Módulo basado en tecnologías semánticas para realizar búsquedas de documentos en la plataforma Signar.

María Laura Caliusco[1], Agustín Martínez[2] and Graciela Brusa[2]

[1] CIDISI-UTN – Facultad Regional Santa Fe, 3000, Santa Fe, Argentina
[2] Lyris IT SAS, 3000, Santa Fe, Argentina
mcaliusco@frsf.utn.edu.ar, mrtnz.agustin@gmail.com, gbrusa@-lyris.com.ar

Resumen. El objetivo del presente trabajo es mostrar la incorporación de tecnologías semánticas para realizar búsquedas de documentos. En este caso, esa búsqueda de documentos es realizada en la plataforma Signar de la empresa Lyris IT S.A.S. Este trabajo muestra cómo investigaciones que se llevan a cabo en una Universidad aportan valor agregado a empresas para potenciar sus productos y hacerlos más competitivos en el mercado internacional, incorporando tecnologías innovadoras.

Palabras claves: gestión documental, tecnologías semánticas

1 Introducción

Lyris IT SAS es una PyME de base tecnológica de la ciudad de Santa Fe especializada en tecnologías de la información y la comunicación (TIC), específicamente, en procesos de gestión documental con valor jurídico. Lyris IT desarrolló una plataforma de software, denominada SIGNAR, que permite la gestión de documentos digitales que fluyen en los diferentes procesos administrativos y de gestión, tanto en organizaciones públicas como privadas, mediante la aplicación de firma digital/electrónica.

En el marco de la pandemia, la necesidad de realizar tareas a distancia aumentó fuertemente debido a las medidas de cuidado y distanciamiento. En este sentido, el servicio de firma ofrecido por Lyris IT SAS se volvió de suma utilidad. La firma digital es una tecnología que permite firmar documentos electrónicos con las mismas propiedades que tiene un documento firmado en papel. Sus principales ventajas son la no presencialidad, la reducción de tiempos y costos vinculados al transporte de la documentación y la despapelización efectiva que surge al ser equiparada totalmente a la firma manuscrita por la legislación de diferentes países en la materia.

La plataforma SIGNAR contempla la gestión de certificados de firma electrónica, recibos digitales, procesos digitales y documentos electrónicos, proceso de aplicación y verificación de firmas digitales/electrónicas. Dentro de las funcionalidades provistas

por la plataforma SIGNAR se detectó la necesidad de mejorar la funcionalidad de las búsquedas de documentos permitiendo contar con una gestión documental inteligente basada en metadatos semánticos [1].

Para cumplir con el objetivo planteado anteriormente, se formuló un proyecto cuyo aspecto diferencial es que cuenta con la capacidad de realizar búsquedas inteligentes de documentos a partir de la semántica de su contenido. Esta funcionalidad innovadora se provee desde la ingeniería ontológica utilizando herramientas de búsquedas basadas en el contexto y a través del lenguaje convencional interactuando con los vocabularios controlados que existen en los diferentes dominios de aplicación. La Agencia I+D+i consideró estratégico impulsar este proyecto, a través del Fondo Argentino Sectorial (FONARSEC), a partir del aumento en la demanda de los tipos de servicios ofrecidos por la plataforma SIGNAR. Para llevar adelante dicho proyecto se incorporaron investigadores en la temática que trabajan en el Centro de Investigación de la UTN-Facultad Regional Santa Fe.

2 Organización del Proyecto

2.1 Instituciones y Empresas Participantes

Lyris IT SAS[1] es una empresa que ofrece soluciones para la gestión de documentos electrónicos, mediante la implementación de firma digital/electrónica cuando se requiera. Esta empresa brinda soluciones para administrar procesos con documentos digitales que permitan a su vez, contar con un archivo digital seguro y con validez jurídica, lo que deriva en procesos de despapelización efectiva en las organizaciones donde se apliquen.

CIDISI es un centro de Investigación y Desarrollo de Ingeniería en Sistemas de Información perteneciente a la UTN – Facultad Regional Santa Fe. Los integrantes del proyecto poseen experiencia y conocimiento en temáticas de: Gestión de conocimiento, tecnologías semánticas e ingeniería de Software.

2.2 Tipo de Interacción

El tipo de interacción se corresponde con la Colaboración en I+D entre empresa y universidad. Se logró una sinergia de trabajo muy buena con la empresa que estuvo muy receptiva con el proyecto, y por lo tanto el mismo pudo ser llevado a cabo en tiempo y forma. Si bien se pidió una extensión del proyecto, la misma fue para mejorar lo ya realizado.

[1]https://lyris.com.ar

3 Ejecución del Proyecto

3.1 Análisis Preliminar

Durante la etapa de análisis preliminar, que antecede a la definición de los requerimientos, se llevaron a cabo actividades tendientes a realizar un análisis del dominio de trabajo. El objetivo de esta etapa fue la de analizar la Plataforma Signar y sus fuentes de información para comprender el dominio de aplicación del módulo a desarrollar. La segunda actividad que se realizó tuvo como propósito analizar el dominio del anotado semántico y algunas técnicas que se utilizan en dicho dominio y que son importantes para definir el alcance y usos de la aplicación que a desarrollar.

3.1.1. Análisis de la Plataforma Signar

Esta actividad refiere al análisis tanto de tecnologías utilizadas como de funcionalidades presentes en la plataforma Signar. Los objetivos de esta actividad fueron:

- **Relevamiento de tecnologías:** Relevar cuáles son las tecnologías utilizadas en la plataforma para coordinar luego la implementación de estándares, modelos y tecnologías que sean compatibles con las mismas a la hora de especificar requerimientos e implementar soluciones.
- **Relevamiento de funcionalidades:** Detectar los flujos de trabajo presentes en el sistema. Principalmente: cómo interviene el modelado de la información en el almacenamiento de metadatos y cómo se realizan las búsquedas de documentos actualmente en la plataforma.

Durante esta etapa se mantuvieron reuniones periódicas, tanto con el equipo de tecnologías de información y análisis funcional como con el área directiva de Lyris IT para diferentes tareas: 1) Relevamiento de requerimientos generales y específicos para la consultoría, 2) Análisis y corrección de posibles desvíos que se produzcan en la consultoría, 3) Relevamiento de información asociada a la plataforma y al esquema de datos y 4) Solicitud de cambios para adaptabilidad de la plataforma a fines de implementar requerimientos.

Debido a las restricciones impuestas por la pandemia de Covid 19, todas las reuniones se llevaron a cabo en forma virtual. La virtualidad en este caso nos dio otra forma de trabajar sin ser un impedimento para llevar a cabo las tareas programadas. Más aún, generó una comunicación más fluída con la empresa dado que las reuniones se podían organizar en franjas horarias más amplias y con más duración.

Para realizar el análisis de la Plataforma Signar, se solicitó a Lyris IT un dump SQL con datos y esquema asociado a la base de datos del sistema. A partir de la misma se realizó un análisis de: Dónde están ubicados cada uno de los datos/informacio-

nes relevantes para incorporar al modelo semántico y a la especificación de requerimientos; 2) Cómo están relacionadas las diferentes entidades y tablas que representan la información del sistema; 3) Qué restricciones poseen la relaciones; 4) Qué tipos de datos están asociados a cada una de las informaciones asociadas a documentos (cadenas de texto, enteros, flotantes, valores de verdad, etc) y 5) Qué datos de los indicados se actualizan, bajo qué condiciones y cuándo.

De esta etapa de relevamiento se concluyó que si bien la plataforma Signar tenía la capacidad de realizar búsquedas sobre los documentos cargados, se dejaba al usuario de la plataforma que los definiera libremente. Esto causaba que en la implementaciones realizadas no se definieron muchos metadatos. Por lo tanto, se concluyó que era necesario definir un modelo de metadatos que el usuario pueda usar de base y que además aplicando técnicas de IA ya se le pudiera sugerir al usuario un valor por defecto de alguno de ellos.

3.1.2. Estándares, Modelos y Tecnologías Semánticas

Esta actividad refiere al relevamiento y análisis tanto de estándares como de modelos y tecnologías semánticas asociadas y relevantes para el manejo de documentos y de metadatos asociados a los mismos [2].

Dentro de algunas posibilidades se enuncia una lista no taxativa y meramente descriptiva de estándares, modelos y tecnologías que resultaron oportunos para la consultoría y que fueron materia de análisis para actuales y futuras implementaciones: DublinCore[3], CIDOC CRM [4], ElasticSearch, Apache Lucene, Apache Tika, Spacy, GraphDB y BPMN. El análisis de dichas tecnologías fue el primer aporte que hizo el equipo de investigación ya que son tecnologías con las cuales se viene trabajando en los últimos años. Con lo analizado se hizo una capacitación al personal de Lyris IT.

3.2. Construcción de la base de conocimiento y anotaciones semánticas

3.2.1. Construcción del Modelo Semántico

Durante esta etapa de la consultoría se llevó a cabo el desarrollo de un modelo semántico, basado en ontologías, confeccionado a medida donde se representaron las terminologías y modelos que son relevantes para la plataforma Signar. Este es el segundo aporte que hizo el equipo de investigación que tiene basta experiencia en el desarrollo e implementación de modelos semánticos.

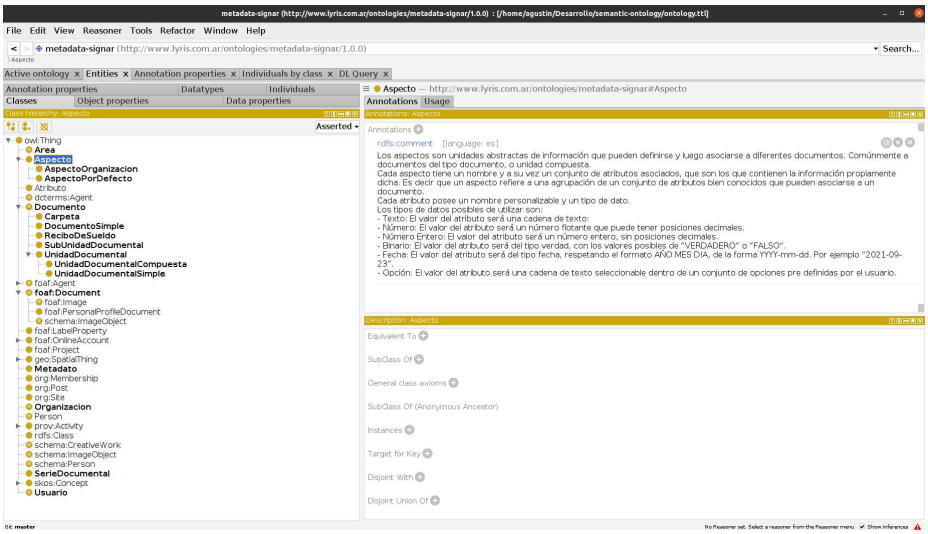
Para el desarrollo del modelo semántico se siguió la metodología NeOn basada en escenarios, definida para el diseño e implementación de redes de ontologías. En particular, se implementó el escenario que propone la reutilización de ontologías. A

partir del relevamiento de los requerimientos se identificaron los términos principales del dominio.

Una vez identificado los términos principales, sus propiedades y sus relaciones se procedió a modelar los mismos en la herramienta de edición de ontologías Protégé (https://protege.stanford.edu/). Dicha herramienta ofrece la posibilidad de guardar la ontología modelada en diferentes lenguajes, como RDF, OWL y Turtle.

Una vez definida la estructura principal de la ontología, se decidió enriquecerla con las ontologías estándares que se habían analizado. De esta forma, tenemos un modelo interoperable. Ese enriquecimiento se muestra a continuación.

Con el objetivo de enriquecer la ontología base desarrollada se procedió a reutilizar las ontologías estándares recomendadas por la W3C (https://www.w3.org/): FOAF, Organization Ontology y Dublin Core Elements. Para realizar el enriquecimiento, las mismas se importaron en el modelo anterior utilizando la herramienta Protégé. Protégé es un editor de ontologías de código abierto y un sistema de adquisición de conocimiento. Esta herramienta cuenta con el respaldo de una sólida comunidad de desarrolladores y usuarios académicos, gubernamentales y corporativos.



3.3. Pruebas y Ajustes en la búsquedas de interfaces

3.3.1. Arquitectura

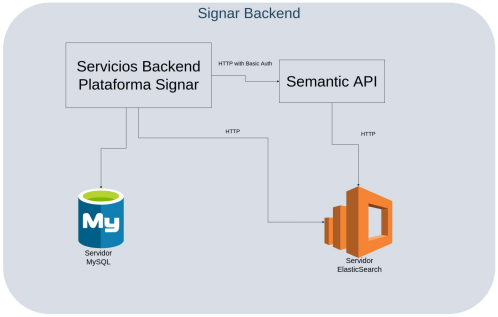
Tomando como base la arquitectura relevada con respecto a tecnologías y servicios de la plataforma Signar, se detalla en esta sección la arquitectura de la solución de la API desarrollada y a su vez cómo se integra en la arquitectura de Signar.

La arquitectura con la que consta la solución desarrollada se integra exclusivamente en la sección de servicios Backend de la presente arquitectura de Signar.

En detalle, la interacción con el API semántico desarrollado en Python, se dará exclusivamente desde el servicio de backend de Signar en una red privada y con seguridad mínima con autenticación básica http (incluída en el desarrollo y habilitada según configuración de entorno).

Por otro lado, si bien el componente API semántico desarrollado será el que se comunique en forma exclusiva con el servicio de ElasticSearch, no se descarta la posibilidad de una interacción directa entre los servicios backend de Signar y el servidor de ElasticSearch ya que pueden realizarse interacciones personalizadas, de una forma similar a la que hoy se interactúa con los servicios de MySQL.

Se deja un detalle a continuación de la arquitectura de la solución.



Para la implementación de la solución se utilizaron las siguientes tecnologías.

Elasticsearch
Elasticsearch es un servidor de búsqueda basado en Lucene. Provee un motor de búsqueda de texto completo, distribuido y con capacidad de multitenencia con una interfaz web RESTful y con documentos JSON.
Elasticsearch está desarrollado en Java y está publicado como código abierto bajo las condiciones de la licencia Apache.

Casos de uso

1. Búsqueda de información en una app o sitio web
2. Motor de almacenamiento para automatizar flujos de negocio
3. Machine learning, obtener información sobre grandes set de datos.
4. Manejar información geoespacial usando elasticsearch como un GIS.
5. Entre otros.

Python

Python es un lenguaje de programación interpretado cuya filosofía hace hincapié en la legibilidad de su código. Se trata de un lenguaje de programación multiparadigma, ya que soporta parcialmente la orientación a objetos, programación imperativa y, en menor medida, programación funcional.

Especificación OpenAPI

La especificación OpenAPI[2], originalmente conocida como la especificación Swagger, es una especificación para archivos de interfaz legibles por máquina para describir, producir, consumir y visualizar servicios web RESTful.

Todos los servicios web implementados en la solución fueron documentados con esta especificación.

Mapeo Índice Elasticsearch

Para la elaboración de los servicios de búsqueda y manipulación de documentos, se realizó un relevamiento de las fuentes de información y una propuesta de mapeo simple.

En la siguiente imagen se puede observar el mapeo definido con los diferentes tipos de datos elasticsearch, para cada uno de los atributos definidos en el mapeo simple.

Más detalles de definición de cada uno de los tipos de datos existentes en elasticsearch puede encontrarse en la siguiente dirección web:
https://www.elastic.co/guide/en/elasticsearch/reference/current/mapping-types.html

Este mapeo es el que habilita luego la inicialización del índice relacionado a los documentos en elasticsearch y a su vez es la clase Python que permitirá luego hacer uso de las funciones que nos brinda la librería de implementación para la creación, eliminación, modificación y búsqueda de documentos.

Servicios Web

Se detallarán en el presente apartado los diferentes servicios web implementados para cumplimentar las actividades:

1. Instanciación de ElasticSeach con dicho mapeo
2. Prueba y ajuste de mapeo Elasticsearch para anotado simple
3. Servicio web para ABM de documentos en índices
4. Desarrollo y prueba de servicio web para búsqueda simple
5. Desarrollo y prueba de servicio web para búsqueda avanzada

[2] https://www.openapis.org/

Inicialización de Índice de Documentos en Elasticsearch

Se utiliza este servicio web para la inicialización/instanciación del índice de documentos a partir de mapeo definido anteriormente.

Esto es necesario realizarlo cuando se instala e inicia elasticsearch por primera vez y el mapeo aún no existe en el servidor. A su vez, es necesario su utilización si por alguna razón se elimina el índice, por ejemplo para su reconstrucción.

Todos los servicios para alta, baja y modificación de documentos respetan una arquitectura API RESTful [5].

4 Discusión de los resultados y lecciones aprendidas

Se destaca el trabajo en equipo logrado a través de una interacción del sector privado con la Universidad, como primer punto relevante de lo aprendido. Este trabajo es posible y potencia las capacidades de las empresas de un modo efectivo. Y desde la óptica de la Universidad, permite plasmar en el medio en el que se inserta, todos aquellos conocimientos que desarrolla y expande de forma continua internamente, motivando a quienes día a día estudian e investigan pensando en sus posibles aplicaciones.

Por otra parte, desde el punto de vista técnico, se han sentado las bases para lograr una mejora inmediata en las búsquedas de documentos a través de Signar Gestor Documental pero también para continuar con otros proyectos derivados, tales como la extracción de información de dichos documentos para realizar un análisis posterior de la misma y generar conocimiento para la toma de decisiones. Queda ahora por parte de la empresa aplicar todo lo aprendido durante el desarrollo de este proyecto para seguir mejorando su producto.

References

1. Cuba Rodríguez, Y., & Olivera Batista, D. (2018). Metadata, search and information retrieval from the Information Science. *E-Ciencias de la Información*, *8*(2), 146-158.
2. Bravo, A. A. (2016). Tecnologías de la web semántica (Tesis doctoral). Universidad Complutense de Madrid, Facultad de Ciencias de la Información, España. Disponible en: https://eprints.ucm.es/id/eprint/41646/1/T38547.pdf
3. Rühle, S. Baker, T. and Johnston, P. Creating Metadata. https://www.dublincore.org/resources/userguide/publishing_metadata/. Accedido 23/06/2022.
4. 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.
5. Fielding, R. T. (2000) Architectural Styles and the Design of Network-Based Software Architectures. Ph.D. Dissertation. University of California, Irvine.