Searching over Public Administration Legal Documents using Ontologies

Diego Berrueta Jose E. Labra Luis Polo

CTIC Foundation, R & D Department
University of Oviedo, Spain

General Aim

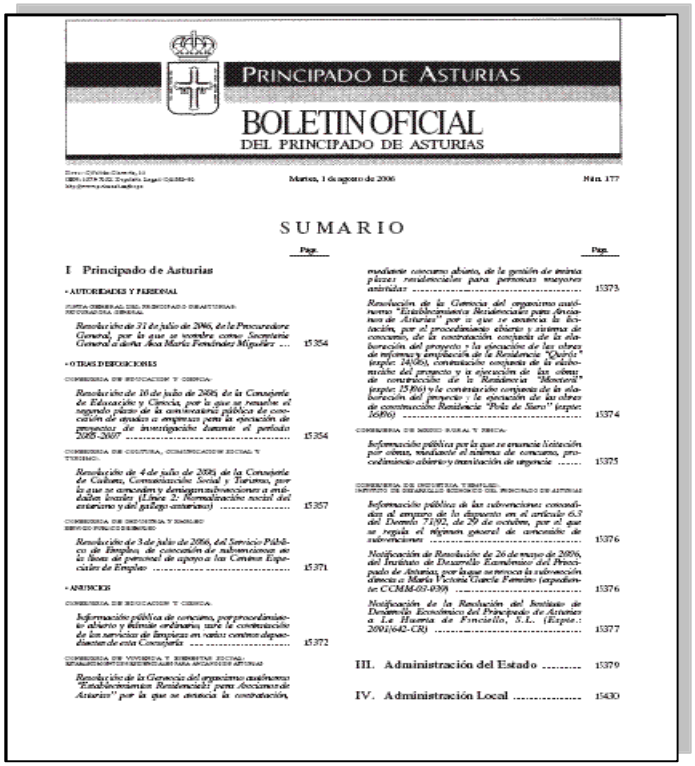
- Features of public administration documents
  - Lots of documents are published everyday
  - Concrete domains: Law, administration, etc.
- Poor effectiveness of syntactic approaches
- Our approach:
  - We applied Semantic web tools and ontologies in this domain
  - A search tool with ontology based query expansion
- General aim:
  - Bridge the gap between public administration and citizens

Motivation
BOPA (1)

- BOPA (**B**oletín **O**ficial del **P**rincipado de **A**sturias)
  - Official Journal of the Principality of Asturias, Spain
  - Asturias:
    - Northern region of Spain
    - Approx. 1,1 million people

Motivation
BOPA (2)

- The BOPA contains every announcement related with Asturias public administration
  - Laws, resolutions, public auctions, etc.
    - General interest to every citizen
    - Specialized jargon
- The previous system:
  - PDF and plain HTML
  - Simple syntactic Search tool
    - Chronological and by title
- Some statistics:
  - 14.314 articles in 23.100 pages (2005)



PRINCIPADO DE ASTURIAS

BOLETIN OFICIAL
DEL PRINCIPADO DE ASTURIAS

Martes, 1 de agosto de 2006 — Núm. 177

SUMARIO

Pág.

I Principado de Asturias

• AUTORIDADES Y PERSONAL

JUNTA GENERAL DEL PRINCIPADO DE ASTURIAS
SECRETARÍA GENERAL

Resolución de 31 de julio de 2006, de la Procuradora General, por la que se nombra como Secretaria General a doña Ana María Fernández Miguélez ... 15354

• OTRAS DISPOSICIONES

CONSEJERÍA DE EDUCACIÓN Y CIENCIA

Resolución de 10 de julio de 2006, de la Consejería de Educación y Ciencia, por la que se resuelve el segundo plazo de la convocatoria pública de concesión de ayudas a empresas para la ejecución de proyectos de investigación durante el período 2005-2007 ... 15354

CONSEJERÍA DE CULTURA, COMUNICACIÓN SOCIAL Y TURISMO

Resolución de 4 de julio de 2006, de la Consejería de Cultura, Comunicación Social y Turismo, por la que se conceden y deniegan subvenciones a entidades locales (Línea 2: Normalización social del asturiano y del gallego asturiano) ... 15357

CONSEJERÍA DE INDUSTRIA Y EMPLEO
SERVICIO PÚBLICO DE EMPLEO

Resolución de 3 de julio de 2006, del Servicio Público de Empleo, de concesión de subvenciones en la línea de personal de apoyo a los Centros Especiales de Empleo ... 15371

• ANUNCIOS

CONSEJERÍA DE EDUCACIÓN Y CIENCIA

Información pública de concurso, por procedimiento abierto y trámite ordinario, para la contratación de los servicios de limpieza en varios centros dependientes de esta Consejería ... 15372

CONSEJERÍA DE VIVIENDA Y BIENESTAR SOCIAL
REGULARIZACIÓN DE RESIDENCIAS PARA ANCIANOS DE ASTURIAS

Resolución de la Gerencia del organismo autónomo "Establecimientos Residenciales para Ancianos de Asturias" por la que se anuncia la contratación, mediante concurso abierto, de la gestión de treinta plazas residenciales para personas mayores asistidas ... 15373

Resolución de la Gerencia del organismo autónomo "Establecimientos Residenciales para Ancianos de Asturias" por la que se anuncia la licitación, por el procedimiento abierto y sistema de concurso, de la contratación conjunta de la elaboración del proyecto y la ejecución de las obras de reforma y ampliación de la Residencia "Quirós" (expte: 14/06), contratación conjunta de la elaboración del proyecto y la ejecución de las obras de construcción de la Residencia "Moscovi" (expte: 15/06) y la contratación conjunta de la elaboración del proyecto y la ejecución de las obras de construcción Residencia "Pola de Siero" (expte: 16/06) ... 15374

CONSEJERÍA DE MEDIO RURAL Y PESCA

Información pública por la que se anuncia licitación por obras, mediante el sistema de concurso, procedimiento abierto y tramitación de urgencia ... 15375

CONSEJERÍA DE INDUSTRIA Y EMPLEO
INSTITUTO DE DESARROLLO ECONÓMICO DEL PRINCIPADO DE ASTURIAS

Información pública de las subvenciones concedidas al amparo de lo dispuesto en el artículo 6.3 del Decreto 71/92, de 29 de octubre, por el que se regula el régimen general de concesión de subvenciones ... 15376

Notificación de Resolución de 26 de mayo de 2006, del Instituto de Desarrollo Económico del Principado de Asturias, por la que se revoca la subvención directa a María Victoria García Fernández (expediente: CCMM-03-030) ... 15376

Notificación de la Resolución del Instituto de Desarrollo Económico del Principado de Asturias a La Huerta de Fonciello, S.L. (Expte.: 2001/642-CR) ... 15377

III. Administración del Estado ... 15379

IV. Administración Local ... 15430

Motivation
Some restrictions

- We could not interfere with the established process
  - We were allowed to work only with the generated data
- Adding tags to each document manually = not practical
  - Not enough personal and lots of documents
  - Error prone

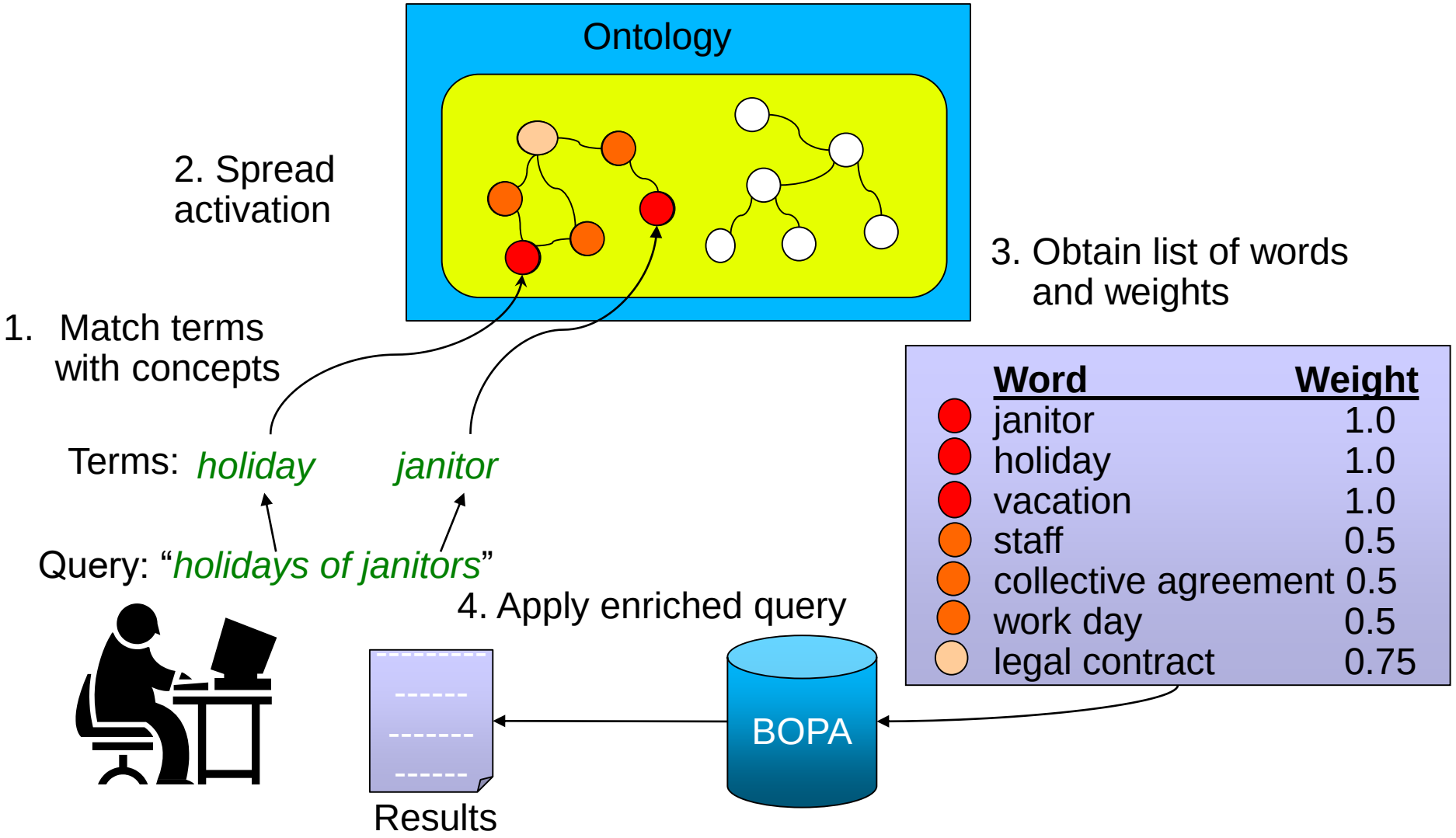
Ontologies

- We defined ontologies for our problem domain
  - A team of 3 domain experts was employed
- Ontologies were developed in a modular way
  - One upper legal & administrative Ontology
  - Several domain ontologies
    - We defined micro-ontologies for several contexts
    - Micro-ontologies import the upper ontology
    - Examples: public-employment, subventions, etc.

Synsets

- Each concept has a synonym set (synset) associated
  - In practice, we use 2 types of synsets:
    - Input synset: Terms used in common language
      - Allow to find a concept from a query term
    - Output synset: Legal and administrative terms
      - Allow to find a document that contains a concept
  - Example:
    - Concept: Holiday
    - Input synsets: "holiday", "vacation", "break"
    - Output synsets: "holiday" "vacation"

Ontology Based Search
Search Process overview



Step 1. Match terms to concepts

- Query sentence is parsed to obtain terms
  - Remove non-content words like “the”, “but”, etc.
  - Remove suffixes of words
- Terms are matched against input synsets of ontology concepts
- A context (micro-ontology) is determined for the query
  - If several concepts belong to different contexts, we take the common one
  - In case of several choices, the user can select one
- Search can be restricted to that context

Ontology Based Search
Step 2. Spreading activation

- We use the spreading activation algorithm
  - It returns a weighted list of related concepts

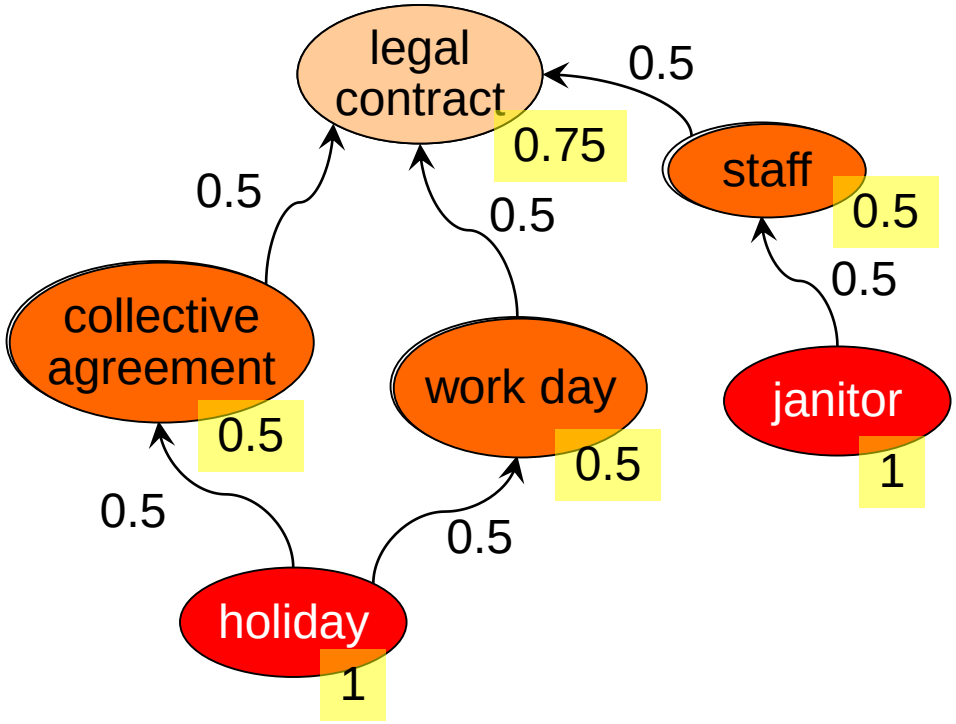
w_{ij} = weight of relation between nodes i and j
Δ = set of activation nodes
Initialized to initial query concepts
Implemented as a queue ordered by activation value
Θ = set of output concepts
N_{min} = Minimum activation level

while $\Delta \neq \varnothing$ and $N_k > N_{min}$
extract a node n_k from Δ
add n_k to Θ
For each node n_i such that $w_{ki} > 0$
$N_i = N_i + w_{ki} N_k$
Add n_i to Δ
endWhile
return Θ

Spreading activation algorithm (*simplified*)

Step 2. Spreading Activation

- Example of Spread activation



Output concepts

Concept	Weight
janitor	1.0
holiday	1.0
staff	0.5
collective agreement	0.5
work day	0.5
legal contract	0.75

Step 3. Obtain words from concepts

- For each concept, we obtain the representative words
  - The words are obtained from the output synsets

Output concepts

Concept	Weight
janitor	1.0
holiday	1.0
staff	0.5
collective agreement	0.5
work day	0.5
legal contract	0.75

Word list

Word	Weight
janitor	1.0
holiday	1.0
vacation	1.0
staff	0.5
collective agreement	0.5
work day	0.5
legal contract	0.75

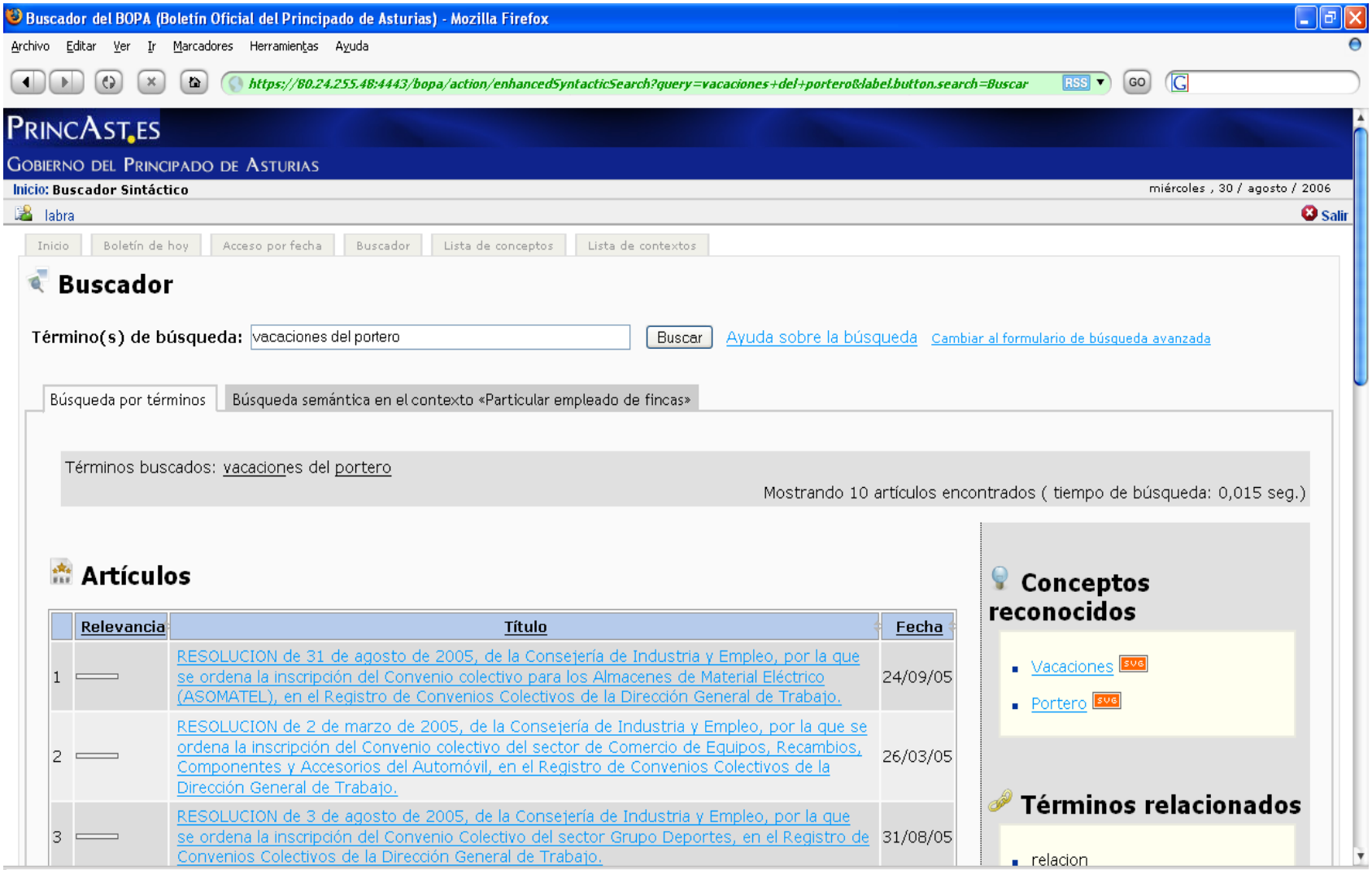
Step 4. Apply syntactic search

- Syntactic Search
  - We used Apache Lucene search engine
  - Search expression:

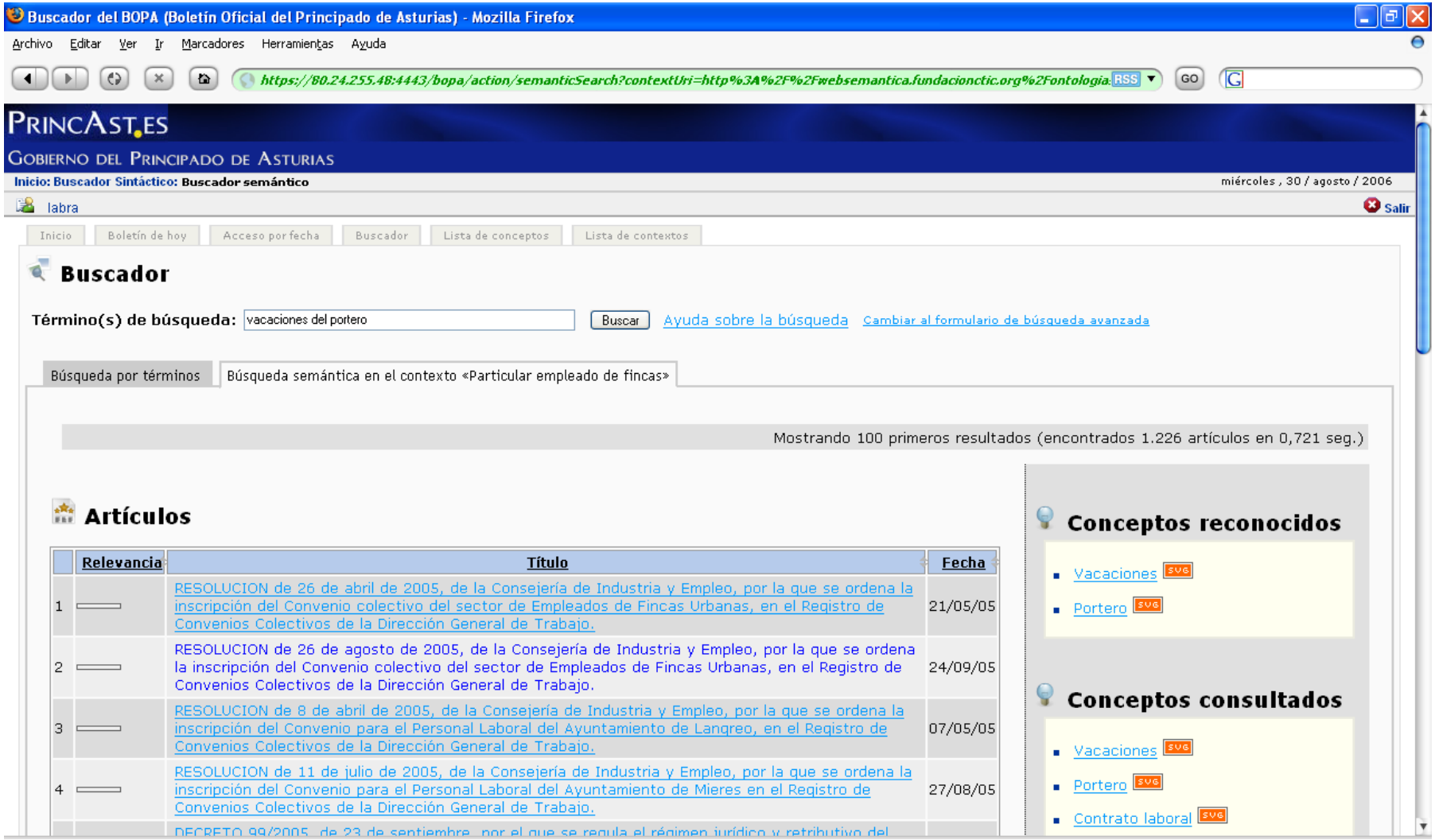
```
"janitor"^1 "holiday"^1 "vacation"^1 "staff"^.5
("collective agreement"~1)^.5 ("work day"~1)^.5
("legal contract"~1)^.75
```

- Search results are ordered according to relevance level

Other Services
Syntactic search



Other Services
Semantic search

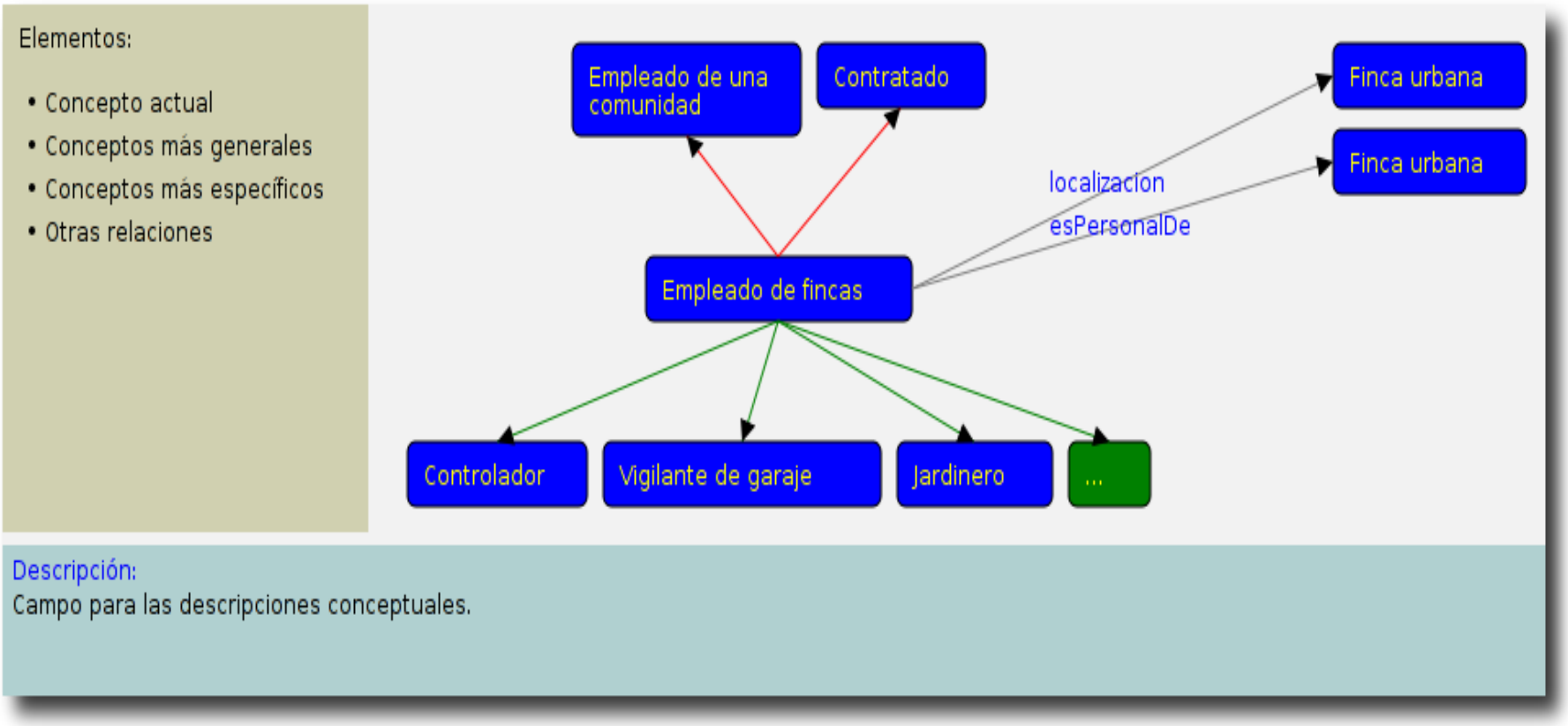


Alert subscription service

- Query subscription
  - Users can subscribe to a query
  - A notification can be send when new results for that query appear
  - Email and SMS alerts can be send to registered users

Other Services
Concept Browser

- User interface developed in SVG
- Navigational representation of the Ontology concepts



Other Services
Simple Explanation module

PRINCAST.ES

GOBIERNO DEL PRINCIPADO DE ASTURIAS

Inicio: Buscador Sintáctico: Buscador semántico — sábado, 26 / agosto / 2006

Inicio | Boletín de hoy | Acceso por fecha | Buscador | Lista de conceptos | Lista de contextos

Explicación de los resultados de la búsqueda semántica

Conceptos reconocidos

Relevancia	Concepto	Explicación	Explicación en SVG
	Portero SVG	=> Portero	SVG
	Vacaciones SVG	=> Vacaciones	SVG
	Contrato laboral SVG	=> Vacaciones => Contrato laboral	SVG
	Elementos de un contrato laboral SVG	=> Vacaciones => Elementos de un contrato laboral	SVG
	Convenio laboral SVG	=> Vacaciones => Elementos de un contrato laboral => Convenio laboral	SVG
	Regulación empleado de fincas SVG	=> Vacaciones => Elementos de un contrato laboral => Regulación empleado de fincas	SVG
	Empleado de fincas SVG	=> Portero => Empleado de fincas	SVG
	Contratado SVG	=> Portero => Empleado de fincas => Contratado	SVG
	Empleado de una comunidad SVG	=> Portero => Empleado de fincas => Empleado de una comunidad	SVG
	Jornada laboral SVG	=> Vacaciones => Elementos de un contrato laboral => Jornada laboral	SVG
	Trabajo de mantenimiento SVG	=> Portero => Trabajo de mantenimiento	SVG
	Finca urbana SVG	=> Portero => Empleado de fincas => Finca urbana	SVG
	Limpiador SVG	=> Portero => Empleado de fincas => Limpiador	SVG
	Conserje SVG	=> Portero => Empleado de fincas => Conserje	SVG
	Jardinero SVG	=> Portero => Empleado de fincas => Jardinero	SVG

System Development
Architecture and Methodology

- J2EE application
  - Data access tier
    - XML Native Database: XIndice
    - Relational database
    - Syntactical indexes managed by Lucene
    - Original web server where BOPA is still published
  - Business tier
    - Several vertical subsystems (low coupling)
      - Search engines, syntactical analysis, ontology processing, etc.
    - Functionality exported through web services
  - Main User interface developed with the Struts framework
- Development methodology
  - We used some agile development techniques
    - Pair programming

Recovering information from HTML

- A web crawler fetches HTML pages
  - The pages are not valid HTML
- JTidy transforms bad formed HTML to well formed XHTML
- A chain of XSLT transformations allows to extract information in XML format
  - We use a custom XML vocabulary
- Static HTML data is completed with database data

Application integration

- Interoperability has been a design goal
- All the functionality can be accessed by web services
- 5 clients have been implemented
  - Standard Web application front end
  - VoiceXML interface
  - Desktop .Net client
  - Desktop Python client
  - Java client using SWT libraries for Eclipse
- We export a RSS channel which is updated daily with the contents of new bulletins
- Bulletin contents are exported in HTML, PDF and RDF

Evaluation of search results

- At this moment, we are logging user activity and we collect different statistics
  - How many times the first result is selected, etc....
- Developed framework for testing quality of results
  - Difficulties to define good search results
  - We apply it mainly for regression tests
- We are planning to do medium scale tests with real users in short term

Conclusions and Future work

- Practical application using a knowledge based approach
  - The prototype has been adopted by the Public Administration department
    - It will be part of a production environment
- Future work
  - Extend the search to the whole public administration site
  - Practical evaluation of the approach
    - Usability, search results, etc.
  - Extend Context-Theory
    - Allow more flexibility in context definition
    - Relationship between contexts and spread-activated nodes
    - Allow users feedback (tags)
  - Ontology development and reusability
    - Incorporate other ontologies (DOLCE, SUMO, etc.)
    - Develop a general legal and administrative ontology
  - Digitization project of past documents
    - BOPA started to be published before 1900

The end