Integración de fuentes de datos bibliográficas utilizando tecnologías de Linked Data - Caso de uso: Biblioteca de la Universidad de Cuenca

Freddy Sumba, José Ortiz, José Segarra, Víctor Saquicela

Departamento de Ciencias de la Computación, Universidad de Cuenca, Av. 21 de Abril y Agustín Cueva, Cuenca, Ecuador.

Autores para correspondencia: freddy.sumbao@ucuenca.ec, jose.ortizv@ucuenca.ec, jose.segarraf@ucuenca.ec, victor.saquicela@ucuenca.edu.ec

Fecha de recepción: 4 de junio del 2017 - Fecha de aceptación: 12 de julio de 2017

RESUMEN

Actualmente, las bibliotecas de todo el mundo almacenan su información bibliográfica en múltiples fuentes digitales. Dependiendo del tipo de contenido: libros, trabajos de pregrado, contenido de terceros (servicios externos tales como Scopus y Springer) y otros materiales suelen ser administrados por diferentes herramientas de software. Sin embargo, el almacenamiento independiente de la producción académica y científica dificulta el acceso y la exploración de datos de manera unificada. En este contexto, se propone un enfoque de integración centralizado para aprovechar el conocimiento que tienen las instituciones bibliotecarias. El enfoque presentado aplica el Ciclo de Vida de Datos Enlazados sobre todas las fuentes bibliográficas con el fin de estandarizar la información tanto en formato como en vocabulario, y para enlazar todos los recursos de las fuentes entre sí. Además, todos los datos son publicados en un terminal único que permite acceso y búsqueda centralizada sobre la información almacenada. Esta propuesta se ha aplicado con éxito en la integración de las fuentes bibliográficas en la biblioteca de la Universidad de Cuenca.

Palabras clave: Bibliotecas, datos enlazados, ontologías.

ABSTRACT

Nowadays, libraries around the world store their bibliographic information in multiple digital sources. Depending on the type of content: books, undergraduate works, third party content (i.e. external services such Scopus and Springer) and other material are usually managed by different software tools. However, the independent storage of academic and scientific production causes difficulties in accessing and exploring data in a unified manner. In this context, we propose a centralized integration approach to exploit the knowledge available at library institutions. The presented approach applies the Linked Data Life Cycle over bibliographic sources in order to standardize the information in both format and vocabulary and to link all the sources' resources among them. Moreover, all the data is published on a single endpoint, which allows centralized access and search through the stored information. This approach has been successfully applied for the integration of the bibliographic sources of the library of the University of Cuenca.

Keywords: Libraries, linked data, ontologies.

1. INTRODUCCIÓN

En la actualidad, la adopción de tecnologías de *Linked Data* ha crecido de manera considerable, puesto que aumentan las posibilidades de descubrir, usar y reutilizar la información. Las bibliotecas juegan un papel importante dentro de este escenario, pues exponen grandes colecciones de datos que pueden llegar a ser de interés general. Debido a este motivo, muchas instituciones bibliotecarias han adoptado este

enfoque como mecanismo para gestionar e integrar la información bibliográfica, utilizando para ello ontologías y vocabularios especializados que describen los recursos bibliográficos en formato en *RDF* (Ríos-Hilario, Martín-Campo & Ferreras-Fernández, 2012). Mediante la aplicación de tecnologías de Linked Data es posible aumentar la visibilidad del contenido bibliográfico a través de la navegación entre datos vinculados y no ligarse a tecnologías que no permiten escalabilidad acorde a las nuevas necesidades de los usuarios en la Web (Harper & Tillett 2007).

La aplicación de Datos Enlazados en Bibliotecas (*LLD - Library Linked Data*), tiene la misión de colaborar al incremento de la interoperabilidad de la información que disponen las bibliotecas en la Web (Baker, Bermes, Coyle, Dunsire, Isaac *et al.*, 2005). Esto facilita la creación de vínculos entre conjuntos de datos, vocabularios y otros elementos a través de *RDF* que especifica las relaciones entre los recursos sobre repositorios que integran la información de múltiples fuentes con el objetivo de la construcción de una base de conocimiento común. Según *W3C Library Linked Data Incubator Group*, la aplicación de *Linked Data* tiene varios beneficios como se menciona en Baker *et al.* (2005), los cuales son:

- *LLD* proporciona una mejor navegación y descubrimiento de información bibliográfica;
- aumenta la visibilidad de los datos bibliográficos en la Web;
- ofrece la integración de información bibliográfica y objetos digitales por medio de estándares web abiertos;
- proporciona un modelo semántico más robusto para la descripción de los recursos en contraste de los formatos de metadatos tradicionales;
- facilita la reutilización de los conjuntos de datos sobre de diferentes dominios;
- permite a los desarrolladores y proveedores evitar ser vinculados a datos específicos de las bibliotecas, tales como *MARC*[1] y *Z39.50*[2].

El objetivo del presente artículo es aplicar los conceptos de *LLD* a través de un caso real del Centro de Documentación Regional Juan Bautista Vázquez (Biblioteca de la Universidad de Cuenca), el cual dispone de contenido bibliográfico almacenado en diferentes fuentes de información, que varían de acuerdo con el tipo de contenido, como la producción científica de la institución, el catálogo literario, artículos científicos externos, etc. Por otra parte, el contenido es almacenado utilizando una estructura de datos heterogénea a pesar de tratarse de contenido bibliográfico en general. El acceso a estas fuentes de información no está centralizado y se dispone de herramientas de software comerciales que permiten una búsqueda sintáctica por cada una de las fuentes de forma individual, lo cual minimiza la visibilidad del contenido que dispone la biblioteca.

Por lo tanto, en este trabajo se presenta una alternativa para el acceso unificado al contenido bibliográfico almacenado en las diferentes fuentes de datos de la biblioteca de la Universidad de Cuenca, utilizando tecnologías de *Linked Data*. La propuesta está enfocada en un modelo centralizado, en el cual, un repositorio semántico almacena toda la meta información de los recursos bibliográficos de cada fuente de datos. Además, se adopta un buscador que aprovecha las ventajas de las tecnologías semánticas, ofreciendo resultados integrados de las diferentes fuentes bibliográficas.

El trabajo descrito a continuación está organizado de la siguiente manera: en la sección 2 se describen los trabajos relacionados que presentan enfoques similares al implantado en este trabajo. La sección 3 describe el caso de estudio en la biblioteca de la Universidad de Cuenca para la integración de fuentes bibliográficas. Finalmente, en la sección 4 se presentan los resultados y en la sección 5 las conclusiones obtenidas en base al proceso realizado, los resultados obtenidos y los trabajos futuros.

2. TRABAJOS RELACIONADOS

Como se menciona en Torre-Bastida, González-Rodríguez & Villar-Rodríguez (2015), se han desarrollado diferentes análisis e implantaciones de tecnologías semánticas sobre fuentes bibliotecarias, los cuales describen casos de uso similares al presentado en este trabajo. Por ejemplo Kruk *et al.* (2005) describen *MarkOnt*, que es un proyecto que intenta definir una ontología basada en *Marc21*, *Bibtex* y

[1] https://www.loc.gov/marc/
[2] http://www.Loc.gov/z3950/agency/

Dublin Core, con el objetivo de generar meta data de recursos bibliotecarios que sean interoperables entre diferentes herramientas. Sin embargo, la ontología se limita a utilizar como base *MARC21* debido que es un estándar de transmisión de metadatos, no un estándar de contenido. Adicionalmente dicha propuesta no aborda el problema de la aplicación práctica de la ontología sobre fuentes de datos reales.

El trabajo presentado por Vila-Suero & Gómez-Pérez (2013), detalla el método y la aplicación de *Linked Data* a bibliotecas con fuentes bibliográficas en formato *MARC 21*. Para esto, los autores tomaron, como caso de uso real, los datos de la Biblioteca Nacional Española y presentaron la herramienta *MARiMba*[3], la cual realiza el proceso de generación a *RDF* en base a una metodología de generación de datos enlazados. En contraste, se propone un enfoque que abarca fuentes de datos con información de diferentes dominios y tecnologías, no limitadas solo a fuentes de datos *MARC21*.

Por otro lado, Ríos Hilario & Martín Campo (2012), presentaron un caso de estudio similar, en el cual se aplicaron los principios de *LLD* sobre *Europeana*[4] y para lo cual propusieron *Europeana data model* (*EDM*) para la descripción de recursos culturales. Este trabajo no estuvo regido por una metodología que soporte el proceso de publicación de *Linked Data*. Dichos autores definieron vocabularios específicos para cada problemática, lo que perjudica la interoperabilidad de la información (Heflin & Hendler, 2000).

Sobre este mismo tema, Malmsten (2008) describió la herramienta *Libris* que integra varias bibliotecas que parten desde fuentes *Marc21*. *Libris* utiliza principios de *Linked Data* para el proceso de publicación de recursos bibliográficos. Sin embargo, el proceso y las herramientas utilizadas no son de código abierto, por lo cual el servicio está limitado a licencias comerciales.

La aplicación de *Linked Data* sobre fuentes de información bibliográficas (*LLD*) ha sido caso de estudio de múltiples trabajos, sin embargo, la mayoría de estos tienen ciertas limitaciones principalmente relacionadas con su aplicación sobre entornos heterogéneos, debido a que la mayoría de los planteamientos están basado en un número limitado de estándares para modelar la información (*MARC21* principalmente). Por otro lado, la falta de definición de guías metodológicas y la utilización de herramientas de software cerradas dificultan su aplicación en ambientes de bibliotecarios reales. En este contexto, el presente trabajo busca superar las limitantes antes mencionadas mediante una arquitectura flexible para la aplicación de *LLD*.

3. INTEGRACIÓN DE RECURSOS BIBLIOGRÁFICOS

La biblioteca de la Universidad de Cuenca dispone de varias fuentes de información bibliográfica, las cuales almacenan los datos de forma aislada y proveen a los usuarios herramientas de búsqueda a estos datos de forma independiente. Esto genera una brecha entre la información que contienen y dificulta su explotación de manera unificada. En este trabajo se presenta una propuesta de integración de fuentes bibliotecarias mediante el empleo de tecnologías de *Web Semántica y Linked Data*, siguiendo la metodología definida por Villazón-Terrazas, Vilches-Blázquez, Corcho & Gómez-Pérez (2011) para la publicación de datos enlazados. En este caso se utiliza un enfoque centralizado, debido a que todas las fuentes bibliográficas son gestionadas por la misma institución. La arquitectura planteada para resolver este problema se ilustra en la Figura 1.

Los componentes en conjunto permiten tener un canal de acceso centralizado a través de un buscador semántico al contenido digital de una institución bibliotecaria, obteniendo de esta forma una mejora a la visualización del contenido bibliográfico almacenado en las fuentes mencionadas y por ende los resultados esperados acorde a la búsqueda del usuario.

[3] http://marimba4lib.com
[4] http://www.europeana.eu/portal/es

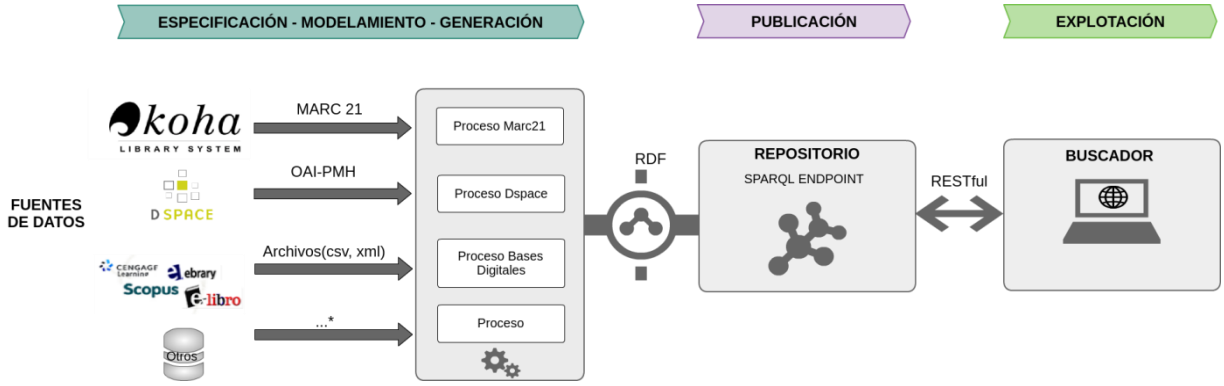

Figura 1. Arquitectura para la integración de recursos bibliográficos.

La arquitectura implementada consta de cinco etapas, tal como se define en (Villazón-Terrazas *et al.*, 2011). Para lo cual, las etapas de *especificación, modelamiento y generación* están contenidas en procesos definidos que cumplen con el enfoque de la metodología acorde al tipo de fuente (Koha, Dspace, E-libro, EBSCO, Ebracy-AC, etc). En la etapa de *publicación de datos*, se establece el proceso de almacenamiento de datos en un repositorio centralizado, utilizando el formato *RDF*. Finalmente, en la etapa de *explotación* el resultado del proceso es aprovechado a través de un buscador semántico que permite acceder al contenido de forma centralizada.

A continuación, se describe el proceso realizado en las 5 etapas que define la metodología de publicación de datos a través de *Linked Data* acorde a cada fuente de datos bibliográficos.

3.1. Especificación

Esta etapa está enfocada en determinar tanto los requerimientos como las fuentes de datos sobre las cuales se aplicará el proceso de publicación de datos enlazados. Además de la definición de las URIs y el licenciamiento de los recursos bibliográficos. Por lo cual, esta tarea se ha descompuesto en tres partes:

Identificación y análisis de las fuentes de datos

Las fuentes bibliográficas que dispone la biblioteca de la Universidad de Cuenca están enmarcadas en un conjunto de servicios que ofrece a sus usuarios para el acceso al material digital que produce la institución o contratan a terceros. En específico, la institución dispone de tres tipos de fuentes bibliográficas: sistemas bibliotecarios, repositorios institucionales y fuentes bibliográficas externas.

a) Sistemas bibliotecarios

Los sistemas bibliotecarios almacenan gran cantidad de datos bibliográficos tales como: libros, revistas científicas, producción académica, etc., de tal manera que estos datos puedan ser visualizados por los usuarios a través de diferentes formas de acceso (Lencinas, 2001). La diversidad de sistemas bibliotecarios, en cuanto a estructuración y almacenamiento de la información, genera varias problemáticas al momento de integrar la información de otras fuentes. Por lo que se tienen diferentes estándares y herramientas para facilitar la integración de fuentes bibliográficas. Un ejemplo es el formato *MARC*, el cual es un estándar digital internacional de descripción de información bibliográfica (Estivill-Rius, 2011). Actualmente, se dispone de varias versiones de *MARC*, entre ellas *Marc21*, el cual es el formato predominante en la mayoría de bibliotecas españolas y a nivel mundial. Este formato fue creado en 1999 como resultado de la armonización de los formatos *MARC* de Estados Unidos y Canadá (Est11). *MARC21* soporta la descripción de materiales textuales impresos y manuscritos, archivos digitales, mapas, música, etc. Las partes de un registro en *MARC21* constan de cabecera, directorio, campos variables de control y campos variables de datos que pueden ser extendidos con sub campos, de esta manera aumenta la capacidad de descripción de un registro.

De manera similar, se dispone del formato *UNIMARC*, organizado a través de etiquetas, indicadores y sub campos que se asignan a los registros bibliográficos en formato legible por

herramientas informáticas. Este formato cubre monografías, materiales cartográficos, música, grabaciones de sonido, gráficos, materiales proyectados y de vídeo, libros y archivos electrónicos (Hopkinson, 2016).

De acuerdo a una investigación realizada en este trabajo en 26 instituciones bibliotecarias del Ecuador se han identificado cuatro tipos de sistemas bibliotecarios: 1) *Koha*, 2) *ABCD*, 3) *Siabuc*, 4) *WinISIS*. Estos sistemas difieren en las técnicas y herramientas utilizadas para almacenar el contenido bibliográfico, utilizan bases de datos relacionales y en otros utilizan almacenamiento basado en texto como el formato *MARC21*. En la biblioteca de la Universidad de Cuenca se utiliza *WinISIS y Koha,* y los datos están disponibles en formato *MARC21*.

b) Repositorios digitales institucionales

Los repositorios institucionales almacenan la producción bibliográfica de una institución como artículos científicos, ponencias o comunicaciones a congresos, revistas electrónicas editadas por la institución, material de docentes, elaborados por los profesores e investigadores. Las bibliotecas disponen de su propia implementación de repositorios institucionales y en otros casos utilizan herramientas comerciales u *Open Source*, tales como: *Dspace*[5], *EPrints*[6] o *Fedora Commons*[7]. En la biblioteca de la Universidad de Cuenca se utiliza el repositorio *Dspace* en la versión 5.1, en el cual se almacena el contenido bibliográfico que dispone la institución.

Las herramientas de software de los repositorios institucionales generalmente utilizan el protocolo *OAI-PMH*[8], el cual es un protocolo propuesto por la *Open Archives Initiative*[9] para la extracción de metadatos desde repositorios institucionales digitales. La descripción de los recursos digitales en los repositorios utiliza varios formatos (MARC[10], Dublin Core[11]), los cuales varían de acuerdo a las necesidades específicas de cada repositorio institucional.

c) Bases digitales bibliográficas

Las bases digitales ofrecen material multidisciplinario a texto completo de las más importantes revistas y editoriales científicas: artículos revisados por pares académicos, reseñas, ebooks, tesis, videos, imágenes, estadísticas, etc. (García Álvarez de Toledo & Fernández Sánchez 2011). El contenido que ofrecen las bases digitales es adquirido por la institución a través de una subscripción anual, cuyo costo es proporcional al tipo de contenido y a la cantidad de material contratado. Las bases digitales más utilizadas en la Universidad de Cuenca son:

- Scopus
- E-Libro
- ProQuest
- Ebrary
- Britannica Enciclopedia Moderna
- UNWTO Elibrary
- Taylor & Francis
- Gale Virtual Reference Library

Generalmente las bases digitales proporcionan los metadatos del contenido contratado por la institución, por lo cual es necesario aprovechar esta información a través de herramientas de búsqueda accesibles al usuario. Estas fuentes ofrecen dos maneras de extraer la información: a) APIs bajo limitaciones de cuota, b) Archivos; los cuales proporciona la fuente acorde al material bibliográfico contratado por la institución.

5 http://www.dspace.org/
6 http://www.eprints.org/
7 http://fedorarepository.org/
8 Open Archives Initiative - Protocol for Metadata Harvesting
9 https://www.openarchives.org/
10 https://www.loc.gov/marc/bibliographic/
11 http://dublincore.org/

Diseño de URIs

Esta tarea define las URIs que son utilizadas para identificar a los recursos bibliográficos en la Web. Por lo cual se ha definido dos tipos de URIs:

- URIs acorde a Vocabulario: identifica la terminología de clases, propiedades y relaciones del vocabulario utilizado.
- URIs de acuerdo a la fuente bibliográfica: identifica a los recursos bibliográficos de acuerdo a la fuente de procedencia.

Por ejemplo, los vocabularios empleados en este trabajo se identifican a través de namespaces como: *http://purl.org/ontology/bibo*, *http://purl.org/dc/terms/*, *http://xmlns.com/foaf/0.1/Person*, los cuales son parte de los identificadores de recursos bibliográficos que se describen semánticamente. Por otra parte, las URIs formadas acorde a las fuentes están formadas de la siguiente manera, para el recurso de la fuente *Koha* número 27186: *http://sgb.ucuenca.edu.ec:- 3000 /library/koha/recurso/27186*. La definición de URIs adoptada permite acceder a los recursos de forma rápida e interpretable para el usuario, de tal manera que resulte visible identificar al tipo de recurso que se accede y a la fuente bibliográfica que procede.

Definición de licenciamiento de los recursos bibliográficos

Diferentes instituciones como: *Europeana*, *CENL*, *Harvard Library* (*http://openmetadata.lib*), han optado por publicar sus datos bajo licenciamientos abiertos (Vila-Suero & Gómez-Pérez, 2013). Por ejemplo: Creative Commons' Public Domain license (*http://creativecommons.org/public-domain/zero/1.0/*) soporta la publicación de datos a través de *LLD*, por lo cual los datos de este trabajo siguen los lineamientos acordes a los principios de *Linked Open Data* (LOD), permitiendo reutilizar la información y mejorar la visibilidad del contenido.

3.2. Modelamiento

En esta etapa se selecciona vocabularios de datos que serán utilizados para describir semánticamente los recursos en base a las ontologías. De esta manera, es posible representar los recursos en formato *RDF* siguiendo los principios de *Linked Data*. Los recursos bibliográficos son descritos utilizando ontologías como: *BIBO (Bibliographic Ontology)*[12], *FOAF (Friend Of a Friend)*[13] y *DCTERMS (Dublin Core)*. La selección de las ontologías mencionadas se basa en las siguientes razones:

- las ontologías permiten la descripción de documentos bibliográficos, personas y relaciones entre los recursos y hacia recursos externos;
- la interoperabilidad entre vocabularios ontológicos, debido a que facilita la reutilización de los datos en la Web.

El modelo de datos definido se ilustra en la Figura 2. El modelo define los recursos principales como: documentos, personas, colecciones y sus propiedades asociadas. Cada recurso bibliográfico es por lo tanto descrito con el modelo de datos definido, indistintamente de la fuente que provenga. Por ejemplo, los datos provenientes de fuentes bibliotecarias están modelados de acuerdo a la especificación *MARC21*, mientras que los datos provenientes de repositorios *Dspace* utilizan *Dublin Core*. Por esto, es necesario establecer la correspondencia entre los modelos de datos específicos de cada fuente bibliográfica a un modelo común que represente entidades, propiedades y relaciones entre los recursos.

[12] http://bibliontology.com/
[13] http://xmlns.com/foaf/spec/

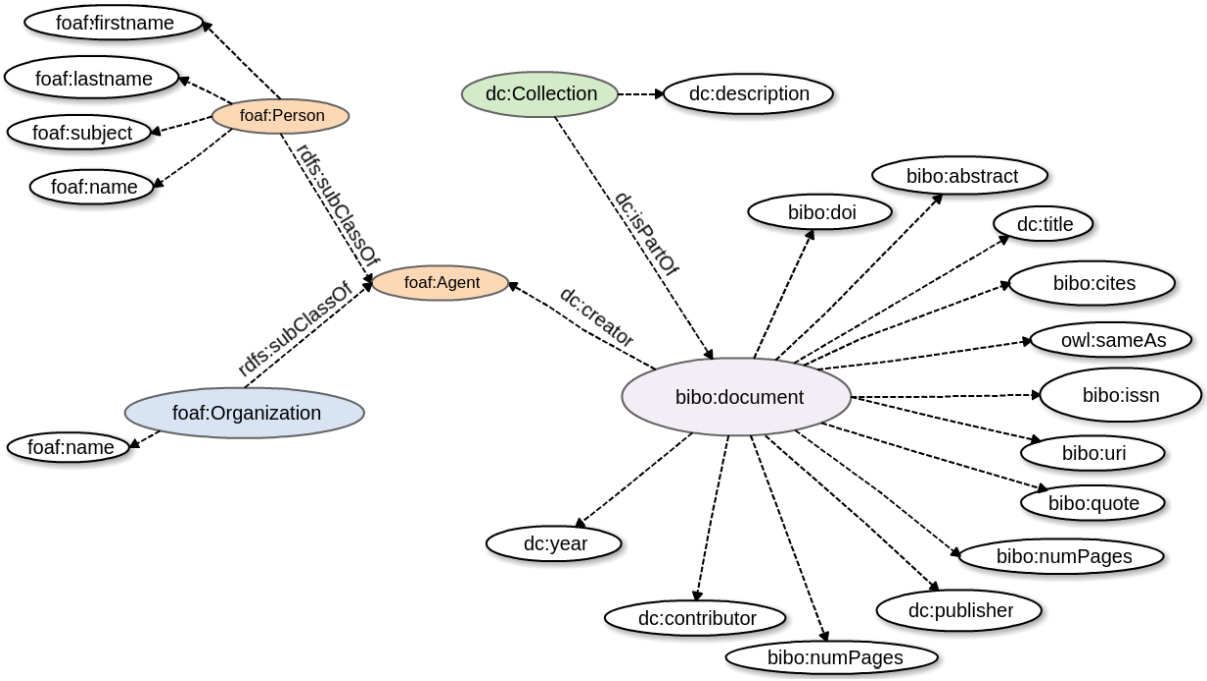

Figura 2. Modelo de datos bibliográfico.

3.3. Generación

La etapa de generación consiste en la transformación de datos de distintas fuentes al formato RDF, acorde a las consideraciones presentadas en la etapa de *especificación*, y el modelo de datos definido en base a los vocabularios utilizados en la etapa de *modelamiento.* Para lo cual, se ha definido diferentes procesos como se observa en la Figura 1, acorde a cada fuente de datos; los cuales tienen las siguientes tareas:

- extraer los datos desde las fuentes bibliográficas utilizadas (*Koha*, *Dspace*, bases digitales externas);
- transformar los recursos bibliográficos desde el modelo de datos de la fuente al modelo de datos definido, en formato RDF.

Para este proceso se ha utilizado un *framework* para la publicación de *Linked Data* que está basado en Pentaho Data Integration[14]. Este *framework* permite cubrir las etapas de publicación de datos, con la ventaja de poder modelar los procesos que dispone acorde al tipo de problemática que se presente. A continuación, se describen ciertas particularidades que se han presentado acorde al tipo de fuente de datos bibliográfica:

i). fuentes bibliotecarias,
ii). repositorios Dspace,
iii). bases digitales bibliográficas externas.

Fuentes bibliotecarias

En este caso se realiza la generación de *RDF* a partir de recursos en formato MARC21. En la Tabla 1, se presenta el mapeo correspondiente entre los atributos del formato MARC21 al modelo de datos ilustrado en la Figura 2. El procedimiento de generación de RDF, utilizando el modelo común, se efectúa sobre los 166,620 registros que dispone la biblioteca de la Universidad de Cuenca. El resultado es almacenado en el repositorio central.

[14] https://ucuenca.github.io/lodplatform/

Tabla 1. Mapeo entre MARC21 y el modelo de datos definido.

Campos de MARC21	Entidad/Propiedad/Relación al modelo
100a	dcterms:creator
245a	dcterms:title
20a	biboidentifier
250a	bibo:edition
260a	bibo:Editor
490anpvx	bibo:isbn
520a	bibo:Abstract
650a	dcterms:subject
653a	dcterms:subject
260abc	dcterms:issud
300a	bibo:numPages
326a	dcterms:Lotacion
41a	dcterms:Languaje
245abchn	bibo:volume

Repositorios Dspace

En el trabajo presentado en Segarra, Ortiz, Espinoza & Saquicela (2016) se describe el proceso a seguir para la publicación de datos enlazados de fuentes *Dspace*. Por lo cual, se ha utilizado el mismo procedimiento para la generación de datos desde este tipo de fuentes bibliográficas. El protocolo utilizado para la cosecha de meta datos es *OAI-PMH*. En la Tabla 2, se observa el mapeo entre los campos *OAI* definidos en los repositorios *Dspace*, correspondientes al modelo de datos definido.

Tabla 2. Mapeo entre campos OAI y el modelo de datos definido.

Campos OAI	Entidad/Propiedad/Relación al modelo
Date accessioned	dcterms:dateSubmitted
Date available	dcterms:available
Date issued	dcterms:issued
Abstract	bibo:abstract
Provenance	dcterms:provenance
Subject	dcterms:subject
Title	dcterms:title
Language	dcterms:language
License	dcterms:license
URI	bibo:uri
Handle	bibo:handle

Bases digitales bibliográficas externas

Los problemas identificados con bases bibliográficas externas radican en la validez del enlace establecido con la fuente original, y la cantidad de meta datos obtenido de cada recurso. Generalmente el contenido bibliográfico de bases digitales se contrata anualmente, por lo cual los enlaces proporcionados a los recursos caducan y son inservibles fuera del periodo contratado. Por otra parte, la cantidad y calidad de la meta data proporcionada por las bases digitales es precario en algunos casos y diferente en su estructura, debido a que cada base digital utiliza modelos de datos diferentes. Por lo cual, se define las correspondencias de entidades, propiedades y relaciones al modelo común por cada base digital. Por ejemplo, en Tabla 3 se especifica la correspondencia entre el modelo de datos de la base digital Ebrary-AC con respecto al modelo de datos común que se ha definido, en la cual se observa la diferencia existente entre las características de los modelos de datos. El procedimiento de mapeo se realiza con las bases digitales especificadas en la sección 192, el cual es utilizado para el proceso de generación de RDF de las fuentes digitales bibliográficas.

Tabla 3. Mapeo entre el modelo de datos de Ebrary-AC y el modelo de datos común.

Ebrary-AC	Entidad/Propiedad/Relación al modelo
ebrary DocID	bibo:identifier
ISBN print	-------
ISBN electronic	bibo:isbn
ISBN other	-------
ISSN	bibo:issn
OCLC number	bibo:oclcnum
Title	dcterms:title
Author	dcterms:creator
Publisher	bibo:Publisher
Imprint	-------
Year Published	dcterms:issued
Edition	bibo:edition
LC Call	-------
MARC Available	-------
Document Type	bibo:Document
Document Pages	bibo: numPages
Document URL	bibo:handle
Cover URL	bibo:Image

Linking

Relacionar recursos entre distintas fuentes de información es posiblemente el objetivo principal de *Linked Data*. Estas relaciones permiten enriquecer la información y así descubrir nuevo conocimiento que de otra forma permanecería oculto si los recursos estuvieran aislados. Encontrar recursos equivalentes es la forma más simple y usada de relacionar fuentes de datos. Estas equivalencias son llamadas enlaces y por convención son anotadas usando la ontología *OWL*[15], con el concepto *owl:sameAs*. Estos enlaces entre recursos son descubiertos usando el *RDF* generado desde las fuentes de información o inclusive servicios *SPARQL* de fuentes externas como *Dbpedia*.

En el presente proyecto se realizó un proceso de *Linking* interno, es decir, entre las fuentes de datos bibliográficas de la Universidad de Cuenca. Para esto se empleó la información generada como *RDF* proveniente de dichas fuentes. En este proceso se utilizó el *framework Silk*[16], mediante el cual es posible estimar la similitud entre recursos bibliográficos a través de comparaciones entre atributos y métricas de similitud sintáctica y semántica. Los recursos enlazados entre las fuentes corresponden a registros de libros y personas que son equivalentes, pero se encuentran repetidos en varias fuentes. De esta manera se permite navegar a través de recursos bibliográficos relacionados sobre distintas fuentes bibliográficas, lo posibilita la elaboración de consultas más complejas que involucren varias fuentes de información a la vez y que hagan explícito nuevo conocimiento.

3.4. Publicación

El objetivo de esta etapa es visibilizar el contenido bibliográfico, de tal manera que esté disponible y accesible en la web. Por lo cual, los datos transformados son almacenados en un repositorio semántico basado en tripletas. Las herramientas soportadas son *Apache Fuseki*[17] *o Apache Marmotta*[18], que son servidores *SPARQL*, que permiten el acceso a los datos a través de servicios web. Los datos son almacenados en grafos, acorde al origen de los datos. Por lo cual se ha establecido tres grafos: 1) Library: recursos provenientes de sistemas bibliotecarios. 2) *Dspace*: grafo para almacenar recursos provenientes de repositorios institucionales *Dspace*. 3) dSources: recursos provenientes de bases digitales

[15] http://www.w3.org/2002/07/owl#
[16] http://silkframework.org/
[17] http://marmotta.apache.org/
[18] https://jena.apache.org/documentation/fuseki2/index.html

contratadas. En la Figura 3, se ilustra el proceso de publicación de datos bibliográficos implementado, el cual se enfoca en un almacenamiento centralizado de información.

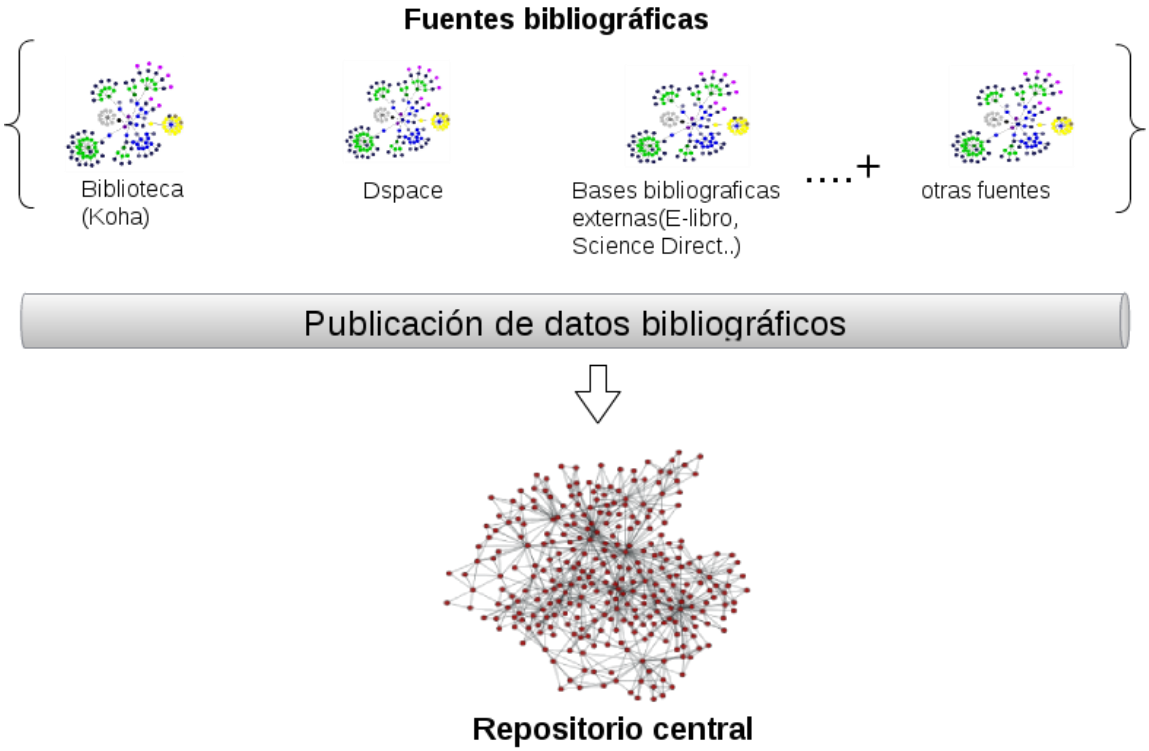

Figura 3. Publicación de recursos bibliográficos.

3.5. Explotación

Acorde con la guía metodológica de publicación de datos enlazados, en la etapa de explotación de datos deben implementarse mecanismos de acceso a la información para usuarios finales. En el presente caso de estudio se optó por la implementación de un buscador de recursos bibliográficos basado en tecnologías semánticas, el cual explota las potencialidades de los datos enlazados. Específicamente, se escogió a *FEDQuest*[19], un buscador diseñado para consumir datos enlazados provenientes de repositorios digitales. En la Figura 4, se ilustra la manera con la que se ha adoptado esta herramienta, que permite acceder al contenido bibliográfico de diferentes fuentes de forma centralizada. Esta herramienta tuvo que ser adaptada para trabajar con la información generada en este proyecto debido a pequeñas incompatibilidades con los modelos de datos y cambios de apariencia específicos. Este buscador junta varias funcionalidades para la explotación de información bibliográfica dentro de una estructura modular, funcionalidades que utilizan tanto enfoques puramente sintácticos como semánticos.

FEDQuest es una aplicación web que se conecta a un repositorio semántico a través de interacciones *RESTful* y utilizando el lenguaje de consulta *SPARQL*. Esta aplicación contiene varios módulos o componentes que implementan distintas funcionalidades para la búsqueda y exploración de recursos como búsquedas por texto, consultas gráficas y navegación por grafos. Además de características adicionales como control de usuarios, estadísticas y recomendación de contenidos.

Este nuevo servicio de búsqueda centralizada de la Universidad de Cuenca está disponible para su utilización por la comunidad universitaria en la siguiente página web: http://sgb.ucuenca.edu.ec/buscador.

[19] https://github.com/f35/sparql-fedquest

Figura 4. Arquitectura del buscador implementado.

4. EVALUACIÓN Y RESULTADOS

Actualmente la Universidad de Cuenca posee varios servicios para el acceso a la información bibliográfica de la institución. Servicios que son accedidos diariamente por la comunidad universitaria en sus diversos niveles: estudiantes, docentes e investigadores. En el presente trabajo se aplicó un proceso de integración sobre las fuentes de datos de la Universidad de Cuenca que ha permitido el surgimiento de un nuevo servicio de acceso centralizado a la información bibliográfica de esta institución. Este nuevo servicio posee notables ventajas con respecto al estado actual de los servicios bibliotecarios de la universidad. En la Tabla 4 se comparan los dos enfoques de acceso a la información a través de varios parámetros, tanto el que se utiliza actualmente y el propuesto en este trabajo. Los parámetros han sido seleccionados en base a los criterios presentados tanto por personal de la biblioteca como por técnicos de la institución.

Tabla 4. Comparación de los servicios bibliotecarios.

Funcionalidades	Propuesta	Estado actual
Modelo de datos	Ontologías bibliográficas.	MARC21, DCTERMS y modelos Ad-Hoc.
Identificación de recursos	URIs, identificadores únicos para la Web.	Varios sistemas de identificación, dependientes de las fuentes de datos.
Búsqueda y acceso a la información	Un solo buscador para todas las fuentes.	Un buscador por cada fuente.
Tipos de recursos independientes	Documentos, autores y colecciones.	Solo documentos.
Desambiguación de recursos.	Enlaces entre recursos equivalentes tanto inter como intra fuentes de datos.	Ninguno (Redundancia de recursos).
Consultas dependientes entre repositorios	Soporte para consultas complejas dependientes de varias fuentes	Consultas independientes.
Acceso a los datos	SPARQL Endpoint centralizado.	Dependientes de las fuentes de información: OAI, CSV, BDR, ...

- *Modelo de datos:* Los datos obtenidos de las fuentes bibliográficas se han descrito utilizando un modelo de datos bibliográfico en base a vocabularios y ontologías estándar en la web, de forma que se disponga de una forma común de representar la información proveniente de

fuentes bibliográficas independientes, las cuales utilizan diferentes tecnologías como MARC21, DCTERMS y modelos Ad-Hoc para representar la información.

- *Identificación de recursos:* Los datos proceden de varias fuentes bibliográficas, por lo cual se ha asignado a cada recurso (documentos, colecciones y autores) identificadores únicos en la web de forma unívoca, de tal manera que el recurso pueda ser ubicado fácilmente ya que en este identificador se especifica el dominio y el tipo de recurso al que se está accediendo, por ejemplo: http://sgb.ucuenca.edu.ec:3000/resource/documento/cdjbv%2Fcce7fa3f1af8d0d8dd88213f1faecc7a_1955.
- La *búsqueda y acceso a la información* con el enfoque propuesto en este trabajo permite acceder a los recursos bibliográficos a través de un solo buscador de forma centralizada, en contraste al procedimiento que utiliza la biblioteca actualmente, en el cual ofrece al usuario varios buscadores independientes por cada fuente de información.
- *Tipos de recursos independientes* que permita buscar por documentos, colecciones de documentos como se lustra en la Figura 5 y por autores, en contraste al procedimiento que actualmente se emplea en la biblioteca, en el cual se dispone únicamente de búsqueda por documentos.

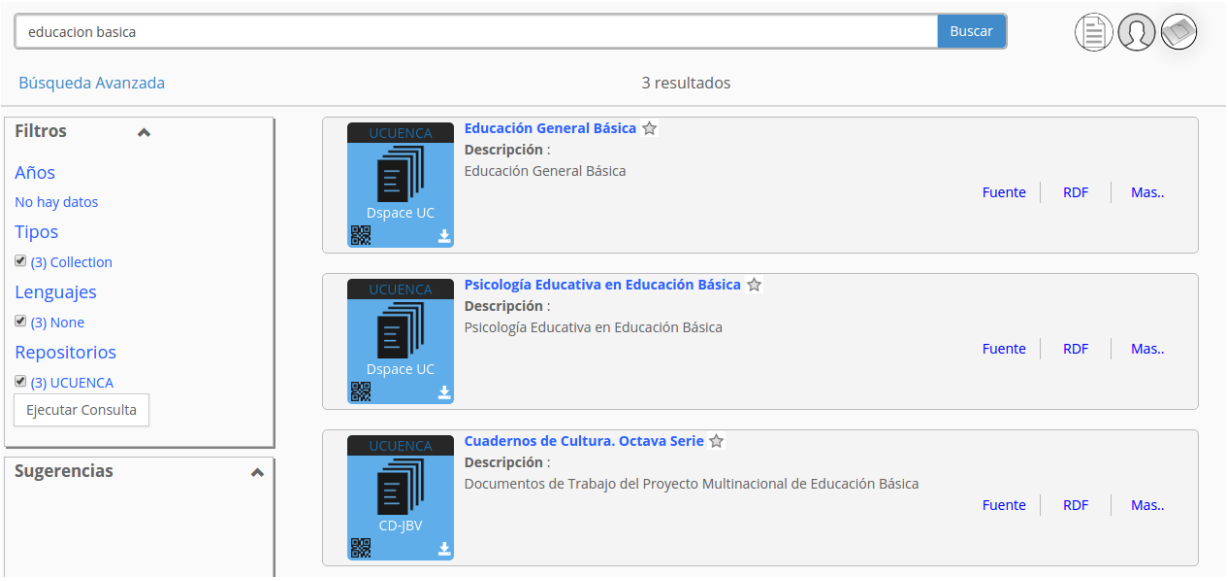

Figura 5. Búsqueda por colecciones.

- *Desambiguación*: la ambigüedad de recursos es un problema recurrente en las bibliotecas principalmente debido a la utilización de múltiples y diversas fuentes de información que en muchos casos contienen información repetida. La Universidad de Cuenca no es la excepción, puesto que se pueden encontrar documentos y autores repetidos en las diversas fuentes de información disponibles. Este problema aún no ha sido tratado por la institución debido a la falta de mecanismos para enlazar recursos equivalentes entre las fuentes. En contraste, en el presente trabajo se presenta una solución a este problema que utiliza la infraestructura de datos enlazados para desambiguar recursos a través de creación enlaces inter e intra fuentes.

  Una de las formas más comunes para aprovechar estos enlaces es la navegación y visualización de información de forma unificada. En la Figura 6 se presenta una captura de esta funcionalidad dentro del buscador *FEDQuest*, donde se despliega información de recursos relacionados a la obra '*Nine modern muses: Women and the nobel prize for literature*'. La información presentada incluye autores, colecciones y otros documentos relacionados presentados en forma de un grafo expandible, información que es extraída de varias fuentes y relacionada a través de enlaces.

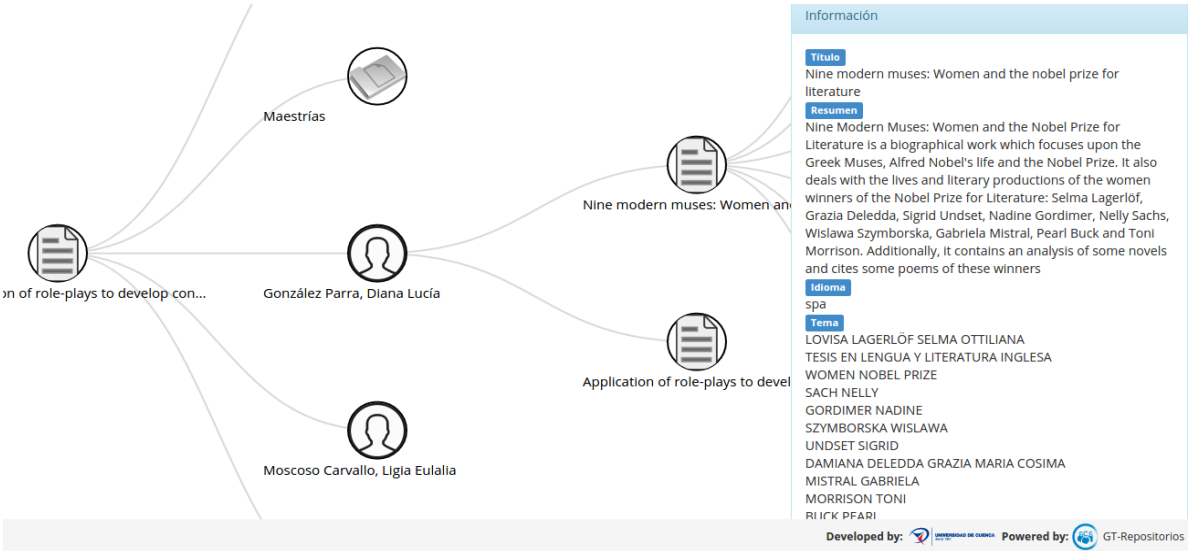

Figura 6. Navegación a través de grafos.

- *Consultas dependientes*: la ejecución de consultas complejas sobre fuentes de información distribuidas es una necesidad que cada vez cobra mayor fuerza en el ámbito bibliotecario. Esto, debido a que pueden ser utilizadas en la elaboración de estudios bibliométricos o de minería de datos que ayuden al descubrimiento de nuevo conocimiento. Actualmente, la Universidad de Cuenca no posee mecanismos que permitan realizar este tipo de consultas de forma automática debido a que las fuentes de información poseen modelos de datos heterogéneos y manejan formatos distintos. Esto impide que se puedan realizar consultas que extraigan información de varias fuentes a la vez (consultas dependientes) o que realicen operaciones complejas de procesamiento de la información.
  En el presente trabajo se abre una ventana a este tipo de requerimientos puesto que además del buscador Web, se tiene acceso a la información a través de consultas *SPARQL*. Estas consultas proveen una gran flexibilidad para acceder a información de varias fuentes simultáneamente y también permite la interconexión con repositorios externos de datos enlazados como *Dbpedia*. Además, existe una considerable cantidad de herramientas para el análisis y minería de datos sobre datos enlazados que abstraen la complejidad del lenguaje *SPARQL* a través de interfaces gráficas más simples. Esto facilitaría en gran medida trabajos futuros relacionados con análisis de información o funcionalidades de búsqueda más avanzadas para usuarios y bibliotecarios.
- *Acceso a los datos*: El acceso a la información bibliotecaria para desarrolladores es una consideración necesaria en los sistemas modernos, en especial con el continuo mejoramiento de *software* de gestión académica. En la Universidad de Cuenca como en otras instituciones se están desarrollando herramientas para facilitar el manejo de la información académica e investigativa como: sistemas de portafolios para docentes, sistemas de sílabos o repositorios de investigadores. En este contexto, y con el fin de generar sistemas integrados, resulta útil enlazar la información de estos sistemas de gestión con la información bibliotecaria de la institución. Por ejemplo, relacionar los textos guías a ser usados en los cursos (sistemas de sílabos) con los libros disponibles en la biblioteca.
  Actualmente, no ha podido realizarse una relación entre los diferentes tipos de recursos de diversas fuentes debido a la gran variedad de sistemas de identificación y métodos de acceso a la información de las fuentes bibliotecarias. Esto dificulta en gran medida el incorporar valiosa información de fuentes adicionales dentro de los sistemas de gestión de la universidad u otro tipo de sistemas. En el presente trabajo se ha abordado esta problemática proveyendo un sistema de identificación único de recursos a través de URIs, además de brindar un acceso homogenizado a los datos a través de SPARQL. Todo esto facilita el acceso a la información para ser accedida desde otros sistemas o para propósitos diferentes.

5. CONCLUSIONES Y TRABAJO FUTURO

La diversidad de fuentes en el ámbito bibliotecario continúa siendo un reto actual, tanto para los administradores como para los usuarios que comúnmente tienen que recurrir a múltiples herramientas para la gestión y acceso de los recursos disponibles. En este trabajo se presentó una propuesta para superar la heterogeneidad de fuentes bibliográficas mediante la integración de estas a través de un enfoque centralizado de *Linked Data*. Esta propuesta fue aplicada en la biblioteca de la Universidad de Cuenca. Para la aplicación de este enfoque, se han seguido las etapas definidas por la metodología de publicación de *Linked Data* a través de diferentes herramientas disponibles. Mediante este procedimiento se ha logrado mejorar la visibilidad del contenido bibliográfico que ahora puede ser accedido a través de una sola herramienta, lo que ayuda a aprovechar el conocimiento académico y científico disponible en diversas fuentes. El enfoque utilizado permite también conseguir un grado más alto escalabilidad e interoperabilidad frente a la adición de nuevas fuentes bibliográficas, debido a que comparten un modelo de identificación y de representación común (ontologías).

El trabajo futuro se centrará en aumentar el soporte para diferentes tecnologías de *SPARQL Endpoints*, además de mejorar los tiempos de respuesta en las consultas *SPARQL*. Por otra parte, es necesario enriquecer los recursos bibliográficos con enlaces a otras fuentes externas, logrando aumentar la visibilidad de la producción académica y científica de la institución y de otras instituciones.

AGRADECIMIENTOS

Este trabajo fue realizado en el Departamento de Ciencias de la Computación de la Universidad de Cuenca[20] y ha sido patrocinado por RED-CEDIA[21] como parte del Grupo de Trabajo de Repositorios[22] y del proyecto 'Repositorio Semántico de Investigadores del Ecuador' (REDI).

REFERENCIAS

Baker, T., Bermes, E., Coyle, K., Dunsire, G., Isaac, A., Murray, P., Panzer, M., Schneider, J., Singer, R., Summers, E., Waites, W., Young, J., & Zeng, M. (2011). *Library Linked Data Incubator Group Final Report*. Disponible en https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/

Estivill-Rius, A. (2011). *Resource description and access, RDA.* Un nuevo retraso para preparar mejor el cambio. *El Profesional de la Informacion*, *20*(6), 694-700.

García Álvarez de Toledo, J., & Fernández Sánchez, R. (2011). *Difusión y divulgación científica en internet.* Gobierno del Principado de Asturias, Cienciatec, Asturias, 114 pp. Disponible en http://blogs.ujaen.es/cienciabuja/wp-content/uploads/2013/06/Difusion-y-divulgacion-cientifica-en-Internet.pdf

Harper, C.A., & Tillett, B.B. (2007). Library of congress controlled vocabularies and their application to the semantic web. *Cataloging & Classification Quarterly*, *43*(3-4), 47-68.

Heflin, J., & Hendler, J. (2000). *Semantic interoperability on the web*. Extreme Markup Languages 2000. Department of Computer Science, University of Maryland, College Park, MD, USA, 15 pp.

Hopkinson, A. (2016). *UNIMARC Manual - Bibliographic Format* (3rd ed.). IFLA Series on Bibliographic Control; Nr 36.

[20] https://www.ucuenca.edu.ec/la-oferta-academica/oferta-de-grado/facultad-de-ingenieria/dptos/dcc/sobre-el-dpto

[21] http://www.cedia.edu.ec/

[22] http://gtrepositorios.cedia.org.ec/

Kruk, S., Synak, M., & Zimmermann, K. (2005). *MarcOnt: integration ontology for bibliographic description formats*. DC-2005: Proc. Int. Conf. on Dublin Core and Metadata Applications, Madrid, Spain, 4 pp.

Lencinas, V. (2001). *Software bibliotecario - abierto y gratuito*. III Jornadas De Bibliotecas De La Provincia de Cordoba y I Jornadas de Profesionales de la Información - Córdoba, Argentina, 44 pp. Disponible en http://eprints.rclis.org/19994/1/software%20libre.pdf

Malmsten, M. (2008). *Making a library catalogue part of the semantic web*. Proc. of the 2008 International Conference on Dublin Core and Metadata Applications, pp. 146-152. Disponible en http://eprints.rclis.org/19994/1/software%20libre.pdf

Ríos-Hilario, A., Martín-Campo, D., & Ferreras-Fernández, T. 2012. Linked data and linked open data and its implementation in a digital library: the case of Europeana. *El Profesional de la Informacion*, *23*(3), 292-297.

Segarra, J., Ortiz, J., Espinoza, M., & Saquicela, V. (2016). *Integration of digital repositories through federated queries using semantic technologies*. Computing Conference (CLEI), 2016 XLII Latin American, Valparaiso, Chile.

Torre-Bastida, A., González-Rodríguez, M., & Villar-Rodríguez, E. (2015). Datos abiertos enlazados (lod) y su implantación en bibliotecas: iniciativas y tecnologías. *El Profesional de la Información*, *24*(2), 113-120.

Vila-Suero, D., & Gómez-Pérez, A. (2013). datos.bne.es and MARiMbA: an insight into library linked data. *Library hi tech*, *31*(4), 575-601.

Villazón-Terrazas, B., Vilches-Blázquez, L.M., Corcho, O., & Gómez-Pérez, A. (2011). *Methodological guidelines for publishing government linked data*. In: Wood, D. (Ed.). Linking Government Data. New York, NY: Springer New York, pp. 27-49.