

Linked Open Data en la Biblioteca Digital Semántica Académica

EDER ÁVILA BARRIENTOS
Universidad Nacional Autónoma de México

INTRODUCCIÓN

El objetivo que persigue este documento recae en describir y analizar la aplicación de Linked Open Data (LOD: Datos Abiertos Vinculados) en el entorno de las bibliotecas digitales semánticas.

La Biblioteca Digital Semántica Académica (BDSA) es el resultado del desarrollo de las Tecnologías de la Información y la Comunicación (TIC); su relación con el entorno digital se ha consagrado principalmente con su presencia en la Web y, con ello, se ha propiciado una serie de desarrollos en las actividades bibliotecarias, educativas, sociales, culturales y de investigación.

La BDSA se puede concebir como un sistema accesible a través de la Web semántica, integrado por una serie de componentes y que forma parte o es patrocinado por una institución académica de nivel superior. Por lo tanto, apoya los planes y programas de estudio de dicha institución y, como objetivo principal, busca satisfacer las necesidades de información de la comunidad académica. Para ello, ofrece colecciones y servicios digitales diseñados para estar comprendidos en un abanico muy grande de posibilidades y opciones, lo cual deriva de la amplia variedad de perfiles informativos presentes en la cada vez más demandante comunidad académica.

Este documento presenta un marco general sobre los atributos que se ven inmersos en la aplicación del Linked Open Data, en el entorno de las bibliotecas semánticas académicas.

LA BIBLIOTECA DIGITAL SEMÁNTICA ACADÉMICA (BDSA)

En la Biblioteca Semántica se reflejan las tendencias en el manejo de la información de la Web 3.0. Ryszard señala que la biblioteca digital semántica es la siguiente generación de biblioteca digital, y que muestra las siguientes características:

- Integra información de diferentes tipos de metadatos.
- Ofrece interoperabilidad con otros sistemas (no sólo entre bibliotecas digitales), en un nivel de comunicación entre metadatos.
- Permite una búsqueda más robusta y amigable, gracias a los recursos semánticos.

El enriquecimiento de las tecnologías de la Web semántica se refleja de forma más clara en el área de servicios de información. Morales del Castillo menciona que en la biblioteca digital las tecnologías de la Web semántica se aplican para desarrollar servicios de valor añadido en la definición de una descripción enriquecida, en el establecimiento de ontologías que favorecen la interoperabilidad de los recursos, en el desarrollo de interfaces de usuarios, y en mejoras en la navegación, visualización y presentación de contenidos, entre otras cuestiones. La Web semántica supone una extensión de la Web actual, donde la información está dotada de un significado bien definido, puede ser procesada por las máquinas e interpretada por los humanos, lo que permite una mejor cooperación.

Efectivamente, lo que se busca es conseguir una Web donde sea posible explicitar y explotar la interconexión de los datos que circulan por ella. En este sentido, las tecnologías de la Web se-

mántica posibilitarían mejores búsquedas de información a través de lineamientos, propuestas y protocolos que fomenten la relación y el intercambio de información.

En la Web semántica han surgido propuestas para la integración de datos enlazados. *Open Data* (datos abiertos) es un término que se refiere a la manera de publicar datos de organizaciones públicas y privadas, en formatos que sean reutilizables por personas e instituciones.

LOS DATOS EN EL ÁMBITO DE LA BIBLIOTECA DIGITAL ACADÉMICA SEMÁNTICA

La generación de datos en el entorno digital es una realidad en los campos científicos y académicos. Los datos son representaciones numéricas, alfabéticas, algorítmicas, etcétera, que permiten identificar un determinado recurso de información. La riqueza de datos y las herramientas generadas para su búsqueda, análisis, recuperación y visualización, permite a los investigadores académicos interpretar cada vez mayores cantidades de datos. Los conjuntos de datos (datasets) representan a los productos finales de determinadas investigaciones, lo cual complementa notablemente el papel tradicional de las publicaciones académicas. Es decir, dichos datos les dan un valor agregado a estas publicaciones mediante la posibilidad de establecer relaciones estructuradas, significativas y que contribuyan a la colaboración en el entorno académico. Así, si los datos y recursos pueden enlazarse mediante una infraestructura de información, la investigación colaborativa y multidisciplinaria en el contexto digital será posible.

Las BDSA seleccionan, recopilan, organizan y hacen accesibles publicaciones (recursos) de diversa índole temática. Estas actividades generan una serie de datos o conjuntos de datos que son representados mediante esquemas, formatos, modelos y normas que describen los atributos físicos e intelectuales de dichas publicaciones, dando como resultado una serie de registros que son ac-

cesibles a la comunidad usuaria mediante el uso de herramientas para el acceso y control de la información documental, como es el caso de catálogos en línea, repositorios, bases de datos y descubridores de información. Una infraestructura normalizada de información, como es el caso de la que presentan las BDSA, puede mantener vínculos entre datos, publicaciones y otras formas de comunicación científica mediante la generación de mappings.

Los registros semánticos (mappings) son representaciones codificadas de información documental que posibilitan la descripción y vinculación de datos mediante puntos de acceso claramente identificables, como las autoridades y las temáticas que identifican a una determinada publicación. Cuando estos puntos de acceso son vinculados, entonces adquieren un significado que puede favorecer la búsqueda y recuperación de información en el entorno de la investigación académica. El papel de este tipo de datos en la comunicación científica es esencial para la publicación de trabajos académicos, los cuales pueden estar disponibles a través de la biblioteca digital semántica académica.

Al momento de concebir la presencia de datos enlazados en este tipo de bibliotecas, es preciso no perder de vista las recomendaciones que se dan para el acceso, preservación y conservación de dichos datos. Actualmente, los temas de migración de datos, interoperabilidad de formatos y normas para su publicación, han tomado gran importancia debido a la búsqueda de compatibilidad que se desea alcanzar mediante medidas de preservación y conservación de los recursos digitales. Estas medidas están encaminadas a propiciar el acceso permanente a los recursos y sus respectivos datos en el entorno digital.

“La búsqueda de publicaciones académicas y sus respectivos datos puede requerir navegar a través de muchas bibliotecas digitales, motores de búsqueda, catálogos e índices”. Los investigadores dentro de un campo específico del conocimiento manejan múltiples herramientas para buscar, recuperar y acceder a la información que necesitan. En ocasiones, la localización del recurso deseado por el investigador resulta ágil y exitosa debido, primordialmente, a la normalización metodológica utilizada

en la representación de dichos recursos. En el caso de los datos de publicaciones académicas, su localización en ocasiones se ve obstaculizada por la ausencia de una metodología que permita su pronta y satisfactoria identificación. Actualmente, en el entorno de la Web semántica, han surgido propuestas que pretenden conformar un medio para el acceso significativo a los recursos de información, tal es el caso de la propuesta de los datos enlazados.

PRINCIPIOS BÁSICOS Y COMPONENTES DE LINKED OPEN DATA (LOD)

Berners Lee definió cuatro principios básicos para la publicación de Linked data:

- Usar URI (Uniform Resource Identifiers: identificadores uniformes de recursos) para identificar los recursos de forma unívoca.
- Usar URI http para que la gente pueda acceder a la información del recurso;
- Ofrecer información sobre los recursos usando RDF (Resource Description Frame Work: Marco de Descripción de Recursos).
- Incluir enlaces a otros URI para facilitar el vínculo entre distintos datos distribuidos en la Web.

De esta manera, es viable publicar datos que cumplan sólo los tres primeros principios, pero el hecho de no aplicar el cuarto los convierte en menos visibles y, como consecuencia, menos reutilizables. Linked Open Data forma parte esencial de la concepción de la Web semántica y, por lo tanto, posee una serie de componentes que permiten la construcción de este espacio digital. A continuación presentamos estos elementos:

- XML (Extensible Markup Language: Lenguaje Extensible de Mercado). Sintaxis superficial para documentos estructurados.
- XML Schema. Lenguaje para definir la estructura de los documentos XML.
- RDF. Modelo de datos para los recursos y las relaciones que se puedan establecer entre ellos.
- RDF Schema. Vocabulario para describir las propiedades y las clases de los recursos RDF con una semántica para establecer jerarquías de generalización entre dichas propiedades y clases.
- SPARQL. Lenguaje de consulta sobre RDF que permite hacer búsquedas sobre los recursos de la Web semántica utilizando distintas fuentes de datos.
- OWL. Añade más vocabulario para describir propiedades y clases, como: relaciones entre clases, cardinalidad, igualdad, tipologías de propiedades más complejas, caracterización de propiedades o clases enumeradas.

LAS CINCO ESTRELLAS DE LINKED OPEN DATA

En 2010, Berners Lee redefinió su concepción de LD añadiendo la filosofía de la apertura, de tal forma que los datos “abiertos” enlazados (LOD) se liberan bajo una licencia abierta, no impidiendo de esta manera su libre reutilización. Utilizando la metáfora de las estrellas en los hoteles, estableció cinco niveles para medir la calidad de LOD. Las cinco estrellas de Linked Open Data son una guía para detectar el grado de reutilización de los datos abiertos enlazados:

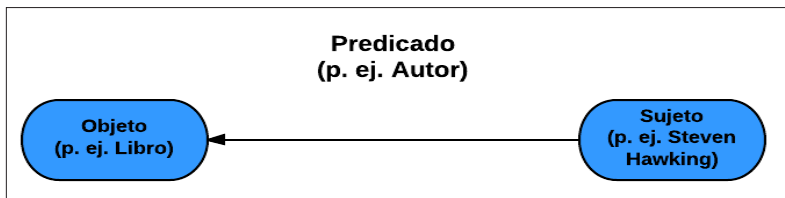
- *Una estrella:* ofrece los datos en cualquier formato, aunque sean difíciles de manipular no desagregar, como un pdf por ejemplo, o una imagen escaneada.
- *Dos estrellas:* entregar los datos de manera estructurada, como en un archivo excel con extensión xls.

- *Tres estrellas*: entregar los datos en un formato que no sea propietario, como csv, en vez de excel, xml, rdf, etcétera.
- *Cuatro estrellas*: usar URI (que es una dirección web de un dato que sirve para enlazarlo con otros datos) para identificar cosas y propiedades; de manera que se pueda apuntar a los datos. Requiere utilizar el estándar RDF.
- *Cinco estrellas*: vincular sus datos con los de otras personas, dotándolos de contexto. En la práctica, permite que la información entregada también apunte a otras fuentes de datos. Por ejemplo, si coloco un país dentro de mis datos, puedo agregar una URI de referencia que apunte a los datos de ese país en Dbpedia o a Geonames.

TRIPLETES RDF

Figura 1

Representación gráfica de un triplete



La construcción de tripletes con RDF se basa en la idea de declarar recursos usando la expresión en la forma sujeto-predicado-objeto. Esta expresión se conoce en la terminología RDF como triple o triplete. Un triplete RDF contiene tres componentes, todos con referencia en un URI (*Figura 1*).

- *Sujeto*. Una referencia URI, una persona, o un nodo, es el ente al cual nos referimos.
- *Predicado*. Es la propiedad o relación que se desea establecer acerca del sujeto.

- *Objeto*. Es el valor de la propiedad o del otro recurso con el que se establece la relación.

El uso de LOD hace posible publicar información migrada desde diferentes fuentes en las que se encuentra almacenada, como por ejemplo bases de datos, repositorios, hojas de cálculo, catálogos en línea, siempre y cuando se utilice el lenguaje RDF para permitir enlazar esta información con otros datos, de forma que aumente la conectividad entre dicha información y se facilite así la búsqueda, la recuperación y la agregación de información relacionada.

Esta construcción de tripletes se logra mediante la aplicación e integración de diversos componentes, que en su conjunto forman parte del modelo integral de los datos abiertos enlazados. Al conjunto de tripletes se les identifica como DataSets (o conjunto de datos). Dichos conjuntos integran un contexto en específico, el cual se relaciona con otros contextos mediante la vinculación de sus atributos. La construcción de estas relaciones tienen gran relevancia al momento de realizar una determinada búsqueda sobre un autor, recurso o tema en específico en el entorno digital.

DATOS ENLAZADOS EN LA BIBLIOTECA DIGITAL ACADÉMICA SEMÁNTICA

El término Linked Data (datos enlazados) se refiere al método con el que se pueden mostrar, intercambiar y conectar datos a través de URIs desreferenciables en la Web. Estas dos propuestas han nacido en el marco general de la Web semántica y han propiciado los mecanismos, métodos y directrices a seguir para conformar un modelo de datos enlazados abiertos (Linked Open Data) que propicie una Web con mayor significado y mejores oportunidades en la búsqueda y acceso a la información. “Los dos conceptos clave son Open data y Linked data o Linked open data (LOD), traducible al español como datos abiertos vinculados (o enlazados). Ambos se derivan de las declaraciones del creador

de la web, Tim Berners-Lee (2006), asesor del gobierno británico. Y además el segundo se relaciona con la familia de estándares del W3Consortium (2010) para la construcción de la web semántica”. El informe sobre datos enlazados bibliotecarios del Grupo incubadora del w3c (Library Linked Data IncubatorGroup: Datasets, ValueVocabularies, and MetadataElement Sets), es relevante para entender la oportunidad que se les abre a las bibliotecas para alcanzar el impacto digital que pretenden, además de concretar nuevos modelos en el acceso y la utilización de recursos de información digital. En este sentido, “[...] a escala internacional, el w3c identificó la comunidad bibliotecaria como uno de sus focos de interés y a través de sus Incubatorgroups –grupos dinámicos que se utilizan para detectar nuevos nichos de actuación–, creó en mayo de 2010 uno orientado a la publicación de datos estructurados en Linked data procedentes de bibliotecas”. El Library Linked Data IncubatorGroup tiene como misión analizar la situación de los modelos y esquemas de metadatos y los estándares y protocolos de interoperabilidad que se deberían usar para la publicación y el uso de Linked Open Data (LOD) con datos de bibliotecas. A su vez anima a la comunidad a publicar sus registros bibliográficos y control de autoridades, como Linked Data.

Las BDSA están constituidas por colecciones y contenidos digitales. Estos elementos tienen la facultad de producir datos que pueden enlazarse, y que pueden ser útiles para la el apoyo de la investigación mediante la generación de nuevos servicios en el entorno digital. En *Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services*, documento realizado por la Biblioteca del Congreso de los Estados Unidos puede apreciarse la aplicación de LOD en la construcción de un marco bibliográfico basado en datos enlazados que permite, a su vez, la generación de núcleos recursos digitales de información enlazados.

Esto supone la integración de dichos recursos y LOD en el contexto de la Web semántica mediante la generación de relaciones que permitan a las bibliotecas digitales semánticas académicas enlazar sus recursos mediante los componentes de dicha Web. Los recursos de información digital presentes en esta modalidad

de bibliotecas deberán ser descritos, organizados y enlazados mediante la adopción de modelos en sus catálogos de autoridades, bibliográficos u otras herramientas para el control y acceso a la información en el entorno digital. Estas fuentes de información pueden ser de utilidad en la generación de servicios innovadores de investigación en las bibliotecas. Así, nos referimos a la Web semántica como el contexto en donde la biblioteca digital académica y Linked Open Data pueden integrarse, y por ende, potencializar la generación de nuevos servicios de investigación mediante la innovación abierta.

SERVICIOS DE INNOVACIÓN ABIERTA EN LA BDSA

El término *innovación abierta* hace referencia a la apertura del proceso de investigación de una empresa a terceros. Supone delegar en agentes externos a la propia organización parte de las tareas de desarrollo de innovaciones, y compartir con ellos nuestros datos y nuestra información. En este sentido, la BDSA, mediante la incorporación de Linked Open Data en su estructura de información, propicia la apertura de procesos de investigación a sus usuarios. Este proceso resulta innovador desde el momento en que dicho usuario puede visualizar las relaciones temáticas y de autoridades que provoca la vinculación de los datos en la BDSA. La visualización de estas relaciones propiciará la conformación de redes que podrían facilitar al investigador su búsqueda y recuperación de información mediante la aplicación de la BDSA en sus procesos.

La incorporación de los servicios de la Web semántica en los de la BDSA puede ser una latente alternativa para alcanzar tal cometido. Los servicios de la Web semántica son una aplicación de la semántica y las ontologías. Utilizan la estructura XML para describir y contener aplicaciones que pueden hacerse accesibles a través de Internet.

Mediante el uso de estas tecnologías en la BDSA, sería viable la posibilidad de implementar nuevos servicios de búsqueda y recuperación de información con una arquitectura de información

basada en datos enlazados. Actualmente, pueden observarse algunos esfuerzos por parte de las bibliotecas nacionales por incorporar nuevos servicios mediante la implementación de los datos enlazados en sus arquitecturas.

La propuesta *datos.bne.es* es un proyecto conjunto del OntologyEngineeringGroup (OEG) y la Biblioteca Nacional de España, destinado al enriquecimiento de la Web Semántica con datos bibliográficos provenientes de su catálogo. Esta iniciativa se ha puesto en marcha con la publicación, conforme a los principios de Linked Data, de información procedente de los catálogos bibliográficos y de autoridad, haciéndolos disponibles como bases de conocimiento RDF (Resource Description Framework). Además, estos datos se interrelacionan con otras bases de conocimiento existentes en la iniciativa Linking Open Data.

De esta manera, el proyecto de datos enlazados de la Biblioteca Nacional de España se suma a los proyectos que otras instituciones, como la Biblioteca Británica y la Biblioteca Nacional de Alemania, que han comenzado recientemente. Estos proyectos nos permiten apreciar la innovación presente en la implementación de datos enlazados en las arquitecturas de grandes bibliotecas. En el caso de la BDSA, será necesario estudiar, analizar y comprender los atributos y variables que estarán inmersos en la aplicación de los datos enlazados en su arquitectura.

TENDENCIAS ACTUALES DEL USO DE DATOS ENLAZADOS COMO SERVICIO EN LAS BIBLIOTECAS DIGITALES

La aplicación de datos enlazados en los servicios de las bibliotecas digitales ha empezado a marcar una determinada tendencia. Se observa que estos servicios se caracterizan por la implementación de nuevas plataformas tecnológicas que son accesibles a través del portal Web de las bibliotecas. Estas plataformas permiten llevar a cabo búsquedas y recuperación de información en el entorno de datos enlazados que dichas bibliotecas han creado

En el modelo de datos para monografías de la Biblioteca Británica, podemos apreciar el ejemplo de un recurso de información (RESOURCE BL URI) y las múltiples relaciones que pueden establecerse a través de sus atributos. Cada una de las relaciones permite descubrir nueva información a través de los atributos de un solo recurso. A su vez, puede apreciarse la categorización de estas relaciones mediante núcleos de tema, título, autor, datos de publicación y series. Cada uno de estos núcleos responde a los atributos de la monografía ejemplificada en el modelo.

SERVICIO DE DATOS ENLAZADOS DE LA BIBLIOTECA NACIONAL DE ALEMANIA

A largo plazo, la Biblioteca Nacional de Alemania está planeando ofrecer un servicio de datos enlazados que permitirá a la comunidad de la Web semántica utilizar todas las existencias de sus datos bibliográficos nacionales, incluyendo sus datos de autoridad. Uno de los objetivos de este servicio es atraer a nuevos grupos de usuarios a la biblioteca; por lo tanto, el proyecto también contempla la identificación de las necesidades de información de estos usuarios. En la actualidad, se puede buscar en el portal de la Biblioteca Nacional de Alemania y se obtienen los datos enlazados a través de los vínculos de representación RDF / XML.

PRESENCIA DE LINKED OPEN DATA EN EL CONTEXTO MEXICANO

En el contexto mexicano, la presencia de Linked Open Data se manifiesta a través de una serie de proyectos y propuestas, sobre todo de carácter gubernamental. De esta forma, se propone que los datos estén disponibles de forma libre para el ciudadano. Linked Open Data se ha convertido en el mecanismo ideal para promover la Reutilización de Información del Sector Público

(RISP), ya que este sector produce una gran variedad de información (económica, social, geográfica, meteorológica, turística, educativa) que puede ser reutilizada tanto por ciudadanos como por la iniciativa privada. Con la finalidad de mostrar los avances de Linked Open Data en nuestro país, se presentan ejemplos de los dataset (conjunto de datos) creados con información de las dependencias del país. Los dataset pueden apreciarse en el sitio Web de Datos Abiertos del Gobierno Mexicano.¹

En el ámbito de desarrollo informático, en México se ha comenzado a desarrollar software de utilidad semántica, como es el caso de *Semantic Web Builder*, una suite de productos que sirven como plataforma para desarrollar aplicaciones y portales semánticos, creada por el Fondo de Información y Documentación para la Industria, INFOTEC, Centro de Investigación y Desarrollo Tecnológico del CONACYT.

La suite Semantic Web Builder está conformada por una serie de herramientas que permiten la evolución de los sitios Web convencionales (sin significado), hacia los portales que cumplan con la visión de la Web Semántica (con significado), lo que les permite exponer en formatos estándar como RDF todo el conocimiento generado, además de contar con un Modelo (Ontología) que permite el intercambio de esta información entre los diferentes sistemas.²

Como puede apreciarse, Linked Open Data tiene una considerable presencia en los ámbitos gubernamentales e informáticos de nuestro país. Sin embargo, carece de presencia en el entorno de las bibliotecas digitales.

1 Disponible en: <http://datosabiertos.gob.mx/>

2 Semantic Web Builder: Quiénes somos. Disponible en: http://www.semanticwebbuilder.org.mx/en/swb/SWB_Platform

CONSIDERACIONES FINALES

En este estudio se ha presentado el marco teórico que sustenta la presencia de la BDSA. Se ha abordado la aplicación de los datos enlazados abiertos en los servicios de dicha biblioteca. Sin embargo, es necesario profundizar en el estudio de los elementos de la tecnología semántica que permiten la aplicación de estos datos en los servicios de las bibliotecas.

La innovación abierta se presenta a través de nuevas alternativas para la publicación, búsqueda y recuperación de información; los repositorios institucionales y las publicaciones académicas de acceso abierto son ejemplos de su aplicación en las bibliotecas.

Como hemos apreciado en este documento, los datos enlazados abiertos y su adaptación en las bibliotecas permiten la generación de servicios innovadores, basados en búsquedas y recuperación de información que tienen un mayor significado y trascendencia para el usuario.

Las relaciones entre recursos de información permitirán visualizar las redes de conocimiento construidas por diversas comunidades científicas y académicas. La visualización de redes de conocimiento a través de las plataformas de datos enlazados abiertos puede ser un apoyo para la generación de futuras investigaciones.

REFERENCIAS BIBLIOGRÁFICAS

Berners Lee, Tim, Hendler, James and Lassila, Ora, "The semantic web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities", en: *The scientific American*, Vol. 284, Núm. 5, 2001, p. 35 (35-43).

Biblioteca Nacional de España, Datos enlazados en la BNE, disponible en: <http://www.bne.es/es/Inicio/Perfiles/Bibliotecarios/DatosEnlazados/index.html>.

Biblioteca Digital Académica en Bibliotecología...

Borgman, Christine L., *Scholarship in the digital age: information, infrastructure and the Internet*, United States of America: MIT Press, 2007, p. 138.

Chesbrough, Henry, *Open Business Models: How to Thrive in the New Innovation Landscape*, Boston, Massachusetts: Harvard Business School, 2006, p. 75.

Coyle, Karen, “Linked data tools: connecting on the web”, en: *Library Technology Reports*, May/june, 2012.

Isaac, Antoine, Waites, William, Young, Jeff *et al.*, *Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets*, w3c IncubatorGroupReport 25 October 2011, disponible en: <http://www.w3.org/2005/Incubator/ld/XGR-ld-vocabdataset-20111025/>

Library of Congress, *Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services*, Library of Congress: Washington, DC, 2012.

Morales del Castillo, José Manuel, *Hacia la biblioteca digital semántica*, Gijón: TREA, 2001, p. 80.

Pesset, Fernanda, Ferrer-Sapena, Antonia y Subirats Coll, Imma, “Open data y linked open data: su impacto en el área de bibliotecas y documentación”, en *El profesional de la información*, Vol. 20, Núm. 2, 2011, p. 169.

Ryszard Kruk, Sebastian, “Goals of semantic digital libraries”, en: Sebastian Ryszard Kruk, Bill McDaniel (eds.), *Semantic digital libraries*, Berlin: Springer, 2009, p. 72.

Semantic Web Builder, Quiénes somos, disponible en: http://www.semanticwebbuilder.org.mx/en/swb/SWB_Platform