

# Servicios Semánticos para la Documentación de la Administración Pública del Principado de Asturias

Luis Polo and Diego Berrueta

Fundación CTIC, Gijón, Asturias, España,  
{luis.polo,diego.berrueta}@fundacionctic.org,  
Página WWW: <http://www.fundacionctic.org/>

**Resumen** El Departamento de I+D+i de la Fundación CTIC (Centro para el desarrollo de las Tecnologías de la Información y Comunicación en Asturias), dirigido por Antonio Campos, ha iniciado este año su actividad en el ámbito de la Web Semántica. Nuestro equipo, compuesto por tres informáticos y dos lingüistas, en colaboración con los profesores Jose E. Labra, Agustín Cernuda, Enrique del Teso, Guillermo Lorenzo y Roger Bosch de la Universidad de Oviedo, está actualmente trabajando en el desarrollo de servicios semánticos para el Principado de Asturias.

## 1. Introducción

El Boletín Oficial del Principado de Asturias (BOPA) es una publicación diaria emitida por el gobierno autonómico. La legislación presupone que cualquier ciudadano está capacitado para acceder a la información que aparece en los boletines oficiales: subvenciones, notificaciones, leyes, convocatorias de exámenes, etc. Sin embargo, el volumen de textos y la opacidad del lenguaje técnico administrativo y jurídico dificultan a menudo esta comunicación entre el ciudadano y la Administración Pública.

La primera aplicación en la que trabaja nuestro equipo va dirigida a facilitar el acceso a esta información sin necesidad de tener conocimientos jurídicos ni del lenguaje específico. Para su desarrollo nos basamos en herramientas *open source* como Protégé 3.1. (editor de ontologías), Lucene 1.4[1] (buscador sintáctico), OWL API (API para tratar y manejar ontologías OWL[2]) y Racer 1.7 (razonador).

## 2. Desarrollo y Estado del Prototipo

Nuestro primer paso consistió en extraer la información (artículos del boletín) de distintas bases de datos y representarla en XML. A partir de aquí comienza la labor semántica con un enfoque basado en ontologías. El modelo que estamos desarrollando consiste en usar las ontologías como bases de conocimiento complejas que puedan servir de interfaz entre las intenciones de búsqueda del usuario

y la base documental que conforma el BOPA. Actualmente ya disponemos de varias ontologías y hemos desarrollado, sobre éstas, tres servicios: un buscador semántico, una suscripción a alertas y un navegador de conceptos.

## 2.1. Las Ontologías

Estamos construyendo dos tipos de ontologías OWL-DL[3] con propósitos distintos:

- Ontología administrativa: formaliza la estructura básica del BOPA y de la Administración Pública regional.
- Ontologías de dominio: se corresponden con ámbitos concretos y reconocibles por el usuario. Para el prototipo nos hemos centrado en el dominio de las *comunidades de vecinos*.

La arquitectura de la ontología de dominio es modular, construida a partir de micro-ontologías reutilizables, cada una de las cuales abarca un ámbito de conocimiento que es competencia de un organismo de la Administración. De este modo, el diseño es más escalable y permite acometer futuras ampliaciones. Por otro lado, la ontología administrativa sirve como filtro de la ontología de dominio, en nuestro caso, la ontología de *comunidades de vecinos*. Cada tipo de disposición del BOPA (leyes, resoluciones, notificaciones, decretos, etc.) se ajusta a un entorno conceptual particular: un reglamento regula, mientras que un anuncio sólo informa.

## 2.2. Relación entre los Artículos del BOPA y las Ontologías

Una aproximación convencional, utilizando etiquetado de los artículos con conceptos de la ontología, presenta dos problemas fundamentales: el tamaño de la base documental, que asciende a centenas de miles de artículos, y el mantenimiento conforme se vayan incorporando nuevas ontologías de dominio. Finalmente, utilizamos un enfoque distinto basado en ontologías para acceder a los documentos a través de un complejo híbrido de reglas de inferencia y un algoritmo de activación de conceptos (*Spreading Activation algorithm*[4,5]).

De este modo, cuando un usuario realiza una búsqueda, primero se reconocen los reconocen los conceptos subyacentes a la cadena de texto introducida (palabras), y después discurren dos procesos paralelos:

1. *Spreading Activation algorithm*, mediante el cual se puntúa la proximidad semántica de los conceptos de la ontología en función del tipo de relación que guarden entre ellos y la distancia que los separa.
2. Inferencia del organismo competente: a partir de los conceptos activados se infiere el organismo competente en esa materia. Trabajamos actualmente en conseguir que también deduzca el posible tipo de disposición que se asocia con la consulta.

Con estos datos (los conceptos activados, su relevancia, y el organismo competente) y utilizando conocimiento lingüístico (morfología y sinonimia) se prepara una consulta sintáctica sobre los artículos indexados con el motor de búsqueda Lucene. Aunque en último término se utilice un buscador sintáctico, el valor semántico del buscador radica en que la consulta está guiada por las relaciones entre los conceptos de la ontología.

De esta forma, una consulta por «accesibilidad» nos remite a artículos vinculados con el concepto *Accesibilidad* y con conceptos semánticamente próximos, como *Minusvalía*, *Discapacidad*, *Barrera arquitectónica* o *Sordera*. Se obtienen resultados muy interesantes cuando el usuario introduce varios conceptos.

### 2.3. Relación entre los Usuarios y las Ontologías

Nuestra aplicación dispone, además del habitual interfaz en XHTML, de una interfaz gráfica creada con SVG que permite al usuario navegar por los conceptos de las ontologías. En aras de construir una herramienta multimodal y completamente accesible, estamos contemplando la integración de un interfaz de voz usando VoiceXML.

### 2.4. Líneas de Investigación

Actualmente nos encontramos en un proceso de ajuste del prototipo y refinamiento de las búsquedas. Estamos trabajando en puntos tales como:

1. Mapeo de las palabras sobre los conceptos.
2. Mapeo de los conceptos sobre los artículos.
3. Ajuste del algoritmo de activación de conceptos.
4. Incremento de la accesibilidad y usabilidad de la aplicación.
5. Evaluación de la herramienta mediante métricas de calidad y realimentación de los usuarios, a través de un periodo de pruebas en la red de Telecentros del Principado de Asturias.

En el campo de las ontologías, nuestra arquitectura modular implica la necesidad de poder operar con ellas mediante su ensamblaje y comparación. En este sentido, avanzamos en dos direcciones:

1. Hemos construido nuestras ontologías conforme al modelo semántico que propone la ontología de alto nivel DOLCE[6], desarrollada en el proyecto WonderWeb. De esta forma, tenemos un marco de integración para ontologías de distinto tipo. El alineamiento entre DOLCE y las categorías de WordNet da pie a pensar que es posible llegar a crear un triángulo entre las tres bases de conocimiento: nuestras ontologías, DOLCE y WordNet (EuroWordNet[7]).
2. El primitivo estado de las actuales herramientas de integración de ontologías ha motivado en el seno del equipo la idea de desarrollar un *ensamblador* propio.

## Referencias

1. Gospodnetić, O., Hatcher, E.: Lucene in Action. Manning (2005)
2. McGuinness, D.L., van Harmelen, F.: Owl web ontology language overview. Technical report, W3C (2004)
3. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F., eds.: The Description Logic Handbook: Theory, Implementation, and Applications. In Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F., eds.: Description Logic Handbook, Cambridge University Press (2003)
4. Rocha, C., Schwabe, D., de Aragão, M.P.: A hybrid approach for searching in the semantic web. In: WWW. (2004) 374–383
5. Crestani, F.: Application of spreading activation techniques in information retrieval. Artificial Intelligence Review (1997) 453–482
6. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A.: The WonderWeb Library of Foundational Ontologies (D18). Laboratory for Applied Ontology - ISTC-CNR. (2003)
7. Vossen, P.: Euowordnet, general document. Technical report, University of Amsterdam (1999)