# Uso de RDF y bases de datos de metadatos nativas dentro del proyecto Omnipaper

Cesar Ariza and Ana Alice Baptista

Universidade do Minho
{cariza,analice}@dsi.uminho.pt

**Resumen** Este artículo describe el trabajo realizado para la creación de un prototipo para la búsqueda de información en archivos digitales distribuidos utilizando la tecnología RDF y una base de datos nativa. El articulo reseña los prerrequisitos para la descripción y normalización de la información de los archivos distribuidos, luego los criterios para la selección de la base de datos nativa, muestra las funcionalidades del prototipo creado y al final tiene una síntesis de las lecciones aprendidas y el trabajo futuro.
**Palabras clave**: Digital Libraries, Distributed Information Retrieval, Wordnet, Metadata

## 1. Introducción

En las últimas décadas, la cantidad de información digital, así como el número de ordenadores y conexiones a Internet, ha crecido a un ritmo exponencial. Cada vez es más y más frecuente acceder a la información en formato electrónico, posibilidad que se ve altamente potenciada gracias a la red. Asimismo, cada vez resulta más fácil comparar información procedente de lugares geográficamente distintos, de ahí la creciente importancia de que esta información se encuentre relacionada entre sí a nivel semántico.

El proyecto OmniPaper (Smart Access to European Newspapers) patrocinado por la Comisión Europea IST investiga técnicas para acceder fuentes de información distribuidas con base en tecnologías de IA y XML (SOAP, RDF, XTM). La arquitectura de la solución propuesta por OmniPaper tiene en su base un conjunto de archivos digitales distribuidos, todos dentro de ambientes, bases de datos y mecanismos de indexación diferentes, los cuales pueden ser accedidos de una manera uniforme mediante una interfaz SOAP. Los archivos contienen recursos (artículos de periódicos) catalogados mediante RDF y XTM ("*local layer*") lo que permite una búsqueda inteligente gracias a una capa superior (*"knowledge layer"*) que contiene conceptos relacionados entre si con ocurrencias en diferentes idiomas [1].

Fueron creados prototipos en RDF y XTM para comparar las dos tecnologías. Este artículo describe la fase inicial del prototipo RDF y la incorporación de WordNet [5] para facilitar la navegación dentro del mismo.

## 2. RDF

RDF (del ingles *"Resource Description Framework"*), como su nombre lo indica es un marco de trabajo para describir e intercambiar metadatos. RDF se divide en dos partes, una de ellas el "Modelo RDF y especificación de sintaxis" que tiene un modelo para representar metadatos y la sintaxis para codificarlos y la segunda parte Especificación de Esquemas RDF para la ceración de vocabularios de metadatos, los vocabularios permiten describir objetos de un negocio o área de conocimiento. RDF esta construido en base a las siguientes reglas[2]:

a) *Un recurso* es cualquier cosa que puede tener un URI, esto incluye todas las páginas web, todos los elementos individuales de cada documento XML y mucho más;

b) *Una propiedad* es un recurso que tienen un nombre y que puede usarse como una propiedad, por ejemplo autor o titulo. En muchos casos todo lo que nos importa en realidad es el nombre, pero una propiedad necesita ser un recurso de forma tal que pueda tener sus propias propiedades,

c) *Una sentencia* consiste en la combinación de un recurso, una propiedad y un valor .

Estas tres partes son conocidas como el sujeto, predicado y el objeto de la sentencia, un conjunto de las tres partes es llamado un triple. RDF se representa en XML, de otra manera RDF es una aplicación XML, los triplos RDF pueden ser almacenados en bases de datos normales o en bases de datos especializadas.

## 3. Prototipo RDF

Los objetivos del trabajo con RDF y el desarrollo del prototipo se derivan de los objetivos del proyecto OmniPaper. El principal objetivo es explorar la tecnología RDF para la descripción, búsqueda y recuperación de información en el contexto de un archivo digital de noticias distribuido.

### 3.1 Trabajo previo

El trabajo comienza con formalización de la descripción de los artículos de noticias; para describir los artículos se definió un vocabulario de metadatos común para tres archivos de noticias (los archivos pertenecen a tres empresas proveedoras de noticias en línea); se estudiaron varios vocabularios estándar principalmente dentro del sector de noticias. Finalmente fueron seleccionados los elementos que mejor se ajustaron a las necesidades y como resultado el vocabulario de metadatos de OmniPaper contiene elementos Dublin Core, NIFT y algunos elementos propios de OmniPaper, además, fue creado un perfil de aplicación ("*Application profile*") para el vocabulario. El vocabulario cuenta con 25 elementos pertenecientes a 6 categorías [3]. La categoría "Clasificación del Artículo" tiene el elemento "keylist", que es una lista pesada de palabras llave contenidas en cada artículo, las cuales son extraídas mediante métodos de inteligencia artificial.

Con base en el vocabulario creado se describió un conjunto de noticias; dichas descripciones son documentos RDF/XML, los documentos RDF/XML fueron creados en dos pasos, el primero fue la transformación de los artículos, que originalmente son ficheros XML, por medio de plantillas de transformación XSL, se convirtieron en documentos RDF/XML, el proceso de transformación, mas que transformar mapeo los metadatos existentes dentro de los ficheros XML con los del vocabulario creado, se corrigieron algunos formatos de datos como el de la fecha; el contenido de los artículos no fue *mapeado*; el segundo paso es la inclusión de la lista pesada de palabras llave del artículo que previamente han sido extraídas por un proceso desarrollado para tal fin.

### 3.2 Base de datos nativa: RDF Gateway

La selección de la base de datos se hizo bajo la premisa de ser orientada a XML (nativa); inicialmente se exploro el producto TAMINO de SoftwareAG, después fue explorado RDF Gateway de Intellidimension; debido a que RDF-Gateway fue creado para manipular información en RDF/XML y ofrecía funcionalidades de manejo fue seleccionado, a pesar de ser un producto nuevo.

RDF Gateway además de manipular documentos RFD/XML, tiene un servidor web con un ambiente de programación que soporta paginas activas (RSP), facilidad de conexión con otros programas, y características que lo diferencian de las demás bases de datos normales, como reglas de inferencia, soporte para esquemas y la creación automática de índices con base en las raíces de las palabras (“steeming”) contenidas en los campos indexados.

### 3.3 Descripción y funcionamiento del prototipo

El primer prototipo fue desarrollado en el ambiente de páginas activas (RSP) que soporta RDF Gateway; debido a que el conjunto de sentencias que ofrece RDF Gateway es limitado [4]. El segundo prototipo (que es el descrito en este articulo) fue desarrollado en un ambiente de desarrollo VisualStudio de Microsoft y esta compuesto por varias paginas ASP. El acceso al motor de base de datos de RDF Gateway se hace vía ODBC.

El prototipo RDF es una aplicación que envía consultas inteligentes a la metabase. Una consulta inteligente tiene como base un consulta normal con dos diferencias, la primera, el enriquecimiento de la consulta con sinónimos, la segunda, gracias al motor de base de datos se realiza la búsqueda con la raíces de las palabras. La consulta inteligente tiene ventajas sobre una consulta normal de texto completo (“*full text search*”), principalmente en los tiempos de respuesta y en la calidad de la información encontrada. La consulta enviada al motor de base de datos se hace en RDFQL, que es un el leguaje de consultas parecido a SQL.

La interfaz con el usuario del prototipo RDF tiene cuatro componentes, la primera es una búsqueda inteligente, que permite la entrada de varias palabras llave para realizar una consulta, la segunda, permite la búsqueda en cada uno de los elementos de

vocabulario de metadatos, la tercera es una búsqueda orientada a la navegación, desde una palabra llave inicial, permite recorrer las palabras relacionadas y a su vez muestra los artículos relacionados a esas palabras, los resultados aparecen en el cuarto componente de la interfaz. El componente del interfaz orientado a navegación utiliza WordNet para ayudar al usuario a encontrar mejores resultados.

## 4. Lecciones aprendidas

Se comprobaron algunas premisas planteadas en otras áreas, las cuales se pueden resumir en:

- Los productos, como RDF Gateway, que ofrecen muchas funcionalidades, no tienen la suficiente calidad en todas las funcionalidades, de ahí que se cambio el ambiente de desarrollo a VisualStudio y a la tecnología ASP.
- La indexación que tienen los datos almacenados en RDF Gateway, produce buenos tiempos de repuesta, pero ralentiza la actualización e inserción de información.
- El diseño de las consultas y la base de datos es difícil debido a los conocimientos previos de SQL, normalmente se intenta hacer lo mismo que se realiza en una base de datos típica, olvidando que es una filosofía totalmente diferente.

## 5. Conclusiones y trabajo futuro

El principal beneficio de utilizar RDF y RDF Gateway es la y facilidad uso debido a que los conceptos que se necesitarían para realizar el mismo trabajo en otras tecnologías, ya están encapsulados en RDF. Esta facilidad se ve disminuida porque las tecnologías envueltas son muy recientes, lo cual no permite alcanzar resultados esperados.

Como trabajo futro se pretende la integración del prototipo RDF con el prototipo XTM en una sola aplicación, lo cual permitirá potenciar las dos tecnologías.

## Referencias

[1] OmniPaper – Smart Access to European Newspapers, EU project IST 2001-32174. http://www.omnipaper.org/, 2002.

[2] Resource Description Framework. http://www.w3.org/RDF/

[3] Yaginuma, T., Pereira, T. and Baptista, A. A. (2003) Design of Metadata Elements for Digital News Articles in the OmniPaper Project. From the proceedings of the 7th International Conference on Electronic Publishing, Portugal.

[4] Pereira, T. and Baptista, A. A. (2003) The OmniPaper Metadata RDF/XML Prototype Implementation. From the proceedings of the 7th International Conference on Electronic Publishing, Portugal.

[5]WordNet. http:// www.cogsci.princeton.edu/~wn/