

Use Case: Enhancing Web Searches within the Principality of Asturias

Diego Berrueta and Luis Polo, CTIC Foundation, University of Oviedo, and the Principality of Asturias

August 2007



General Description

The Problem

The Principality of Asturias, like other Public Administrations in Spain, publishes lots of documents every day that citizens are supposed to read. These documents include new laws, announcements, notifications or opportunities for public funding. Historically this information was published in printed bulletins, but it is increasingly being published on the web.

Public Administration documents are typically written in legal and administrative jargon, which is far from ordinary language. This represents a hindrance for the communication between citizens and Public Administration. In particular, this language barrier renders the traditional syntactic search (Information Retrieval) almost useless, as terms that are considered synonyms by citizens have a clearly different meaning for the expert lawmaker.

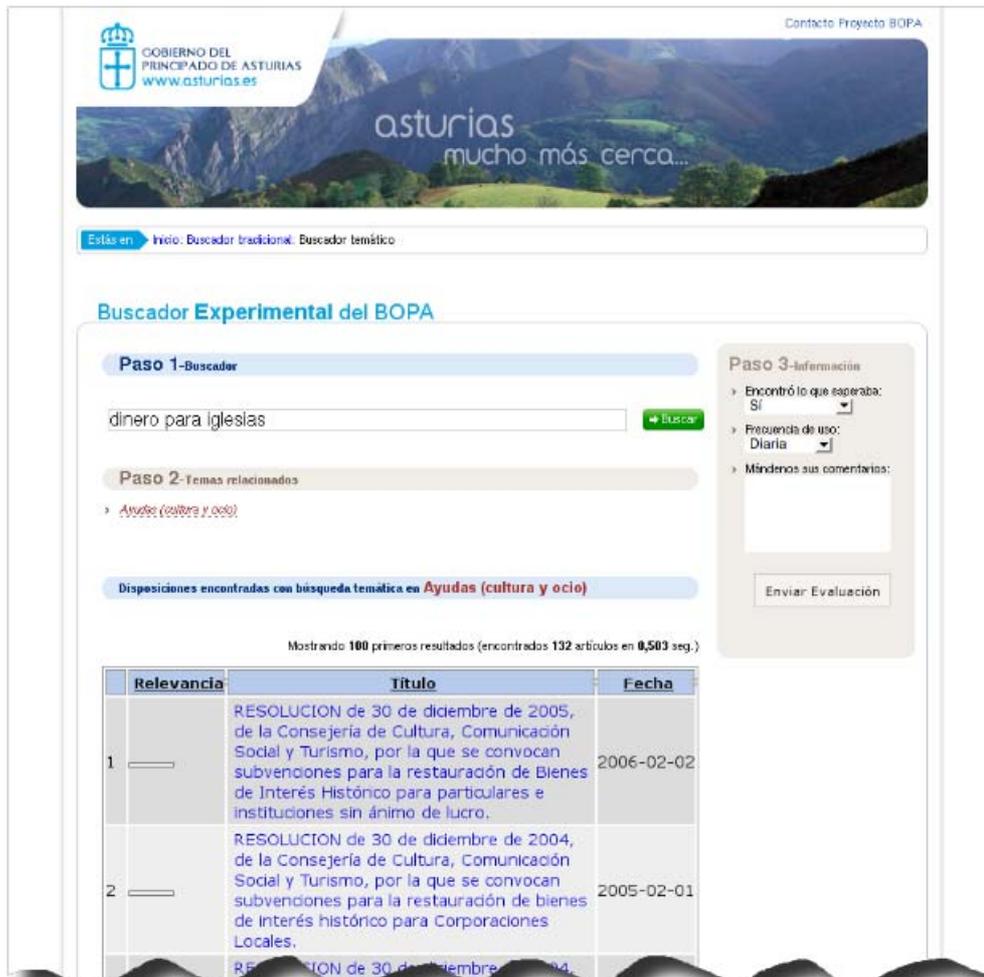


Figure 1: Screen snapshot of the Web site showing the Semantic search capability and results.

The Solution

CTIC Foundation and the University of Oviedo have developed a semantic-based extension for the search engine which is used for the Bulletin of the Principality of Asturias (BOPA). This capability allows users to express queries in their own words. The system recognizes the underlying concepts of the terms, transforms the query, and puts the query into a common context. Some examples of available contexts are public procurement, funding for cultural activities, etc. When the context cannot be automatically determined, the user picks from a list of alternatives. This solves the problem of lexical ambiguity, because terms acquire meaning only within a context.

Like most search engines, the result is a list of documents sorted by relevance with respect to the query.

A screenshot of the application is shown in [Figure 1](#).

Technical details

The application extends an information retrieval product (Apache Lucene) with a query-rewriting module based on semantic technologies. We have developed OWL based ontologies that include up to 10,000 concepts. The ontologies cover both the search domains from user's point of view and the structure of the documents (metadata such as author, document type...). Due to the wide range of possible queries, we have restricted our efforts to the most frequently queried domains, including public procurement, job openings at the Public Administration and public funding. We have re-used international and national classifications and concept schemes, such as CPV (Common Procurement Vocabulary) and CNO (*Clasificación Nacional de Ocupaciones*, National Classification of Occupations).

We have assembled two SKOS thesauri linked to the concepts in the ontologies. The first one contains the terms known by the final users. The second one is designed to match the specialized vocabulary used by the publishers of the bulletin. Therefore, the ontologies are the bridge between the input search terms and the terms that are actually used in the final query.

The application is written in Java using the Jena API to access the ontologies and the thesauri.

The development version of the application (feature-complete) (<http://bopa.fundacionctic.org/>) and the production version (semantic features are not enabled yet) (<http://www.asturias.es/>) are available for use.

Benefits of using Semantic Web Technologies

The benefits for the citizens are:

- They can use their own words and still obtain results even if there is no direct match between terms in the query and terms in the document.
- They can discover new documents that are related to their initial search.
- They get suggestions on how to improve their queries (e.g.: semantically-related concepts).
- They can use a more powerful search tool without a significant increase in the complexity of the user interface.