# ESTUDIO PRELIMINAR PARA EL DESARROLLO DE UNA HERRAMIENTA DE BÚSQUEDA EN REPOSITORIOS ACADÉMICOS

María del Pilar Gálvez, Sergio L. Martínez, Nélida R. Cáceres, Ana C. Tolaba, José R. Quispe, Felipe F. Mullicundo, Laura R. Villarrubia

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

## RESUMEN

A través de repositorios digitales de Acceso Abierto, instituciones académicas buscan exponer su producción científica/académica. No obstante, los datos publicados pueden resultar insuficientes ya sea porque no cuentan con la terminología adecuada para su descripción o bien 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. Todo esto limita la realización de búsquedas más integrales y eficaces de forma de obtener mejores resultados.

Ante esta situación, el presente proyecto plantea en primer lugar, el estudio de repositorios digitales que utilizan las instituciones académicas con el fin de definir todos los conceptos relacionados al mismo y que sean adecuados para el repositorio de la Facultad de Ingeniería, para definir estrategias de búsquedas, técnicas y herramientas que incorporen tecnologías de web semántica y sistemas NoSQL. Este artículo presenta el estado de avance alcanzado en este proyecto, los resultados y la formación de recursos humanos concretada en el marco del mismo.

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

## CONTEXTO

La línea de investigación que se presenta en este trabajo corresponde al proyecto "*Desarrollo de Herramienta de Búsqueda utilizando Web Semántica y Sistemas NoSQL*" aprobado por la Secretaría de Ciencia y Técnica de la Universidad Nacional de Jujuy. Este proyecto es categoría A (código D/0168) y se encuentra bajo incentivo.

## 1. INTRODUCCIÓN

La tendencia de exponer la producción de las instituciones a través de repositorios digitales de Acceso Abierto [1] ha incrementado la cantidad de repositorios digitales notablemente. El Ministerio de Ciencia y Tecnología, en Argentina, impulsó la creación de repositorios de Acceso Abierto mediante el Sistema Nacional de Repositorios Digitales en C y T (SNRD). Además, elaboró un proyecto de ley para "Creación de repositorios Digitales Institucionales de Acceso Abierto, Propios o Compartidos" que fue aprobado a fines de 2013. La ley 26899 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].

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 [3]. Para lograr estos beneficios es necesario considerar la organización de la información disponible en los repositorios de forma tal que el conocimiento implícito que favorece la descripción de nuevas relaciones entre los datos explicitados sea aprovechado.

En este sentido, 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 [4].

La web semántica, es otra tecnología a ser considerada para obtener los beneficios indicados anteriormente. La web semántica permite el acceso inteligente y preciso a grandes repositorios de datos, favoreciendo a la difusión del contenido de los repositorios [5]. 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 [6]. También ofrece ontologías, estas son estructuras más completas que permiten una representación formal de un concepto, además de la representación semántica y sintáctica del mismo [7].

Estas tecnologías permitirán la recuperación de información mediante búsquedas semántica. Las cuales se refieren 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 [8].

La implementación de una herramienta para la realización de búsquedas semánticas comprende el empleo de modelos de datos que incluyen información semántica la cual puede ser gestionada mediante sistemas NoSQL [9], [10]. 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.

## 2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

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 y Recuperación de la información. En la actualidad se trabaja en:

- Estudio de los repositorios digitales implementados en la actualidad para definir las características del repositorio a realizar, como así también el buscador adecuado.
- Definición de los metadatos para los trabajos finales de grado y la estrategia para la recopilación de los mismos.
- Definición de las características del repositorio digital de la Facultad de Ingeniería con información de trabajos finales de grado.
- Definición de la estrategia de búsqueda a desarrollar utilizando web semántica y bases de datos NoSQL.

## 3. RESULTADOS OBTENIDOS/ESPERADOS

Este proyecto tiene estipulado cuatro años de duración, y se establecieron los siguientes objetivos.

Como objetivo general, el proyecto de investigación tiene como propósito desarrollar una herramienta de búsqueda que facilite el análisis y comprensión de los datos almacenados en el repositorio digital de trabajos finales de grado de la Facultad de Ingeniería de la UNJu. La herramienta propuesta combinará para su desarrollo, tecnologías 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.

Además se establecieron como objetivos particulares:

- Realizar un estudio de los repositorios digitales.
- Realizar un estudio de web semántica y sistemas NoSQL.
- Efectuar un análisis respecto de los metadatos y la forma de acceder al contenido de los repositorios institucionales, por ejemplo, análisis de motores de búsqueda empleados.
- Ejecutar pruebas mediante distintos tipos de consultas, con los datos de repositorios digitales institucionales de libre acceso, que permitan realizar un análisis de los resultados obtenidos.
- Realizar un estudio de la implementación de búsquedas semánticas en bases de datos NoSQL.
- Definir el tipo de bases de datos NoSQL y los motores de búsqueda adecuados para el repositorio digital propuesto.
- Realizar el relevamiento de los datos de los proyectos finales de grado de la Facultad de Ingeniería de la UNJu.
- Analizar la estructura de metadatos de los proyectos finales relevados.
- Desarrollar una herramienta de búsqueda mediante la combinación de web semántica y bases de datos NoSQL.
- Generar distintos tipos de búsquedas utilizando la herramienta desarrollada.
- Evaluar los resultados obtenidos por la herramienta de búsqueda a través de pruebas de aceptación.
- Comparar los resultados obtenidos tanto por la herramienta propuesta como por el sistema actual de consulta SIBUNJU.

Considerando los objetivos descritos anteriormente durante el año 2020 se obtuvo como resultado el trabajo "*Herramienta de búsqueda en repositorios académicos basada en web semántica y sistemas NoSQL*". Gálvez Díaa, María del Pilar; Martinez, Sergio L.; Cáceres, Nélida R.; Tolaba, Ana C.; Villarrubia, Laura R.; Mullicundo, Felipe F.; Quispe, José R.; Sanguezo Ballon, Marcelo R.; Sandoval, Iván L.; Quispe, Jairo J.M.; Lamas, Daniel A., XXII Workshop de Investigadores en Ciencias de la Computación (WICC 2020, El Calafate, Santa Cruz). ISBN: 978-987-3714-82-5. Páginas: 425-429.

Mediante la realización de un curso de posgrado se avanzó en la identificación del modelo de metadatos para el repositorio propuesto para la facultad de ingeniería. Se prevé continuar con la incorporación de web semántica para el desarrollo de una herramienta que búsqueda que permita mejorar los resultados de búsquedas exhaustivas de antecedentes sobre trabajos concluidos en la unidad académica a la cual pertenece para dotar a su trabajo final de originalidad.

## 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:

- Mg. Ing. Nélida Raquel Cáceres. Categoría de Investigación IV.
- 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.

Con la realización de este proyecto de investigación se espera la consolidación de los miembros del grupo, además de la formación de jóvenes investigadores principalmente alumnos avanzados de las carreras afines de la facultad de ingeniería. Se espera formar nuevos trabajos finales de grado y participación en becas, cuyas temáticas serán propias del mencionado proyecto de investigación.

Los integrantes de este proyecto de investigación participaron en:

- Curso de postgrado "*Bibliotecas y Repositorios Digitales. Tecnologías y Aplicaciones*" dictado por la Universidad Nacional de la Plata, llevado a cabo los días 1 al 5 de marzo de 2021, a cargo de la docente Dra. Marisa De Giusti, correspondiente al Doctorado en Ciencias Informáticas.
- Curso de capacitación denominado "*Administrador de base de datos no relacionales*" dictado por la Fundación Carlos Slim en modalidad online. Disponible en https://capacitateparaelempleo.org/pages.php?r=.tema&tagID=4066
- Dirección de Trabajo Final de Carrera denominado "*Coopertino: Aplicación web que asiste a la formación de grupos de estudio y el trabajo en equipo*". Alumnos: Eduardo Andrés Albornoz y Matías Ramón Ruiz de la Carrera Ingeniería en Informática de la UNJu. Res. FI N° 352/2021.

## 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] 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.

[4] 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.

[5] 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.

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

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

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

[9] 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.

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