LBD LOCAL: Un Sistema para la Recuperación de Documentos con Referencias Geográficas

Miguel R. Luaces, José R. Paramá, Oscar Pedreira, Diego Seco

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

Organización

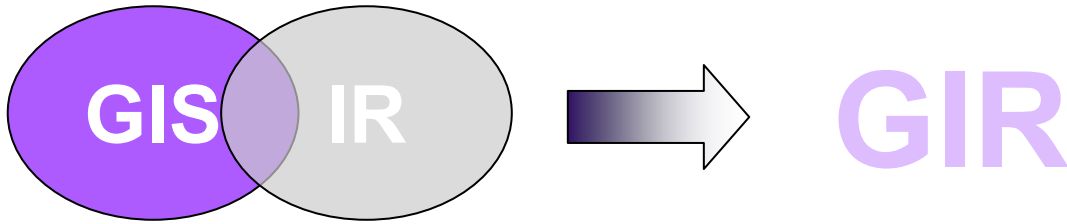
- Introducción
- Motivación
- Trabajo relacionado
- Arquitectura
- Estructura de indexación
- Tipos de consultas soportadas
- Demo
- Conclusiones y futuros desarrollos

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 SIG Libre celebradas en Girona en marzo de 2007..."*
- 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

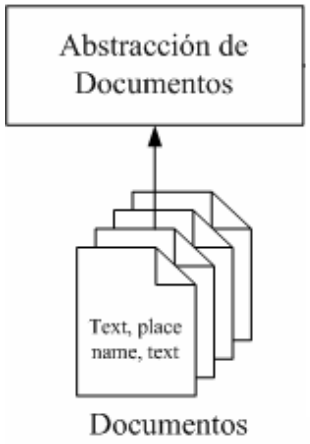
Trabajo relacionado

- Indexación de texto: Índice Invertido
  - Ignoran completamente las referencias geográficas
- 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

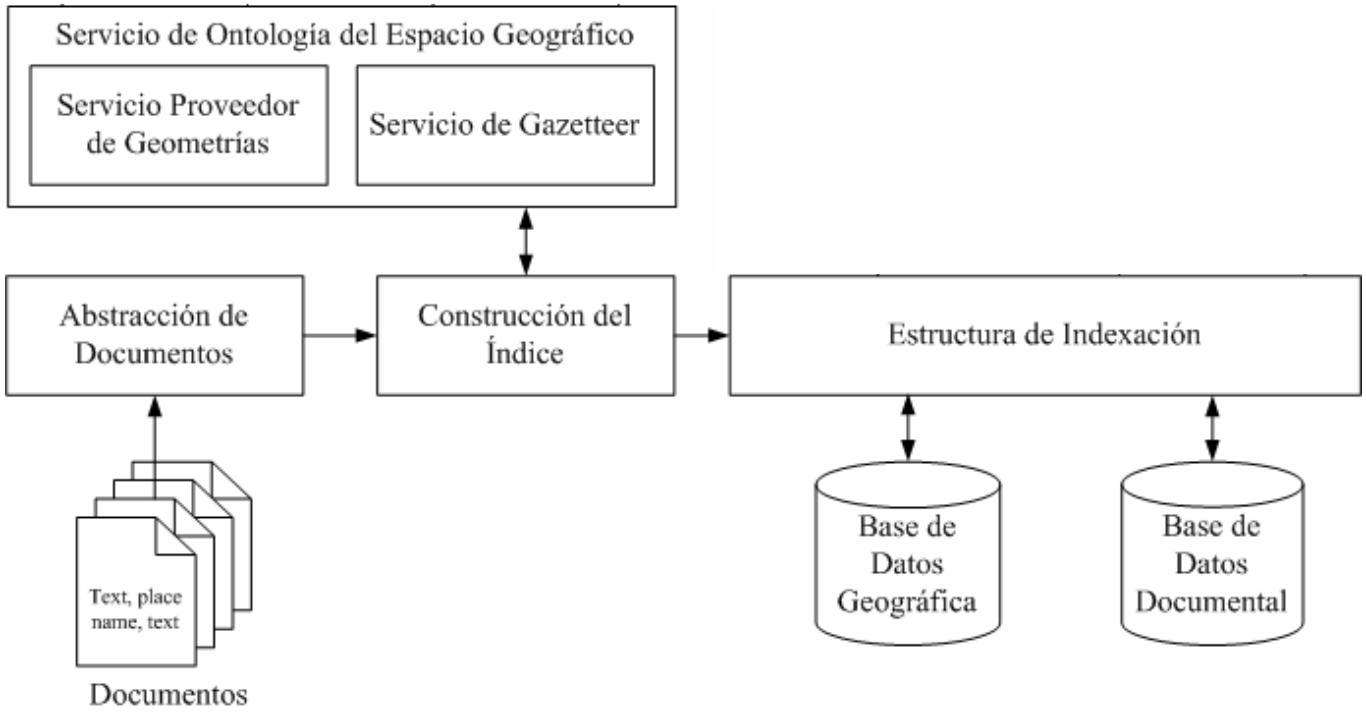
Arquitectura



Arquitectura

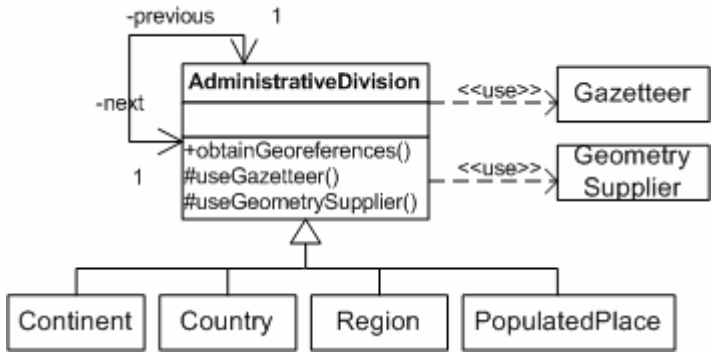
- Abstracción de documentos
  - Documentos diferentes:
    - Diferentes formatos de archivo (texto plano, XML, etc.)
    - Diferentes esquemas de contenido
  - Representación abstracta de esos documentos:
    - *Documento* representado como agregado de *Campos*
    - Idea similar en el motor de búsqueda textual *Lucene*
    - Posibilidad de indexación espacial

Arquitectura

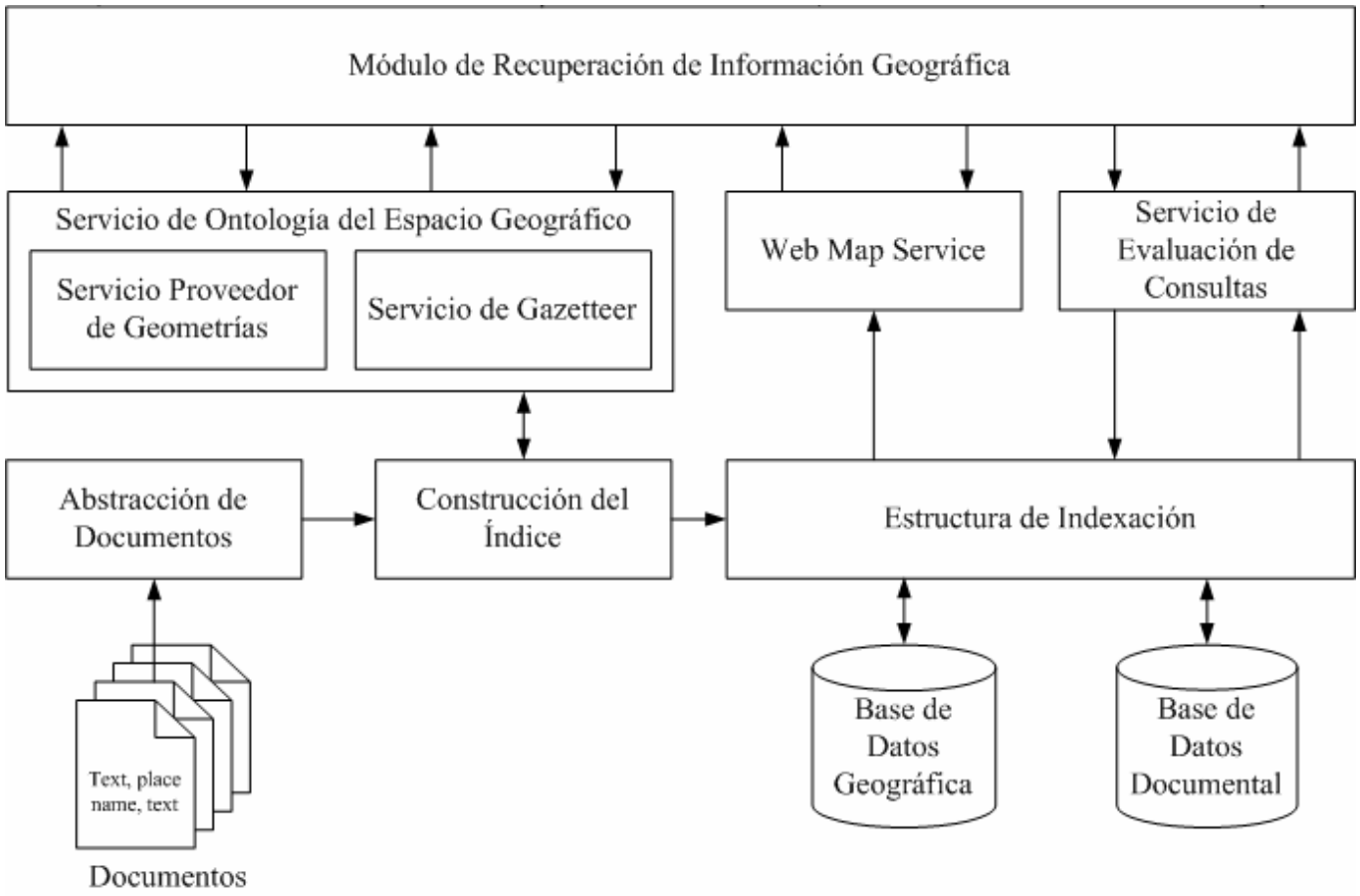


Arquitectura

- Construcción de la estructura de indexación
  - Indexación textual
    - *Lucene*
  - Indexación espacial
    - Obtención de posibles topónimos
      - Análisis Lingüístico: *Named-Entity Recognition*
    - Geo-referenciación de esos topónimos
      - *Servicio de Ontología del Espacio-Geográfico*



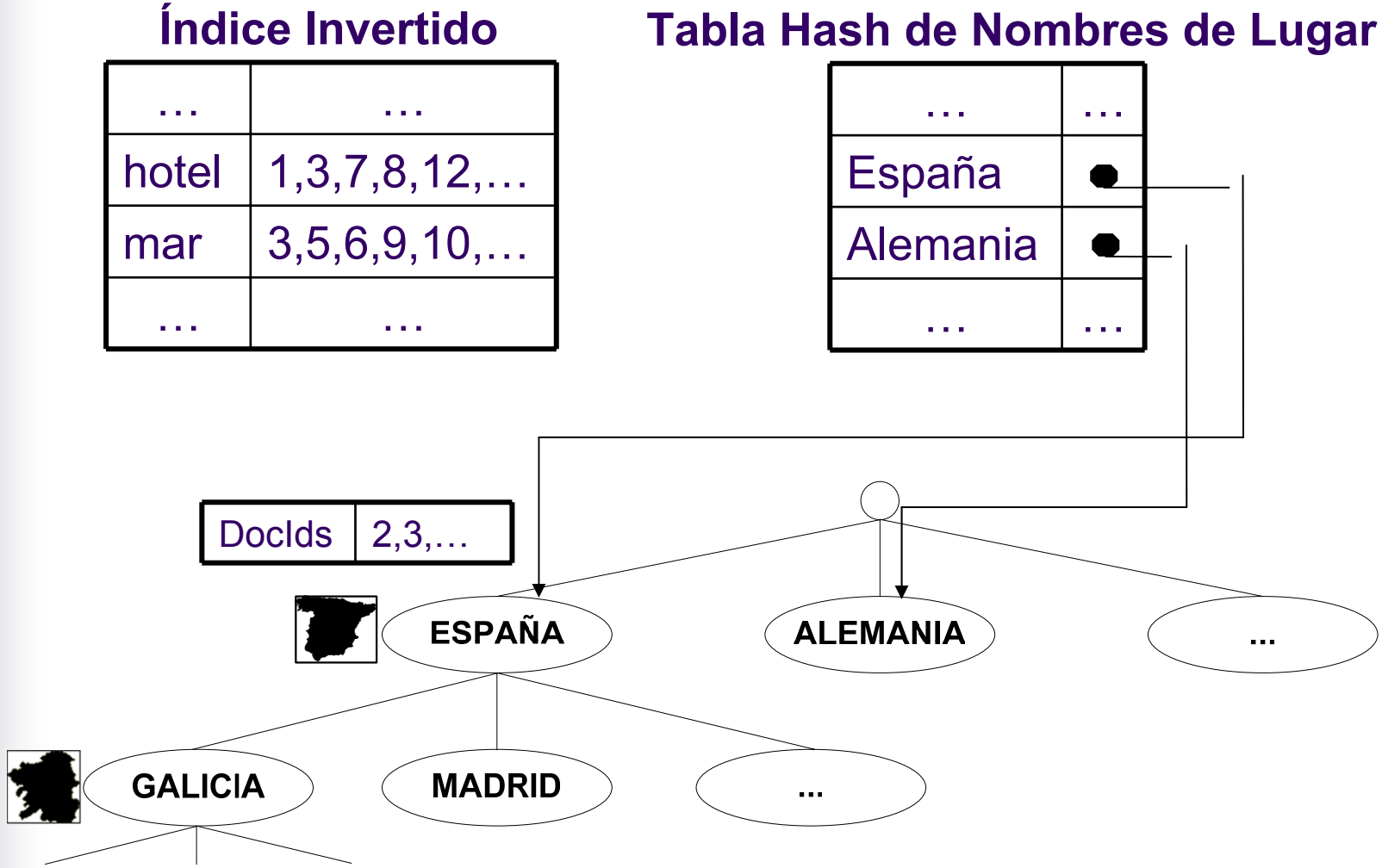
Arquitectura



Organización

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

Estructura de indexación



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

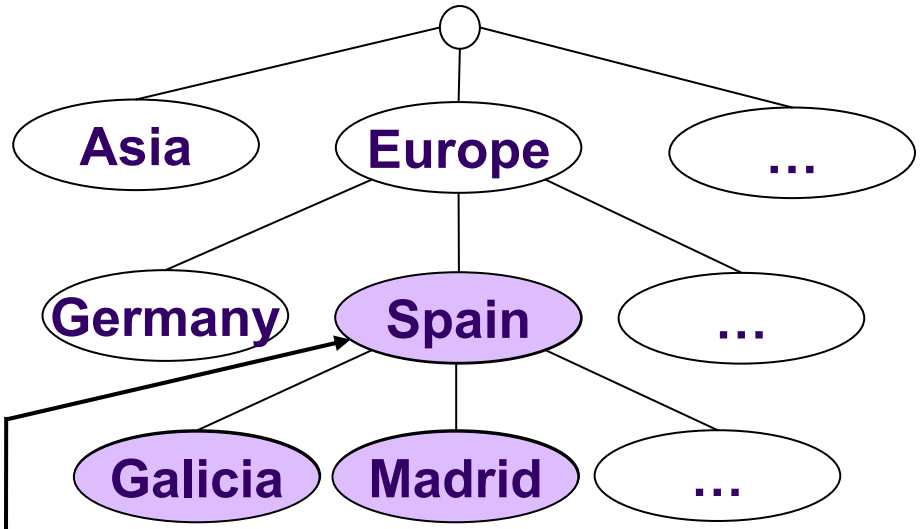
- Consultas textuales con nombres de lugar
  - *"recuperar todos los documentos con la palabra* hotel *referidos a España"*
  - ¿Cómo las resolvemos?
    - Ejemplo
  - Ahorro de tiempo evitando parte del recorrido en el árbol

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	2,3,12,14,…

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

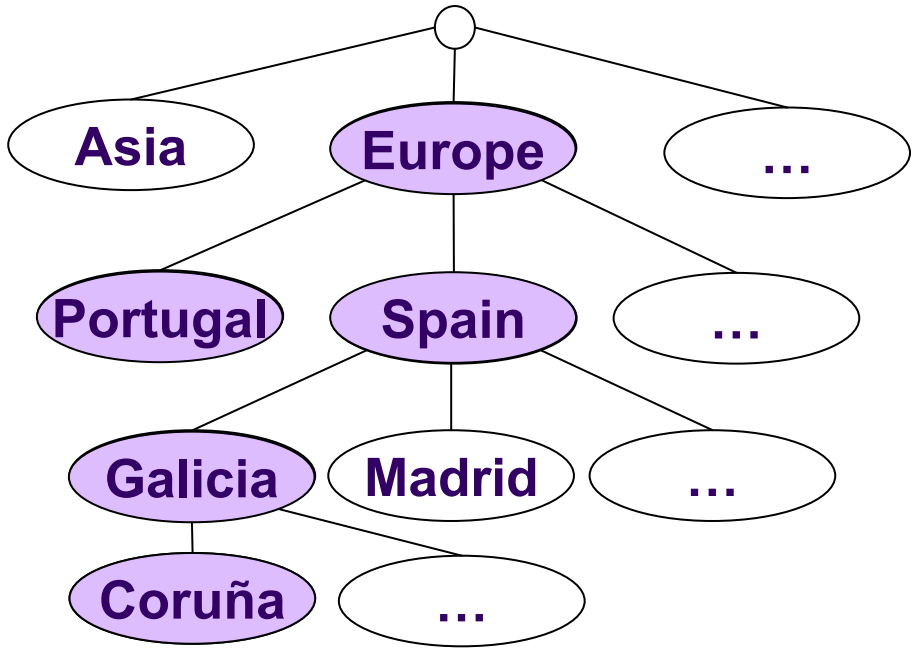
Tipos de consultas soportadas

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,…

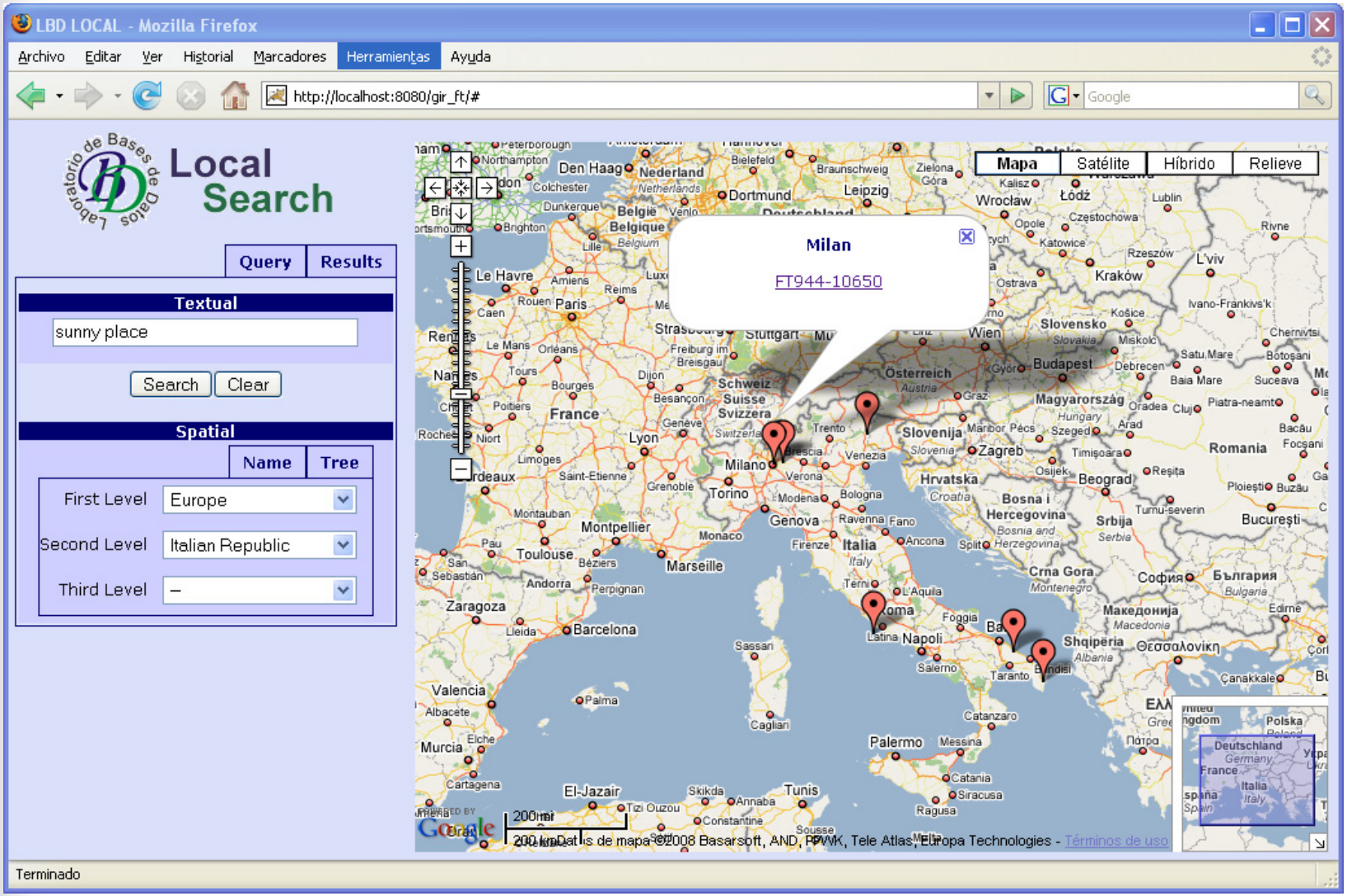
Tipos de consultas soportadas

- 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_ft/#
Local Search
Query
Results
Textual
sunny place
Search
Clear
Spatial
Name
Tree
First Level
Europe
Second Level
Italian Republic
Third Level
–
Mapa
Satélite
Híbrido
Relieve
Milan
FT944-10650
Terminado

Organización

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

Conclusiones y futuros desarrollos

- Conclusiones:
  - Arquitectura de sistema 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:
  - OpenLayers + WMS
  - Evaluación del prototipo
  - Desambiguación de topónimos
  - Implementación de algoritmos de ranking
  - Inclusión de otros tipos de relaciones (ej. Adyacencia)
  - Liberar el código

GRACIAS POR SU ATENCIÓN