Prototipo de un Sistema de búsqueda de documentos basado en anotado semántico

Azul Rossini[1], Mauro Juarez[1], Emiliano Reynares[1], María Rosa Galli[1,2], María Laura Caliusco[1]

[1] CIDISI-UTN – Facultad Regional Santa Fe, 3000, Santa Fe, Argentina
[2] INGAR-UTN -CONICET, 3000, Santa Fe, Argentina
mazulrossini@gmail.com, mauroojuarez@gmail.com, ereynares@frsf.utn.edu.ar, mrgalli@santafe-conicet.gov.ar, mcaliusc@frsf.utn.edu.ar

Resumen. Gran parte del conocimiento presente en las empresas se encuentra almacenado en objetos de conocimiento. Estos objetos deben ser administrados por una plataforma de gestión de conocimiento en su representación original a fin de facilitar su correcto acceso. Para ello, se debe capturar la semántica de su contenido y proveer herramientas para realizar búsquedas inteligentes. En este contexto, se estableció una vinculación entre el CIDISI, la Fundación Sadosky y la empresa Accion Point S.A. para desarrollar una plataforma prototipo que facilite la búsqueda de documentos asociados a los productos que comercializan. Este trabajo describe la experiencia de un proceso de Desarrollo de un Sistema de gestión de documentos basado en una red de ontologías.

Palabras Claves: Gestión de Conocimiento, Red de Ontologías

1 Caracterización General del Proyecto

1.1 Instituciones y Empresas Participantes

El proyecto se desarrolla en el marco de la convocatoria Fase Cero - 2019, entre el Área de Vinculación Tecnológica de la Fundación Dr. Manuel Sadosky de Investigación y Desarrollo en las Tecnologías de la Información y Comunicación, la empresa Acción Point S.A. y el Centro de Investigación CIDISI de la Univ. Tecnológica Nacional, Fac. Regional Santa Fe.

Accion Point S. A. (AP) es un grupo de empresas con más de 150 profesionales especializados en el desarrollo de soluciones de negocios. AP es partner de DL&A, desarrolladores del core bancario BanTotal[@], y de Genexus[@], herramienta de desarrollo multiplataforma. Pertenece al Polo Tecnológico Rosario y la cámara CECCI. AP cuenta con un área de I+D+I (coordinada por un doctor en ingeniería) cuyos principales objetivos incluyen brindar capacitación en nuevas tecnologías, incorporar tecnologías innovadoras a sus desarrollos y optimizar sus servicios a través de desarrollos internos y proyectos de vinculación con organismos de Investigación y Desarrollo.

CIDISI es un centro de Investigación y Desarrollo de Ingeniería en Sistemas de

Información con sede en la ciudad de Santa Fe. Los integrantes del proyecto poseen experiencia y conocimiento en temáticas de: Gestión de conocimiento, tecnologías semánticas e ingeniería de Software.
La Fundaciòn Dr. Manuel Sadosky es una institución pública privada cuyo objetivo es favorecer el nexo entre la estructura productiva y el sistema científico–tecnológico en todo lo referido a la temática de las TICs. El instrumento de Financiamiento de Fase Cero tiene el objetivo de promover la realización de primeras experiencias de proyectos de I+D+i en colaboración entre grupos de I+D y empresas TIC nacionales.

1.2 Personas Participantes

El grupo de investigación está formado por 5 integrantes del CIDISI: 3 docentes-investigadores (UTN-CONICET) - Dra. María Laura Caliusco, Dra. María Rosa Galli y Dr. Emiliano Reynares-, y dos alumnos avanzados de la carrera Ingeniería en Sistemas de Información (UTN-FRSF) Azul Rossini y Mauro Juarez. Por el lado de ACCION POINT S.A., el grupo está integrado por 2 personas: el Dr. Carlos Toledo y el Ing. Sergio Lembo.

1.3 Tipo de Interacción

El tipo de interacción se corresponde con la Colaboración en I+D entre empresa y universidad.

2 Detalles de Ejecución del Proyecto

2.1 Objetivos

Este proyecto tiene como propósito elaborar un prototipo de una plataforma tecnológica de Gestión de Conocimiento que incluya estrategias para el anotado semántico y la búsqueda de objetos de conocimiento relativos al sistema BanTotal. BanTotal es un core bancario formado por sistemas base para la operación de Instituciones Financieras (clientes, contabilidad, impuestos, transacciones, etc.). Es adaptable a diferentes requerimientos del negocio considerando que debe reflejar toda la normativa cambiaria y financiera que establece el banco central de cada país en el cual se implementa.
Como BanTotal es un sistema muy amplio se decidió, para esta fase, sólo considerar el subdominio de ANSES. BanTotal en Argentina debe ser adaptado para incorporar las cuentas bancarias que otorga ANSES en sus programas de ayuda económica. El sistema de consulta a desarrollar será usado por los stakeholders involucrados en la implementación del Módulo ANSES durante el desarrollo de cualquier proyecto de implementación de BanTotal en una entidad bancaria de Argentina. En general, los posibles usuarios serán expertos en ingeniería de software quienes implementan BanTotal y probablemente cuenten con escaso conocimiento en el subdominio ANSES

2.2 Actividades Realizadas

El proyecto se dividió en tres etapas con sus correspondientes entregables entre las

cuales se distribuyen cuatro actividades principales:
Actividad 1: Selección de la documentación relevante que contiene el conocimiento del dominio y caracterización de las fuentes de información.
Actividad 2: Construcción de un modelo semántico del dominio.
Actividad 3: Construcción de una base de conocimiento compuesta por el modelo semántico desarrollado y el anotado semántico de los documentos seleccionados.
Actividad 4: Implementación de la plataforma para la búsqueda de documentos.

3 Resultados del Proyecto

Esta sección resume los resultados obtenidos en cada actividad. Cabe aclarar que el proyecto aún se encuentra en desarrollo.

3.1 Resultados de cada Actividad

Actividad 1: AP compartió un repositorio con alrededor de 100 documentos relativos al dominio ANSES de Bantotal. Algunos documentos estaban almacenados dentro de carpetas que les daban una cierta clasificación, pero en su mayoría eran documentos aislados, sin una relación establecida, ni una clasificación determinada y con diversos formatos (.doc, .pdf, .txt, .zip, .rar, .jpg). Algunos archivos contenían archivos comprimidos con a su vez más archivos comprimidos de forma anidada. Algunos de dichos archivos tenían extensión desconocida y contenido incomprensible para el ser humano. Como resultado de una primera selección, se obtuvieron cerca de 70 documentos que tenían una extensión ordinaria o conocida. Luego, se analizó el contenido de estos documentos y se seleccionaron cerca de 50 considerados relevantes. Simultáneamente se realizó una clasificación de los documentos, según su contenido se refiriera al dominio ANSES o BanTotal. A tal efecto, se crearon localmente dos directorios con sub-carpetas que referenciaban a características específicas de los dominios ANSES y BanTotal donde se almacenaron los documentos analizados. Esto facilitó notoriamente las actividades posteriores.

Actividad 2: Del análisis de los documentos se observó que los mismos no seguían una estructura establecida, sino que cada uno tenía su propia forma de estructurar la información. Además, se detectó que muchos de los documentos referenciaban una fase específica del desarrollo de software. Entonces, con el fin de enriquecer la especificación de los documentos y la clasificación de los mismos, se decidió construir un modelo semántico compuesto por un conjunto de ontologías interrelacionas:
1. Ontología ANSES: formada por los conceptos propios del MODULO ANSES de BANTOTAL. 2. Ontología BANTOTAL: formada por los conceptos básicos de toda operatoria financiera que se ve reflejada en los requerimientos funcionales que soporta BANTOTAL. 3. Ontología de Ingeniería de Software: con los conceptos propios de un proceso de desarrollo de software. 4. Ontología de Documentos: con los conceptos que describen la estructura de los documentos involucrados.
Las dos primeras ontologías se desarrollaron analizando los conceptos incluidos en los documentos, principalmente orientado a determinar la temática específica de los mismos. Así se identificaron las particularidades de los respectivos dominios y se

crearon sendas redes de conceptos a partir de los cuales se desarrollaron las estructuras taxonómicas que formaron cada ontología. Para la ontología de Ingeniería de Software, si bien en la literatura existen diferentes ontologías que describen semánticamente los distintos procesos involucrados en el desarrollo de un sistema de información [1], se decidió crear una ontología nueva, dado que el dominio de trabajo es único y poco similar a otros dominios que se pueden considerar "generales". De esta manera el diseño acabó siendo más sencillo y específico.
La ontología de Documentos se obtuvo a partir del modelo de Datos México [2] basado en el estándar CIDOC-CRM [3], eliminando aquellos conceptos no relevantes para el dominio del trabajo.

*Actividad 3***:** El proceso de anotado semántico en la red de ontologías se realizó en forma manual. Los documentos, se describieron mediante la ontología de Documentos como instancias de las clases Documento y Objeto Digital. Para cada concepto (clase) de cada una de las restantes ontologías, se instanciaron individuos que corresponden a las mismas y se instanciaron propiedades que relacionarán a futuro la instancia de cada concepto, con los documentos propiamente dichos.
Dado que en muchos documentos la información no está organizada en forma estandarizada ni sigue un patrón determinado, no existe una herramienta automática que se adapte, y el proceso de anotado se realizó en forma manual asistida utilizando la herramienta OpenSemanticSearch (https://www.opensemanticsearch.org/).

Actividad 4: Para el diseño e implementación del algoritmo de búsqueda se utilizó la herramienta OpenSemanticSearch, dado que siendo una tecnología Open Source puede adaptarse al dominio particular del presente proyecto. Esta actividad está aún en etapa de implementación y prueba para lograr los objetivos propuestos.

3.2 Evaluación de los Resultados y Lecciones Aprendidas

El prototipo desarrollado constituye la fase inicial de un proyecto a mayor escala que involucre los distintos módulos del core BanTotal. La principal lección aprendida con este trabajo ha sido comprobar que la etapa crítica en este tipo de desarrollos consiste en el proceso previo de análisis de documentos y la posterior construcción de la/las ontologías adecuadas. Esta etapa requiere de una especial habilidad y capacidad de análisis. Existen herramientas adaptables a diferentes dominios de conocimiento que permiten implementar sistemas de recuperación semántica de documentos, pero la efectividad de los sistemas resultantes depende casi exclusivamente de la calidad del modelo semántico construido y del anotado semántico de los objetos de conocimiento.

Reference

[1] Bhatia, M. P. S. ; Kumar, A.; Beniwa, R. (2016) Ontologies for Software Engineering: Past, Present and Future. Indian Journal of Science and Technology,Vol 9(9).
[2] Secretaría de Cultura, Dirección General de TICs. (2018). MEXICANA. Repositorio del Patrimonio Cultural de México. Accesible en: https://mexicana.cultura.gob.mx/
[3] Le Boeuf, P., Doerr, M., Emil Ore, C., Stead, S (2018). CIDOC Conceptual Reference Model. Produced by the ICOM/CIDOC. Accesible en: http://www.cidoc-crm.org/Version/version-6.2.3