Repositorio Semántico para la Universidad Nacional de Chilecito

Jose Texier – *Universidad Nacional de Chilecito*. Marisa De Giusti, Gonzalo Villarreal, Ariel Lira – *Universidad Nacional de La Plata (SEDICI)*. Jusmeidy Zambrano – *Universidad Nacional de Chilecito*

Abstract—El Ministerio de Ciencia, Tecnología e Innovación Productiva y el Consejo Interinstitucional de Ciencia y Tecnología implementaron el Programa de Grandes Instrumentos y Bases de Datos. Estos programas contienen recursos digitales junto con los metadatos estandarizados del Sistema Nacional de Repositorios Digitales, el Sistema Nacional de Datos Biológicos y el Sistema Nacional de Datos Genómicos. Para la interoperabilidad semántica entre esos repositorios heterogéneos, se pueden aplicar enfoques basados en tecnologías Web (por la variedad de tipos de datos) con el protocolo de interoperabilidad OAI-PMH para adaptarlas a la filosofía de la Web Semántica a partir del Linked Open Data. Este trabajo propone un repositorio semántico para la Universidad Nacional de Chilecito acorde con la ontología que unifique los recursos en los Sistemas Nacionales, con base en la metodología de desarrollo de ontologías propuesta por Heath y Bizer [11]. Los investigadores y comunidad en general de la Universidad contarán con un sistema que será de gran aporte para la sistematización de sus estudios y la búsqueda de información.

Index Terms—repositorio semántico, UNdeC, Linked Open Data, ontologías.

I. INTRODUCCIÓN

En la actualidad se convive en un entorno abierto con gran cantidad de información compartida proveniente de diversas fuentes de información confiables: repositorios institucionales, repositorios de datos, revistas científicas, entre otros. Este amplio (y creciente) conjunto de recursos digitales se caracteriza por la falta de un modelo de organización y clasificación unificado, originado por la ausencia de un esquema de metadatos estandarizado para describir cada recurso, lo que finalmente impide o desfavorece el aprovechamiento de toda esa información, tanto por humanos como por computadoras. Un sistema que permita establecer directrices de interoperabilidad a nivel semántico entre estas fuentes de información sería de gran ayuda, porque conectará recursos muy diferentes, realizará búsquedas, clasificará la información e incluso podrá inferir o derivar nuevos datos/información. Una interoperabilidad semántica entre diversas fuentes o repositorios heterogéneos, se puede garantizar mediante modelos basados en la Web Semántica (por la variedad de tipos de datos) mediante el Linked Open Data (LOD) con soporte en el protocolo OAI-PMH (protocolo que promueve estándares de interoperabilidad de contenidos en Internet). La interoperabilidad semántica es necesaria porque el uso de sistemas de interoperabilidad "sintáctica" (como OAI PMH) no es suficiente.

Por tal razón, esta propuesta involucrará, de manera muy general, los siguientes procesos: extracción de metadatos desde fuentes confiables, la generación y publicación de datos enlazados. Todo esto con el propósito de mejorar la integración e interoperabilidad de recursos almacenados en distintas redes nacionales, es decir, promover el reuso y generar servicios de valor agregado al contexto de la Universidad Nacional de Chilecito (UndeC), a partir de la información nacional recolectada tales como: visualizaciones, extracción de información, reportes estadísticos, entre otros. De acuerdo con lo planteado, se propone un repositorio semántico para la UNdeC (en Chilecito, La Rioja, Argentina), que permita trabajar con: Sistema Nacional de Repositorios Digitales (SNRD), Sistema Nacional de Datos Biológicos (SNDB) y Sistema Nacional de Datos Genómicos (SNDG). Cabe destacar que ninguno de estos repositorios está enfocado específicamente a necesidades particulares de instituciones, pero tienen su base en la política de datos abiertos de la Argentina, ya que los recursos se encuentran bajo una licencia abierta y usan un formato estandar abiertos [8, 16–18]. Además, los Sistemas Nacionales tiene su fundamentación legal en la Ley 26.899/2013 sobre los repositorios digitales de acceso abierto de la Argentina y la Ley 27.275/2016 sobre el derecho de acceso a la información pública.

La propuesta del repositorio semántico con base en una ontología, permitirá la incorporación de nuevas potencialidades a la comunidad de la UNdeC en procesos de representación, organización, diseminación y recuperación de información, usando métodos diferentes a los disponibles en la actualidad, por ejemplo, localizar términos en fuentes de información incorporadas al repositorio, detectar términos genéricos y equivalentes, eliminar ambigüedades en las búsquedas, identificar hipónimos, etc. [10]. Las principales condiciones de este desarrollo tecnológico son: seguir las directrices OPENAire, usar el software DSPACE v.6., generar un vocabulario propio a partir de la ontología desarrollada sobre la base del Linked Open Data y de acuerdo con la metodología propuesta por Heath y Bizer en el 2011 [11].

Artículo recibido el 28 de abril

II. Objetivos

El Servicio de Difusión de la Creación Intelectual (SEDICI) junto con la UNdeC están desarrollando un proyecto con el propósito de desarrollar un repositorio semántico para la Universidad Nacional de Chilecito que unifique los recursos de los Sistemas Nacionales de datos biólogicos, genómicos y repositorios digitales, pero enfocados en forma específica, a las necesidades de la Universidad, de tal manera que cualquier persona puede acceder y vincularse con las diferentes áreas que tiene la UNdeC. El **objetivo general** del proyecto es brindar un conjunto de herramientas/aplicaciones web a la comunidad de la UNdeC, sobre la base de una ontología que se utilizará en la búsqueda de recursos digitales (académicos, científicos, datos, etc) almacenados en los diferentes Sistemas Nacionales. Gracias a esto, se fomenta la cultura de compartición de recursos, vinculación de personas en actividades e investigaciones académicas y científicas similares, contribución a la generación de nuevos conocimientos y disminución de las brechas que dificultan el acceso de las personas a los materiales. La investigación propuesta contempla tanto aspectos de investigación aplicada (beneficio directo de la sociedad) y de campo (recolección y disponibilidad de recursos digitales), como también una fase de transferencia, ya que este proyecto se podrá desarrollar en otros contextos y/o instituciones.

III. Fundamentos

La Web Semántica, mejor conocida como la Web Extendida, ha estado regulada por el Consorcio de la World Wide Web y se basa en metadatos semánticos y en ontologías, con el objetivo de mejorar la interoperabilidad entre los sistemas informáticos que gestionan la información sobre la Web [5]. Los principales componentes de la Web Semántica son los metalenguajes y los estándares de representación XML, XML Schema, RDF, RDF Schema y OWL, y el lenguaje de consulta de datos RDF, llamado SPARQL [5]. La forma que tiene la Web Semántica de vincular los datos de diferentes fuentes en la Web es a través de los Datos Enlazados (en inglés *Linked Data*), mediante el cual los datos se referencian de la misma forma que lo hacen los enlaces de las páginas web. Linked Data se fundamenta en cuatro principios básicos de diseño [4]: 1) Un identificador de recursos (lugares, personas, eventos, etc.) en la Web. 2) Enlaces de los recursos para que los usuarios puedan localizar y consultarlos. 3) Información sobre los recursos usando RDF para describir recursos. d) Enlaces a otros identificadores de recursos de la Web para relacionarlos con los datos contenidos en el recurso.

El enfoque de Linked Data ofrece ventajas significativas sobre las prácticas actuales de publicación de datos, donde la Web es la infraestructura de transporte de datos y metadatos, mientras que desde Linked Data, los datos y su semántica son parte de la misma Web. La principal diferencia entre la Web del hipertexto (páginas Web) y la Web Semántica es que mientras la primera vincula páginas Web o documentos, la segunda se centra por ir más allá del concepto documento y enlaza datos estructurados. Por ello, se hace necesario para enlazarlos, definir una jerarquía de conceptos a partir de los datos estructurados. Esta definición es una ontología [9], ya que proporciona un vocabulario de clases y relaciones para describir un dominio [6].

La filosofía de la Web Semántica [4] establece a RDF (Resource Description Framework) como un lenguaje para la definición de ontologías y metadatos en la web, es decir, RDF sirve como marco de descripción de recursos para metadatos en la Web [14]. El lenguaje se basa en declarar recursos usando la expresión en la forma sujeto, predicado y objeto, conocida como tripleta.

El RDF se puede encontrar en sistemas de software para repositorios, como es el caso de DSpace, entendiendo a los Repositorios Digitales como sistemas de información interoperables que alojan recursos científicos, académicos y administrativos, descriptos por medio de un conjunto de datos específicos (metadatos). Los repositorios tienen como propósito recopilar, catalogar, gestionar, acceder, difundir y preservar tales recursos [20]. DSpace es una herramienta de código abierto desarrollada por el Instituto Tecnológico de Massachusetts en colaboración con la empresa Hewlett-Packard para la implementación del repositorio de dicha institución. Fue liberada en el 2002 y se presenta como una solución que proporciona toda la funcionalidad necesaria de un repositorio digital, lo que incluye la administración de colecciones digitales tales como libros, artículos, fotos, vídeos, tesis y entre otros [20].

En el contexto de la Web Semántica existe una enorme oportunidad con los Repositorios Digitales. Se pueden destacar algunos esfuerzos: el Repositorio de Investigadores de Ecuador [2], Semantic Scholar, Semantic Data Repository EU, Upapers, PaperCube. Muchos sistemas de software ya cuentan con soporte para RDF y SPARQL, lo que sirve para establecer conexiones entre los metadatos con diferentes bases de conocimiento (*endpoints SPARQL*), haciendo así más rica la información que se presenta al usuario final y consolidando la iniciativa de Linked Open Data, que se refiere a datos publicados (*Open Data*) y enlazados (*Linked Data*) mediante la estructura de la Web Semántica, es decir, cuando ambos se unen. Por tanto, la propuesta se plantea desde la filosofía de LOD, que indica que son datos abiertos en RDF, ya que el usuario puede enlazar datos provenientes de diversas fuentes, instituciones u organizaciones, explorar y combinar estos datos de manera libre y sin restricciones de acceso ni de derechos [1, 3, 15].

IV. Metodología

La investigación será proyectiva, puesto que se describirá, caracterizará y propondrá un conjunto de estrategias del Linked Open Data para colocar al servicio de la UNdeC un repositorio semántico a partir de los recursos ofrecidos por los Sistemas Nacionales para el acceso a la información

académica y científica, es decir, identificar un proceso para mejorarlo [12]. El plan propuesto se desarrollará en un computador de la UNdeC (junto con apoyo técnico de SEDICI), desde el cual se hará la cosecha mensual de las diferentes fuentes (SNRD, SNDB Y SNDG) usando el software DSpace. En este software, en primera instancia se cargaran estos metadatos comunes de cualquier recurso de los tres Sistemas Nacionales: autor, fecha de creación, identificador, tipo de recurso, descripción, idioma, título, tipo de licencia. Luego, DSpace ofrecerá una opción distinta para cada uno de ellos. Para los recursos del SNRD [7] se solicitarán estos metadatos: páginas, fulltext, peer review, revista, volumen, issue y archivo digital del recurso (un PDF usualmente). Para datos del SNDB [13], el software solicita estos metadatos propios: colección, número de catálogo, base de registro, datasets asociados, taxonomía, geolocalización y registro de calidad de datos. Finalmente, con datos del SNDG[19], los metadatos no comunes son: procedencia, secuenciación ADN, tipo de organismo, tipo de comunidad, datasets asociados.

Para el desarrollo de la ontología, se recomienda seguir la metodología de Heath y Bizer [11], que permite poder aplicar el concepto de Linked Open Data. Los procedimientos y técnicas a realizar en diferentes etapas son [3]:

- Determinación de la población a examinar: sistemas nacionales del Ministerio de Ciencia, Tecnología e Innovación Productiva.
- Recolección y selección de datos: se realizará usando el protocolo de interoperabilidad OAI-PMH. Además, se realizará una verificación de duplicados por las cosechas mensuales. De esta manera, se garantiza mantener la ontología actualizada.
- Validación para asegurar el control de calidad de los datos: se analizará la estructura del conjunto de datos cosechados y reutilizables.
- Conversión de los datos: los datos obtenidos y depurados se prepararán para su posterior análisis (conversión a RDF).
- Almacenamiento centralizado de datos: los datos se almacenarán bajo el esquema RDF.
- Uso del software: se pone en marcha el módulo de RDF en DSpace que proporcione un patrón para publicar Linked Open Data.
- Aplicación del módulo RDF y visualización de los datos.
- Análisis de los resultados: a partir de la aplicación del prototipo se revisará y se comparará (usando una matriz FODA o técnicas de Benchmarking por ejemplo) con otras posibles soluciones, como buscadores semánticos, Europeana, Biblioteca Digital de los EEUU, entre otros [10].

V. Referencias

1. Allemang, D., Hendler, J.: Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL. Elsevier (2011).
2. Baculima, F. et al.: Repositorio Semántico de Investigadores del Ecuador. (2016).
3. Barber, E.E. et al.: 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. Inf. Cult. Soc. 33, 89–96 (2015).
4. Berners-Lee, T.: Design Note: Linked Data. (2006).
5. Berners-Lee, Ti.: Tejiendo la red: el inventor del World Wide Web nos descubre su origen. (2000).
6. Castells, P.: La web semántica. Sist. Interact. Colab. En Web. 195–212 (2003).
7. De Giusti, M.R. et al.: SeDiCI - Desafíos y experiencias en la vida de un repositorio digital - Metadatos. E-Colab. 1, no. 2, (2011).
8. De Giusti, R.: Estado del AA: caso Argentina. Presented at the BIREDIAL-ISTEC 2015 (Barranquilla, Colombia, 17 al 21 de noviembre de 2015) (2015).
9. Guarino, N. et al.: What Is an Ontology? In: Staab, S. and Studer, R. (eds.) Handbook on Ontologies. pp. 1–17 Springer Berlin Heidelberg (2009).
10. Hallo, M. et al.: Current state of Linked Data in digital libraries. J. Inf. Sci. 42, 2, 117–127 (2016).
11. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synth. Lect. Semantic Web Theory Technol. 1, 1, 1–136 (2011).
12. Hurtado, J.: Cómo formular objetivos de investigación. Fundacion Sypal, Caracas, Venezuela (2008).
13. Levatich, T.: EOD - eBird Observation Dataset, http://datos.sndb.mincyt.gob.ar/collectory/public/show/dr370, (2016).
14. Méndez, E.: RDF: un modelo de metadatos flexible para las bibliotecas digitales del próximo milenio. Jornades Catalanes Doc. 487–498 (1999).
15. Michelan, G. et al.: Integration of Scientific Information through Linked Data. Presented at the II Simposio Argentino de Ontologías y sus Aplicaciones (SAOA) - JAIIO 45 (2016).
16. MINCYT: SISTEMAS NACIONALES - República Argentina, http://sistemasnacionales.mincyt.gob.ar/.
17. Naser, A., Rosales, D.: Panorama regional de los datos abiertos: avances y desafíos en América Latina y el Caribe. (2016).
18. Silvestri, L.C., Cirvini, S.: Datos abiertos en instituciones culturales en Argentina. Rev. Ph. 0, 0, (2017).
19. SNDG: Aves de Argentina, http://datos.sndg.mincyt.gob.ar/datos/4.
20. Texier, J.: Los repositorios institucionales y las bibliotecas digitales: una somera revisión bibliográfica y su relación en la educación superior. Presented at the 11th Latin American and Caribbean Conference for Engineering and Technology - 2013 , Cancun, Mexico October 14 (2013).