propuesto está formado por guías metodológicas y una plataforma informática que siguen el ciclo de vida de los proyectos de datos enlazados en el contexto de los metadatos bibliográficos diseminados por el protocolo OAI-PMH.

3. El caso de estudio y el pre experimento desarrollado demostraron la aplicabilidad del enfoque propuesto en un escenario real de utilización, obteniéndose buenos resultados.

Recomendaciones

1. Generalizar el marco de trabajo propuesto no solo para el protocolo OAI-PMH sino a otras fuentes de datos heterogéneas y distribuidas.

2. Incorporar a la plataforma BM2LOD funcionalidades que proporcionen servicios de valor añadido al usuario, tales como: búsqueda de expertos, detección y visualización de comunidades, integración con redes sociales y gestión de perfiles.

Glosario de términos

EIF European Interoperability Framework

KML Keyhole Markup Language

SIG Sistemas de Información Geográfica

TIC Tecnologías de la Información y las Comunicaciones

XBRL eXtensible Business Reporting Language

FRAD Functional Requirements for Authority Data

FRBR Functional Requirements for Bibliographic Records

IFLA International Federation of Library Associations and Institutions

RDA Resource Description and Access

SKOS Simple Knowledge Organization System

FOAF Friend of a Friend

DCMI Dublin Core Metadata Initiative

IEEE Institute of Electrical and Electronics Engineers

CSV Comma-Separated Values

HTTP Hyper Text Transfer Protocol

URI Uniform Resource Identifier

URL Uniform Resource Locator

HTML Hyper Text Marckup Language

RDF Resource Description Framework

W3C World Wide Web Consortium

SPARQL SPARQL Protocol and RDF Query Language

OWL Ontology Web Language

RDFS RDF Schema

XML eXtensible Markup Language

XSLT eXtensible Stylesheet Language Transformations

MARC21 MAchine-Readable Cataloging

MIT Massachusetts Institute of Technology

OAI-PMH Open Archives Initiative Protocol for Metadata Harvesting

Referencias bibliográficas

Luis Alonso-Sierra and Yusniel Hidalgo-Delgado. Desambiguación del nombre de los autores en metadatos bibliográficos publicados como datos enlazados. In *Proceedings of 1st Cuban Workshop on Semantic Web*, pages 60–71, La Habana, Cuba, 2014. URL http://ceur-ws.org/Vol-1219/paper6.pdf.

Ana Lupe Cristán and Elena Escolano Rodríguez. *RDA: Descripción y Acceso al Recurso. Un código de catalogación para el siglo 21*. RDA-JSC, 2009. URL http://www.rda-jsc.org/docs/rdabrochure-spa.pdf.

Asunción Gómez-Pérez, Mariano Fernández-López, and Oscar Corcho. *Ontological Engineering*. Advanced Information and Knowledge Processing. Springer, 2004. ISBN 1-85233-551-3.

Bernadette Hyland and David Wood. The Joy of Data - A Cookbook for Publishing Linked Government Data on the Web. In David Wood, editor, *Linking Government Data*, pages 3–25. Springer New York Dordrecht Heidelberg London, New York, 2011. ISBN 978-1-4614-1766-8.

Bernadette Hyland, Ghislain Atemezing, and Boris Villazón-Terrazas. Best Practices for Publishing Linked Data, 2014. URL http://www.w3.org/TR/ld-bp/.

Tim Berners-Lee. *Linked Data - Design Issues*. 2006. URL http://www.w3.org/DesignIssues/LinkedData.html.

Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American Magazine*, pages 29–37, 2001. URL http://www.scientificamerican.com/article.cfm?id=the-semantic-web.

Boris Villazón-Terrazas, Luis. M. Vilches-Blázquez, Oscar Corcho, and Asunción Gómez-Pérez. Methodological Guidelines for Publishing Government Linked Data. In *Linking Government Data*, pages 27–49. Springer New York Dordrecht Heidelberg London, New York, 2011. ISBN 978-1-4614-1767-5.

Ricardo Casate and Jose Senso. The Landscape of Open Access Journals in Cuba: the Strategy and Model for its Development. In *Open Access and Digital Libraries*, pages 89–111. De Gruyter

Saur, Berlin, Germany, 2013. ISBN 978-3-11-028102-6. URL http://www.degruyter.com/viewbooktoc/product/181804.

Christian Bizer and Andy Seaborne. D2rq – Treating Non-RDF Databases as Virtual RDF Graphs. In *Proceedings of the 3rd International Semantic Web Conference*, volume 3298 of *Lecture Notes in Computer Science*, Japan, 2004. Springer. ISBN 978-3-540-23798-3.

European Commission. European Interoperability Framework for European Public Services. Technical Report 744, Bruxelles, 2010. URL http://ec.europa.eu/isa/documents/isa_annex_ii_eif_en.pdf.

Sam Coppens, Erik Mannens, and Rik Van de Walle. Disseminating heritage records as linked open data. *International Journal of Virtual Reality*, 8(3):39–44, 2009. URL https://biblio.ugent.be/publication/789837?embed=1&hide_info=1&hide_options=1.

Carlos Cordoví-García, Claudia Hernández-Rizo, Yusniel Hidalgo-Delgado, and Liudmila Reyes-Álvarez. Using Search Paradigms and Architecture Information Components to Consume Linked Data. In *Proceedings of 1st Cuban Workshop on Semantic Web*, volume 1219, pages 13–30, Cuba, 2014. CEUR Workshop Proceedings. URL http://ceur-ws.org/Vol-1219/paper2.pdf.

Richard Cyganiak and Chris Bizer. Pubby – A Linked Data Frontend for SPARQL Endpoints, 2011. URL http://wifo5-03.informatik.uni-mannheim.de/pubby/.

Daniel Vila-Suero, Mathieu d'Aquin, and Asunción Gómez-Pérez. *Marimba: unlocking your library data*. 2013. URL http://marimba4lib.com/.

John Domingue, Dieter Fensel, and James A. Hendler. *Handbook of Semantic Web Technologies*, volume 1. Springer Science & Business Media, New York, 2011. ISBN 978-3-540-92912-3. URL https://www.google.com/books?hl=en&lr=&id=sdEFvSb9WNsC&oi=fnd&pg=PR1&dq=handbook+of+semantic+web+technologies&ots=aympSWAmP_&sig=VZaiWPrHitZ7w4P4QKmIdSai--s.

Avigdor Gal. Uncertain Entity Resolution: Re-evaluating Entity Resolution in the Big Data Era: Tutorial. *Proc. VLDB Endow.*, 7(13):1711–1712, August 2014. ISSN 2150-8097. URL http://dl.acm.org/citation.cfm?id=2733004.2733068.

Lise Getoor and Ashwin Machanavajjhala. Entity Resolution: Theory, Practice & Open Challenges. *Proceedings of the VLDB Endowment*, 5(12):2018–2019, 2012. URL http://dl.acm.org/citation.cfm?id=2367564.

Hesamedin Hakimjavadi, Mohamad Noorman Masrek, and Shah Alam. SW-MIS: A Semantic Web based model for integration of institutional repositories metadata records. *Science Series Data Report*, 4(11):57–75, 2012. URL http://www.researchgate.net/profile/Mohamad_Masrek/publication/233855032_SW-MIS_A_Semantic_Web_Based_Model_for_Integration_ofInstitutional_Repositories_Metadata_Records/links/09e4150c28690c74fd000000.pdf.

Bernhard Haslhofer and Antoine Isaac. data.europeana.eu: The Europeana Linked Open Data Pilot. *International Conference on Dublin Core and Metadata Applications*, 0:94–104, September 2011. ISSN 1939-1366. URL http://dcpapers.dublincore.org/pubs/article/view/3625.

Bernhard Haslhofer and Bernhard Schandl. The OAI2lod Server: Exposing OAI-PMH metadata as linked data. In *International Workshop on Linked Data on the Web*, Beijing, China, 2008. URL http://eprints.cs.univie.ac.at/284.

Bernhard Haslhofer and Bernhard Schandl. Interweaving OAI-PMH data sources with the linked data cloud. *International Journal of Metadata, Semantics and Ontologies*, 5(1):17–31, 2010. URL http://inderscience.metapress.com/index/t770153631427210.pdf.

Tom Heath and Christian Bizer. *Linked Data. Evolving the Web into a Global Data Space.* SYNTHESIS LECTURES ON THE SEMANTIC WEB: THEORY AND TECHNOLOGY. Morgan & Claypool, first edition edition, 2011. ISBN 9781608454310.

Yusniel Hidalgo-Delgado, Rafael Rodríguez-Puentes, Ernesto Ortíz-Muñoz, and Luis Alonso-Sierra. Herramienta para la recolección de metadatos bibliográficos mediante el protocolo OAI-PMH. In *II Conferencia Internacional de Ciencias Computacionales e Informáticas*, La Habana, Cuba, 2013. ISBN 978-959-7213-02-4.

Yusniel Hidalgo-Delgado, Liudmila Reyes-Álvarez, Amed Leiva-Mederos, María del Mar Roldán, and José F. Aldana-Montes. BM2LOD: Platform For Publishing Bibliographic Data As Linked

Open Data. In *Proceedings of 7th IADIS International Conference on Information Systems*, pages 27–34, Madrid, 2014. IADIS Press. ISBN 978-989-8704-04-7. URL http://www.researchgate.net/profile/Yusniel_Hidalgo_Delgado/publication/260554972_BM2LOD_Platform_For_Publishing_Bibliographic_Data_As_Linked_Open_Data/links/0c96053191f6c677d8000000.pdf.

Ian Davis and Richard Newman. *Expression of Core FRBR Concepts in RDF*. 2005. URL http://vocab.org/frbr/core.html.

Institute of Electrical and Electronics Engineers. IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries. New York, 1990.

Anja Jentzsch, Robert Isele, and Christian Bizer. Silk–Generating RDF Links While Publishing or Consuming Linked Data. In *9th International Semantic Web Conference (ISWC'10)*. Citeseer, 2010. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.231.4262&rep=rep1&type=pdf#page=63.

Graham Klyne and Jeremy J. Caroll. *Resource Description Framework (RDF): Concepts and Abstract Syntax*. 2004. URL http://www.w3.org/TR/rdf-concepts/.

M. A. Manso, M. Wachowicz, M. A. Bernabé, A. Sanchez, and A. F. Rodriguez. Modelo de Interoperabilidad Basado en Metadatos (MIBM). *Proceedings JIDEE 2008, Adeje (Tenerife), Spain, November 14-15 2008*, 2008. URL http://latingeo.upm.es/intranet/CCD/Lists/DI_Publicaciones/Attachments/130/037.pdf.

Martin Malmsten. Making a library catalogue part of the semantic web. page 146, Berlin, Germany, 2008. Universitätsverlag Göttingen. URL http://www.google.com/books?hl=en&lr=&id=kUEGX-tiGYUC&oi=fnd&pg=PA146&dq=making+a+library++catalogue+part+of+the+semantic+web&ots=uxsPhITh9e&sig=umt-Mdtwg2hO0kThNM1rwJqaseg.

Larry Masinter, Tim Berners-Lee, and Roy T. Fielding. *Uniform Resource Identifier (URI): Generic Syntax*. 2005. URL http://tools.ietf.org/html/rfc3986.

R. Studer, R. Benjamins, and D. Fensel. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1-2):161–198, 1998.

Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. Pellet: A practical owl-dl reasoner. *Web Semantics: science, services and agents on the World Wide Web*, 5(2):51–53, 2007. URL http://www.sciencedirect.com/science/article/pii/S1570826807000169.

Stefano Mazzocchi. *OAI-PMH RDFizer - SIMILE*. 2007. URL http://simile.mit.edu/wiki/OAI-PMH_RDFizer.

Steffen Staab and Rudi Studer. *Handbook on Ontologies*. International Handbooks on Information Systems. Springer, second edition, 2009. ISBN 978-3-540-92673-3.

Ed Summers, Antoine Isaac, Clay Redding, and Dan Krech. LCSH, SKOS and linked data. *Proc. Int'l Conf. on Dublin Core and Metadata Applications*, pages 25–33, 2008. ISSN 1939-1366. URL http://arxiv.org/abs/0805.2855.

Dmitry Tsarkov and Ian Horrocks. FaCT++ description logic reasoner: System description. In *Automated reasoning*, pages 292–297. Springer, 2006. URL http://link.springer.com/chapter/10.1007/11814771_26.

Ruben Verborgh, Olaf Hartig, Ben De Meester, Gerald Haesendonck, Laurens De Vocht, Miel Vander Sande, Richard Cyganiak, Pieter Colpaert, Erik Mannens, and Rik Van de Walle. Querying Datasets on the Web with High Availability. In *Proceedings of 13th International Semantic Web Conference*, volume 8796 of *Lecture Notes in Computer Science*, pages 180–196, Italy, 2014. Springer. ISBN 978-3-319-11964-9. doi: 10.1007/978-3-319-11964-9_12. URL http://link.springer.com/chapter/10.1007/978-3-319-11964-9_12.

Victor de Boer, Jan Wielemaker, Judith van Gent, Michiel Hildebrand, Antoine Isaac, Jacco van Ossenbruggen, and Guus Schreiber. Supporting linked data production for cultural heritage institutes: the amsterdam museum case study. In *The Semantic Web: Research and Applications*, pages 733–747. Springer, 2012. URL http://link.springer.com/chapter/10.1007/978-3-642-30284-8_56.

Daniel Vila-Suero, Boris Villazón-Terrazas, and Asunción Gómez-Pérez. datos. bne. es: a Library Linked Data Dataset. *Semantic Web*, 2012. ISSN 2210-4968. doi: 10.3233/SW-120094. URL http://iospress.metapress.com/index/FK52210L2G10823H.pdf.

Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Silk-A Link Discovery Framework for the Web of Data. In *LDOW*, volume 538, 2009. URL http://vsr-mobile.informatik.tu-chemnitz.de/svnproxy/download/publications/doc/2009/06.pdf.

William Denton. *FRBR: Object-Oriented Definition and Mapping to FRBRER. draft 0.9*. 2008. URL http://www.frbr.org/2008/02/07/frbroo-09-draft.

Xavier Agenjo and María Luisa Martínez-Conde. *Requisitos Funcionales de los Registros Bibliográficos*. IFLA, Holanda, 1998. ISBN 84-8181-213-7.