ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 33 (diciembre 2015)

Aplicación de *Linked Open Data* para la realización de un modelo conceptual que permita diseñar un mapa de las investigaciones académicas y científicas de la Argentina

Elsa E. Barber, Silvia Pisano, Sandra Romagnoli, Verónica Parsiale, Gabriela de Pedro, Carolina Gregui, Nancy Blanco, María Rosa Mostaccio

Instituto de Investigaciones Bibliotecológicas. Facultad de Filosofía y Letras. Universidad de Buenos Aires / elsabarber@grebyd.com.ar

Resumen

El proyecto UBACyT 773BA tiene por objetivo contribuir con la elaboración de un modelo conceptual para la creación de un mapa de las investigaciones académicas y científicas en Argentina basado en las nuevas tendencias tecnológicas de Datos Abiertos Enlazados *(Linked Open Data,* LOD). Sostiene como hipótesis que esta tecnología facilita el diseño de un mapa sobre dichas investigaciones que pueda ser compartido y reutilizado a nivel nacional e internacional. Para alcanzar el objetivo planteado prevé aplicar la metodología propuesta por Heath y Bizer (2011).

Palabras clave

Datos abiertos enlazados
Ontologías
Investigaciones científicas
Argentina

Abstract

Application of Linked Open Data to create a conceptual model to design a map of academic and scientific research in Argentina. Project UBACyT 773BA aims to contribute to a conceptual model for the creation of a map of academic and scientific research in Argentina based on the new technological trends of Linked Open Data (LOD). The proposed hypothesis considers that this technology facilitates the design of a map of these investigations that can be shared and reused at a national and international level. The methodology proposed by Heath and Bizer (2011) will be implemented to achieve the stated objective.

Keywords

Linked Open Data
Ontologies
Scientific researches
Argentina

Artículo recibido: 09-09-2015. Aceptado: 06-11-2015

1. Antecedentes del tema

Según Berners-Lee y Fischetti (2000) estamos inmersos en la web semántica, la cual tiende hacia un futuro de información en la internet global y organizada. Aunque la web hizo posible vincular y conectar los documentos, en lo conceptual se concentró en el documento con enlaces no semánticos sino hipertextuales. Al implementarse los formatos técnicos tales como HTML, la consecuencia fueron páginas web optimizadas para el procesamiento humano en lugar del procesamiento por máquina. Si bien la Web 2.0 fue eficiente para compartir documentos y crear medios para la colaboración, se necesita la intervención humana para que se comprenda la semántica de un documento. En pocas palabras, las máquinas por sí solas no pueden dar sentido a los documentos (Alemu et al., 2012)

La Web 3.0 o web semántica, que desarrolla su conjunto de actividades mediante la *World Wide Web Consortium* (W3C), busca vincular metadatos semánticos y ontológicos que describen el contenido, el significado y sus relaciones. Su idea es permitir enlazar información que ha sido tratada por separado e integrarla en la Web actual con la posibilidad de añadir datos relacionados entre sí, ya sea semánticamente o por atributos que los determinan (W3C Incubator Group, 2011).

Por otra parte, con relación al movimiento de acceso abierto, Katsirikou (2011) sostiene que no solamente brinda apoyo a las obras de estudiantes e investigadores sino que además, y principalmente, sustenta el desarrollo de los ciudadanos, las industrias y los profesionales. En este sentido, la Web se constituye en una plataforma abierta como medio para comunicar, administrar y compartir información en el ámbito de la academia, de la comunidad y de la sociedad en general (Miller, 2005). Así hoy en día la discusión se centra en nuevas nociones como Datos Abiertos (Open Data), Datos Enlazados (Linked Data) y Datos Abiertos Enlazados (Linked Open Data). La Conferencia Regional de Datos Abiertos (2013) considera que estos presuponen la "... publicación y difusión de información en la Internet, sin limitaciones de acceso ni de uso, compartida en formato electrónico y abierto. El formato abierto permite la combinación de conjuntos de datos de diferentes orígenes, su reutilización y difusión, libremente y de forma automatizada."

Alemu et al. (2012) definen *Linked Data* (LD) como un meta-modelo de datos que identifica, describe, enlaza, vincula y relaciona la estructura de elementos de datos en la web. Este modelo plantea un marco para definir, diseñar, desarrollar y mantener tanto esquemas como vocabularios de cualquier clase y tamaño en un dominio determinado. La idea principal de LD se sustenta en una red distribuida a nivel de datos antes que a nivel de presentación de documentos. La adopción de LD ofrece una apertura a sistemas interactivos con vínculos externos y se basa en cómo estos se pueden comunicar, administrar y compartir con el fin de descubrir, usar y reutilizar los recursos de las instituciones culturales (Miller, 2005; Peset, Ferrer-Sapena y Subirats-Coll, 2011; Heath y Bizer, 2011; Ríos-Hilario, Martin-Campo y Ferreras-Fernández, 2012).

Berners-Lee y Fischetti (2000, citados por Guerrini y Possemato, 2013) identifican cuatro reglas para la creación de *Linked Data* en la web:

1. Utilizar los URIs para identificar objetos, esto significa que cada recurso en la web (un sitio, una página dentro de un sitio, un documento, un objeto) debe ser identificado por un URI para ser encontrado, utilizado, enlazado o vinculado por otros sistemas.
2. Utilizar HTTP URI para que los objetos puedan ser consultados por personas y por agentes de software.

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)

3. Adoptar los estándares *RDF* (modelo que permite describir todo dato), *RDF Schema* (creación de vocabularios y de conjuntos de términos descriptivos), *OWL* (*Web Ontology Language*, Lenguaje de Ontologías Web), y *SPARQL* (lenguaje estandarizado de consulta para las bases de RDF).
4. Incluir enlaces a otros URIs, para que puedan descubrir y enlazar otros datos relacionados con estos.

Por su parte, *Linked Open Data* se refiere a datos publicados y enlazados mediante la estructura de la web semántica. Posee la ventaja de conectar los datos independientemente de dónde estos residan. Puede vincular o enlazar datos utilizando como referencia global el Identificador Uniforme de Recurso, denominado *Uniform Resource Identifier* (URI) (Allemnag y Hendler, 2008).

Otra ventaja importante de reutilizar URIs es que estos permiten a los proveedores de datos contribuir con partes de sus datos como declaraciones. En la web actual, basada en los documentos, la información se intercambia siempre en forma de registros completos. En cambio, en un sistema basado en el modelo de datos (grafo) RDF una organización puede proporcionar declaraciones individuales acerca de un recurso y todas las declaraciones proporcionadas acerca de un recurso identificado unívocamente pueden ser reunidas en un gráfico global. Al utilizar URIs para designar recursos como obras, lugares, personas, actividades, temas, y otros objetos o conceptos de interés, las bibliotecas permitirán que sus recursos de información sean citados a través de una amplia variedad de fuentes y en consecuencia, hacer sus metadatos descriptivos más accesibles y visibles.

En su informe final, el grupo de trabajo *W3C Library Linked Data Incubator Group* (2011), caracterizó el estado actual de la gestión de datos bibliotecarios. Delineó los beneficios potenciales de publicar los metadatos y la información bibliográfica como *Linked Data*, tanto para la comunidad académica y usuarios en general, como para las organizaciones, bibliotecarios, archivistas, desarrolladores y proveedores. Además recomendó, entre otras cuestiones, que los líderes bibliotecarios identificaran conjuntos de datos como posibles candidatos para una primera aproximación a *Linked Data*; que bibliotecarios y archivistas preservaran los conjuntos de elementos de *Linked Data* y sus vocabularios y aplicaran la experiencia bibliotecaria en la curación y preservación a largo plazo de los conjuntos de datos enlazados.

2. La investigación

Desde el año 1995 este grupo de investigación ha indagado sobre el acceso a la información a través de distintos proyectos. En una primera instancia, ha estudiado los niveles de automatización de las bibliotecas argentinas (universitarias, públicas, entre otras). Posteriormente, extendió su análisis a los catálogos en línea de acceso público (*Online Public Access Catalogs*, OPACs, según su sigla en inglés) del Mercosur y en una segunda etapa, de los países de Latinoamérica.

En el proyecto actual el interés se centra en las nuevas corrientes de acceso a la información, basadas en la web semántica. En los últimos años los movimientos vinculados con el acceso abierto surgen como objeto de estudio en el ámbito bibliotecario y se los relaciona principalmente con el libre acceso a la información y al conocimiento a través de internet, sin barreras económicas ni restricciones derivadas de los derechos de copyright (Ferreras-Fernández, 2011). En este contexto, Guerrini y Possemato (2013) mencionan que los datos que se producen en los catálogos de las bibliotecas no se encuentran integrados en la Web. Surge entonces la pregunta: ¿cómo modificar los catálogos y la estructura de los datos para posibilitar su interoperabilidad en

la web? Esta es la filosofía subyacente bajo la tecnología de *Linked Data* que ofrece un punto de partida para lograr que los datos de los catálogos sean reutilizados y compartidos en la web.

En este marco, para contribuir con un modelo conceptual que permita representar y describir la producción científico-técnica del país, la investigación plantea los siguientes objetivos:

1. Contribuir con un modelo conceptual para la construcción de un mapa de las investigaciones académicas y científicas en la Argentina basado en las nuevas tendencias tecnológicas de *Linked Open Data*.
2. Explorar los sitios web de organismos gubernamentales y repositorios institucionales académicos de las universidades nacionales del país, para obtener el conjunto de datos a utilizar e interrelacionar.
3. Aplicar los principios de *Linked Data* a datos abiertos utilizando los estándares OWL, *RDF* y *SPARQL*, entre otros, para la construcción del modelo conceptual.
4. Verificar que los datos cumplan con los requerimientos exigidos para la aplicación de *Linked Open Data* y que proporcionen la meta-información pertinente para su integración con otras fuentes.

Dado que es fundamental el acceso a las investigaciones que se realizan en los diferentes ámbitos de creación y acumulación del conocimiento para el avance de la investigación y el desarrollo (I+D) en el país, así como para garantizar su visibilidad en el contexto regional e internacional, y dado que los datos relacionados mediante LOD pueden proporcionar la forma de vincular a las descripciones y publicaciones, tanto como a sus autores y contribuyentes a través de vocabularios, esta investigación considera relevante plantear como hipótesis que la aplicación de *Linked Open Data* posibilita la realización de un modelo conceptual para la construcción de un mapa de las investigaciones académicas y científicas en Argentina que pueda ser compartido y reutilizado a nivel nacional e internacional.

En función de las características del tema, de los objetivos planteados y de la hipótesis expuesta se adoptará la metodología propuesta por Heath y Bizer (2011). De acuerdo con ella, se llevarán a cabo una serie de procedimientos y técnicas a cumplimentar en diferentes etapas:

a) Determinación de la población a examinar: se delimitará la población conformada por los repositorios académicos y científicos nacionales.
b) Recolección y selección de datos: se definirán las palabras clave a partir de las cuales se recolectarán los datos de los repositorios académicos y científicos con el protocolo OAI-PMH / OAI-ORE. Se revisará el ranking generado por el Research Webometrics Info. Se generará una muestra representativa del total de los repositorios cosechados.
c) Validación para asegurar el control de calidad de los datos: se analizará la estructura del conjunto de datos cosechados y reutilizables. Se aplicarán procedimientos y algoritmos para detectar errores (duplicados, errores de tipeo, entradas dobles, etc.).
d) Conversión de los datos: los datos obtenidos y depurados se prepararán para su posterior análisis, mediante herramientas de conversión a *RDF*. Se verificará que los datos cumplan con los estándares de la web semántica.
e) Almacenamiento centralizado de datos: una vez convertidos, los datos se almacenarán en un servidor *RDF*.
f) Construcción de un prototipo: se generará un prototipo que proporcione un patrón para publicar *Linked Data*. Para ello se contextualizarán los datos,

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)

se definirá el modelo mediante la especificación de los tipos de normas, las relaciones, los metadatos de acuerdo con SKOS, LOD, OAI-ORE, RDF, RDF/XML y OWL.

g) Aplicación del prototipo y visualización de los datos.
h) Análisis de los resultados: a partir de la aplicación del prototipo se revisará y ajustará el modelo.

3. Contribuciones esperadas

El contenido semántico puede representarse a través de conceptos de vocabularios skosificados. Con vocabularios abiertos y enlazados (*Linked Open Vocabularies*, LOV) se despliega un enorme potencial para aprovecharlos en el ámbito global de la Web, vinculando la información en un entorno abierto, ya que ayudan a la adquisición de conocimiento a través de un control estricto y de una contextualización de los datos (Méndez y Greenberg, 2012). Las ventajas que ofrecen han favorecido la aparición de proyectos de implementación basados en este modelo (Doerr et al, 2010; Kramer et al, 2012).

Al respecto, en Argentina, el desarrollo científico y tecnológico ha seguido un proceso signado por numerosas rupturas, estrechamente relacionadas con los vaivenes del contexto político e institucional del país (Albornoz; Estébanez y Luchilo, 2004). No obstante, hoy las tecnologías aplicables a la información disponible en la Web nos permiten vislumbrar la posibilidad de reunir el resultado intelectual de este proceso, más allá de las circunstancias tecnológicas o socio-políticas de su génesis.

La construcción de un mapa de las investigaciones académicas y científicas en la Argentina mediante la aplicación del modelo *Linked Open Data* podrá contribuir a nuevos desarrollos en la reutilización de los datos como un mecanismo de uso, valor y evaluación de la información. Permitirá fomentar la construcción de comunidades científicas abiertas a nivel nacional, regional y global difundiendo la producción intelectual de las universidades argentinas y de los organismos dedicados a la investigación científico-técnica. A futuro ayudará a sentar las bases para un modelo de ciudadanía abierta, cuyo libre acceso a la información y al conocimiento constituirá la plataforma de una sociedad inteligente y sustentable a nivel económico, político y social.

4. Referencias Bibliográficas

» Albornoz, M.; M. E. Estébanez y L. Luchilo. 2004. La investigación en las Universidades nacionales: actores e instituciones. En Barsky, O.; M. Dávila y V. Sigal, comp. *Los desafíos de la Universidad argentina*. Buenos Aires: Universidad de Belgrano: Siglo Veintiuno.

» Alemu, G.; B. Stevens; P. Ross y J. Chandler. 2012. Linked Data for libraries: benefits of a conceptual shift from library-specific record structures to RDF-based data models. En *New Library World*. Vol. 113, no. 11-12, 549-570.

» Allemnag, D. y J. Hendler. 2008. *Semantic web for the working ontologist: effective modelling in RDFS and OWL*. Amsterdam: Morgan Kaufmann.

» Berners-Lee, T. y M. Fischetti. 2000. *Weaving the web: the original design and ultimate destiny of the world wide web by its inventor*. New York: Harper.

» Conferencia Regional de Datos Abiertos para América Latina y el Caribe (2013: Montevideo). ¿Qué son los datos abiertos? <http://confdatosabiertos.uy/inicio/datos-abiertos/que+son+los+datos+abiertos> [Consulta: 7 Septiembre 2015].

» Doerr, M.; S. Gradmann; S. Hennicke; A. Issac; C. Meghini y H. van de Sompel. 2010. El Modelo de Datos de Europeana (EDM). En World Library and Information Congress (76th: 2010: Gothenburg) IFLA General Conference and Assembly. <http://conference.ifla.org/past-wlic/2010/149-doerr-es.pdf> [Consulta: 7 Septiembre 2015].

» Ferreras-Fernández, T. 2011. Open Access en España: los repositorios institucionales. En Jornadas de E-learning en la Formación para el Empleo en las Administraciones Públicas (5a: 2011: Valladolid). <http://eprints.rclis.org/16355/1/E-LIS.pdf> [Consulta: 7 Septiembre 2015].

» Guerrini, M. y T. Possemato. 2013. Linked data: a new alphabet for the semantic web. En *JLIS.it*. Vol. 4, no. 1, 67-90. <http://leo.cilea.it/index.php/jlis/article/view/6305/7891> [Consulta: 7 Septiembre 2015].

» Heath, T y C. Bizer. 2011. *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Florida: Morgan & Claypool.

» Katsirikou, A., ed. 2011. *Open access to STM information: trends, models and strategies for libraries*. Berlin: De Gruyter.

» Kramer, S.; A. Leahey; H. Southall; J. Vompras y J. Wackerow. 2012. *Using RDF to describe and link social science data to related resources on the web*. (DDI Working Paper Series - Semantic Web, no. 1) <http://www.ddialliance.org/system/files/UsingRDFToDescribeAndLinkSocialScienceDataToRelatedResourcesOnTheWeb.pdf> [Consulta: 7 Septiembre 2015].

» Méndez, E. y J. Greenberg. 2012. Linked data for open vocabularies and HIVE's global framework. En *El profesional de la información*. Vol. 21, no. 3, 236-244. <http://www.elprofesionaldelainformacion.com/contenidos/2012/mayo/03_eng.pdf> [Consulta: 7 Septiembre 2015].

» Miller, P. 2005. Web 2.0: building the new library. En *Ariadne*. Vol. 45. <http://www.ariadne.ac.uk/print/issue45/miller> [Consulta: 7 Septiembre 2015].

» Peset, F.; A. Ferrer-Sapena y I. Subirats-Coll. 2011. Open data y Linked open data: su impacto en el área de bibliotecas y documentación. En *El profesional de la*

información. Vol. 20, no. 2, 165-173. <http://www.elprofesionaldelainformacion.com/contenidos/2011/marzo/06.pdf> [Consulta: 7 Septiembre 2015].

» Ríos-Hilario, A.; D. Martín-Campo y T. Ferreras-Fernández. 2012. Linked data y linked open data: su implantación en una biblioteca digital. El caso de Europeana. En *El profesional de la información*. Vol. 21, no. 3, 292-297. <http://gredos.usal.es/jspui/bitstream/10366/115842/1/DBD_Rios_Martin_Ferreras_LinkedOpenData.pdf> [Consulta: 7 Septiembre 2015].

» W3C Incubator Group. *Library Linked Data Incubator Group final report: W3C Incubator Group report 25 October 2011*. < http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/> [Consulta: 7 Septiembre 2015].