

Integración Semántica de Recursos Educativos Abiertos cosechados con OAI-PMH. Proceso aplicado al servicio de búsqueda de OERs en la Red ESVIAL

Nelson Piedra¹, Janneth Chicaiza¹, Pricila Quichimbo¹

¹ Tecnologías Avanzadas de la Web y Sistemas Basados en el Conocimiento
Departamento de Ciencias de la Computación y Electrónica

Universidad Técnica Particular de Loja
1101608 San Cayetano Alto S/N (Loja-Ecuador)

nopiedra@utpl.edu.ec, jachicaiza@utpl.edu.ec, pvquichimbo@utpl.edu.ec

Resumen. En un contexto de interoperabilidad y re-uso de contenidos educativos abiertos existen ciertas circunstancias que deben darse para lograr la integración. Básicamente deben garantizarse ciertas condiciones de apertura y autonomía local y ciertas condiciones de apertura y autonomía global. Interoperabilidad es la capacidad de dos entidades para trabajar conjuntamente con el propósito de realizar alguna tarea. Se trata de ampliar la apertura, visibilidad, participación y acceso a los sistemas de información. El mayor reto es no requerir, para cooperar o reusar, acuerdo previo entre las entidades que gestionan los sistemas de información, en cuanto a la gestión y el tratamiento de los recursos digitales que se comparten. Para lograr la interoperabilidad entre repositorios de Librerías Digitales, la Iniciativa de Archivos Abiertos (OAI) propone el protocolo para la exposición y recolección de metadatos denominado OAI-PMH. Actualmente, para mejorar la interoperabilidad de datos en la Web y mejorar el acceso a datos de repositorios distribuidos, se aplican tecnologías semánticas. En este trabajo, se presenta el proceso de extracción de metadatos, generación de RDF y publicación de datos enlazados con el propósito de mejorar la integración e interoperabilidad entre recursos almacenados en Librerías Digitales; bajo este enfoque, cada Librería Digital puede conservar sus cualidades locales específicas y no requerirá resignarlas para poder normalizar el intercambio o la cosecha de recursos digitales. La propuesta descrita facilita la existencia de diversidad de métodos y estándares en los procesos de cada proveedor de recursos digitales, que serán integrados al servicio de búsqueda de Recursos Educativos Abiertos de la Red ESVI-AL.

Palabras clave: repositorios digitales, datos enlazados, Web Semántica, RDF, OAI-MPH, ontología, ESVI-AL.

1 Introducción

Internet genera un escenario global en el que las condiciones de interoperabilidad son las que garantizan el descubrimiento, la distribución y re-uso de recursos digitales, más allá de condiciones de gestión local, de modelos tecnológicos y de herramientas de cualquier tipo. La interoperabilidad se define como la capacidad de dos o más sistemas o componentes de intercambiar información para su posterior uso [1]. Para lograr la

interoperabilidad entre repositorios de Librerías Digitales la Iniciativa de Archivos Abiertos (OAI) propone el protocolo para la recolección de metadatos denominado OAI-PMH¹, el cual está basado en estándares abiertos. En el contexto de este trabajo, los estándares abiertos garantizan la interoperabilidad entre emisores y receptores de los recursos digitales, con independencia del software utilizado e incentivando la neutralidad tecnológica y la innovación.

La interoperabilidad e integración de metadatos bibliográficos puede proporcionar a la comunidad académica y científica una cantidad enorme de servicios [12], por ejemplo: recursos digitales que pueden reusarse y adaptarse a nuevas necesidades educativas, búsqueda de expertos en un dominio determinado, detección de nuevos temas de investigación, análisis de redes científicas, entre otros. Un escenario de integración hace posible la combinación de recursos de información existentes en diversas fuentes, esto proporciona al usuario una vista unificada de dichos recursos y también puede actuar como una fuente de datos para diversas aplicaciones [2].

En este trabajo, se presenta el proceso de publicación de los datos bibliográficos extraídos de los repositorios digitales que usan OAI-PMH, siguiendo los principios de Linked Data y el ciclo de vida para la publicación de datos enlazados, contribuyendo así al enriquecimiento de la Web de datos. El piloto se aplica a repositorios de universidades latinoamericanas que usan tecnologías como Eprints² y DSpace³ porque ofrecen la posibilidad de acceder a su información a través del protocolo OAI-PMH, y representan los metadatos de los recursos utilizando el estándar Dublin Core⁴. Este proceso de cosecha es importante para el servicio de búsqueda de Recursos Educativos Abiertos (OER, por sus siglas en inglés *Open Educational Resources*) relacionados con Accesibilidad de la Red ESVAL⁵, y que serán integrados desde el buscador de contenidos abiertos Serendipity⁶, así como repositorios que cuenten con servicios OAI para acceso a recursos digitales. La propuesta emerge sobre una base contextual de cultura libre, una serie de herramientas de software libre para autoría y edición de recursos educativos y acceso a repositorios OER con licencias del tipo creative commons o similares.

1.1 Protocolo OAI-PMH

Desde el punto de vista tecnológico, las librerías digitales son repositorios que almacenan objetos digitales y utilizan OAI-PMH para exponer sus metadatos.

OAI-PMH es un protocolo propuesto por la *Open Archives Initiative*, para poder extraer metadatos (normalmente en formato Dublin Core) de repositorios digitales como DSpace. Para obtener los metadatos de los repositorios digitales se utilizan los servidores de datos, que realizan solicitudes conocidas como verbos. Como se puede

¹ OAI-PMH, OAI-PMH, abreviatura de Open Archives Initiative Protocol for Metadata Harvesting www.openarchives.org/OAI/openarchivesprotocol.html

² EPrints, <http://www.eprints.org/>

³ DSpace, <http://www.dspace.org/>

⁴ Dublin Core, <http://dublincore.org/>

⁵ Guía Metodológica y Modelo de Acreditación ESVAL <http://www.esval.org/guia/>

⁶ Serendipity: <http://serendipity.utpl.edu.ec>

ver en la Tabla 1, OAI-PMH soporta 6 tipos de solicitudes.

Tabla 1. Lista de verbos soportados por OAI-PMH

| Verbo | Descripción |
|---------------------|--|
| Identify | Obtener información del repositorio |
| ListMetadataFormats | Obtiene los formatos Soportados por los repositorios |
| ListSets | Obtiene las colecciones creadas por el repositorios |
| ListRecords | Obtiene todos los registros del repositorio |
| ListIdentifiers | Obtiene los identificadores de cada registro. |
| GetRecord | Obtiene la información detalladas de una registro específico |

OAI-PMH es compatible con muchas herramientas que permiten la creación de repositorios institucionales, entre las principales se encuentran Eprints, Dspace, Fedora, entre otros. OAI-PMH divide este fenómeno en proveedores de datos y proveedores de servicios, en donde los primeros son repositorios que exponen sus metadatos a través de OAI-PMH, los segundos también llamados “harvesters o recolectores” desarrollan servicios de valor agregado basados en los metadatos obtenidos de los proveedores. En OAI-MPH cada repositorio almacena sus objetos digitales de manera independiente.

1.2 Esquema de Metadatos Dublin Core

Dublin Core o la Iniciativa de Metadatos Dublin Core (DCMI⁷) es el esquema de metainformación más utilizado a nivel mundial.

La norma ISO15836 define el conjunto de elementos Dublin Core, o lo que se conoce habitualmente como "DC simple". Esos 15 elementos básicos para describir cualquier objeto de información, se presentan habitualmente divididos en tres grupos que indican la clase o alcance de la información incluida en ellos, y que responden, en cierta medida, a las expectativas que tiene el usuario cuando se enfrenta a la información de la Red. En la Tabla 2, se enlista un subconjunto de metadatos que DC propone.

Tabla 2. Elementos Dublin Core

| Contenido | Propiedad intelectual | "Instantiation" /Ejemplo |
|------------------|------------------------------|---------------------------------|
| Title | | |
| Subject | Creator | Date |
| Description | Publisher | Type |
| Source | Contributor | Format |
| Language | Rights | Identifier |
| Relation | | |
| Coverage | | |

⁷ www.dublincore.org, página oficial

Además de estos elementos básicos (ninguno obligatorio y todos repetibles) existen otros mecanismos que sirven para adaptar DC a las necesidades concretas de información y que hacen que este modelo de metadatos sea aplicable a cualquier proyecto de sistema o servicio de información digital. El conjunto de elementos DC se ha convertido en una infraestructura operacional del desarrollo de la Web Semántica.

1.3 Hacia la integración de repositorios a través de Datos Enlazados

La publicación de datos enlazados se fundamenta en cuatro principios básicos de diseño propuestos por Tim Berners-Lee en el 2006:

- (1) usar URIs⁸ para identificar los recursos de la Web,
- (2) usar el protocolo HTTP de la URI para que los usuarios puedan localizar y consultar estos recursos,
- (3) proporcionar información útil acerca del recurso cuando la URI haya sido consultada, utilizando RDF⁹ para describir recursos y SPARQL¹⁰ para consultarlos, e
- (4) incluir enlaces a otras URIs relacionadas con los datos contenidos en el recurso, de forma que se potencie el descubrimiento de información en la Web.

El enfoque de Linked Data ofrece ventajas significativas sobre las prácticas actuales de publicación de datos, pues mediante el uso de identificadores únicos las bibliotecas permitirán que los recursos sean más fácilmente accesibles.

El grupo de la W3C, *Incubator Group* [4] describe los siguientes beneficios de utilizar este enfoque:

- beneficios para los investigadores, estudiantes y clientes, debido a que es posible realizar búsquedas federadas mediante el uso de enlaces para ampliar los índices y es posible acceder a un conjunto más rico de información ya que los datos de la biblioteca estarán integrados en documentos de investigación y bibliografías;
- beneficios para las organizaciones, al aumentar la presencia en la web pues se mantienen mejores vínculos entre los recursos y sus descripciones;
- beneficios para bibliotecarios y archivistas, quienes pueden tener acceso a datos compartidos que describen los recursos con una cantidad limitada de esfuerzos redundantes;
- beneficios para los desarrolladores y proveedores, los desarrolladores se benefician al no estar atados a los formatos de datos de una biblioteca específica y los vendedores serán capaces de comercializar sus productos fuera del mundo de las bibliotecas.

⁸ URI.- Uniform Resource Identifier, permite identificar los recursos de forma unívoca.

⁹ RDF.- Resource Framework Description, es un lenguaje para representar y publicar datos estructurados en la Web.

¹⁰ SPARQL.- lenguaje para consultar grafos RDF.

2 Determinación del alcance y selección de fuentes de datos

2.1 Descripción del Proceso general

Las Librerías Digitales son proveedoras de datos además de repositorios de objetos digitales en formato Dublin Core y utilizan OAI-PMH para exponer sus metadatos. A continuación, se detallan las actividades de extracción de metadatos usando OAI-PMH y su publicación como datos enlazados (ver figura 1):

1. Visionamiento
2. Cosecha de metadatos desde repositorios:
 - 1.1. Ejecución de Harvester (cosechadores de OAI-PMH), uso de librería Harvester 2.0 para la extraer los metadata de los repositorios, a través del protocolo OAI-PMH.
 - 1.2. Almacenamiento de datos cosechados en un repositorio relacional y en formato de tripletas OAI.
3. Diseño de URIs y reuso de recursos ontológicos.
4. Limpieza de datos cosechados.
5. Conversión a RDF:
 - 5.1 Limpieza de datos generados: actividades que buscan reducir ambigüedad, purgar la información extraída y generada durante el proceso de conversión.
 - 5.2 Enlazado de datos: Idiomas, Organismos, Conceptos.
6. Publicación y Explotación.

2.2 Identificación, análisis y selección de fuentes de datos

La Federación de Biblioteca Digitales define a las bibliotecas como, organizaciones que proporcionan los recursos, y el personal especializado para seleccionar, estructurar, ofrecer acceso intelectual, interpretar, distribuir, preservar la integridad y asegurar la persistencia en el tiempo de las colecciones de obras digitales, de tal manera que sean de fácil acceso y económicamente disponibles para su uso por una comunidad o por un conjunto de comunidades [10].

En este trabajo, se usó el protocolo OAI-PMH como modelo técnico para la cosecha de recursos educativos, cuyos registros están descritos mediante Dublin Core. El material incluido en estos repositorios comprende: atlas, CDs, DVDs, ebooks, enciclopedias, folletos, juegos, libros, memorias, revistas, tesis; esto representa una oportunidad inmejorable para analizar los datos contenidos.

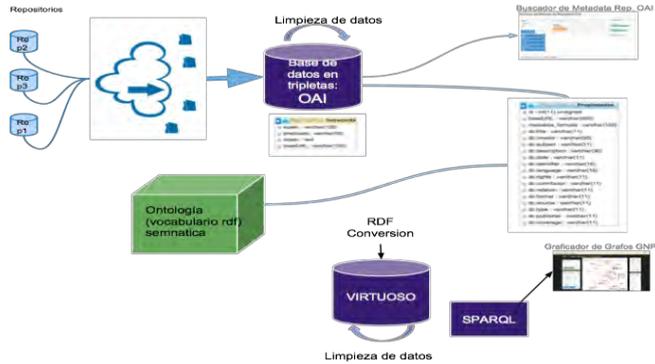


Fig. 1. Visión General de Funcionamiento

2.3 Extracción de información usando Harvester2

Se ha realizado la extracción de metadatos desde repositorios digitales que han habilitado el protocolo OAI. La aplicación de cosecha de metadatos a través de OAI-PMH, está basada en Harvester2. Se uso el verbo “listRecords”. Como ejemplo, el URL para el repositorio de la Universidad Técnica Particular de Loja (UTPL) es <http://dspace.utpl.edu.ec/oai/request>. Los metadatos extraídos se almacenan en una base de datos relacional en forma de tripletas.

2.4 Limpieza de datos extraídos

Se realiza la limpieza de datos con el objetivo de detectar y corregir datos corruptos y erróneos. El proceso consistió en analizar patrones de error en los datos y ejecutar un esquema de limpieza. Por ejemplo se encontraron variaciones al momento de describir ciertos metadatos, como lenguaje. En la Figura 2, la columna “o” tiene el patrón a corregir, y la columna “o_syn” el valor o expresión con el que fue remplazado el valor erróneo.

| o | o_syn |
|-------|---------|
| en | English |
| es | Spanish |
| en_US | English |
| spa | Spanish |
| eng | English |

Fig. 2. Limpieza de Lenguaje.

También se detectaron errores en la descripción de autores. Se consideró como patrón válido a los nombres con la secuencia: <Apellidos, Nombres>; la cadena debía ser atómica, es decir admitir un autor. A continuación se describen ejemplos de errores detectados y corregidos.

| o | o_syn |
|--|--------------------------------|
| Velasco Vaca.Grijalva Rodas, Angelica Marisol.Ruth Cecilia | Velasco Vaca, Angelica Marisol |
| Velasco Vaca.Grijalva Rodas, Angelica Marisol.Ruth Cecilia | Grijalva Rodas, Ruth Cecilia |
| Aguilar Salazar, Byron Gloria Ximena Bustamante | Aguilar Salazar, Gloria Ximena |
| Aguilar Salazar, Byron Gloria Ximena Bustamante | Bustamante, Byron |
| Cueva Rivera Ximena Edditta | Cueva Rivera, Ximena Edditta |
| Alvarez V. Achig A., Jacqueline A. Vickys.S | Alvarez V., Jacqueline A. |
| Alvarez V. Achig A., Jacqueline A. Vickys.S | Achig A., Vickys.S |

Fig. 3. Limpieza de Autores.

Finalmente, se eliminaron problemas de ambigüedad en los tópicos (dc:subject) que describen a cada recurso digital, usando un esquema de corrección semi-automática de los casos encontrados, en base a patrones detectados.

3. Modelamiento de Vocabulario

Las ontologías y los vocabularios abiertos constituyen el esquema base a partir del cual se describen los recursos y entidades de la Web. Las actividades que comprende el modelamiento de la ontología son: (A) mapeo de conceptos, (B) búsqueda de recursos ontológicos y no ontológicos a reusar y (C) diseño de URIs persistentes.

En esta sección se presenta el modelado del vocabulario utilizado para describir los recursos en formato RDF. De forma concreta, se describen los conceptos, las propiedades y las relaciones entre los recursos bibliográficos de la UTPL y los vocabularios reutilizados.

3.1 Mapeo de Conceptos

En esta actividad de creación de ontologías, se mapean los principales conceptos del dominio, en este caso los recursos bibliográficos con los que cuenta el repositorio de la UTPL. Los conceptos identificados son los siguientes:

- Recurso bibliográfico: entidad bibliográfica.
- Persona: creador o contribuyente de la entidad bibliográfica.
- Corporación: organización que realizó la publicación de la entidad bibliográfica.
- Tópicos o temas: tópicos relacionados a la entidad bibliográfica.

Los metadatos extraídos describen cada recurso bibliográfico a través de Dublin Core. En el contexto de este trabajo, estos metadatos se clasificaron en tres grupos. Cada grupo indica la clase o el ámbito de la información que se guarda en ellos: elementos relacionados al contenido del recurso, elementos relacionados al recurso cuando es visto como propiedad intelectual y elementos relacionados con la instanciación del recurso.

Tabla 3. Mapeo de metadatos Dublin Core extraídos

| Contenido | | |
|------------------------------|----------------|--|
| Título | dc:title | Nombre del recurso |
| Tópico | dc:subject | Frasas que describen el título o contenido del recurso |
| Descripción | dc:description | Resumen o descripción del contenido |
| Fuente | dc:source | Secuencia de caracteres usados para identificar unívocamente un trabajo a partir del cual proviene el recurso actual |
| Tipo | dc:type | Categoría del recurso |
| Relación | dc:relation | Es un identificador de un segundo recurso y su relación con el recurso actual |
| Cobertura | dc:coverage | Característica de cobertura espacial o temporal del contenido intelectual del recurso |
| Propiedad intelectual | | |
| Autor o creador | dc:creator | Persona responsable de la creación del contenido intelectual del recurso |
| Editor | dc:publisher | Entidad responsable de hacer que el recurso se encuentre disponible en la red en su formato actual |
| Otros colaboradores | dc:contributor | Persona u organización que haya tenido una contribución intelectual significativa, pero que esta sea secundaria |
| Derechos | dc:rights | Nota sobre los derechos de autor |
| Instanciación | | |
| Fecha | dc:date | Fecha en la cual el recurso se puso a disposición del usuario |
| Formato | dc:format | Es el formato de datos de un recurso |
| Identificador | dc:identifier | Secuencia de caracteres para identificar unívocamente al recurso, por ejemplo: URN, URL, ISBN, etc. |
| Lenguaje | dc:language | Lenguaje del contenido intelectual del recurso |

En cuanto a la clasificación temática de los materiales bibliográficos, se mapearon estos conceptos mediante el vocabulario SKOS¹¹. La Tabla 5, indica los elementos de SKOS que se utilizaron en este trabajo.

Tabla 4. Mapeo de conceptos con elementos del vocabulario SKOS.

| Conceptos y esquemas de conceptos | | |
|--|----------------|---|
| Concepto | skos:Concept | Para describir la estructura conceptual o intelectual de un sistema de organización del conocimiento |
| Etiquetas léxicas | | |
| Etiqueta | skos:preflabel | “preferred label” útiles al generar o crear representaciones legibles de un sistema de organización del conocimiento; proporcionan los indicios más fuertes en cuanto al significado de un concepto |

3.2 Búsqueda de recursos ontológicos y no ontológicos a reusar

El reuso de recursos ontológicos y no-ontológicos reduce el tiempo de desarrollo y los costes asociados en esta fase, también contribuye a la calidad de la ontología [5]. Debido

¹¹ SKOS modela conceptos y esquemas de conceptos, etiquetas léxicas, relaciones semánticas, documentación, colecciones de conceptos, propiedades de mapeado y notaciones.

a que existen varios vocabularios que permiten modelar recursos bibliográficos, y que esto favorece la integración e interoperabilidad de datos en la Web, se han seleccionado los siguientes:

- RDF Schema¹², y OWL [10]¹³ para describir conceptos del vocabulario.
- *Simple Knowledge Organization System (SKOS)*¹⁴ para establecer un modelo de organización del conocimiento, taxonomías, y otras jerarquías temáticas. El vocabulario SKOS permite representar conceptos o temas tratados por el recurso.
- FOAF¹⁵, desarrollado para representar a personas y organizaciones con sus atributos y relaciones hacia otros conceptos.
- *Dublin Core (DC)* proporciona un vocabulario de características “base”, capaces de proporcionar información descriptiva básica sobre cualquier recurso.
- *Dublin Core Metadata Initiative (DCMI) Metadata Terms*¹⁶ usado para representar documentos, así como sus atributos como título, creador, y relaciones con otras entidades.
- *Bibliographic Ontology Specification (BIBO)*¹⁷ provee los principales conceptos y propiedades para describir citas y referencias bibliográficas.
- VIVO¹⁸ que permite describir diferentes conceptos en el dominio académico y científico.
- *Open Provenance Model Vocabulary*¹⁹ para describir la procedencia de datos en la Web.
- *Schema*²⁰ permite describir conceptos como materiales y sus propiedades de datos y relaciones entre clases.

Los vocabularios mencionados describen semánticamente conceptos relacionados en el dominio de trabajo. Los atributos, propiedades y entidades que no son considerados en estos vocabularios, han sido diseñados y creados como parte del vocabulario OAR-UTPL. La Figura 4, presenta cómo se han reutilizados estos recursos en el dominio de este trabajo.

3.3 Diseño de estrategia para identificar a los recursos a través de URIs persistentes

Los URIs son utilizados para identificar de manera unívoca a los recursos Web, por tal razón se han diseñado dos tipos de URIs: 1) para identificar los componentes del vocabulario (clases, propiedades y relaciones); y, 2) para describir el material bibliográfico. Para crear los URIs se han utilizado URIs HTTP tomando en cuenta los

¹² RDF Schema 1.1 (2014). W3C Recommendation: <http://www.w3.org/TR/rdf-schema/>

¹³ OWL Web Ontology Language. W3C Recommendation: <http://www.w3.org/TR/owl2-overview/>

¹⁴ SKOS Specification: <http://www.w3.org/2004/02/skos/>

¹⁵ FOAF Vocabulary Specification: <http://xmlns.com/foaf/spec/>

¹⁶ CDMI Metadata Terms: <http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=terms#H3>

¹⁷ BIBO: <http://purl.org/ontology/bibo/>

¹⁸ VIVO <http://vivoweb.org/ontology/core>

¹⁹ Open Provenance Model Vocabulary Specification <http://open-biomed.sourceforge.net/opmv/ns.html#Process>

²⁰ Schema project <http://schema.org>

principios para la publicación de datos propuestos por Tim Berners-Lee, siguiendo los patrones mencionados a continuación:

- Prefix: oar-utpl
- URI base: <http://data.utpl.edu.ec/serendipity/oar/>
- Schema: <http://data.utpl.edu.ec/serendipity/oar/schema#>
- Resources: <http://data.utpl.edu.ec/serendipity/oar/resource/>
- Properties: <http://data.utpl.edu.ec/serendipity/oar/property/>
- Categories: <http://data.utpl.edu.ec/serendipity/oar/category/>
- Graph: <http://data.utpl.edu.ec/serendipity/oar/>
- SPARQL endpoint: <http://data.utpl.edu.ec/serendipity/oar/sparql>

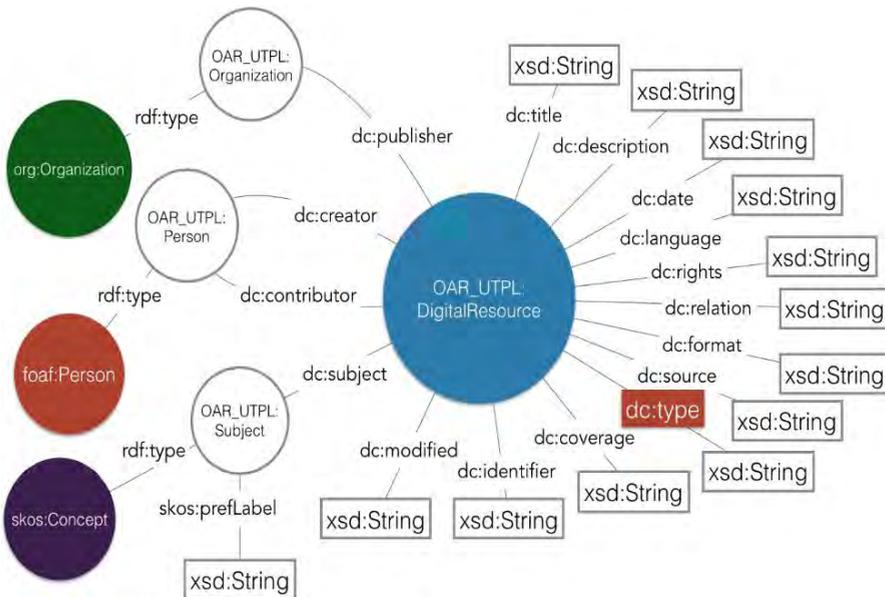


Fig. 4. Extracto de vocabulario UTPL

4 Generación y publicación de los datos enlazados

Para la generación en RDF de los datos cosechados, se desarrolló un generador propio basado en Jena. Un paso importante en el proceso de generación de datos RDF es asignar URIs a los textos extraídos; en este sentido, los metadatos de los recursos bibliográficos se mapearon con los URIs de los términos más apropiados, esto permite que los recursos puedan interoperarse e integrarse con otros conjuntos de datos.

A continuación se muestra un ejemplo del mapeo de texto correspondiente a tipo de material bibliográfico con su correspondiente URI.

Tabla 5. Extracto de mapeo de tipos de material bibliográfico

| Texto extraído | URI equivalente |
|------------------|---|
| Article | http://purl.org/ontology/bibo/AcademicArticle |
| Article | http://schema.org/Article |
| bachelorThesis | http://purl.org/ontology/bibo/Thesis |
| Book | http://purl.org/ontology/bibo/Book |
| Book | http://schema.org/Book |
| Learning Object | http://schema.org/learningResourceType |
| masterThesis | http://purl.org/ontology/bibo/Thesis |
| Other | http://purl.org/ontology/bibo/Document |
| Presentation | http://purl.org/ontology/bibo/Slideshow |
| Presentation | http://vivoweb.org/ontology/core#Presentation |
| Technical Report | http://purl.org/net/nknouf/ns/bibtex#Techreport |
| Technical Report | http://purl.org/ontology/bibo/Report |
| Thesis | http://purl.org/ontology/bibo/Thesis |
| Thesis | http://purl.org/ontology/bibo/ThesisDegree |
| Video | http://purl.org/ontology/bibo/AudioVisualDocument |
| Video | http://schema.org/VideoObject |
| Video | http://vivoweb.org/ontology/core#Video |

Una vez generados los datos, y con el objetivo de que las personas puedan acceder a esta información, se pueden ofrecer algunas interfaces o aplicaciones web que se conecten al repositorio o RDF store y presenten los datos recuperados en un formato entendible para las personas, como puede ser una tabla HTML o una gráfica. Una de las herramientas más populares para mostrar datos RDF en formato tabular es Pubby²¹, aplicación java que es instalada por algunos triple stores, como DBpedia.

Luego de realizar la generación de datos enlazados se procede a almacenarlos en un triplestore. A continuación se describen algunas consultas en SPARQL que pueden ser probadas usando los siguientes datos:

- SPARQL endpoint: <http://data.utpl.edu.ec/serendipity/oar/sparql>
- Grafo: <http://data.utpl.edu.ec/serendipity/oar>

Consulta 01: Material bibliográfico relacionado a temáticas de accesibilidad

Para obtener los recursos académicos relacionados a alguna temática de accesibilidad, se utilizan diferentes filtros sobre el título, la descripción y los tópicos de cada obra.

```
SELECT DISTINCT ?bibresource ?title
WHERE {
  ?bibresource a <http://purl.org/dc/terms/BibliographicResource> .
  ?bibresource <http://purl.org/dc/terms/title> ?title .
  ?bibresource <http://purl.org/dc/terms/subject> ?subject .
  ?bibresource <http://purl.org/dc/terms/description> ?desc .
  FILTER ( REGEX(?title, "Accesibilidad", "i") || REGEX(?desc, "Accesibilidad", "i") || REGEX(?subject, "Accesibilidad", "i") )
}
```

²¹ Pubby está disponible en: <http://wifo5-03.informatik.uni-mannheim.de/pubby/>

Tabla 6. Extracto de resultados obtenidos con la consulta SPARQL

| bibresource | title |
|---|---|
| http://data.utpl.edu.ec/serendipity/oar/resource/41219a41e42cec29b7ff332ff2dd6cea | "Análisis de la atención a las personas con discapacidad a través de los programas y servicios de accesibilidad y recreación, para grupos de atención prioritaria en la Provincia de Cotopaxi en el año 2012" |
| http://data.utpl.edu.ec/serendipity/oar/resource/0fc11a4209b97e27186c9b70aaf0a12b | "Medios de comunicación y el trabajo en la red con herramientas web 2.0" |

Consulta 02: Tipos de entidades y recursos publicados en el dataset

En Web Semántica, cada entidad es asociada a uno o varios tipos de cosas, como: Persona, Recurso Digital, Producto, Localización, etc. En este caso el objetivo es determinar la cantidad de entidades semánticas que están publicadas por cada categoría. Como se puede ver en los resultados, los conceptos SKOS y las personas que han creado un recurso son los tipos de entidades más frecuentes.

```
SELECT DISTINCT ?category, count(*) AS ?num_category
WHERE {
  [] a ?category . }
```

Consulta 03: Cantidad de Recursos Bibliográficos

En este caso se trata de refinar los resultados anteriores, y filtrar los recursos que sean de tipo BibliographicResource.

```
SELECT count(DISTINCT ?s) AS ?bibresources
WHERE {
  ?s a <http://purl.org/dc/terms/BibliographicResource> .
  ?s ?p ?o . }
```

Consulta 04: Top 20 de los temas más referidos

Cada recurso digital es asociado a uno o varios tópicos (la relación es establecida a través del predicado: <http://purl.org/dc/terms/subject>). En este caso, se trata de responder la pregunta de cuáles son los 20 temas más populares entre los recursos digitales; a través de la función *count*, es posible llegar a la respuesta.

```
SELECT DISTINCT ?subject count(*) AS ?freqSubject
WHERE {
  ?bibresource a <http://purl.org/dc/terms/BibliographicResource>.
  ?bibresource <http://purl.org/dc/terms/subject> ?subject .
} ORDER BY DESC(2)
LIMIT 20
```

Consulta 05: Recursos bibliográficos relacionados a un determinado tema

Una de consultas frecuentes que los usuarios pueden realizar en colecciones de recursos, es saber qué recursos existen acerca de un determinado tema. En la consulta que se presenta a continuación, se obtienen ciertos metadatos de cada recurso bibliográfico y se filtra a aquellos que tienen una relación con la categoría. Además se intenta determinar el tipo de recurso que existe relacionado al tema de interés.

Consulta:

```
SELECT DISTINCT *
WHERE {
  ?bibresource a <http://purl.org/dc/terms/BibliographicResource>.
  ?bibresource <http://purl.org/dc/terms/subject>
<http://data.utpl.edu.ec/serendipity/oar/resource/Category:Accessibility>.
  OPTIONAL{
    ?bibresource <http://purl.org/dc/terms/title> ?titulo;
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ?tiporecurso;
```

```
<http://purl.org/dc/terms/creator> ?autor . . }
FILTER(?tiporecurso != <http://purl.org/dc/terms/BibliographicResource> &&
?tiporecurso != <http://purl.org/ontology/bibo/ThesisDegree> )
```

5. Enriquecimiento y reconciliación de datos

Para el enlazado de datos con repositorios externos se usó SILK²² y un componente a medida que permite el enlazado de conjuntos de datos que generan links en base a ciertas condiciones mediante un lenguaje de especificación. El objetivo es crear enlaces RDF hacia recursos en la nube de Linked Open Data; en este caso se realiza un enlazado a DBpedia en inglés, DBpedia en español y DBpedia Latinoamérica.

Tabla 7. Ejemplo de enlazado de Lenguajes

| Texto a enlazar | URI del recurso externo relacionado Fuente: DBPedia & Library of Congress of USA (http://id.loc.gov) |
|-----------------|---|
| English | http://dbpedia.org/resource/English_language |
| Portuguese | http://dbpedia.org/resource/Portuguese_language |
| Spanish | http://dbpedia.org/resource/Spanish_language |
| English | http://id.loc.gov/vocabulary/iso639-1/en |
| Spanish | http://id.loc.gov/vocabulary/iso639-1/es |

Tabla 8. Ejemplo de enlazado de universidades con DBpedia y Dbpedia-LatAm

| Texto a enlazar | URI Resultante |
|---|---|
| Universidad Técnica Particular de Loja | http://dbpedia.org/resource/Universidad_T%C3%A9cnica_Particular_de_Loja |
| Universidade Federal do Rio Grande do Sul | http://dbpedia.org/resource/Universidade_Federal_do_Rio_Grande_do_Sul |
| Universidade de São Paulo | http://es-la.dbpedia.org/resource/Universidade_de_S%C3%A3o_Paulo |
| Universidad Técnica Particular de Loja | http://es-la.dbpedia.org/resource/Universidad_T%C3%A9cnica_Particular_de_Loja |

6 Trabajos Futuros

La Red ESVIAL²³ ha establecido la creación de un servicio para la búsqueda de Recursos Educativos Abiertos relacionados con Accesibilidad, y que serán integrados desde el buscador de contenidos abiertos Serendipity, así como repositorios que cuenten con servicios OAI para acceso a recursos digitales. La búsqueda será a través de navegación por facetas, que consiste en el proceso de explorar el contenido a través del filtrado con refinadores que se asocian a los recursos con categorías. Este tipo de navegación permite especificar refinadores distintos para las páginas de categorías, incluso cuando es la misma página subyacente la que muestra las categorías. La arquitectura macro se puede ver en la Figura 5.

²² SILK – Linke Discovery Framework, una herramienta que permite encontrar relaciones entre entidades dentro de diferentes fuentes de datos, ver: <http://www4.wiwiw.fu-berlin.de/bizer/silk/>

²³ Guía Metodológica y Modelo de Acreditación ESVI-AL <http://www.esvial.org/guia/>

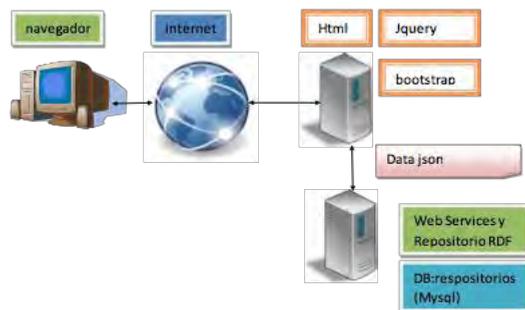


Fig 5. Arquitectura funcional del buscador de OERs

7 Conclusiones

Las librerías digitales tienen el objetivo de hacer accesibles los contenidos, para conseguirlo, se usan diferentes protocolos y tecnologías, sin embargo, aún existen problemas para poder inter-operar e integrar los datos, debido a la heterogeneidad de las plataformas y tecnologías subyacentes. En este trabajo se presenta un marco tecnológico y un método para la publicación y enlazado de datos bibliográficos digitales, siguiendo las buenas prácticas de publicación de datos enlazados en un espacio global. Con el objetivo de mejorar la capacidad de integración y acceso distribuido a los repositorios digitales de diferentes instituciones, se requiere llevar un proceso para generar y publicar datos RDF extraídos a partir del protocolo OAI-PMH. RDF es el modelo de datos recomendado para la Web Semántica.

En un contexto de acceso abierto, interoperabilidad e integración de contenidos compartidos a través de librerías digitales, los proveedores de recursos digitales abiertos deben adoptar un modelo que mejore la integración de repositorios, de manera que puedan soportar estándares de metadatos formales y abiertos para la descripción de recursos educativos, y cuyo nivel de especificidad, granularidad y complejidad sea realizable. El modelo a adoptar debe respetar y garantizar condiciones de autonomía local en un marco de interoperabilidad tecnológica global. El enfoque de Linked Data potencia la interoperabilidad e integración en un contexto como el descrito: alta heterogeneidad y distribución.

Agradecimientos

Este trabajo ha sido financiado por la Unión Europea en el marco del Proyecto ALFA ESVIAL, Educación Superior Virtual Accesible para América Latina con contrato DCI-ALA/19.09.01/11/21526/279-146/ALFAIII(2011)11, y por el Consorcio Ecuatoriano para Desarrollo de Internet Avanzado CEDIA. Agradecemos el soporte brindado por el Grupo de Repositorios de CEDIA y el Laboratorio de Ciencia de Datos de UTPL.

Referencias Bibliográficas

1. IEEE Standards Board. (1990). *IEEE Standard Glossary of Software Engineering Terminology*. New York.
2. Lenzerini, M. (2002). *1. Data Integration: A Theoretical Perspective, Twenty-first ACM SIG- MOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*.
3. Piedra, N. Tovar, E. López, J. Chicaiza, J. "Consuming and producing linked open data: The case of Opencourseware", Emerald EarlyCite, 2014. DOI: 10.1108/PROG-07-2012-0045
4. Baker, T., Bermés, E., Coyle, k., Dunsire, G., Isaac, A., Murray, P., y otros. (25 de octubre de 2011). *Library Linked Data Incubator Group Final Report*. Obtenido de http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/#Benefits_of_the_Linked_Data_Approach
5. Villazón-Terrazas, B. (2011). *A Method for Reusing and Re-engineering Non-ontological Resources for Building Ontologies*. Obtenido de <http://mayor2.dia.fi.upm.es/oeg-upm/files/pdf/BorisVillazonTerrazas.pdf>
6. Berners-Lee, T (Septiembre 1998). Web Architect and Semantic Web Roadmap. Obtenido de <http://www.w3.org/DesignIssues/RDFnot.-html>
7. The Open Arhives Initiative Protocolo for Metada Harvesting, Implementation GuideLines, (14 de Junio del 2008) Obtenidos de: <http://www.openarchives.org/OAI/openarchivesprotocol.html>.
8. Dublin Core Metadata Element Set, Version 1.1 (14 de Junio del 2014) Obtenido de: <http://dublincore.org/documents/dces>
9. OAI2RDF Library Introduction v2.0, (14 de Agosto del 2008), Obtenido de <http://inkdroid.org/journal/2006/08/24/oai2rdf>
10. Chinwe, V. A., & Majesty, I. E. (2011). Digital library deployment in a university. *Library Hi Tech*, 29(2), 373-386. doi:<http://dx.doi.org/10.1108/07378831111138233>
11. Alipour-Hafezi, M., Horri, A., Ali, S., & Ghaebi, A. (2010). Interoperability models in digital libraries: An overview. *The Electronic Library*, 28(3), 438-452. doi:<http://dx.doi.org/10.1108/02640471011052016>
12. Ian Rowlands and David Bawden, "Digital Libraries: a conceptual framework," *Libri Journal*, 1999, vol. 49, pp. 192-202.