# MASKANA: un gestor de conocimiento para recuperación y búsqueda inteligente de trabajos de grado en la Universidad de Nariño

Jimmy Guerrero Restrepo[a], Ricardo Timarán Pereira[a]

[a] Departamento de Sistemas, Facultad de Ingeniería, Universidad de Nariño, Ciudad Universitaria Torobajo, Pasto, Colombia
jimaguere@hotmail.com, ritimar@udenar.edu.co

**Resumen.** En este artículo se presenta uno de los resultados del proyecto de investigación que tiene como objetivo la implementación de un gestor de conocimiento soportado en una ontología dinámica débilmente acoplado con el sistema gestor de base de datos PostgreSQL, denominado MASKANA. Esta herramienta fue desarrollada, bajo software libre, en el laboratorio KDD del grupo de investigación GRIAS del Departamento de Sistemas de la Facultad de Ingeniería de la Universidad de Nariño (Colombia). MASKANA permite la recuperación y búsqueda inteligente de trabajos de grado, que se encuentran almacenados en un repositorio digital en la biblioteca "Alberto Quijano Guerrero" de la Universidad de Nariño. La arquitectura de la herramienta MASKANA la conforman tres módulos: el módulo de interfaz gráfica que permite al usuario interactuar con la herramienta de manera amigable, facilitando su uso, el módulo kernel, en el cual se realizan los procesos de búsqueda y recuperación de los documentos digitales de los trabajos de grado solicitados y el módulo de conexión que permite la comunicación con el repositorio digital de trabajos de grado, la interacción con la ontología SAWA y el índice de búsqueda denominado LUCENE. Las pruebas de funcionalidad realizadas con trabajos de grado del programa de Ingeniería de Sistemas, demostraron que la herramienta funciona correctamente. Esta herramienta, podrá ser utilizada en cualquier biblioteca de una Institución de Educación Superior de Colombia o de Latinoamérica para la búsqueda y recuperación de documentos digitales.

**Palabras Clave:** Gestor de Conocimiento, Ontología, Repositorio Digital, Trabajos de Grado

## 1. Introducción

En el proceso investigativo realizado en el Grupo de Investigación Aplicada en Sistemas - GRIAS, del departamento de Sistemas, de la facultad de Ingeniería de la Universidad de Nariño, en la línea de investigación de Herramientas y Sistemas de Gestión de Conocimiento y Recuperación de Información, se han desarrollado dos proyectos de investigación, uno de tipo estudiantil denominado "Construcción de una Ontología de Aplicación que Soporte la Búsqueda Inteligente sobre los Trabajos de Grado de la Universidad de Nariño, utilizando la herramienta de software libre Protégé" [1][2] y otro de continuación de este proyecto en la modalidad de trabajo de grado denominado "UMAYUX: Un Modelo de Software

de Gestión de Conocimiento Soportado en una Ontología Dinámica Débilmente Acoplado con un Gestor de Bases de Datos para la Universidad de N,i;U99UIi Variño" [3].

Como resultado de estos proyectos se han obtenido: un modelo de gestión de conocimiento soportado en una ontología dinámica débilmente acoplado con un gestor de base de datos denominado UMAYUX (inteligente en quechua), una ontología para trabajos de grado del programa de Ingeniería de Sistemas denominada SAWA (enlace en quechua) y un prototipo de gestor de conocimiento para recuperación de información relacionada con los trabajos de grado almacenados en formato digital denominado MASKANA (buscador en quechua).

MASKANA únicamente se ha probado y evaluado su funcionamiento con los trabajos de grado que se encuentran en formato digital del programa de Ingeniería de Sistemas.

Teniendo en cuenta estos resultados, se propuso un nuevo proyecto de investigación, el cual fue aceptado y financiado por el programa de jóvenes investigadores de la gobernación del departamento de Nariño (Colombia), para la puesta en producción de MASKANA en la Biblioteca "Alberto Quijano Guerrero " de la Universidad de Nariño que permita la búsqueda inteligente y recuperación eficiente de la información y documentos relacionados con los trabajos de grado de la institución y que se encuentren almacenados en un repositorio digital.

Una vez se termine esta fase, esta herramienta podrá ser utilizada en cualquier biblioteca de una Institución de Educación Superior de Colombia o de Latinoamérica. Además, con la ontología apropiada, MASKANA, podrá gestionar la búsqueda y recuperación eficiente de cualquier documento digital y sus relacionados en diferentes organizaciones, que permita aumentar su competitividad.

El resto del artículo está organizado en secciones. En la sección 2, se dan los fundamentos teóricos básicos. En la sección 3 se describe la arquitectura de la herramienta MASKANA y sus diferentes módulos. En la sección 4, se abordan los aspectos de diseño e implementación de la herramienta. En la sección 5, se muestran las pruebas de funcionalidad y finalmente, en la sección 6, se presentan las conclusiones y futuros trabajos.

## 2. Fundamentos Teóricos

Según Medina José Luis [4], la Gestión del Conocimiento es la disciplina que se ocupa de la identificación, captura, recuperación, compartimiento y evaluación del conocimiento organizacional. Ha sido identificada como un nuevo enfoque gerencial que reconoce y utiliza el valor más importante de las organizaciones: el hombre y el conocimiento que éste posee y aporta. Esto quiere decir que gestionar el factor humano, y sus competencias contribuye a la realización de una gestión del conocimiento eficiente y orientada a la toma de decisiones y objetivos de la organización.

Por otra parte, la falta de significado que maneja la web actual definida en HTML (lenguaje de marcas de hipertexto) dificulta la búsqueda eficiente de información.

La web semántica también conocida como la "web de los datos", se fundamenta principalmente en que a partir de un conjunto de datos se describe a otros datos, tanto semánticos como ontológicos para ser evaluados de manera automática por diferentes equipos de procesamiento, asignándoles características de inteligencia, para que puedan realizar búsquedas deseadas sin ser operados por personas [5].

Para encontrar un significado claro, la web semántica hace uso de ontologías. Una ontología según Gruber [6] es una especificación explícita de una conceptualización, en donde una conceptualización es una visión abstracta y simplificada del mundo que se quiere representar para algún propósito, construida mediante la identificación de los conceptos relevantes a esa representación (normalmente un dominio del conocimiento).

Una ontología es dinámica si esos conceptos relevantes de ese dominio se actualizan con el nuevo conocimiento [2]. Una Ontología se compone de elementos tales como clases (representan los conceptos del dominio), subclases (elementos derivados de las clases que heredan las características de la clase del padre), Propiedades (permiten establecer relaciones entre clases), rango (define el objeto que es afectado por la propiedad), Dominio: define el Sujeto que será definido por la propiedad e instancias (son los elementos que conforman una determinada clase).

Para la construcción ontologías existen varios lenguajes, el más reciente desarrollado por la W3C es el OWL (Web Ontology Language) el cual provee muchas facilidades y añade mayor semántica que el RDF (Resource Description Framework). OWL es un lenguaje desarrollado y recomendados por la organización W3C para la construcción de ontologías. Este lenguaje provee un mayor conjunto de primitivas para representar el significado de los elementos y sus relaciones del dominio de la ontología [7]. El OWL se puede formular en RDF, por lo que se considera una extensión de este. Además, OWL añade más vocabulario para describir propiedades y clases: entre otros, las relaciones entre las clases (por ejemplo, disyunción), cardinalidad (por ejemplo, "exactamente uno"), la igualdad, más rico escribiendo propiedades, características de las propiedades (simetría, por ejemplo), clases enumeradas e incluye toda la capacidad expresiva mediante lógica descriptiva generada automáticamente por un razonador [8].

Finalmente, una herramienta es débilmente acoplada con un gestor de bases de datos (SGBD) cuando todos los componentes se encuentran en una capa externa al SGBD, y su integración con este se hace a partir de una interfaz de conexión, cuya función, en la mayoría de los casos se limita a los comandos "leer de "y "escribir en" [9]. Mientras el SGBD provee el almacenamiento persistente, la mayoría de procesamiento de datos se realiza en la herramienta por fuera del motor del SGBD [9].

## 3. Arquitectura

MASKANA es una herramienta de gestión de conocimiento soportada en una ontología dinámica, débilmente acoplada con un gestor de base de datos que le permite la búsqueda inteligente y recuperación eficiente de los documentos de los trabajos de grado los cuales se encuentran almacenados en un repositorio general en la biblioteca "Alberto Quijano Guerrero" de la Universidad de Nariño.

La arquitectura de MASKANA es modular, compuesta por tres grandes módulos: el módulo de interfaz gráfica de usuario GUI, el módulo KERNEL y el módulo de conexión como se muestra en la figura 1.

### 3.1 Módulo de interfaz gráfica de usuario GUI.

La interfaz Gráfica de Usuario GUI (del inglés *Graphic User Interface*), da soporte grafico a todos los demás módulos que lo requieran y se encarga de presentarle al usuario una interfaz amigable. Permite al usuario hacer uso de la herramienta de búsqueda para realizar consultas relacionadas con el dominio establecido por la ontología. Establece una conexión en doble vía entre el submódulo de Gestor de Información del módulo Núcleo y el GUI, En la figura 2 se muestra la interfaz de búsqueda de MASKANA, y la figura 3 se aprecia la interfaz de presentación de resultados de una búsqueda.

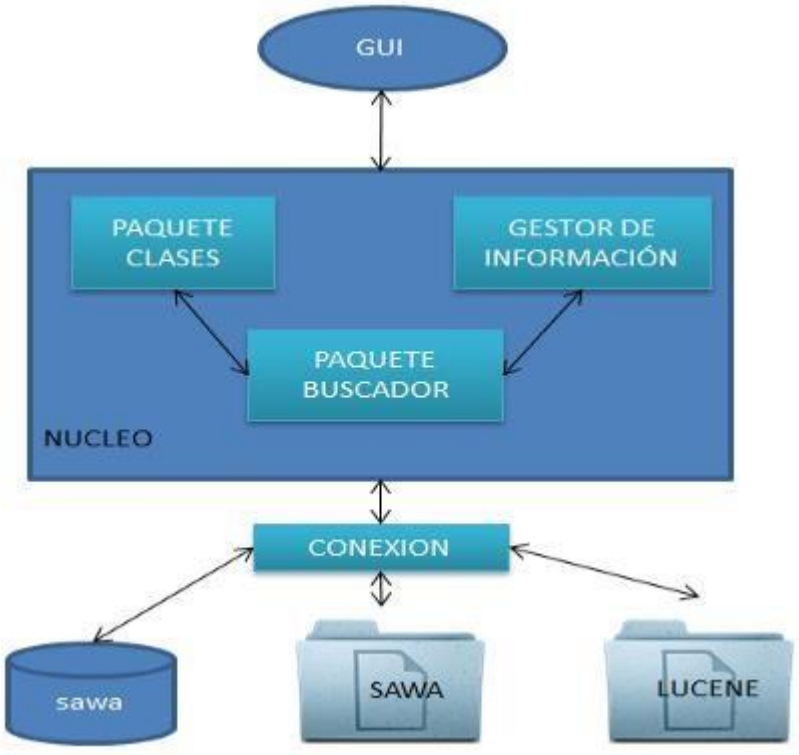


**Fig. 1.** Arquitectura de MASKANA

### 3.2 Módulo núcleo.

En este módulo se encuentran los paquetes principales para la búsqueda inteligente y la recuperación de documentos digitales: paquete de clases, paquete gestor de información y paquete buscador. En ellos están los algoritmos de búsqueda de información, de procesamiento de información, de presentación de resultados, procesos de gestión de información tanto del dominio establecido por la ontología, así como también de usuarios que administraran la herramienta de búsqueda.

- *Entidades gestoras de conocimiento Clases:* Este submódulo contienen las clases que representan la ontología y cada una de las tablas de la base de datos, que posteriormente servirán para gestionar el conocimiento con la ayuda de la herramienta, ofreciendo un orden claro y lógico de la representación de la ontología mediante las clases. En este submódulo se representa la base del conocimiento de las clases del modelo y tiene una relación directa de doble dirección con el paquete de Búsqueda.

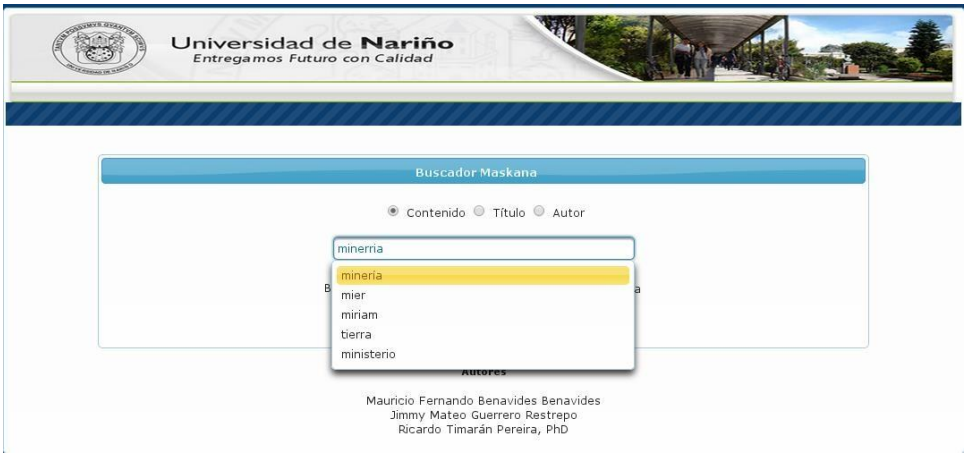


**Fig.2.** Interfaz de búsqueda Maskana


**Fig. 3.** Interfaz de resultados de búsqueda

- *Gestor de información:* Este submódulo se encarga de gestionar información entre la Interfaz Gráfica de Usuario GUI con el submódulo de búsqueda, y de esta manera establecer la comunicación entre el módulo Núcleo y el módulo GUI.
- *Paquete búscador:* Este submódulo se encarga de albergar las clases que realizaran el proceso de la recuperación de información correspondiente a la solicitud de búsqueda generada por el usuario. Este submódulo es de gran importancia ya que se encarga de realizar las consultas sobre la ontología haciendo uso del lenguaje SPARQL [10] o sobre el Índice, dependiendo del tipo de consulta solicitada por el usuario. Este submódulo tiene comunicación directa y de doble sentido con los dos submódulos más que componen el módulo Núcleo, permitiendo así el intercambio de información.

### 3.3 Módulo de conexión

Este módulo es el encargado de realizar la respectiva conexión entre la base de datos, la ontología y el índice de búsqueda y el resto de módulos de la herramienta y de esa manera proveer al usuario información con características de persistencia. Con este módulo se da respuestas eficientes a las diferentes consultas ingresadas por el usuario. Se conecta con los siguientes elementos:

- *Base de datos:* Aquí se almacena el diccionario, vocabulario o glosario de términos que represente la terminología de la ontología, así como la definición de los roles y permisos para los diferentes usuarios de la herramienta.
- *Ontología:* Se almacena la ontología del dominio de los trabajos de grado donde se encuentra el conocimiento y las relaciones del dominio.

- *Índice:* En este elemento se almacena todo el contenido textual indexado en un sistema de archivos para su eficiente recuperación teniendo en cuenta documentos en los diferentes formatos digitales (pdf, doc, txt, entre otros) según sea el caso.

## 4. Aspectos de Diseño e Implementación de MASKANA

La herramienta MASKANA se diseñó utilizando la metodología ágil SCRUM [11], utilizando notación UML, siguiendo el patrón de arquitectura de software MVC (modelo de vista controlador) con el framework *Java Server Faces*, el cual facilita la mantenibilidad de la herramienta a futuros cambios. Su diseño es modular con el fin de permitir el crecimiento continuo de esta herramienta. MASKANA contiene algoritmos de búsqueda basados en el lenguaje SPARQL que se encargan de consultar la ontología Sawa y extraer el conocimiento acerca de una consulta. MASKANA se encuentra en un repositorio de *Gibhub* para el control de versiones. En la figura 4 se presenta el diagrama de paquetes del módulo principal denominado "Clases" de la herramienta.

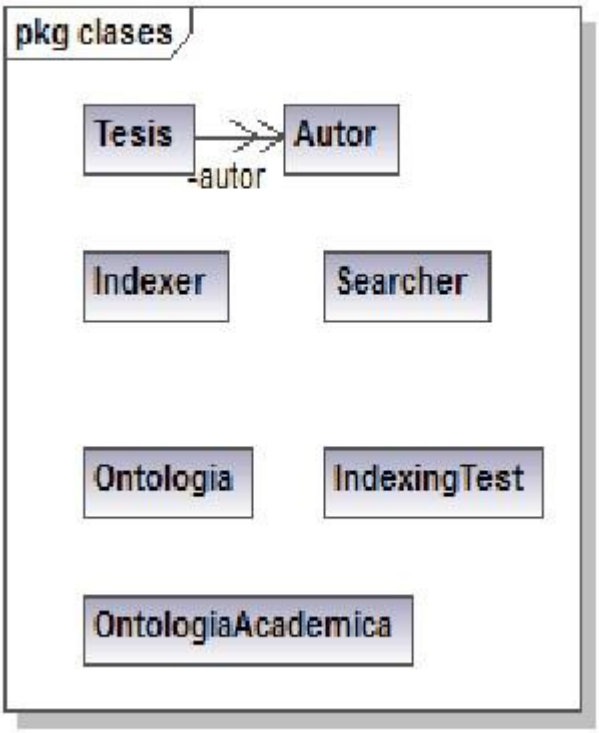


**Fig.4.** Paquete de clases

MASKANA se desarrolló bajo el lenguaje de programación Java™ en su versión 1.7, lo que la convierte en una herramienta independiente a la plataforma donde se ejecute. Para su construcción, se usaron herramientas de software libre tales como el IDE de desarrollo *Netbeans* en su versión 7.2 , *JSF* con la biblioteca *Primefaces* en su versión 5 , *API PDFBOX* para la extracción de archivos con formatos .pdf, el sistema gestor de bases de datos *PostgreSQL* 9 con sus extensiones *PGSIMILARITY* [12] para determinar similitud entre cadenas ,la librería *JENA* para la gestión de ontologías [13] y finalmente se utilizó la librería *LUCENE* para administrar el índice de búsqueda [14].

## 5. Pruebas de Funcionalidad

Para la ejecución de las pruebas se realizaron 7 casos de búsqueda y se analizaron los resultados obtenidos en el gestor de conocimiento MASKANA y en el sistema transaccional de la biblioteca "Alberto Quijano Guerrero" de la Universidad de Nariño. La tabla 1 muestra los casos de prueba aplicados al sistema de la biblioteca Alberto Quijano Guerrero.

**Tabla 1.** Casos de prueba

| Prueba | Biblioteca Udenar |
|---|---|
| **C1** | Búsqueda por título completo |
| **C2** | Búsqueda por autor completo |
| **C3** | Búsqueda por un nombre y un apellido |
| **C4** | Búsqueda por palabras contenidas en el título |
| **C5** | Búsqueda por error de digitación en el título |
| **C6** | Búsqueda por error de digitación en el autor |
| **C7** | Búsqueda por contenido |

Las pruebas se hicieron llevando los siguientes casos de prueba y se calificó como éxito (1) o fracaso (0), teniendo en cuenta que el éxito se lo califica si la búsqueda a realizar esta en los quince primeros resultados.

Para el caso de prueba C1, búsqueda por título completo, en las 5 iteraciones los dos sistemas produjeron resultados concretos obteniendo el 100% de éxito en las consultas y se realizó enviando el titulo exacto (tildes, signos de puntuación y comillas) tal y como se almacena en la base de datos. Los resultados de esta prueba se pueden aprecia en la figura 5.

Para el caso de prueba C2, búsqueda por autor completo, se realizó consultando los nombres y apellidos del autor completos como se encuentran registrados en base de datos, obteniendo como resultado que MASKANA tiene un 100% de éxito, mientras que el sistema de la biblioteca tiene un 60% de éxito, como lo muestra la figura 5 para la prueba C2.

Para el caso de prueba C3, búsqueda por un nombre y un apellido, se envió únicamente un nombre y un apellido tal y como aparecen en la base de datos, dando como resultado que MASKANA arroja un 80% de éxito, mientras que el sistema de la biblioteca tiene un 40% de éxito, como lo muestra la figura 5 para la prueba C3.

Para el caso de prueba C4, búsqueda por palabras contenidas en el título, se realizó enviando la consulta con palabras que estén contenidas en el título, las palabras podían haberse enviado en el orden del título como en desorden, dando como resultado que con MASKANA se obtiene 100% de éxito y el sistema de la biblioteca tiene un 20% de éxito, como lo muestra la figura 5 para la prueba C4.

Para el caso de prueba C5, búsqueda por error de digitación en el título, se realizó enviando la consulta en con uno o dos errores de digitación, dando como resultados que para el sistema MASKANA obtiene un 100% de éxito y para el sistema de la biblioteca hay un 0%, como se muestra en la figura 5 para esta prueba.

Para el caso de prueba C6, búsqueda por error de digitación en el autor, se realizó enviando uno o dos errores de digitación en el autor, dando como resultado que para el sistema MASKANA se obtuvo un 80% de éxito mientras que para el sistema de la biblioteca un 0%, como lo muestra la figura 5 para la prueba C6.

Finalmente, para el caso de prueba C7, búsqueda por contenido, se realizaron las consultas tomando partes textuales del contenidas de los documentos PDF, dando como resultado que MASKANA obtuvo un 100% de éxito y el sistema de la biblioteca 0%, como lo muestra la figura 5 para la prueba C6.

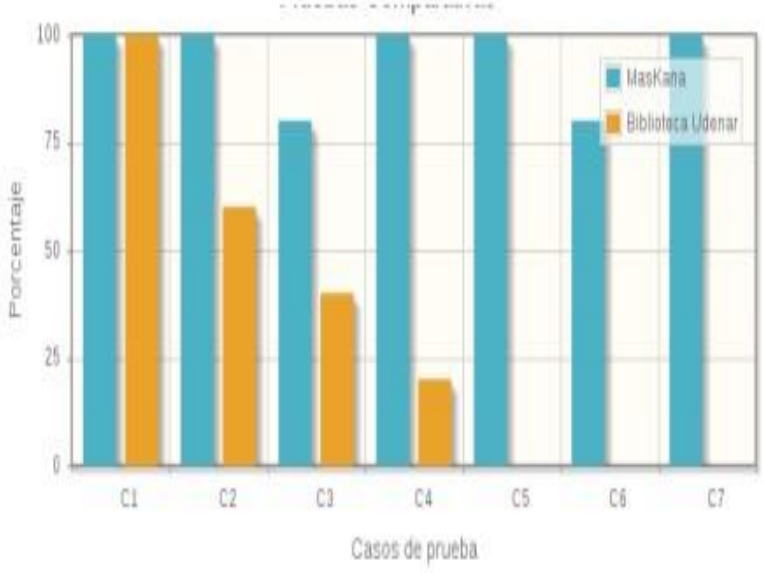


**Fig. 5.** Pruebas de funcionamiento MASKANA

## 6. Conclusiones y trabajos futuros

Como resultado de este proyecto de investigación, la biblioteca "Alberto Quijano Guerrero" de la Universidad de Nariño cuenta con una herramienta que permite la búsqueda inteligente y recuperación de los documentos de trabajos de grado que se encuentran en formato digital, gracias a que esta soportada en una ontología dinámica y débilmente acoplado con el sistema gestor de base de datos PostgreSQL.

Las pruebas realizadas a MASKANA demostraron que ésta herramienta es más eficiente que el sistema transaccional de búsqueda que actualmente posee la biblioteca de la Universidad de Nariño.

Todo el código fuente de la herramienta inteligente MASKANA es libre y se encontrará disponible en el servidor del Grupo de Investigación Aplicado en Sistemas GRIAS, del Departamento de Sistemas de la Facultad de Ingeniería de la Universidad de Nariño.

Como trabajos futuros están el implementar técnicas de minería web a la MASKANA que permitan descubrir patrones acerca de las búsquedas y recuperación de trabajos de grado que realizan los usuarios, con el fin de evitar futuros plagios. También se pretende implementar a esta herramienta otro tipo de ontologías que permita a las organizaciones y especialmente a las PYMES (acrónimo de Pequeñas y Medianas Empresas) y MyPYMES (acrónimo de Micro, Pequeñas y Medianas Empresas), gestionar su conocimiento y de esta manera volverlas más competitivas.

## Agradecimientos

Este proyecto es financiado por la Gobernación del Departamento de Nariño, dentro del programa de Jóvenes Investigadores.

## Referencias

1. Cabrera, O., Guerrero, J., Benavides, M., Timarán, R.: Un acercamiento a la construcción de ontologías con la herramienta libre Protegé. En: Revista Ventana Informática, No. 30 (enejun), Facultad de Ciencias e Ingeniería, Universidad de Manizales, pp. 233-246,ISSN:01239678, Manizales, Colombia (2014).
2. Cabrera, O., Guerrero, J., Benavides, M., Timarán, R.: SAWA: Ontología para la Gestión de
   Conocimiento sobre Trabajos de Grado. En: Revista Ontare, Vol. 1, No. 2, Facultad de Ingeniería, Universidad Escuela de Administración de Negocios, pp. 183-214, Bogotá, Colombia (2013).
3. Benavides, M., Guerrero, J.: UMAYUX: Un modelo de software de gestión de conocimiento soportado en una ontología dinámica débilmente acoplado con un gestor de base de datos para la Universidad de Nariño. Informe final de trabajo de grado, Programa de Ingeniería de Sistemas, Departamento de Sistemas, Facultad de Ingeniería, Universidad de Nariño, Pasto, Colombia (2014).
4. Medina, J. L. La Gestión del Conocimiento en las Organizaciones. Libros En Red, (2001).
5. World Wide Web Consortium (W3C). Guía breve de web semántica. Información sobre la web semántica.
6. Gruber, T. Toward Principles for the Design of Ontologies used for knowledge Sharing. In Formal Ontology in Conceptual Analysis and Knowledge Re-presentation, Kluwer Academic Publishers, In Press. Substantial Revision Of Paper Presented At The International Workshop On Formal Ontology. Kluwer Academic Publishers, (1993).
7. H. A. F. Fernandez, "CONSTRUCCIÓN DE ONTOLOGÍAS OWL," Vínculos, vol. 4, no.
   1, pp. 19–34, Julio. 2013.
8. C. B. Santos and M. Á. R. Duque, Sistemas interactivos y colaborativos en la web. Universidad de Castilla La Mancha, 2004.
9. Timarán, R. Arquitecturas de Integración del Proceso de Descubrimiento de Conocimiento con Sistemas de Gestión de bases de datos: un Estado del Arte. En: Revista Ingeniería y Competitividad. Universidad del Valle. Volumen 3. No. 2. Cali, Colombia, (2001).
10. Segaran, T., Evans, C.,Taylor, J.:Programming the Semantic Web. O'Reilly Media, Inc., (2009).
11. Schwaber, K., Sutherland, J.: La guía de SCRUM: las Reglas de Juego, 2013. Disponible en http://www.scrumguides.org/docs/scrumguide/v1/Scrum-Guide-ES.pdf.
12. Winkler, J.: Algoritmo para Similitud de Palabras en Cadenas de texto, 1990. [En Línea] http://stackoverflow.com/questions/2848807/optimizing-jaro-winkler-algorithm.
13. Carroll, J. J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., Wilkinson, K.: Jena: Implementing the Semantic Web Recommendations. In Proceedings of the 13th

International World Wide Web Conference on Alternate Track Papers & Posters (pp. 74–83). New York, NY, USA: ACM. http://doi.org/10.1145/1013367.1013381, 2004.
14. McCandless, M., Hatcher, E., Gospodnetic, O.: Lucene in Action, Second Edition: Covers Apache Lucene 3.0. Greenwich, CT, USA: Manning Publications Co, 2010.