HERRAMIENTA DE BÚSQUEDA EN REPOSITORIOS ACADÉMICOS BASADA EN WEB SEMÁNTICA Y SISTEMAS NOSQL

María del Pilar Gálvez, Sergio L. Martínez, Nélida R. Cáceres, Ana C. Tolaba, Laura R. Villarrubia, Felipe F. Mullicundo, José R. Quispe, Marcelo R. Sanguezo Ballon, Iván L. Sandoval, Jairo J. M. Quispe, Daniel A. Lamas

Facultad de Ingeniería - Universidad Nacional de Jujuy
Ítalo Palanca 20 San Salvador de Jujuy – 0388 4221576
mdpgalvezdiaz@fi.unju.edu.ar

RESUMEN

Las instituciones académicas buscan exponer su producción científica/académica a través de repositorios digitales en Acceso Abierto. Sin embargo, presentan limitaciones para lograr los objetivos ya que los datos publicados pueden resultar insuficientes, o bien no se cuenta con términos adecuados que puedan ser relacionados para realizar búsquedas más integrales y eficaces de forma de obtener mejores resultados. Los metadatos empleados para la descripción de los datos publicados, al ser semiestructurados no permiten explotar la información de mejor manera porque hay conocimiento implícito que favorece la descripción de nuevas relaciones entre los datos explicitados que no está siendo usado.

Actualmente el sistema de consultas provisto por SIBUNJU permite ver información de los trabajos finales de grados resumida por: Título, autor, idioma, editorial, ISBN y biblioteca depositaria. Este proyecto plantea la definición de una herramienta de búsqueda en un repositorio creado específicamente para la Facultad de Ingeniería, que facilite el acceso y la explotación de los datos residentes en el mismo a través de la incorporación de tecnologías de web semántica y sistemas NoSQL. Esta herramienta será útil tanto para alumnos, egresados, docentes, investigadores y la sociedad puesto que permite crear y compartir conocimiento, además de facilitar su transferencia al sector productivo.

Palabras clave: *NoSQL, Repositorios Digitales, Semántica, Búsquedas Mejoradas.*

CONTEXTO

La presente investigación corresponde al proyecto "*Desarrollo de Herramienta de Búsqueda utilizando Web Semántica y Sistemas NoSQL*". El mismo fue aprobado por la Secretaría de Ciencia y Técnica de la Universidad Nacional de Jujuy como proyecto categoría A (código D/0168) y se encuentra bajo incentivo.

1. INTRODUCCIÓN

Repositorios Digitales

En la actualidad se observa que la cantidad de repositorios digitales se ha incrementado notablemente, debido principalmente a la tendencia a exponer la producción de las instituciones académicas a través de repositorios digitales de Acceso Abierto [1]. En Argentina la creación de repositorios de Acceso Abierto fue impulsada desde el Ministerio de Ciencia y Tecnología que creó el Sistema Nacional de Repositorios Digitales en CyT (SNRD) y elaboró un proyecto de ley que fue aprobado a fines de 2013. La ley 26899 "Creación de repositorios Digitales Institucionales de Acceso Abierto, Propios o Compartidos" establece la obligatoriedad del

acceso abierto a la producción financiada con fondos públicos a nivel nacional a través de repositorios digitales que las instituciones deberán crear, mantener e integrar al SNRD [2].

Los repositorios tienen como propósito recopilar, catalogar, gestionar, acceder, difundir y preservar información, permitiendo entre cosas el acceso libre y gratuito a todos los recursos que los conforman[3]. La implementación de un repositorio digital ofrece diferentes beneficios tanto para investigadores, estudiantes, así como al resto de la sociedad ya que permiten crear y compartir conocimiento, y facilitan la transferencia de conocimiento al sector productivo [4].

Metadatos

Los metadatos son datos que describen otros datos y dan información que permite describir, identificar, recuperar o gestionar recursos de información [5]. Desde el punto de vista de las Ciencias de la Documentación y la Información, los metadatos son un conjunto de atributos de catalogación en los documentos que permite su identificación sin tener que realizar la apertura del documento digital para conocer su contenido [6].

Existen diferentes estándares para la implementación de metadatos como Dublin Core Metadata Initiative (DCMI), Learning Objects Metadata (LOM), entre otros. Es necesario que un repositorio cuente con metadatos precisos, completos y con un formato homogéneo, esto le permitirá interoperar con otros repositorios para realizar intercambio de información además de crear servicios de valor añadido [7].

Web Semántica

Berners-Lee [8] define la Web semántica como una extensión de la web actual, en la que la información tiene un significado bien definido posibilitando a los humanos y las computadoras trabajar en cooperación. La web semántica permite el acceso inteligente y preciso a grandes repositorios de datos, esto favorece a la difusión del contenido de los repositorios [9]. Dentro de las tecnologías de la Web semántica se dispone de RDF (Resource Description Framework) que permiten dotar de significado los datos y transacciones de datos en la Web [10]. Una declaración RDF constituye la forma más simple de expresión de un metadato, mediante grafos de tripletas compuestas por un sujeto, un predicado y un objeto. Donde un sujeto es el recurso que se está describiendo. El predicado es la propiedad o relación que se desea establecer acerca del recurso. Por último, el objeto es el valor de la propiedad de lo que se describe sobre el sujeto o el otro recurso con el que se establece la relación [11].

Otra tecnología que la Web Semántica ofrece son las ontologías, estas son estructuras más completas que permiten una representación formal de un concepto, además de la representación semántica y sintáctica del mismo [12].

Búsqueda Semántica

La recuperación de información es obtenida con a través de motores de búsqueda de propósito general. Estos motores de búsqueda se basan en términos de indexación sin tener en cuenta la semántica de los contenidos y el contexto [13].

La búsqueda semántica se refiere a una búsqueda de conceptos no solo basada por la comparación de palabras (búsqueda sintáctica), sino por deducciones lógicas que consideran la intención y el significado contextual de las palabras empleadas en la búsqueda [14].

Bases de Datos NoSQL

Los sistemas NoSQL se emplean cada vez más para el manejo de datos semánticos, es decir, modelos de datos que incluyen información semántica. El objetivo de los modelos de datos semánticos es capturar el significado de los datos mediante la integración de conceptos relacionales con conceptos de abstracción más poderosos.

Las bases de datos NoSQL se caracterizan por: ser no relacionales, distribuidas, de código abierto y escalables horizontalmente [15] [16].

NoSQL corresponde a una estrategia de persistencia que no siguen el modelo de datos relacional, y que no utilizan SQL como

lenguaje de consulta [17] en otras palabras, no están supeditadas a una estructura de datos en forma de tablas y relaciones entre ellas, permitiendo a los usuarios almacenar información en formatos diferentes a los tradicionales. Algunas aplicaciones de estas bases de datos pueden observarse en [18], [19] y [20].

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Los repositorios digitales se estructuran mediante un modelo de contenidos. Las actuales búsquedas de información en estos repositorios, en términos generales, permiten escribir la consulta por medio de un campo de texto en el que se pueden seleccionar palabras claves o bien discriminadas por autor o institución académica. Estas búsquedas proporcionan resultados que satisfacen las consultas de los usuarios de acuerdo a los valores ingresados, sin embargo, es posible inferir más información de la obtenida hasta el momento.

Por este motivo, este trabajo pretende mostrar la mayor cantidad de información que puede estar representada explícita e implícitamente en el repositorio. Es por ello que la principal línea de investigación involucrada en el proyecto es la Ingeniería de software, mediante la implementación de un repositorio digital junto a una herramienta de búsqueda que permita mejorar los resultados a través de la integración de tecnologías de web semántica y sistemas NoSQL.

El proyecto se adecúa a las líneas prioritarias expuestas por la Facultad de Ingeniería de la UNJu en la Resolución FI Nº 071/98, la cual incluye el área temática "Ingeniería de Software", en la cual se consideran las siguientes líneas de acción:

- Repositorios digitales.
- Gestión de la información y el conocimiento.
- Sistemas de información web y bases de datos.
- Recuperación de la información.

3. RESULTADOS OBTENIDOS Y ESPERADOS

Este proyecto tiene estipulados cuatro años de duración (2020-2023) y su principal objetivo es desarrollar una herramienta de búsqueda que facilite el análisis y comprensión de los datos almacenados en un repositorio digital de trabajos finales de grado de la Facultad de Ingeniería de la UNJu. La herramienta propuesta combinará para su desarrollo, conceptos de web semántica y sistemas NoSQL. La información extraída de estos repositorios será utilizada como apoyo para la toma de decisiones, tanto a nivel administrativo y operativo de los estudiantes de grado ya que les proporciona el conocimiento necesario para llevar a cabo la selección del tema de trabajo final. Además, esta información permitirá que otros usuarios como egresados, docentes, investigadores y agentes externos conozcan las diferentes líneas de investigación de los trabajos desarrollados, logrando de esta forma la transferencia de la UNJu hacia la comunidad.

El desarrollo de este proyecto tiene la intención de reformular el funcionamiento de los repositorios de trabajos finales de la Facultad de Ingeniería de la UNJU, de forma tal que el resultado del proceso de búsqueda información sea utilizado como apoyo en la toma de decisiones en el ámbito académico y social.

En el ámbito académico se observó que los alumnos antes de encarar el inicio de su trabajo final, realizan búsquedas exhaustivas de antecedentes sobre trabajos concluidos en la unidad académica a la cual pertenece para dotar a su trabajo final de originalidad, en muchas ocasiones no lo logran por la falta de información, provocando esta situación demoras en el inicio de la etapa final de su

carrera. En este contexto se pretende que la propuesta facilite al alumno en la elección de un tema, un director o la línea de investigación en el cual desea desarrollar su trabajo final. En el caso de los docentes e investigadores, disminuye la incertidumbre respecto a ciertos temas disciplinares consultados, por ejemplo, aplicación de nuevas tecnologías, nuevos métodos entre otros.

En el ámbito social/laboral, no se identifican a las personas especializadas en ciertos temas que sean de utilidad, por ejemplo, se ignoran los recursos humanos que tienen conocimiento sobre algunos temas como litio, Big Data o desarrollo de sistemas. Este desconocimiento implica que la universidad pierda la oportunidad de realizar la transferencia a la sociedad, y que recursos formados por la UNJu tengan oportunidades laborales. En este sentido agentes externos a la UNJu pueden captar mano de obra regional en áreas que les sean de interés, de esta forma se potenciará los recursos humanos formados en la UNJU.

4. FORMACIÓN DE RECURSOS HUMANOS

El proyecto está siendo desarrollado por un equipo conformado por docentes investigadores del Grupo de Investigación y Desarrollo en Ingeniería de Software (GIDIS) de la Facultad de Ingeniería de la Universidad Nacional de Jujuy. La estructura del equipo de investigación es la siguiente:

- Directora: Mg. María del Pilar Gálvez. Categoría de Investigación III.
- Codirector: Mg. Ing. Sergio Luis Martínez. Categoría de Investigación III.

Investigadores:

- Ing. Nélida Raquel Cáceres. Categoría de Investigación IV. Actualmente realizando tesis de maestría vinculada al área de bases de datos.
- Ing. Ana Carolina Tolaba. Categoría de Investigación V. Actualmente realizando tesis de doctorado vinculada al área de modelado conceptual de datos a través de modelos semánticos.
- Esp. Ing. Laura Rita Villarrubia. Categoría de Investigación IV.
- Lic. Felipe Fernando Mullicundo. Categoría de Investigación V.
- Mg. Ing. José Rolando Quispe.
- APU Marcelo Raúl Sanguezo Ballon.

Participan del proyecto alumnos avanzados de la carrera de Ingeniería Informática:

- Jairo Joel Maximiliano Quispe
- Daniel Alberto Lamas
- Iván Leandro Sandoval.

Con la realización de este proyecto de investigación se espera la consolidación de los miembros del grupo en especial de los alumnos como jóvenes investigadores. Además, el proyecto brindará un marco propicio para la iniciación de trabajos finalesde grado de la carrera Ingeniería Informática.

5. BIBLIOGRAFÍA

[1] Maenza, R., & Darin, S. (2016). Universidades abiertas trabajando en la innovación tecnológica y la transparencia. Revista Internacional Transparencia e Integridad, RITI nro, 2.

[2] Peña, K. I. C. (2014). Modelos de acceso abierto en educación y ciencia. Educación y educadores, 17(2), 8. DOI: 10.5294/edu.2014.17.2.7

[3] Texier, J. (2013). Los repositorios institucionales y las bibliotecas digitales: una somera revisión bibliográfica y su relación en la educación superior. 11th Latin American and Caribbean Conference for Engineering and Technology. Cancún, México. p. 9

[4] Ramírez, M. R., Soto, M. D. C. S., Moreno, H. B. R., Rojas, E. M., Millán, N. D. C. O., &Cisneros, R. F. R. (2019). Metodología SCRUM y desarrollo de Repositorio Digital.

Revista Ibérica de Sistemas e Tecnologías de Informação, (E17), 1062-1072.

[5] Senso, J. A., y De la Rosa, A. (2003). El concepto de metadato: algo más que descripción de recursos electrónicos. Ciência da Informação, 32(2), 95-106. doi:10.1590/S0100-19652003000200011

[6] Testa, P. (2013). Esquemas de metadatos para los repositorios institucionales de las universidades nacionales argentinas.

[7] Delgado, J. C. S., & Alvarado, M. A. C. (2017). Repositorios institucionales digitales: Análisis comparativo entre SEDICI (Argentina) y Kérwá (Costa Rica). e-Ciencias de la Información, 1-32.

[8] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific american, 284(5), 28-37.

[9] Sulé, A., Centelles, M., Franganillo, J., & Gascón, J. (2016). Aplicación del modelo de datos RDF en las colecciones digitales de bibliotecas, archivos y museos de España. Revista española de documentación científica, 39(1), 121.

[10] McBride, B. (2004). Theresource description framework (RDF) and its vocabulary description language RDFS. In Handbook on ontologies (pp. 51-65). Springer, Berlin, Heidelberg.

[11] RDF, Resource Description Framework. W3C recommendation. Disponible en: http://www.w3.org/RDF Acceso: Octubre, 2019.

[12] Gruber, T., Ontology, I. L. L., & Özsu, M. T. (2009). Encyclopedia of database systems. Springer-Verlag, ISBN 978-0-387-49616-0.

[13] Smine, B., Faiz, R., & Desclés, J.-P. (2012). Extracting relevant learning objects using a semantic annotation method. In International Conference on Education and e-Learning Innovations, (pp. 1–6). IEEE.

[14] Portolés Sánchez, M. J. (2010). Búsqueda semántica en repositorios de conceptos biomédicos estandarizados: CT Hunter (Doctoral dissertation).

[15] Venkatraman, S., Fahd, K., Kaspi, S., & Venkatraman, R. (2016). SQL Versus NoSQL movement with big data analytics. Int. J. Inform. Technol. Comput. Sci, 8, 59-66.

[16] "NOSQL Databases," Disponible en: http://nosql-database.org Acceso: Octubre, 2019.

[17] Arévalo, H. H. R., & Cubides, J. F. H. (2013). Un viaje a través de bases de datos espaciales NoSQL. Redes de ingeniería, 4(2), 57-69.

[18] Rodríguez Pérez, A., Rodríguez Hernández, D., & Díaz Martínez, E. (2016). Selección de Base de Datos No SQL para almacenamiento de Históricos en Sistemas de Supervisión. Revista Cubana de Ciencias Informáticas, 10(3), 159-170.

[19] Martín, A. E., Chávez, S. B., Rodríguez, N. R., Valenzuela, A., & Murazzo, M. A. (2013, June). Bases de datos NoSQL en cloud computing. In XV Workshop de Investigadores en Ciencias de la Computación.

[20] Valenzo, M. R., Valencia, R. E. C., & Castro, J. M. M. (2013). Integración de búsquedas de texto completo en Bases de Datos NoSQL. Revista Vínculos, 8(1), 80-92.