Presentación de Trabajo de Fin de Máster

PROPUESTA DE BÚSQUEDA SEMÁNTICA: APLICACIÓN AL CATÁLOGO DE MAPAS, PLANOS Y DIBUJOS DEL ARCHIVO GENERAL DE SIMANCAS

Máster en Lenguajes y Sistemas Informáticos: Tecnologías del Lenguaje en la Web

Universidad de Educación a Distancia

Marzo 2013

Autor: José Alberto Benítez Andrades

Directora: Ana Mª García Serrano

Introducción

- Objetivos
  - Plantear un trabajo de investigación aplicada y una metodología de tratamiento de información estructurada
  - Comparar tres aproximaciones en un dominio concreto
- Estado del arte, 2 perspectivas
  - Tecnológica
  - Teórico-práctica
    - Uso de recursos de la ingeniería lingüística y RI.
    - Uso de ontologías en el área de la Ingeniería del Conocimiento.

Introducción – Trabajo teórico práctico

- La metodología abordada del trabajo realizado es
  - Análisis principal del problema a resolver
  - Estudio de las posibles soluciones existentes para resolver dicho problema
  - Selección de la solución o soluciones
  - Elección de las herramientas necesarias para aplicar dicha solución
  - Desarrollo de herramientas necesarias para abarcar el problema
  - Fase de aplicación de la solución: Experimentación y análisis de resultados.

Parte1: Panorama Tecnológico

- Web Semántica
  - Concepto, herramientas, estándares
  - Definición de ontologías: Protégé
  - Estándares para descripción de contenidos: RDF DC y OWL
  - Método para extraer relaciones semánticas desde la Wikipedia
- Propuesta de un buscador ontológico en Protégé y traductor de RDF DC a OWL necesario para el conjunto de documentos (fichas) estructurados.

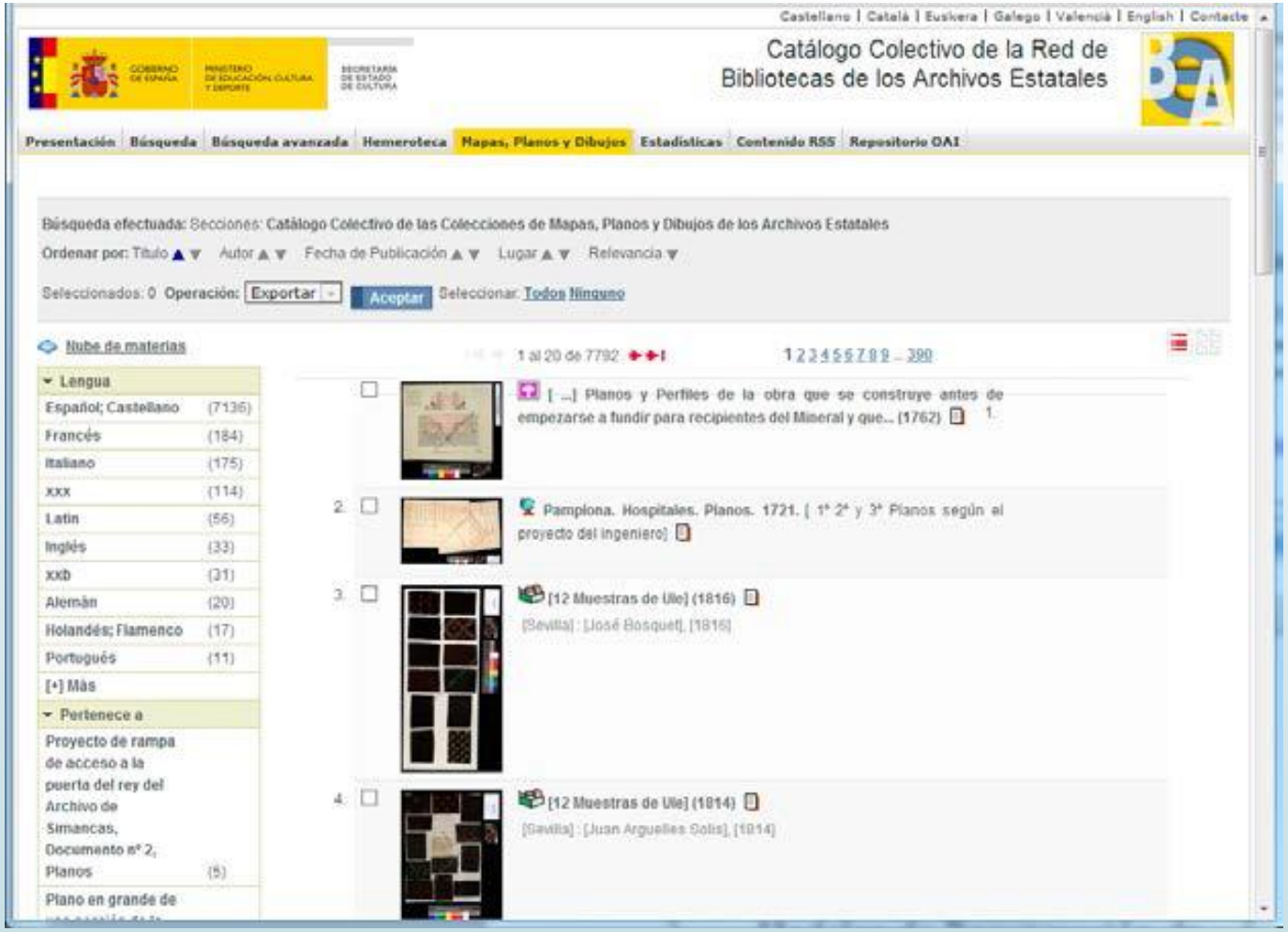
Parte1: Panorama Tecnológico

- Recuperación de información
  - Modelos clásicos de RI
  - Estándares y herramientas relacionadas con Lucene, Solr, Apache y Sparql.
  - Estudio de trabajos relacionados con
    - Los dirigidos por P.Castells sobre búsqueda semántica
    - Sistemas pregunta-respuesta (Question Answering o QA) dirigidos por P. Martínez Barco.
- Propuesta: Metodología para construir una versión de un conjunto de documentos (fichas) estructurados y comparación de un modelo RI facetado y otro textual

Parte 1: Análisis del dominio

- Estructura de las fichas del catálogo **mapas, planos y dibujos del Archivo General de Simancas (AGS)**
  - Poseen 8 campos, facetas:
    - Fecha, Referencias, Creador, Tipo, Idiomas, Temática, Técnica Utilizada
  - El primer paso: descargar las fichas en dos formatos
    - Texto Plano
    - RDF DC

Parte 1: Análisis del dominio



- http://www.mcu.es/ccbae/es/consulta/resultados_busqueda.cmd?tipo_busqueda=mapas_planos_dibujos&posicion=1&id=30485

Parte 1: Análisis del dominio

Parte 2: Propuesta

- Necesidad de realizar un buscador semántico, para mejorar la búsqueda en las fichas.
  - Usuarios no informáticos.
  - Desconocedores de la estructura.
- Planteamiento del trabajo de investigación para conseguirlo:
  1. Desarrollo del entorno para la experimentación
  2. Extracción de información desde el AGS en el formato deseado
  3. Realización de tres aproximaciones de almacenamiento del catálogo
  4. Generación de conjunto de preguntas y comparación entre buscador textual (facetado y no) y ontológico.
  5. Comparación de aproximaciones con TRECEval. No se puedo realizar de forma completa debido a la falta de disponibilidad de juicios de relevancia. Sin embargo se ha realizado una evaluación manual.

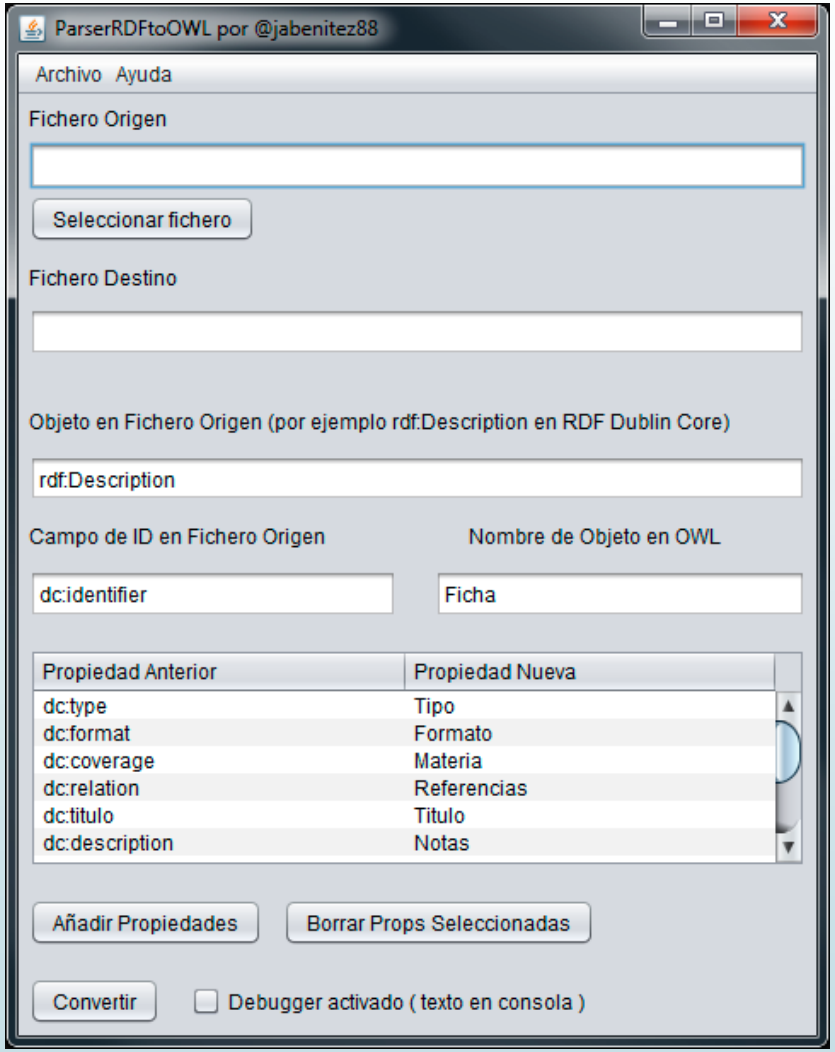
Parte 2: RDF DC a OWL

- Ficha en formato RDF DC

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:dc=
"http://purl.org/dc/elements/1.1/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <rdf:Description>
        <dc:relation>Referencias: Mapas, planos y dibujos (Años 1503-1805). Volumen I : p. 405
        </dc:relation>
        <dc:coverage>-S.XVIII</dc:coverage>
        <dc:title>[ ...] Planos y Perfiles de la obra que se construye antes de empezarse a
        fundir para recipientes del Mineral y que se deshace despues de evacuada la fundición
        que es la que se supone llamarse Crisol [Material gráfico no proyectable]</dc:title>
        <dc:description>AGS. Secretaría de Marina, 00679. Acompaña a carta de don Maximino de
        la Croix a don Ricardo Wall, Chaves, 16 de julio de 1762</dc:description>
        <dc:description>Tinta y colores. Con explicación</dc:description>
        <dc:description>Manuscrito sobre papel.</dc:description>
        <dc:type>Ilustraciones y Fotos</dc:type>
        <dc:language>spa</dc:language>
        <dc:date>1762</dc:date>
        <dc:identifier>http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=178725</dc:identifier>
        <dc:format>image/jpeg</dc:format>
        <dc:subject>Hornos metalúrgicos-Dibujos</dc:subject>
    </rdf:Description>
```

Parte 2: RDF DC a OWL

- Se selecciona el fichero origen
- Se selecciona el fichero de destino
- Se define el nombre del objeto
- Se definen los campos identificadores
- Se agregan las propiedades
- Pulsamos convertir



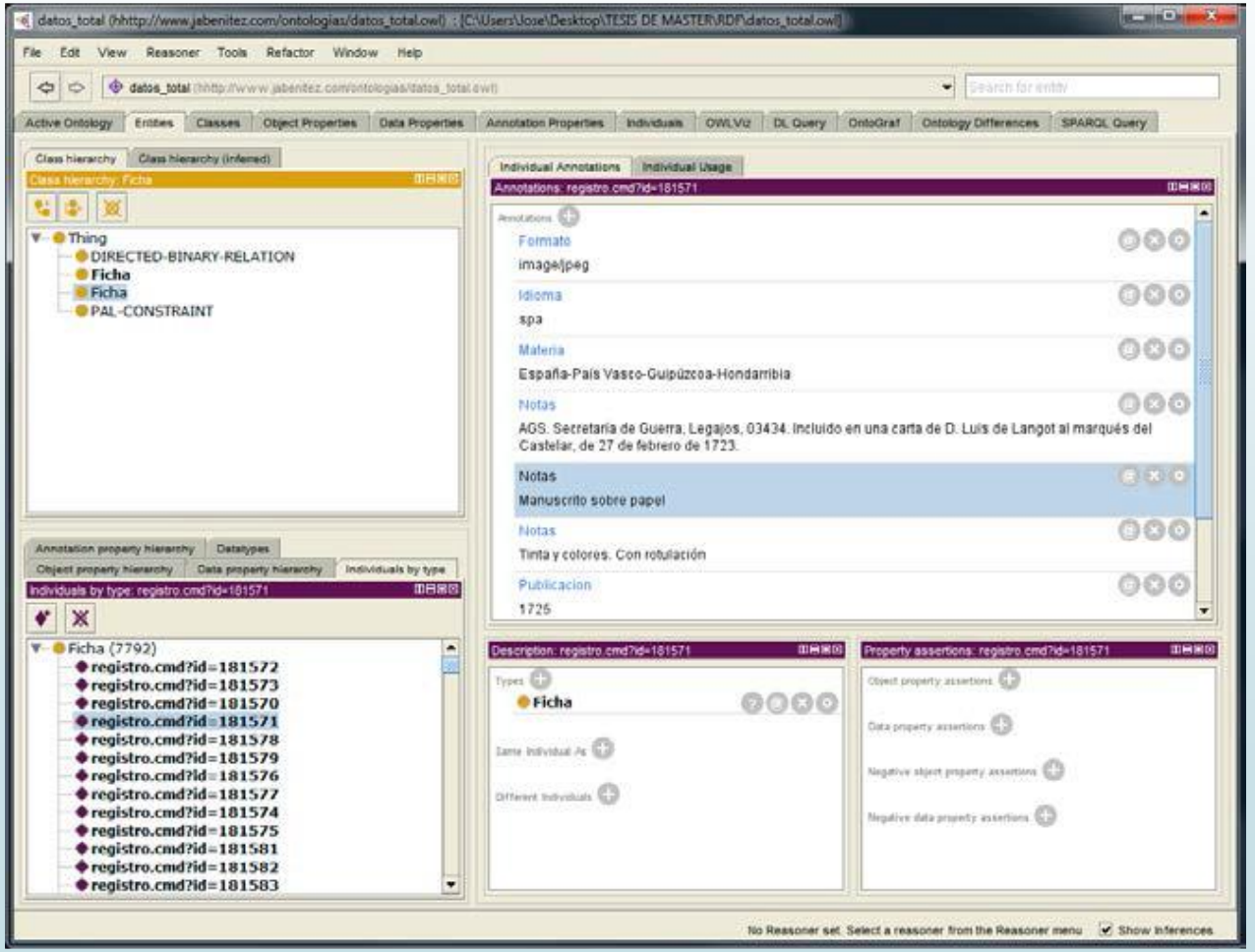
Parte 2: RDF DC a OWL

- Ficha en formato final OWL

```
<Ficha rdf:about="http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=178725">
        <Titulo> Planos y Perfiles de la obra que se construye antes de empezarse a fundir para recipientes del Mineral y que se deshace despues de evacuada la fundición que es la que se supone llamarse Crisol [Material gráfico no proyectable]</Titulo>
        <Referencias>Referencias: Mapas, planos y dibujos (Años 1503-1805). Volumen I : p. 405</Referencias>
        <Materia>-S.XVIII</Materia>
        <Notas>AGS. Secretaría de Marina, 00679. Acompaña a carta de don Maximino de la Croix a don Ricardo Wall, Chaves, 16 de julio de 1762</Notas>
        <Notas>Tinta y colores. Con explicación</Notas>
        <Notas>Manuscrito sobre papel.</Notas>
        <Tipo>Ilustraciones y Fotos</Tipo>
        <Idioma>spa</Idioma>
        <Publicacion>1762</Publicacion>
        <Formato>image/jpeg</Formato>
        <Tematica>Hornos metalúrgicos-Dibujos</Tematica>
</Ficha>
```

Parte 2: Modelo Ontológico con Protégé

- Información de las entidades



Parte 2: Segundo modelo de búsqueda semántica

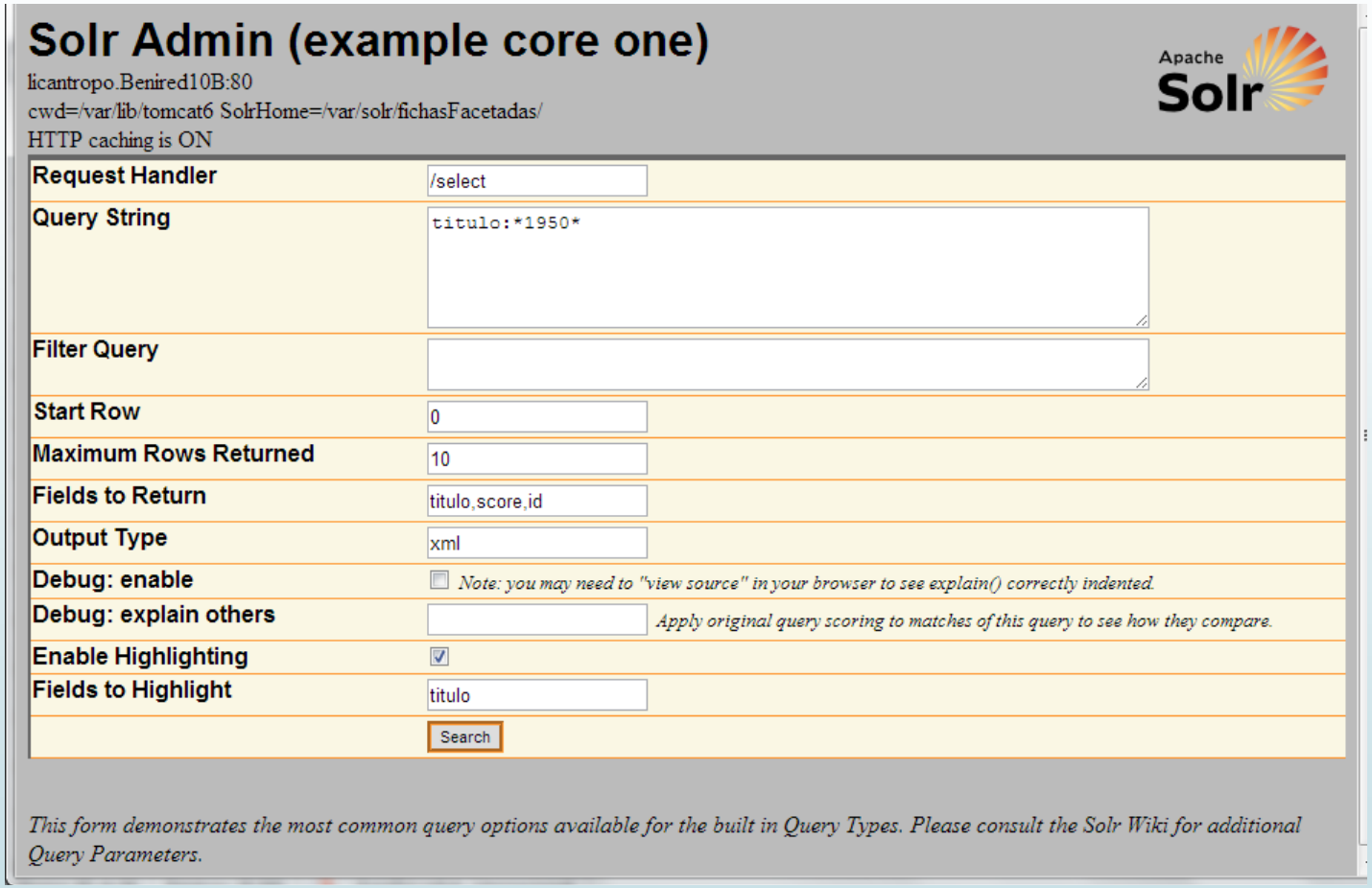
- ¿Ventajas de búsqueda ontológica con Protègè vs otro tipo de buscador?
  - Realización importación de fichas en Apache SOLR
- La fase de aplicación a la solución se subdividió en:
  - Indexación
  - Generación de preguntas de tipo QA
  - Realización de búsquedas de elementos de forma facetada y de forma textual
  - Comparación de los resultados obtenidos

Parte 2: Clasificación de consultas (ej)

- Clases: Qué, quién, dónde, cuándo, cómo

DÓNDE	se realizó / se utilizó / se realizó	El/la/los/ las Algún/ alguno/ algunos/ algunas un/una/ unos/unas *planos/ dibujos mapas*	por el autor **$AUTOR** con el título **$TÍTULO** con los colores **$COLOR** entre las fechas **$FECHA** y **$FECHA** en el idioma **$IDIOMA** en los idiomas **$IDIOMA** y **$IDIOMA** en la época **$ÉPOCA** con la temática **$TEMÁTICA** sobre el soporte **$SOPORTE**	[Ciudad] [País] [Continente]
CUÁNDO	se realizó / se utilizó / se realizó	El/la/los/ las Algún/ alguno/ algunos/ algunas un/una/ unos/unas *planos/ dibujos mapas*	por el autor **$AUTOR** con el título **$TÍTULO** con los colores **$COLOR** entre las fechas **$FECHA** y **$FECHA** en el idioma **$IDIOMA** en los idiomas **$IDIOMA** y **$IDIOMA** en la época **$ÉPOCA** con la temática **$TEMÁTICA** sobre el soporte **$SOPORTE** sobre la ciudad **$CIUDAD** sobre el país de **$PAÍS**	[Época] [Año]

Parte 2: Consultas en Solr



http://casa.jabenitez.com/solr/fichasFacetadas/select?indent=on&version=2.2&q=titulo%3A*1952*&fq=&start=0&rows=10&fl=titulo%2Cscore&wt=&explainOther=&hl=on&hl.fl=titulo

Parte 2: Consultas SPARQL en Protègè

- Ejemplo de una de las consultas
- Resto, ver PDF páginas 111 a 115

Consulta en Protègè para **Qué fichas datan sobre una obra realizada en 1950:**

```
SELECT distinct ?s ?p ?o
WHERE {
        ?s ?p ?o .
        ?s a :Ficha .
        FILTER (regex(?o, "^1950"))}
```

Resultados obtenidos 4:

Results

s	p	o
http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=182452	Publicacion	1950
http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=182451	Publicacion	1950
http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=182449	Publicacion	1950
http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=182450	Publicacion	1950

Parte 2: Buscador textual

- **buscadorTextual.js**
  - Se crea un objeto AJAX: realizará la consulta a SOLR y recibirá la respuesta en JSON
  - Almacenamos los datos en una variable llamada **datos_formulario**
  - Obtenida la cadena, asignamos la propiedad URL a nuestro objeto AJAX con el siguiente valor:
    - http://casa.jabenitez.com/solr/fichas/select?indent=on&version=2.2&q='+datos_formulario+'&fq=&start=0&rows=7500&fl=id%2Cscore&wt=json&explainOther=&hl=on&hl.fl=titulo
  - Un ejemplo práctico: Supongamos que el usuario desea tener la siguiente información: ¿Qué mapas fueron realizados en el año 1950 por el autor Juan Baptista?
  - En el caso de nuestro buscador semántico, recogerá la cadena de texto introducida por el usuario y nos devolverá los resultados de buscar esa cadena completa, no sabiendo interpretar qué tipo de datos necesitaría ni de dónde debe obtenerlos (y tampoco se lo indica el usuario, que no lo conoce).

Parte 2: Buscador facetado

- **buscadorSemantico.js**

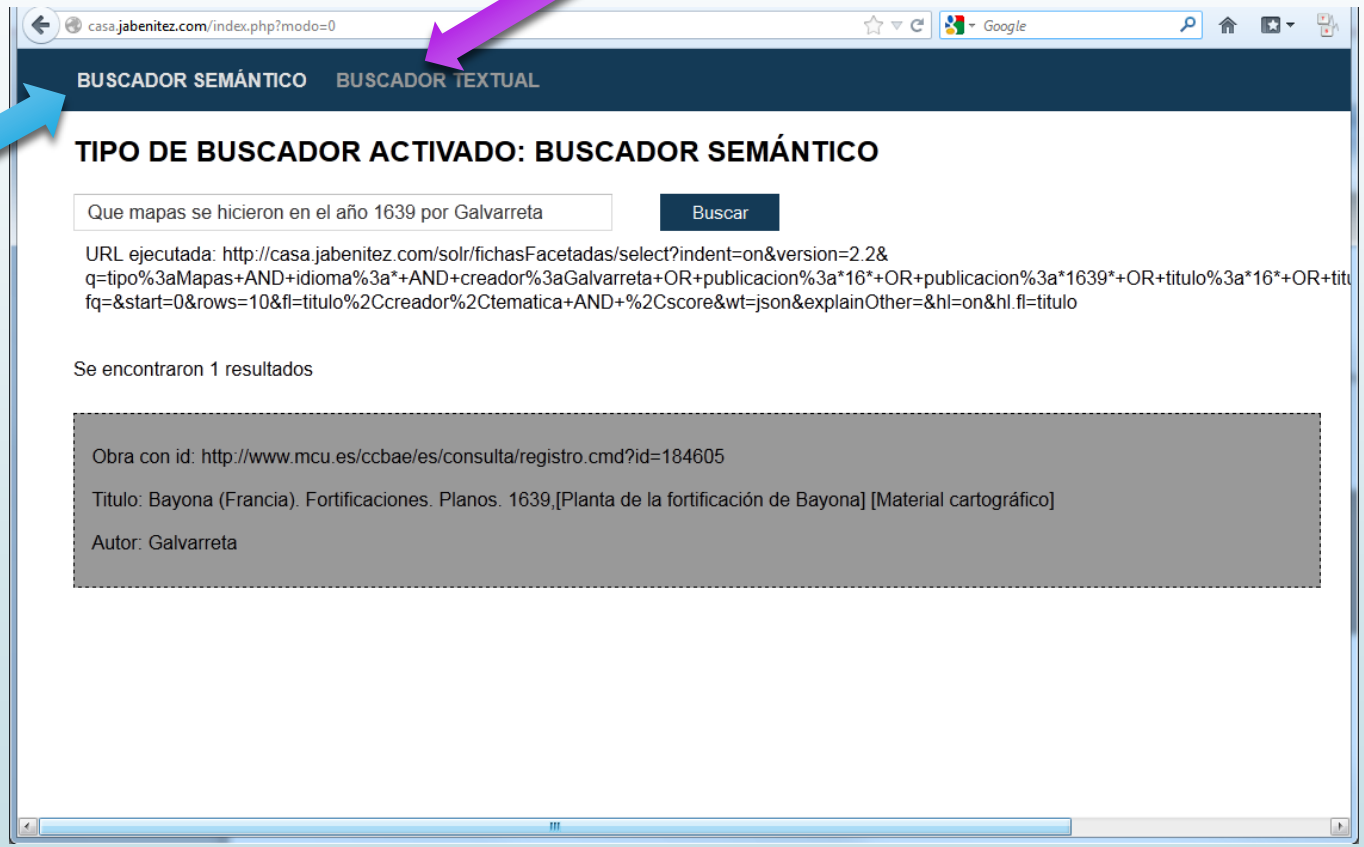
1. Teniendo en cuenta el sistema QA creado, sabemos que existen 5 tipos de pregunta (Qué, Quién, Cómo, Dónde y Cuándo).

   El parser detecta en un primer análisis qué pregunta se encuentra en la cadena insertada por el usuario.

2. En un segundo procesado de la cadena, se analiza el objeto al que se hace referencia: **autor, mapa, obra, dibujo u otro campo.**
3. En una tercera fase, se detecta el verbo de la frase, siendo capaz de detectar **a qué campo** de las fichas facetadas se debe consultar.
4. Finalmente, el objeto AJAX que habíamos creado, recibe la información en formato JSON y la muestra por pantalla.

Parte 2: Parser semántico que traduce consultas a Solr

- Interfaz gráfica del buscador:
  - Buscador semántico
  - Buscador textual



Parte 2: Experimentación y comparación

- Teniendo en cuenta los diferentes buscadores que hemos obtenido tras realizar un análisis específico de nuestro caso particular, a continuación detallaré en una tabla los distintos resultados obtenidos de las consultas realizadas en la base de datos de las fichas sin facetar y facetadas y en el ontológico.
- Por falta de juicios de relevancia, solo se pudo realizar una comparación de forma manual.
- Los análisis realizados, aunque de forma manual, tienen buenas expectativas, tanto en la precisión como en la cobertura.

Parte 2: Experimentación (ejm)

- **Consulta 1** en Solr para ¿Qué fichas datan sobre una obra realizada en 1950?:

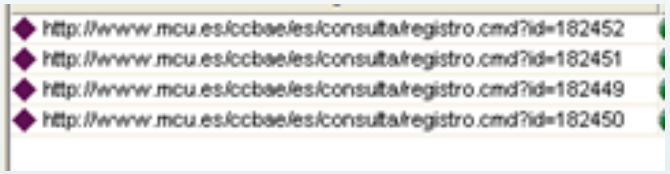
SOLR FACETADA – 4 RESULTADOS

```
▼<response>
 ▼<lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">238</int>
  </lst>
 ▶<result name="response" numFound="4" start="0" maxScore="0.033296965">...</result>
 ▼<lst name="highlighting">
    <lst name="http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=182452"/>
    <lst name="http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=182451"/>
    <lst name="http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=182450"/>
    <lst name="http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=182449"/>
  </lst>
</response>
```

SOLR SIN FACETAR – 9 RESULTADOS

```
▼<response>
 ▼<lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">932</int>
  </lst>
 ▶<result name="response" numFound="9" start="0" maxScore="1.0">...</result>
 ▶<lst name="highlighting">...</lst>
</response>
```

PROTÈGÈ – 4 RESULTADOS



Conclusiones:

- De los 9 resultados en SOLRsin facetar, 4 comunes.
- 5 Que poseen otro campo distinto a la fecha de creación que coincide con la búsqueda 1950

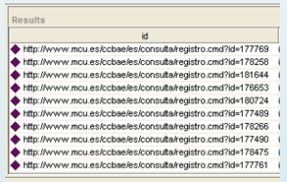
Parte 2: Experimentación (ej)

- **Consulta 2** en Solr para ¿Qué fichas están fabricadas en pergamino?:

SOLR FACETADA – 21 RESULTADOS

```
▼<response>
 ▼<lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">235</int>
   </lst>
 ▶<result name="response" numFound="21" start="0" maxScore="1.0">...</result>
 ▼<lst name="highlighting">
```

PROTÈGÈ – 21 RESULTADOS



SOLR SIN FACETAR – 21 RESULTADOS

```
▼<response>
 ▼<lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">843</int>
   </lst>
 ▶<result name="response" numFound="21" start="0" maxScore="1.0">...</result>
 ▶<lst name="highlighting">...</lst>
 </response>
```

Conclusiones:

- En este caso coinciden los resultados porque la información se obtiene del campo descripción, es una búsqueda muy similar en ambas bases de datos.

Parte 2: Experimentación (ej)

- **Consulta 4** en Solr para ¿Qué fichas fueron creadas por Juan Baptista?:

SOLR FACETADA – 1 RESULTADO

```
▼<response>
 ▼<lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">13</int>
  </lst>
 ▶<result name="response" numFound="1" start="0" maxScore="1.4142135">...</result>
 ▼<lst name="highlighting">
    <lst name="http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=178297"/>
  </lst>
</response>
```

SOLR SIN FACETAR – 3 RESULTADOS

```
▼<response>
 ▼<lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">1828</int>
  </lst>
 ▶<result name="response" numFound="3" start="0" maxScore="1.4142135">...</result>
</response>
```

PROTÈGÈ – 3 RESULTADO

Results

id
http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=180501
http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=178297
http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=176716

Conclusiones:

- De los 3 resultados, 1 de ellos común.
- Los otros dos no son obras de Juan Baptista, pero aparece en las “Notas” de la obra.

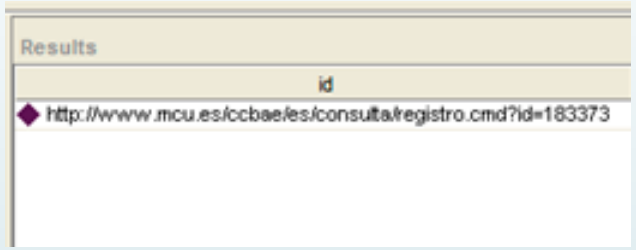
Parte 2: Experimentación (ej)

- **Consulta 6** en Solr para ¿Qué fichas poseen como temática ballenas?:

SOLR FACETADA – 1 RESULTADO

```
▼<response>
 ▼<lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">13</int>
   </lst>
 ▶<result name="response" numFound="1" start="0">...</result>
 ▼<lst name="highlighting">
    <lst name="http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=183373"/>
   </lst>
 </response>
```

PROTÈGÈ – 1 RESULTADO



SOLR SIN FACETAR – 1 RESULTADO

```
▼<response>
 ▼<lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">1117</int>
   </lst>
 ▶<result name="response" numFound="1" start="0" maxScore="0.5">...</result>
 </response>
```

Conclusiones:

- El resultado es el mismo que en los otros casos.

Parte 2: Experimentación (ej)

- **Consulta 7** en Solr para ¿Qué obras tienen las medidas en varas?:

SOLR FACETADA – 467 RESULTADOS

```
▼<response>
 ▼<lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">544</int>
  </lst>
 ▶<result name="response" numFound="467" start="0" maxScore="1.0">...</result>
 ▶<lst name="highlighting">...</lst>
 </response>
```

PROTÈGÈ – 467 RESULTADOS

Results

id
http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=180666
http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=177247
http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=176667
http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=178574
http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=184043
http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=176947
http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=176922
http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=179169
http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=178424
http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=180908

SOLR NO FACETADA – 1683 RESULTADOS

```
▼<response>
 ▼<lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">980</int>
  </lst>
 ▶<result name="response" numFound="1683" start="0" maxScore="1.0">...</result>
 </response>
```

Conclusiones:

- En SOLR no facetada detecta la palabra varas en el campo, pero no están las medidas realmente realizadas en esa unidad.

Parte 2: Experimentación (ej)

- **Consulta 8** en Solr para ¿Qué obras representan un monumento de ceuta?:

SOLR FACETADA – 257 RESULTADOS

```
▼<response>
  ▼<lst name="responseHeader">
     <int name="status">0</int>
     <int name="QTime">18</int>
   </lst>
  ▶<result name="response" numFound="257" start="0" maxScore="1.0">...</result>
```

SOLR NO FACETADA – 3 RESULTADOS

```
▼<response>
  ▼<lst name="responseHeader">
     <int name="status">0</int>
     <int name="QTime">446</int>
   </lst>
  ▶<result name="response" numFound="3" start="0" maxScore="1.0">...</result>
 </response>
```

Conclusiones

- Los resultados en este caso son menos debido a que el campo “materia” en la exportación de fichas en texto plano, no se incluía en dicha exportación.

6.Conclusiones

- El problema de la búsqueda semántica a partir de texto en lenguaje natural.
- La comparación entre el modelo de búsqueda textual no facetado, facetado y ontológico (o clasificación con la ontología) nos permite observar que para alcanzar conclusiones relevantes es necesario seguir investigando en modelos y técnicas semánticas.

6.Futuras líneas de trabajo

- Se proponen dos metodologías para el enriquecimiento semi-automático:
  - Metodología basada en la distancia contextual y ganancia de conocimiento
  - Metodología basada en roles semánticos.
- Ambas abordarían la instanciación automática desde la combinación del análisis lingüístico tradicional y las tecnologías de extracción de conocimiento textual

6.Otras futuras líneas de trabajo

- Otras líneas de trabajo futuro
  - Importancia en el tiempo de respuesta y los resultados obtenidos en las consultas realizadas a las fichas facetadas.
  - Buscador semántico vs buscador textual
    - Análisis de utilización de recursos lingüísticos:
      - Automatización de eliminación de stopwords y otros.
    - Mejora en el algoritmo de funcionamiento del parser semántico.
    - Enriquecedor semántico automatizado

Presentación de Trabajo de Fin de Máster

PROPUESTA DE BÚSQUEDA SEMÁNTICA: APLICACIÓN AL CATÁLOGO DE MAPAS, PLANOS Y DIBUJOS DEL ARCHIVO GENERAL DE SIMANCAS

Máster en Lenguajes y Sistemas Informáticos: Tecnologías del Lenguaje en la Web

Universidad de Educación a Distancia

Marzo 2013

Autor: José Alberto Benítez Andrades

Directora: Ana Mª García Serrano