Un Sistema de Gestión Documental y Workflow con Indexación Temática y Geográfica de los Documentos

Ana Cerdeira-Pena, Miguel R. Luaces, Óscar Pedreira, Diego Seco

Laboratorio de Bases de Datos
Universidade da Coruña
A Coruña, España

Organización

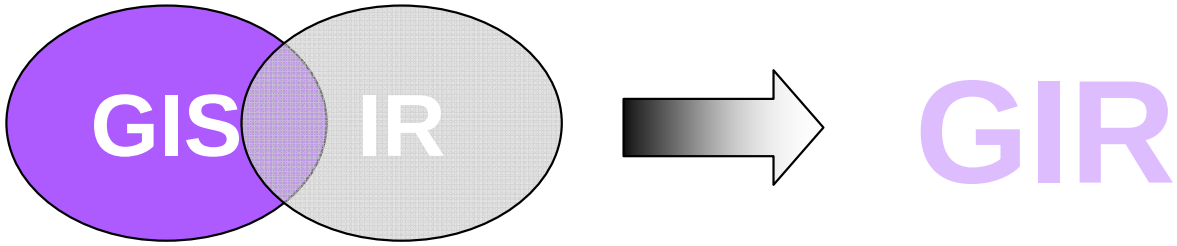
- Introducción
- Motivación
- Trabajo relacionado
- Arquitectura
- Estructura de indexación
- Tipos de consultas soportadas
- Demo
- Conclusiones y futuros desarrollos

Organización

- Introducción
- Motivación
- Trabajo relacionado
- Arquitectura
- Estructura de indexación
- Tipos de consultas soportadas
- Demo
- Conclusiones y futuros desarrollos

Introducción

- Dos campos de investigación muy activos:
  - Geographic Information Systems (GIS)
    - EIEL (http://www.dicoruna.es/webeiel)
  - Information Retrieval (IR)
    - Biblioteca Virtual Galega (http://bvg.udc.es)



Recuperar documentos relevantes temática y geográficamente respondiendo a consultas de la forma <tema, localización>

Organización

- Introducción
- **Motivación**
- Trabajo relacionado
- Arquitectura
- Estructura de indexación
- Tipos de consultas soportadas
- Demo
- Conclusiones y futuros desarrollos

Motivación

- Muchos documentos almacenados en bibliotecas digitales y bases de datos documentales incluyen referencias geográficas
  - Prensa, Web, IDEs, …
    - *"…las Jornadas de la IDE de España celebradas en Tenerife en noviembre de 2008..."*
    - Licencias de obra
- Pocas estructuras de indexación y algoritmos de recuperación explotan las referencias geográficas
- Las propuestas recientes no tienen en cuenta algunas particularidades específicas del espacio geográfico
  - Naturaleza jerárquica del espacio geográfico
  - Relaciones topológicas entre los objetos

Organización

- Introducción
- Motivación
- Trabajo relacionado
- Arquitectura
- Estructura de indexación
- Tipos de consultas soportadas
- Demo
- Conclusiones y futuros desarrollos

Trabajo relacionado

- Indexación de texto: Índice Invertido
  - Las referencias geográficas son *palabras normales*
- Indexación espacial: R-Tree
  - No tienen en cuenta la jerarquía del espacio
- Propuestas para combinarlos (proyecto SPIRIT):
  - *Text-First* (primero filtrado textual y luego espacial)
  - *Geo-First* (primero filtrado espacial y luego textual)
  - No tienen en cuenta las relaciones entre los objetos geográficos que están indexando
- Descripción del espacio geográfico: Ontología
  - Empleadas en GIR para realizar *query expansion*, elaboración de rankings de relevancia y anotación de recursos web
  - Ningún intento de combinarlas con otros tipos de índices para obtener una estructura híbrida

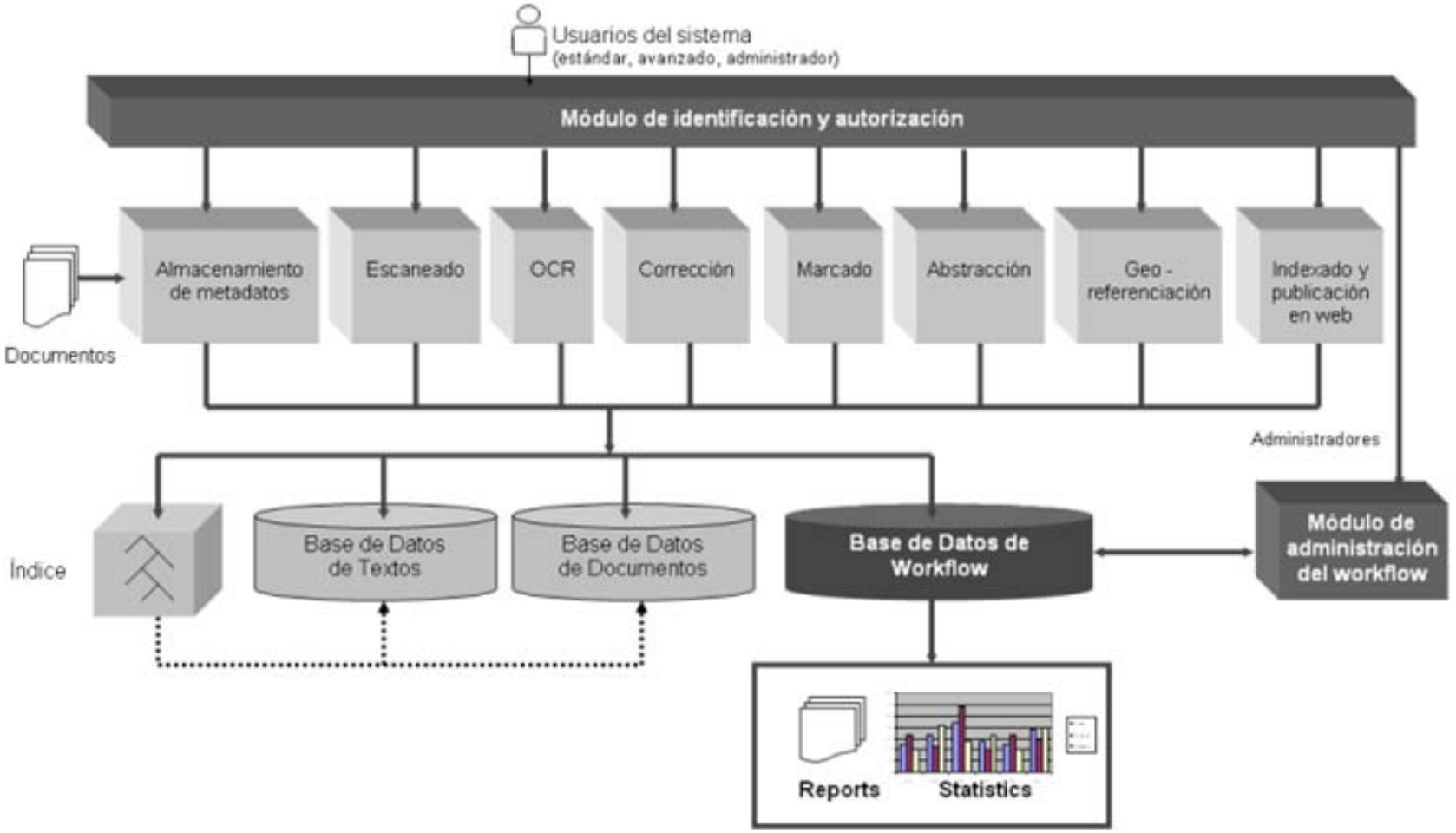
Organización

- Introducción
- Motivación
- Trabajo relacionado
- Arquitectura
- Estructura de indexación
- Tipos de consultas soportadas
- Demo
- Conclusiones y futuros desarrollos

Arquitectura

- Los Sistemas de Gestión Documental y Workflow
  - Definen todas las tareas que se deben realizar sobre los documentos para su gestión y consulta
  - Las personas encargadas de realizarlas
  - Proporcionan las herramientas necesarias para llevar a cabo todo este proceso
  - Por tanto, son imprescindibles en organizaciones donde el número de documentos se incrementa constantemente

Arquitectura



Organización

- Introducción
- Motivación
- Trabajo relacionado
- Arquitectura
- Estructura de indexación
- Tipos de consultas soportadas
- Demo
- Conclusiones y futuros desarrollos

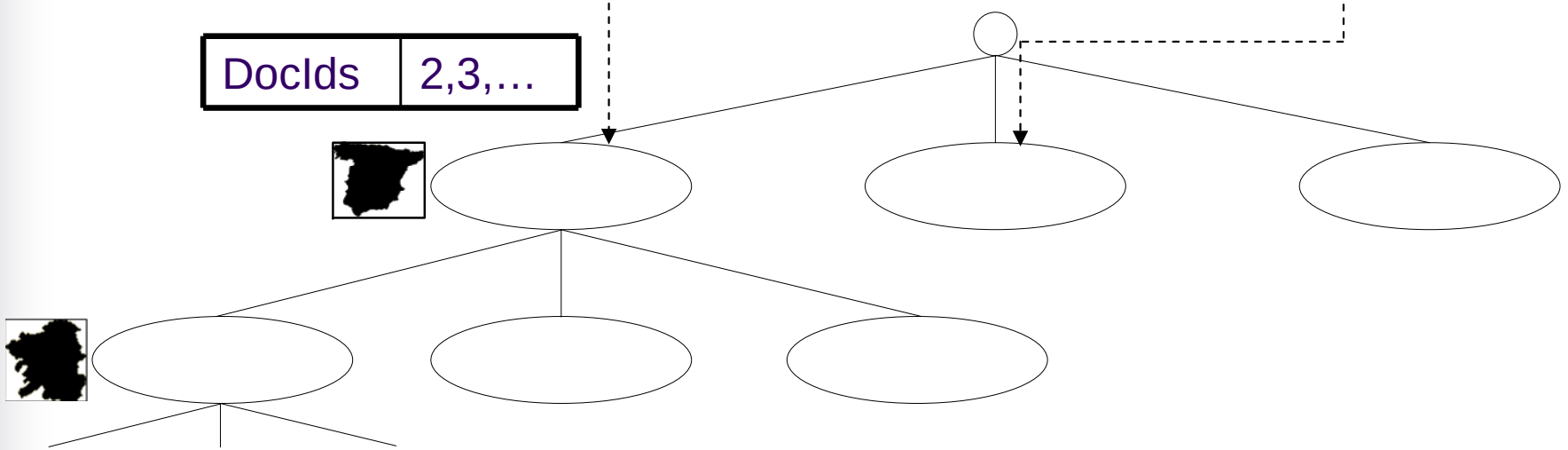
Estructura de indexación

Índice Invertido

…	…
hotel	1,3,7,8,12,…
mar	3,5,6,9,10,…
…	…

Tabla Hash de Nombres de Lugar

…	…
España	●
Alemania	●
…	…



Estructura de indexación

- Toma como base una ontología
- Árbol compuesto por nodos que representan topónimos interconectados por medio de relaciones de contenido
  - Si la lista de nodos hijo es muy larga se emplea un R-Tree
- Estructuras auxiliares:
  - Tabla hash de nombre de lugar a posición en el árbol
  - Índice Invertido tradicional
- Ventajas:
  - Procesado eficiente tanto de consultas textuales como espaciales
  - Soporte para consultas combinadas
  - Actualizaciones y optimizaciones independientes en cada índice
- Inconvenientes:
  - Árbol posiblemente desbalanceado
  - Estructura estática

Organización

- Introducción
- Motivación
- Trabajo relacionado
- Arquitectura
- Estructura de indexación
- Tipos de consultas soportadas
- Demo
- Conclusiones y futuros desarrollos

Tipos de consultas soportadas

- Consultas puramente textuales
  - *"recuperar todos los documentos donde aparezcan las palabras* hotel *y* mar*"*
  - ¿Cómo las resolvemos?
    - Índice textual
- Consultas puramente espaciales
  - *"recuperar todos los documentos que se refieran a la siguiente área geográfica"*
  - ¿Cómo las resolvemos?
    - Descenso en la estructura + refinado del resultado
    - El mismo algoritmo empleado con índices espaciales

Tipos de consultas soportadas

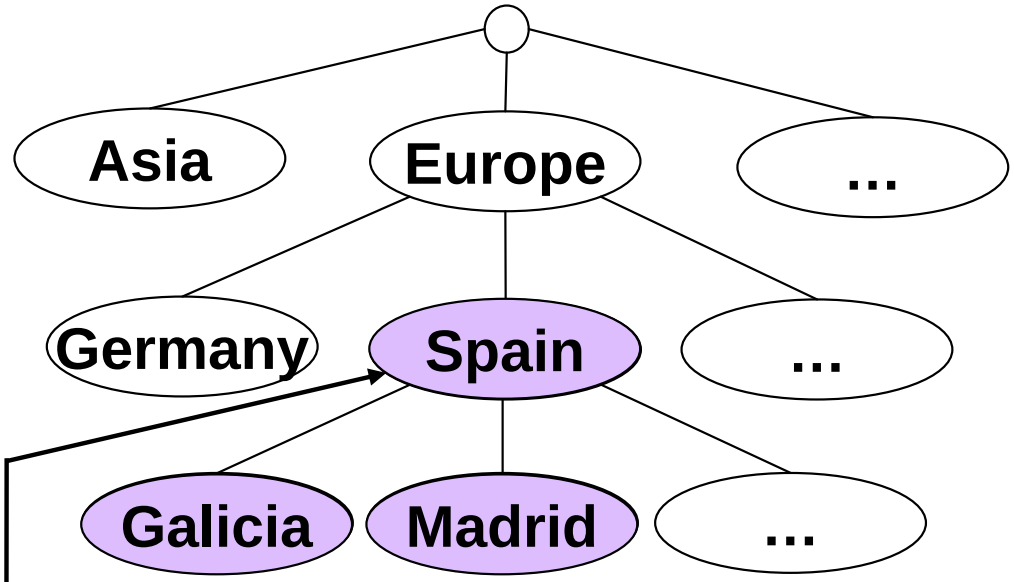
- Consultas textuales con nombres de lugar
  - *"recuperar todos los documentos con la palabra* hotel *referidos a España"*
  - ¿Cómo las resolvemos?
    - Ejemplo
  - Ahorro de tiempo evitando parte del recorrido en el árbol

Tipos de consultas soportadas

Inverted Index

…	…
hotel	1,3,7,8,12,
sea	3,5,6,9,10,
…	… …

Index Structure



Place Name Hash Table

…	…
Spain	●
Germany	
…	…

DocIds	2,3,5,7,12,14,…

Text Result	1,3,7,8,12,…
Spatial Result	2,3,5,7,12,14,…
Query Result	3,7,12,…

Tipos de consultas soportadas

- Consultas textuales sobre un área geográfica
  - *"recuperar todos los documentos con la palabra* hotel *que se refieren a la siguiente área geográfica"*
  - ¿Cómo las resolvemos?
    - Ejemplo

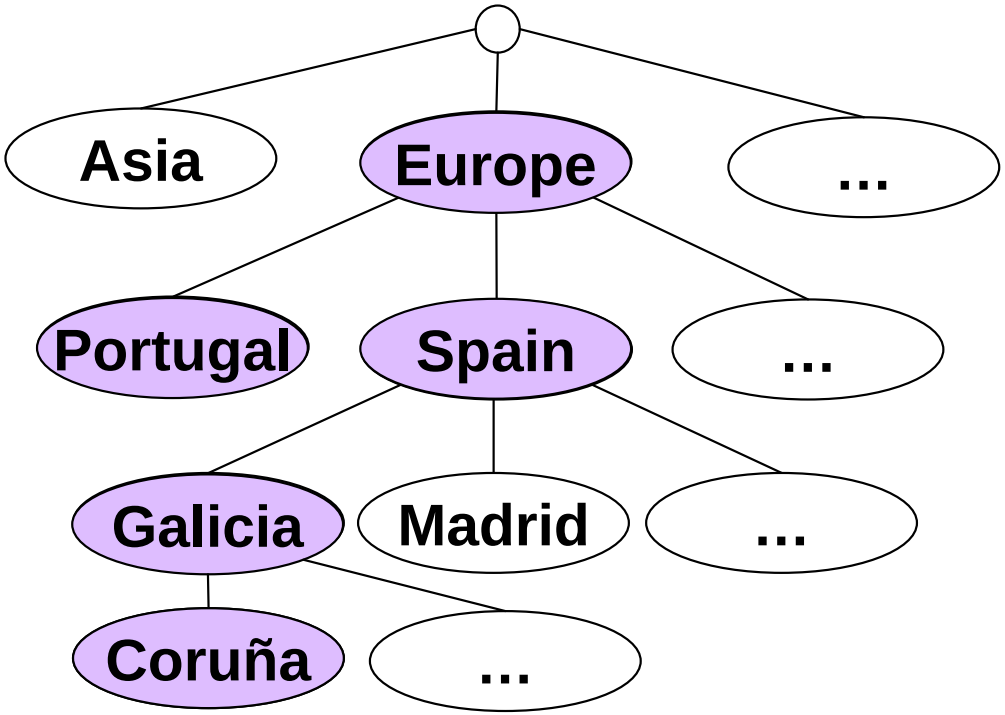
Tipos de consultas soportadas

Inverted Index

…	…
hotel	1,3,7,8,12,
sea	3,5,6,9,10,
…	… …

Query Window

Index Structure



DocIds	12,14,…

Text Result	1,3,7,8,12,…
Spatial Result	12,14,…
Query Result	12,…

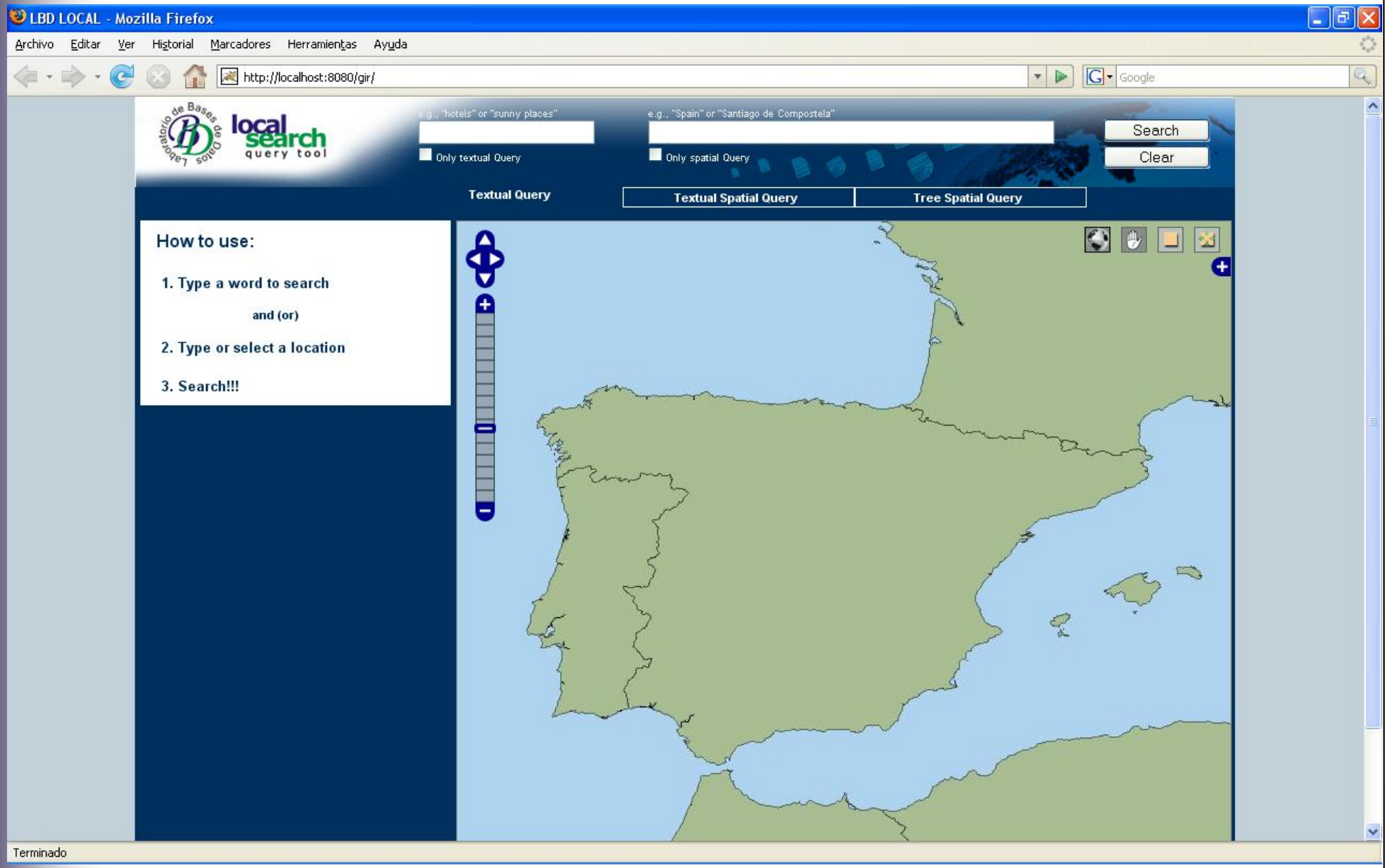
Tipos de consultas soportadas

- Otra ventaja: *EXPANSIÓN DE CONSULTAS*
  - *"recuperar todos los documentos referidos a España"*
  - ¿Cómo las resolvemos?
    - El *Servicio de Evaluación de Consultas* descubrirá que *España* es una referencia geográfica
    - La *Tabla Hash de Nombres de Lugar* localizará rápidamente el nodo interno que representa a *España*
    - Todos los documentos asociados con ese nodo forman parte del resultado
    - Todos los documentos asociados con el subárbol forman parte del resultado
  - El resultado contiene, además de aquellos documentos que incluyen el término *España,* todos los documentos que contienen el nombre de una división administrativa incluida en España

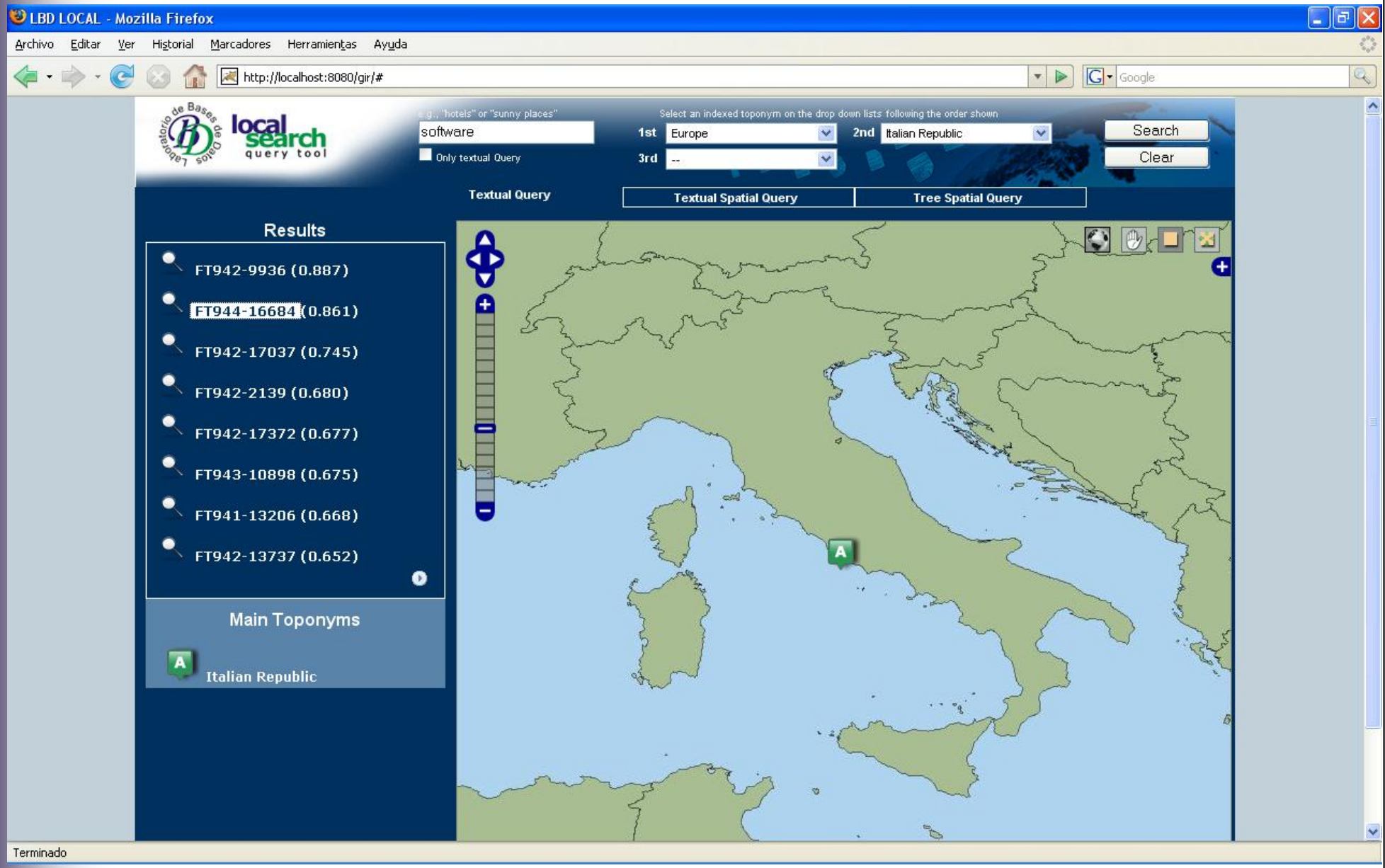
Organización

- Introducción
- Motivación
- Trabajo relacionado
- Arquitectura
- Estructura de indexación
- Tipos de consultas soportadas
- Demo
- Conclusiones y futuros desarrollos

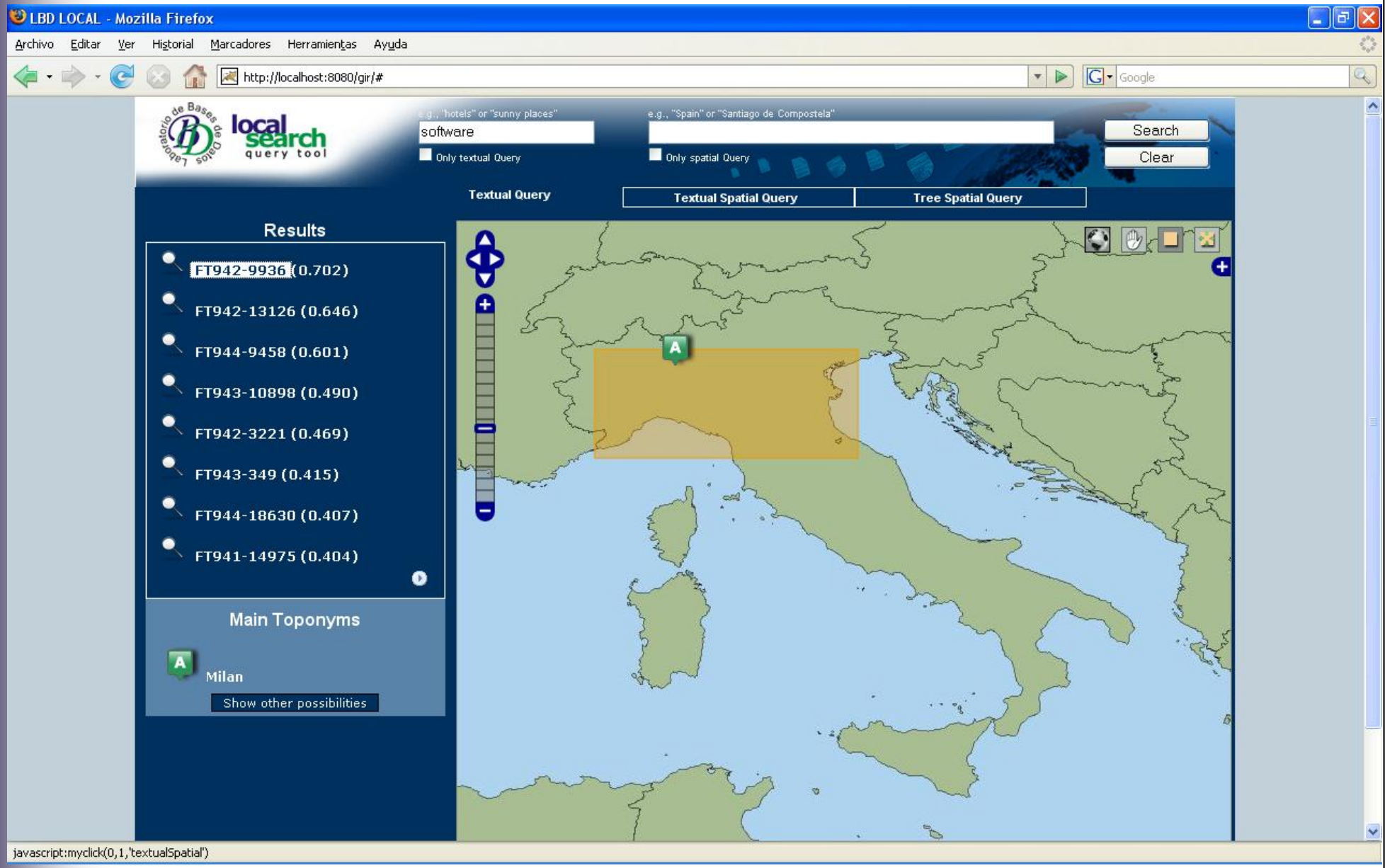
Demo


LBD LOCAL - Mozilla Firefox
Archivo Editar Ver Historial Marcadores Herramientas Ayuda
http://localhost:8080/gir/
Google
local search query tool
e.g., "hotels" or "sunny places"
Only textual Query
e.g., "Spain" or "Santiago de Compostela"
Only spatial Query
Search
Clear
Textual Query
Textual Spatial Query
Tree Spatial Query
How to use:
1. Type a word to search
and (or)
2. Type or select a location
3. Search!!!
Terminado

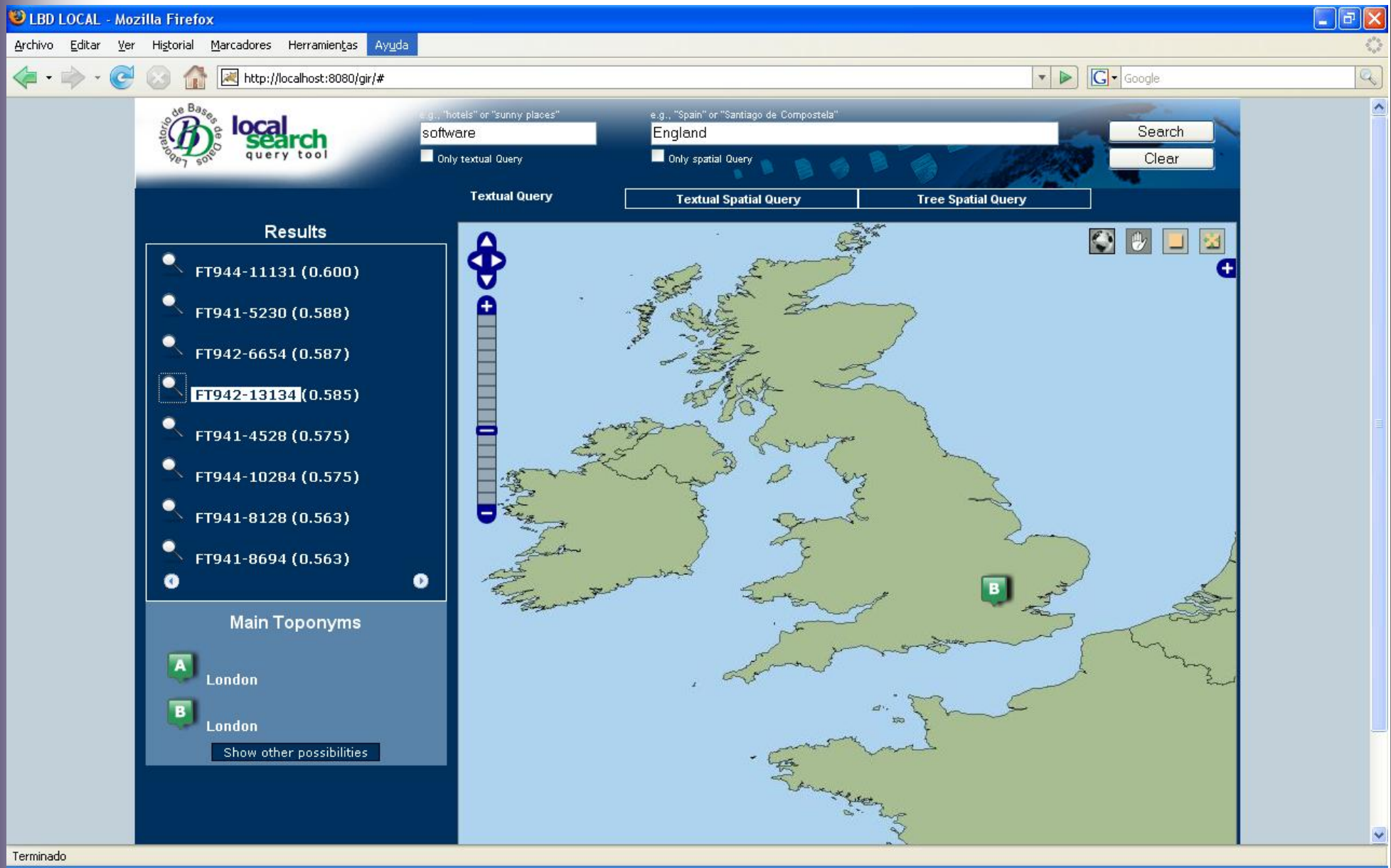
Demo


LBD LOCAL - Mozilla Firefox
Archivo Editar Ver Historial Marcadores Herramientas Ayuda
http://localhost:8080/gir/#
local search query tool
e.g., "hotels" or "sunny places"
software
Only textual Query
Select an indexed toponym on the drop down lists following the order shown
1st Europe
2nd Italian Republic
3rd --
Search
Clear
Textual Query
Textual Spatial Query
Tree Spatial Query
Results
FT942-9936 (0.887)
FT944-16684 (0.861)
FT942-17037 (0.745)
FT942-2139 (0.680)
FT942-17372 (0.677)
FT943-10898 (0.675)
FT941-13206 (0.668)
FT942-13737 (0.652)
Main Toponyms
A
Italian Republic
Terminado

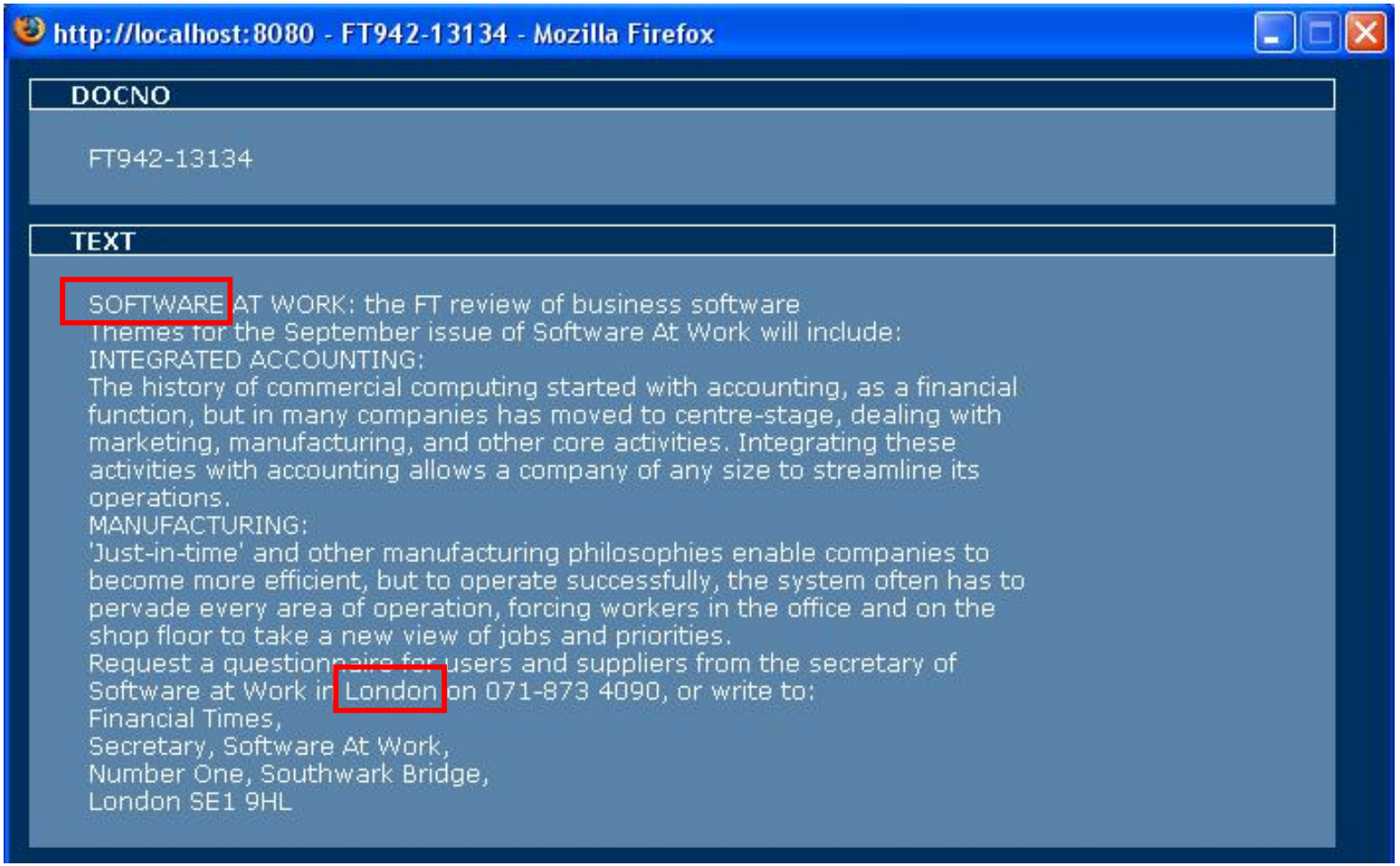
Demo


LBD LOCAL - Mozilla Firefox
Archivo Editar Ver Historial Marcadores Herramientas Ayuda
http://localhost:8080/gir/#
local search query tool
e.g., "hotels" or "sunny places"
software
Only textual Query
e.g., "Spain" or "Santiago de Compostela"
Only spatial Query
Search
Clear
Textual Query
Textual Spatial Query
Tree Spatial Query
Results
FT942-9936 (0.702)
FT942-13126 (0.646)
FT944-9458 (0.601)
FT943-10898 (0.490)
FT942-3221 (0.469)
FT943-349 (0.415)
FT944-18630 (0.407)
FT941-14975 (0.404)
Main Toponyms
A
Milan
Show other possibilities
javascript:myclick(0,1,'textualSpatial')

Demo


LBD LOCAL - Mozilla Firefox
Archivo Editar Ver Historial Marcadores Herramientas Ayuda
http://localhost:8080/gir/#
local search query tool
e.g., "hotels" or "sunny places"
software
Only textual Query
e.g., "Spain" or "Santiago de Compostela"
England
Only spatial Query
Search
Clear
Textual Query
Textual Spatial Query
Tree Spatial Query
Results
FT944-11131 (0.600)
FT941-5230 (0.588)
FT942-6654 (0.587)
FT942-13134 (0.585)
FT941-4528 (0.575)
FT944-10284 (0.575)
FT941-8128 (0.563)
FT941-8694 (0.563)
Main Toponyms
A London
B London
Show other possibilities
Terminado

Demo


http://localhost:8080 - FT942-13134 - Mozilla Firefox
DOCNO
FT942-13134
TEXT
SOFTWARE AT WORK: the FT review of business software
Themes for the September issue of Software At Work will include:
INTEGRATED ACCOUNTING:
The history of commercial computing started with accounting, as a financial function, but in many companies has moved to centre-stage, dealing with marketing, manufacturing, and other core activities. Integrating these activities with accounting allows a company of any size to streamline its operations.
MANUFACTURING:
'Just-in-time' and other manufacturing philosophies enable companies to become more efficient, but to operate successfully, the system often has to pervade every area of operation, forcing workers in the office and on the shop floor to take a new view of jobs and priorities.
Request a questionnaire for users and suppliers from the secretary of Software at Work in London on 071-873 4090, or write to:
Financial Times,
Secretary, Software At Work,
Number One, Southwark Bridge,
London SE1 9HL

Organización

- Introducción
- Motivación
- Trabajo relacionado
- Arquitectura
- Estructura de indexación
- Tipos de consultas soportadas
- Demo
- Conclusiones y futuros desarrollos

Conclusiones y futuros desarrollos

- Conclusiones:
  - Sistema de gestión documental y workflow para recuperación de información geográfica
  - Estructura de indexación formada por un índice textual, un índice espacial y una ontología
  - Resolución de nuevos tipos de consultas

Conclusiones y futuros desarrollos

- Trabajo futuro:
  - Finalización de la herramienta de gestión del workflow
  - Desambiguación de topónimos
  - Mejora de los algoritmos de ranking
  - Inclusión de otros tipos de relaciones (ej. Adyacencia)

Un Sistema de Gestión Documental y Workflow con Indexación Temática y Geográfica de los Documentos

Ana Cerdeira-Pena, Miguel R. Luaces, Óscar Pedreira, Diego Seco

Contacto: dseco@udc.es

Laboratorio de Bases de Datos
Universidade da Coruña
A Coruña, España