View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Servicio de Difusión de la Creación Intelectual

Linked Open Data para la Integración de Información Científica

Gaston Michelan[1] Germán Braun[1,2,3] Laura Cecchi[1]
Pablo Fillottrani[2,4]

email: {gaston.michelan,german.braun,lcecchi}@fi.uncoma.edu.ar, prf@cs.uns.edu.ar

[1] *Grupo de Investigación en Lenguajes e Inteligencia Artificial*
Departamento de Teoría de la Computación - Facultad de Informática
Universidad Nacional del Comahue
[2] *Laboratorio de I&D en Ingeniería de Software y Sistemas de Información*
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur
[3] *Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)*
[4] *Comisión de Investigaciones Científicas de la provincia de Buenos Aires (CIC)*

Resumen

Esta línea de investigación se desarrolla en forma colaborativa entre docentes-investigadores de la Universidad Nacional del Comahue y de la Universidad Nacional del Sur, en el marco de proyectos de investigación financiados por las universidades antes mencionadas.

El objetivo general del trabajo de investigación es relevar y analizar los nuevos lenguajes y tecnologías existentes, para generar y publicar datos abiertos e integrarlos con otras fuentes de datos disponibles en la Web. Asimismo, se proyecta identificar falencias de esta nueva infraestructura, proponer mejoras y evaluar su impacto en usuarios por medio de aplicaciones Web Semánticas. Se prevé utilizar como caso de estudio al grupo y laboratorio de afiliación de los autores.

Palabras Clave: Linked Open Data, Web Semántica, Ingeniería de Software.

Contexto

Este trabajo está parcialmente financiado por la Universidad Nacional del Comahue, en el marco del proyecto de investigación *Agentes Inteligentes en Ambientes Dinámicos (04/F006)* y a través de una beca de Iniciación a la Investigación para alumnos; por la Universidad Nacional del Sur a través del proyecto de investigación *Integración de Información y Servicios en la Web (24/N027)*, y por el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), en el contexto de una beca interna doctoral. Los proyectos de investigación tienen una duración de cuatro años y la beca doctoral una duración de 5 años, finalizando esta última en abril de 2019. La beca de Iniciación a la Investigación tiene una duración de un año.

1. Introducción

Del mismo modo que existen diversas lógicas para representar conocimiento con diferentes niveles de expresividad, en un lenguaje de marcado, también se debe considerar su capacidad para describir datos. En este contexto, HTML [1], por ejemplo, no es lo suficientemente expresivo para permitir que

ciertas entidades sean descriptas en un documento en la Web y que además puedan ser relacionadas entre ellas. Por otra parte, extraer información de fuentes tan heterogéneas disponibles en la Web es un problema complejo debido a que la cantidad de datos es vasta y se incrementa y actualiza constantemente. Una posible solución es estructurar los datos en un formato que permita acceder fácilmente a ellos, sin considerar el estado de las fuentes al extraer información, ya que estaremos siempre consultando la última actualización de los datos.

El objetivo principal de Linked Data [2, 3, 4], propuesto por Tim-Berners Lee, es publicar y conectar datos estructurados en la Web a través de un conjunto de buenas prácticas para tal fin. Así, nuevos documentos serán "entendibles" por las máquinas, tendrán un significado definido explícitamente y serán enlazados con otros, transformando la Web en una colección de tripletas RDF [5] referenciadas por URIs en los diferentes espacios de nombres. Esta capacidad para publicar y conectar datos propuesto por Linked Data es fundamental para la implementación de la Web Semántica [6, 7].

La iniciativa más importante en este campo es el Open Data Movement[1] cuyo objetivo, a través del "Linking Open Data Project", es impulsar la publicación de datos disponibles para todos. En la actualidad existen varios conjuntos de datos abiertos, entre los que podemos nombrar DBpedia[2], WordNet[3] y DBLP[4], con el que estaremos trabajando en esta propuesta. La meta principal del proyecto es extender la Web publicando datos como RDF y enlazando dichos datos con otras fuentes también abiertas. Actualmente, esta iniciativa incluye más de 600 conjuntos de datos publicados, como muestra la última actualización del "LOD cloud diagram" en [8].

[1] https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[2] http://wiki.dbpedia.org/

[3] https://wordnet.princeton.edu/

[4] http://dblp.uni-trier.de/db/

La comunidad científica no queda fuera de esta nueva filosofía y, en esta dirección, han surgido enfoques basados en Linked Open Data aplicados a las actividades de investigación [9, 10].

En el ámbito de este trabajo proponemos analizar diferentes tecnologías y formalismos que se requieren para poder ofrecer una fuente de datos enlazada y abierta, con información científica de grupos de investigación, como así también enfoques para su integración y herramientas para consumir estos datos que se beneficien de su forma estandarizada de representación y de la posibilidad de enlazar nuevas fuentes de datos en tiempo de ejecución. En particular, desarrollaremos una aplicación Web específica para el dominio académico combinando datos desde diferentes fuentes como, por ejemplo, DBLP, entre otras. Dicha aplicación permitirá visualizar la información actualizada asociada a los datos disponibles y realizar consultas en el lenguaje SPARQL [11] para RDF.

Se prevé utilizar como casos de estudio, al grupo de investigación GILIA del Departamento de Teoría de la Computación, de la Facultad de Informática, de la Universidad Nacional del Comahue y al LISSI, del Departamento de Ciencias e Ingeniería de la Computación, de la Universidad Nacional del Sur.

El desarrollo de una herramienta de estas características tendrá un gran impacto sobre nuestros equipos de investigación, ya que incorporarlos a la Web de datos, aumentará la visibilidad, fomentando la colaboración científica entre grupos interdisciplinarios.

La estructura del presente trabajo es la siguiente. En la sección 2 presentamos los objetivos de los proyectos de investigación en los que se enmarca este trabajo y describimos la línea de investigación, el problema que se estudia y los objetivos. En la sección 3 indicamos algunos resultados obtenidos y trabajos futuros. Finalmente, comentamos aspectos referentes a la formación de recursos humanos en esta temática.

2. Línea de Investigación y Desarrollo

El proyecto de investigación *Agentes Inteligentes en Ambientes Dinámicos* tiene varios objetivos generales. Uno de ellos es el de *desarrollar conocimiento especializado en el área de Inteligencia Artificial*. En este sentido, se estudian, entre otras, técnicas de representación de conocimiento y razonamiento aplicadas al desarrollo de agentes y a asistir el modelado conceptual.

Por otro lado, en el proyecto de investigación *Integración de Información y Servicios en la Web* se propone investigar y desarrollar metodologías y herramientas que favorezcan la interoperabilidad semántica de información y de servicios en la Web.

Estas líneas de investigación confluyen en el estudio de formalismos y tecnologías para cubrir las necesidades emergentes de compartir, actualizar e integrar el conocimiento de sistemas computacionales pre-existentes, elementos que se consideran fundamentales dadas las necesidades de interoperabilidad de aplicaciones, tanto a nivel de procesos como de datos. Particularmente, hemos escogido experimentar sobre los lenguajes y tecnologías existentes para generar datos abiertos y enlazados con otras fuentes de datos disponibles en la Web. Asimismo, se proyecta identificar falencias de esta nueva infraestructura, proponer mejoras y evaluar su impacto en usuarios por medio de una aplicación Web Semántica.

Como primer paso se trabajará en la estructuración de los datos de los grupos de investigación siguiendo los principios de Linked Data. Para satisfacer estos principios, se hará un relevamiento de las tecnologías disponibles para generar tripletas RDF a partir de nuestras bases de datos relacionales. Estas herramientas nos debe permitir la generación de URIs para identificar entidades unívocamente, cumpliendo así con el primer principio de Linked Data. Luego, para hacer públicos estos datos, se usará un servidor local que nos permitirá referenciar las URIs usando el protocolo HTTP haciendo accesibles los datos desde fuentes externas, además de proveer la información requerida en el mismo formato RDF. Finalmente, insertaremos enlaces a otras URIs de datos RDF como, por ejemplo, FacetedDBLP[5], con información de publicaciones científicas. Además, incluiremos un motor para consultas SPARQL.

Si bien los datos, al estar disponibles en formato RDF, podrán ser consultados en el formato original con motores ad-hoc, es nuestra intención desarrollar una aplicación Web Semántica que permita visualizarlos y realizar consultas de un modo más amigable. La aplicación propuesta consumirá estos datos y habilitará a los usuarios a consultar la información requerida y navegar la Web de Datos a partir de su consulta inicial, en tiempo de ejecución, accediendo a la versión más reciente de ellos y sin considerar cómo y dónde están físicamente almacenados. De esta manera y, como un efecto directo de este trabajo, lograremos la incorporación de nuestros grupos a la Web de Datos fortaleciendo la integración, a nivel semántico, con otras fuentes de datos de interés.

3. Resultados Obtenidos y Trabajo Futuro

Inicialmente, se realizó un relevamiento de las tecnologías disponibles para nuestra arquitectura. Se analizaron las siguientes plataformas D2R [12], Sesame [13], Talis[6], 4store [14], Jena [15] y Virtuoso [16], entre otras. De este análisis, se determinó que la plataforma más conveniente para un primer prototipo era D2R debido al nivel de complejidad de sus funcionalidades y la posibilidad de publicar datos de bases relacionales como triplas RDF.

Asimismo, a fin de lograr la integración entre los datos, se identificaron un conjunto de ontologías existentes para etiquetar nuestros

[5] http://dblp.l3s.de/?q=&newQuery=yes&resTableName=query_result3kKaKj

[6] https://talis.com/

datos y desambiguarlos. Como primera aproximación usaremos FOAF [17], para etiquetar personas, sus actividades y sus relaciones, y DC (Dublin Core) [18], para etiquetar recursos Web como, por ejemplo, artículos científicos junto con su título, autor, fecha, tipo y formato, entre otros.

Actualmente, se está trabajando en el desarrollo de un prototipo Web para visualizar y consultar estos datos enlazados. Como trabajo futuro, proponemos agregar más URIs de usuarios a otras bases de datos como lo son Bibsonomy[7] y Geonames[8], entre otras. Esto posibilitará el desarrollo de otras aplicaciones Web interdisciplinarias utilizando estos datos abiertos.

Hasta aquí, RDF nos ofrece un modelo basado en grafos para registro de datos, sin embargo, no nos permite agregarles semántica. Por lo tanto, también se espera poder definir un nivel más de abstracción, creando un vocabulario propio para describir entidades y relaciones del dominio. Luego, a partir de la ontología resultante, validar la consistencia de los datos y su relaciones implícitas utilizando técnicas de razonamiento basadas en lógica.

4. Formación de Recursos Humanos

Uno de los autores de este trabajo está inscripto en el Doctorado en Ciencias de la Computación en la Universidad Nacional del Sur(beca interna doctoral CONICET).

Otro de los autores, alumno avanzado de la carrera Licenciatura en Ciencias de la Computación, ha obtenido una Beca de Iniciación en la Investigación para Alumnos Universitarios de la Universidad Nacional del Comahue. Dicho becario realizará su tesis de grado en la temática de la línea de investigación presentada en el marco del GILIA.

Finalmente, nuestro programa de formación de recursos humanos incluye el dictado de los cursos de posgrado "Ontologías y Web Semántica: Interoperabilidad Semántica de la Información", en el marco del Doctorado en Ciencias de la Computación, UNS y "Gestión de contenido y tecnología de las redes sociales", en el marco de la maestría en Bibliotecología y Ciencia de la Información[9], UBA. En ambos casos el disertante es el Prof. Pablo Fillottrani, uno de los autores de este trabajo.

[7] http://www.bibsonomy.org/

[8] http://www.geonames.org/

[9] http://maestriabiblio.blogspot.com.ar/

Referencias

[1] W3C HTML Working Group. HTML, The Web's Core Language, 2008. https://www.w3.org/html/, accedida en Marzo de 2016.

[2] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *Int. J. Semantic Web Inf. Syst.*, 2009.

[3] Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 1st edition, 2011.

[4] Liyang Yu. *A Developer's Guide to the Semantic Web, Second Edition*. Springer, 2014.

[5] Marcelo Arenas, Claudio Gutierrez, and Jorge Pérez. Foundations of rdf databases. In *Reasoning Web*, pages 158–204, 2009.

[6] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, May 2001.

[7] Tim Berners-Lee. Linked Data, 2006. https://www.w3.org/DesignIssues/LinkedData.html, accedida en Marzo de 2016.

[8] Anja Jentzsch Max Schmachtenberg, Christian Bizer and Richard Cyganiak. Linking Open Data cloud diagram,

2014. http://lod-cloud.net/, accedida en Marzo de 2016.

[9] Tomi Kauppinen and Giovana Mira de Espindola. Linked Open Science-Communicating, Sharing and Evaluating Data, Methods and Results for Executable Papers. *Procedia Computer Science*, 2011. Proceedings of the International Conference on Computational Science, ICCS 2011.

[10] Carsten Keßler and Tomi Kauppinen. Linked Open Data University of Muenster—Infrastructure and Applications. In *Demos ESWC*, Heraklion, Crete, Greece, May 2012.

[11] W3C OWL Working Group. *SPARQL 1.1 Query Language*. W3C Recommendation, 21 March 2013. Available at http://www.w3.org/TR/sparql11-query/, accedida en Marzo de 2016.

[12] Christian Bizer and Richard Cyganiak. D2R Server – Publishing Relational Databases on the Semantic Web. Poster at the 5th ISWC, 2006.

[13] Jeen Broekstra, Arjohn Kampman, and Frank van Harmelen. Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In *Proceedings of the First ISWC*, 2002.

[14] 4store.org. 4store - Scalable RDF storage, 2012. http://4store.org/, accedida en Marzo de 2016.

[15] Andy Seaborne. Jena, a Semantic Web Framework, November 2010. http://wiki.apache.org/incubator/JenaProposal, accedida en Marzo de 2016.

[16] OpenLink Software. Virtuoso Universal Server, 2010. http://virtuoso.openlinksw.com/, accedida en Marzo de 2016.

[17] Dan Brickley and Libby Miller. The Friend Of A Friend (FOAF) vocabulary specification, November 2007. http://xmlns.com/foaf/spec/, accedida en Marzo de 2016.

[18] A. Powell, M. Nilsson, A. Naeve, and P. Johnston. Dublin Core Metadata Initiative - Abstract Model, 2005. White Paper.