
esDBpedia:
núcleo de los datos semánticos del español
#DigHumesDBpedia
Mariano Rico, Oct 2015
@marianorico
POLITÉCNICA

Contenido

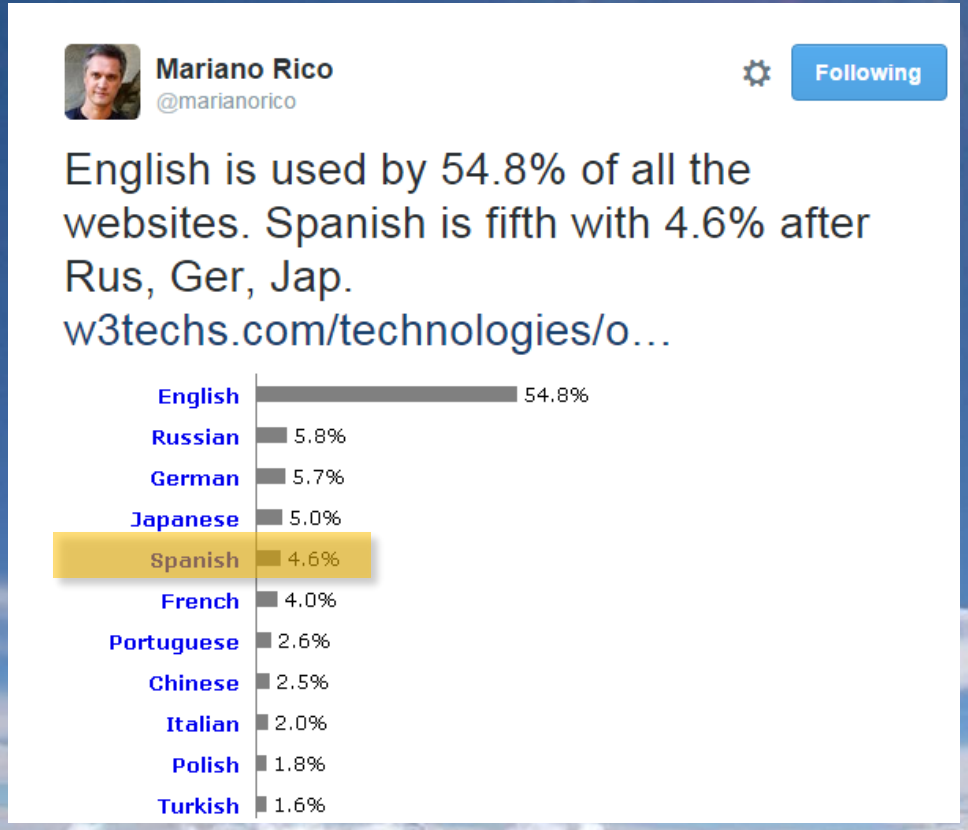
- esDBpedia
- esDBpedia como fuente de datos semánticos
- Lexicalizaciones y multilingüismo
  - Lemonade tools
- Proyectos futuros

¿Qué es ?

ESDBPEDIA

#DigHumesDBpedia

El español en Internet



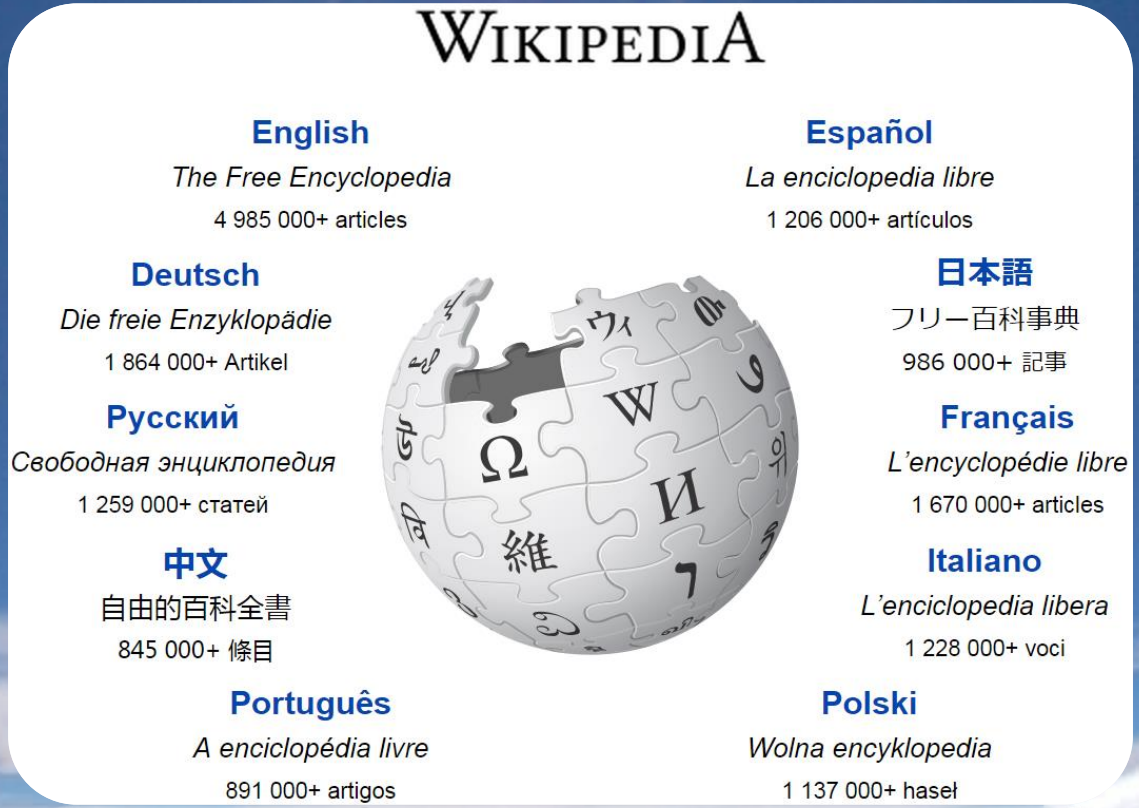
#DigHumesDBpedia

El español en Wikipedia

Ranking:

(Oct. 2015)

1º Inglés (5.0 M)
2º Alemán (1.9 M)
3º Francés (1.7 M)
4º Ruso (1.3M)
5º Italiano (1.2M)
6º Español (1.2M)
7º Polaco (1.1M)
8º Japonés (1.0M)
9º Portugués (0.9M)
10º Chino (0.8M)

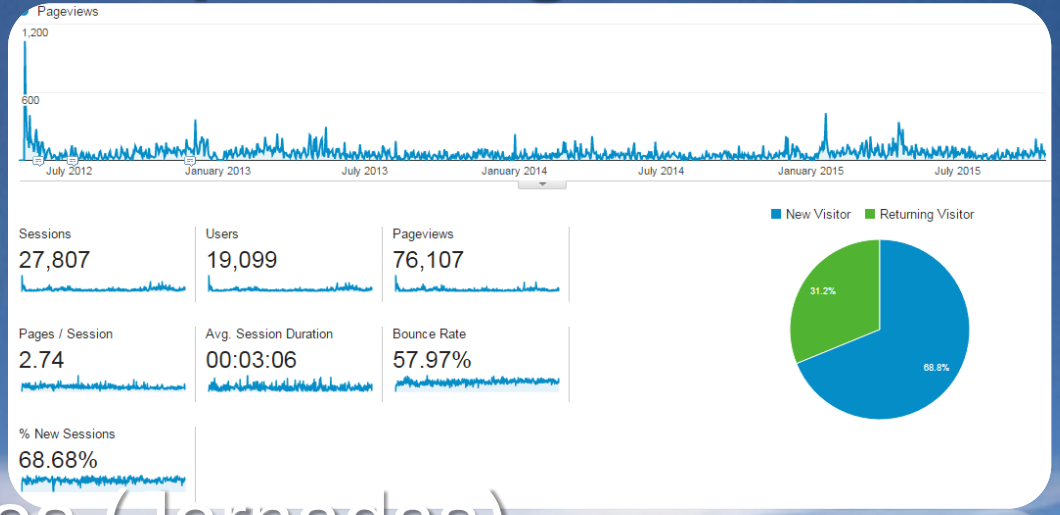


¿Qué es esDBpedia?

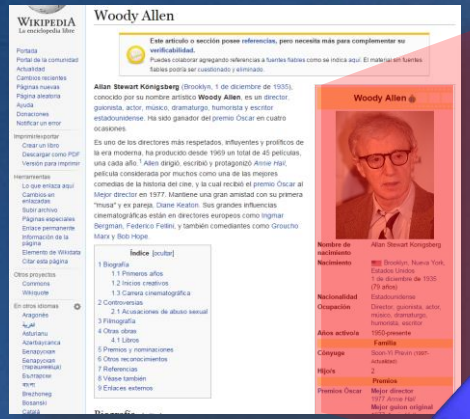
- Es la DBpedia del idioma español
  - No es la DBpedia de un país, sino la de una lengua
- Almacén de datos semánticos obtenidos de la Wikipedia del español

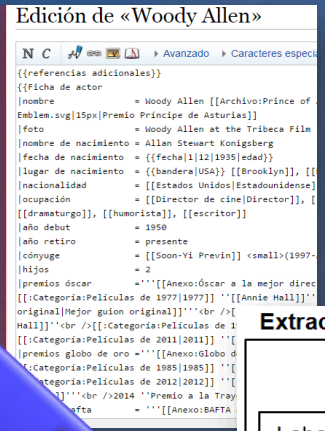
Visibilidad esDBpedia

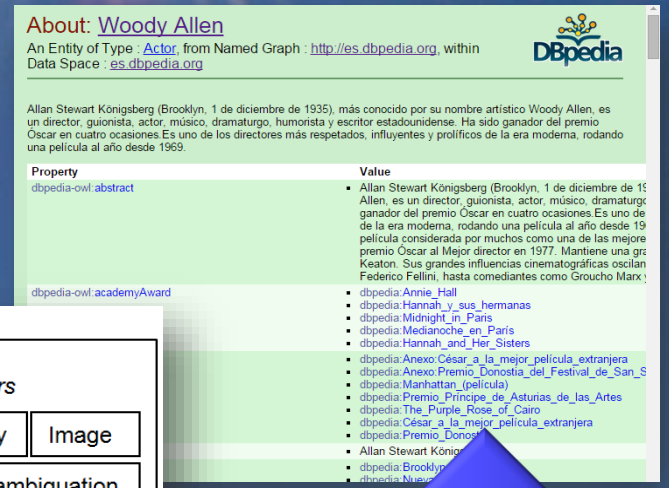
- Sitio web **http://es.dbpedia.org**
- Redes sociales
  - Facebook
  - Twitter
  - Google+
- Eventos periódicos (Jornadas)
- SPARQL EP



#DigHumesDBpedia







infobox

mappings

Extraction Job

Extractors

Label	Category	Image
Redirect	Disambiguation	
Abstract	Geo	Pagelink
Generic Infobox		
Mapping-based Infobox		

Parsers

DateTime	Units	Geo
String-List	Numbers	

RDF

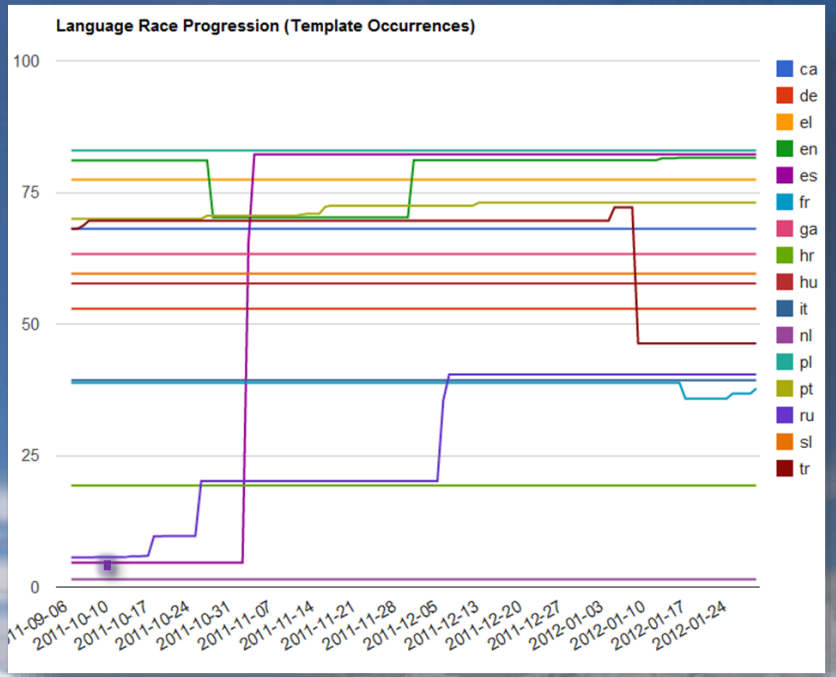
Triple store

Rendering

#DigHumesDBpedia

Mapeos, mapeos y mapeos

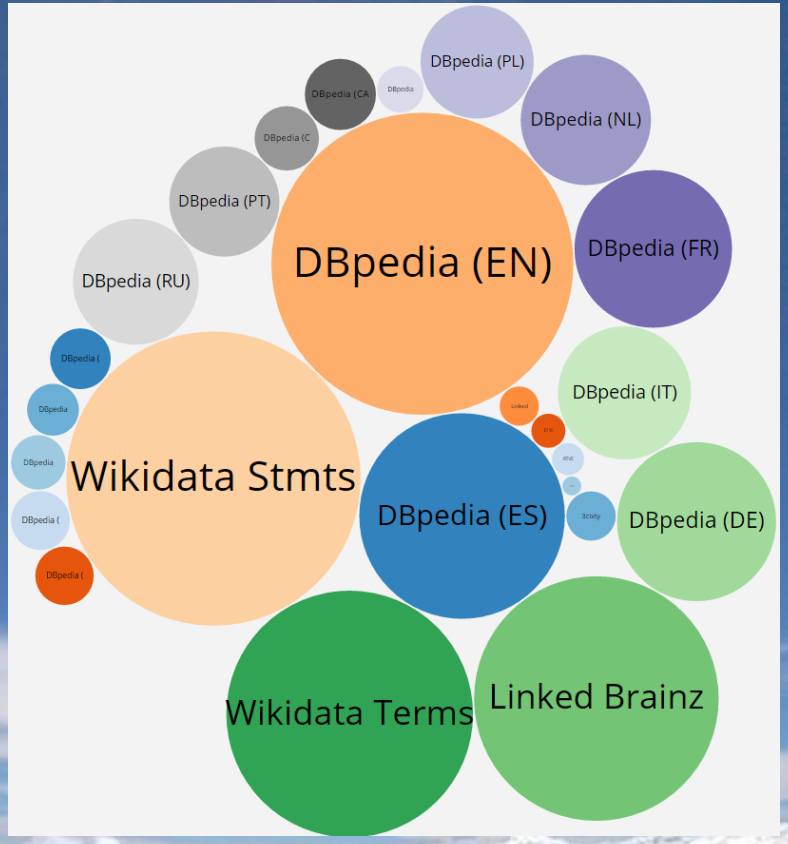
- Jornadas de mapeo (3 ediciones)
- E.g.: Jornadas de mapeos Nov. 2011
  - 15 personas
  - 4h + 4h
  - 101 clases mapeadas (80% instancias)



#DigHumesDBpedia

Datos xxDBpedias

- Primicia: esDBpedia es la **primera** después de la inglesa

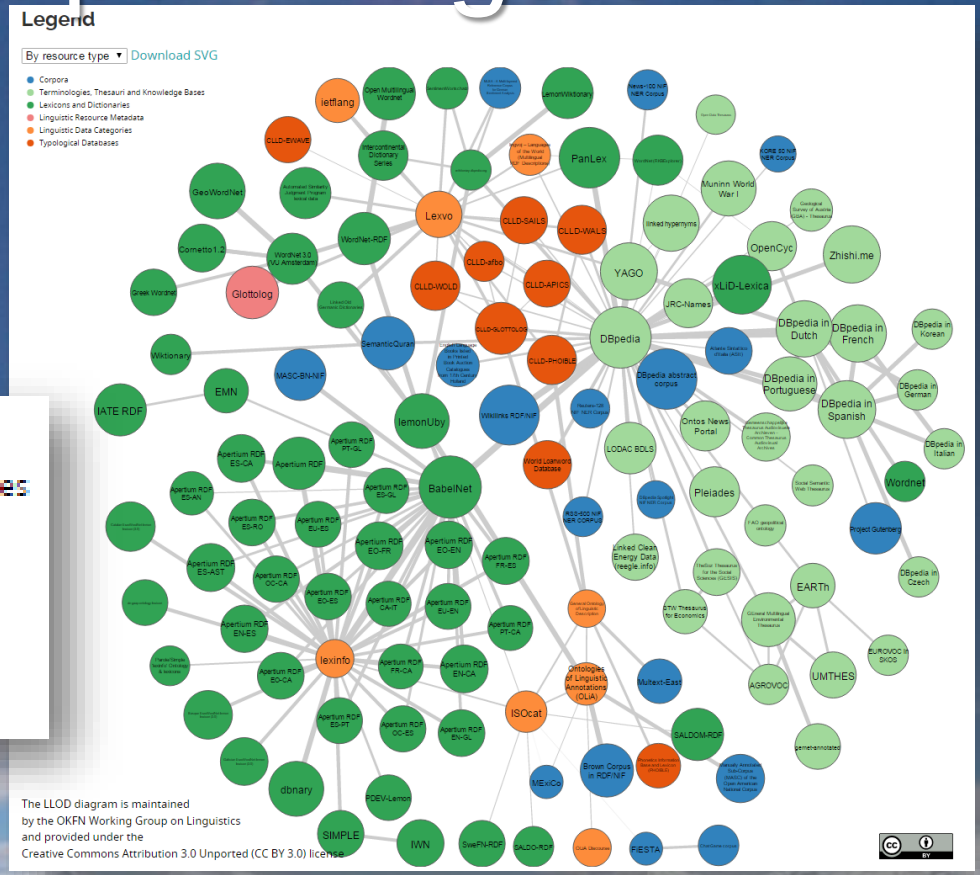


#DigHumesDBpedia

esDBpedia para lingüistas

- Linguistic LOD

- Corpora
- Terminologies, Thesauri and Knowledge Bases
- Lexicons and Dictionaries
- Linguistic Resource Metadata
- Linguistic Data Categories
- Typological Databases



Datos con sentido

ESDBPEDIA COMO FUENTE DE DATOS SEMÁNTICOS

#DigHumesDBpedia

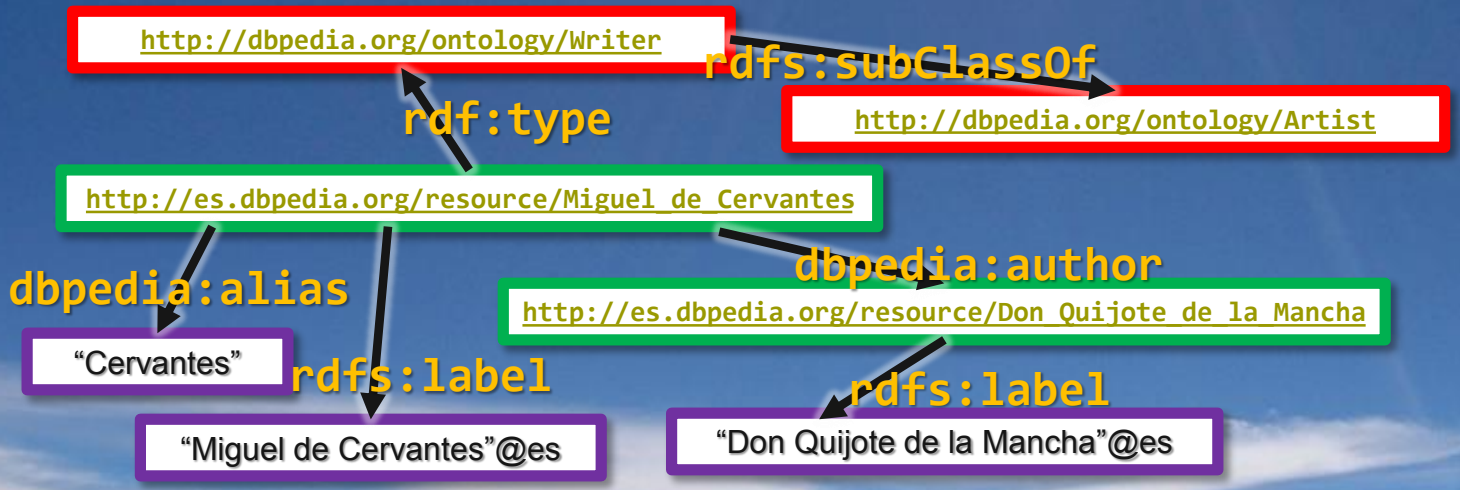
Datos con semántica

Datos con semántica

- Semánticos porque
  - Datos enlazados a una (o varias) ontologías
    - Una ontología es un modelo matemático que permite
      - Razonamiento automático ➔ Inferir nuevos datos
      - Hacer preguntas (e.g. ¿quién es la esposa de Obama?) o saber si algo en cierto (e.g. ¿Michelle es la esposa de Obama?)
  - Sin enlaces a las ontologías los datos no tienen semántica

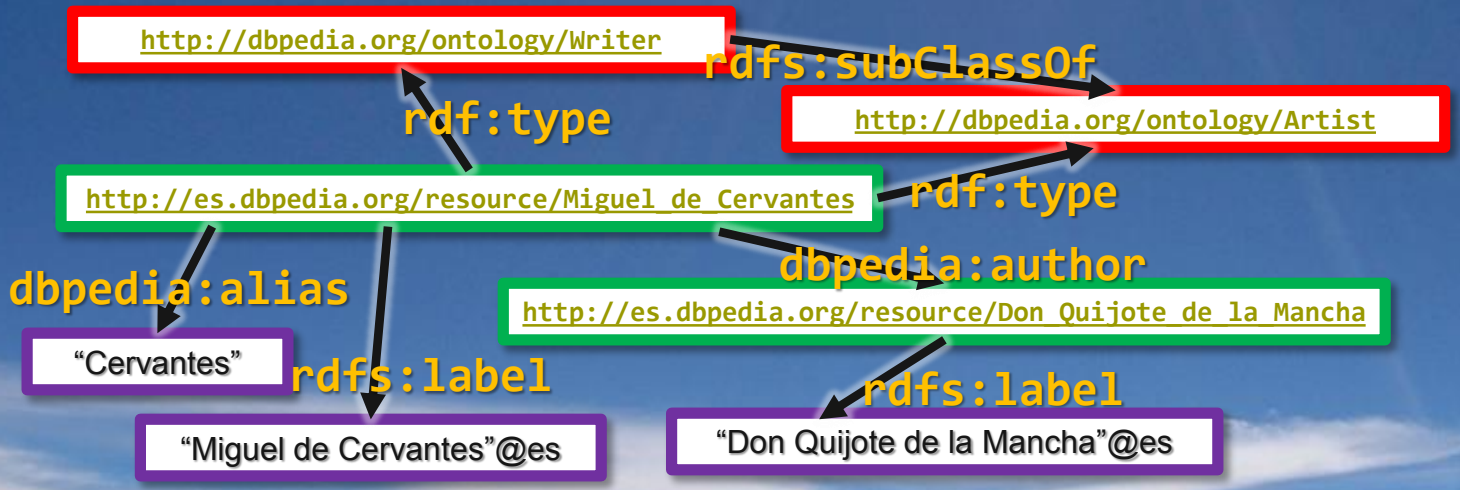
Datos con semántica

- E.g.: “Cervantes escribió Don Quijote”



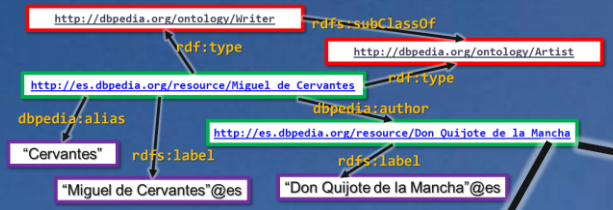
Datos con semántica

- Razonando: “Cervantes es un artista”



Datos con semántica

- Añade más datos enlazados



dbpedia:thumbnail

http://upload.wikimedia.org/Wikipedia/commons/thumbd...

dbpedia:literaryGenre

http://es.dbpedia.org/resource/Novela_de_aventuras

dbpedia:literaryGenreOf

dbpedia:author

http://es.dbpedia.org/resource/La_vida_de_Pi

http://es.dbpedia.org/Yann_Martel

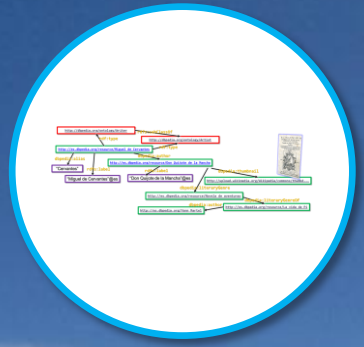
Datos con semántica

- Añade más datos enlazados

#DigHumesDBpedia

Datos con semántica

- Dataset



esDBpedia

- Dataset con datos de la Wikipedia del español

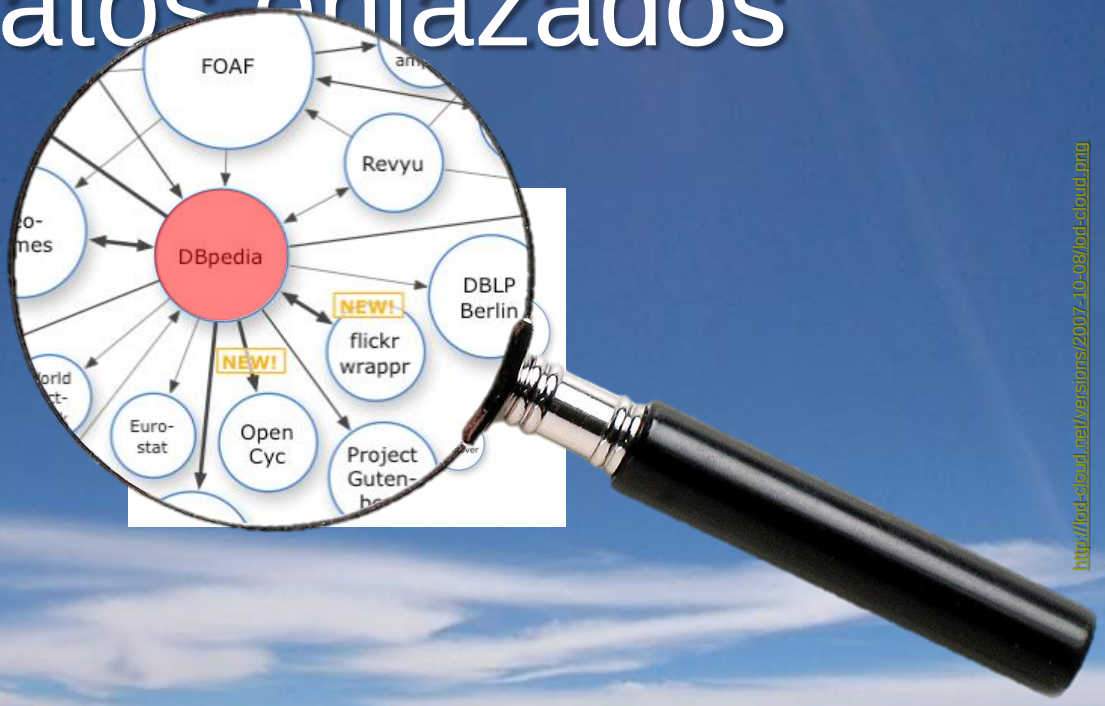
- 160 millones de datos
  de 1 millón de entradas de Wikipedia del español.
- Enlazados a ontología DBpedia
  - Cientos de conceptos
  - Miles de relaciones entre conceptos

#DigHumesDBpedia

DBpedia, núcleo de los datos enlazados

2007

25 datasets



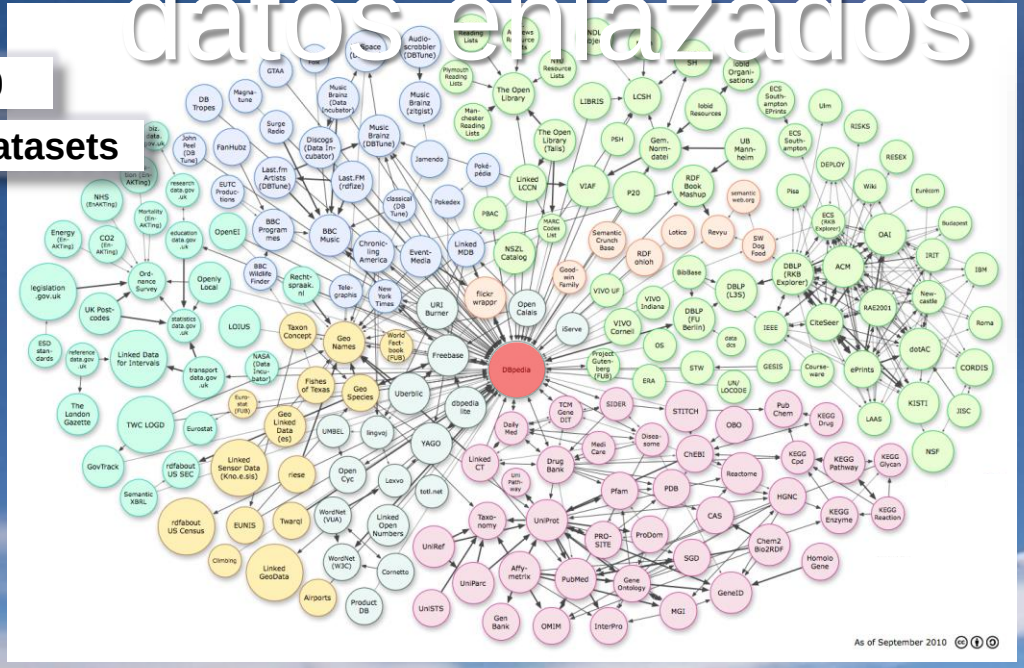
http://lod-cloud.net/versions/2007-10-08/lod-cloud.png

#DigHumesDBpedia

DBpedia, núcleo de los datos enlazados

2010

203 datasets



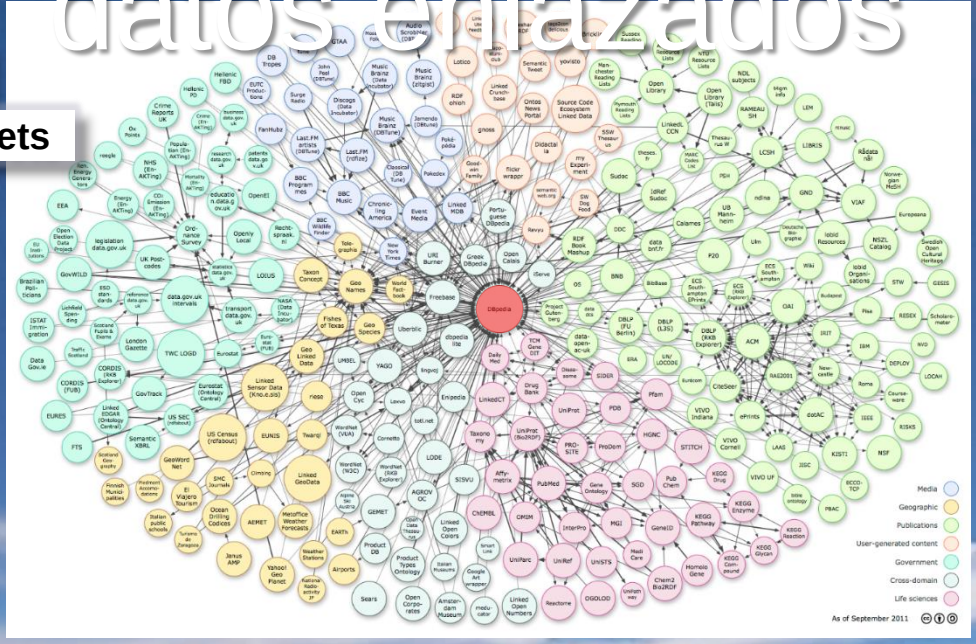
http://lod-cloud.net/versions/2011-09-19/lod-cloud_colored.png

#DigHumesDBpedia

DBpedia, núcleo de los datos enlazados

2011

295 datasets



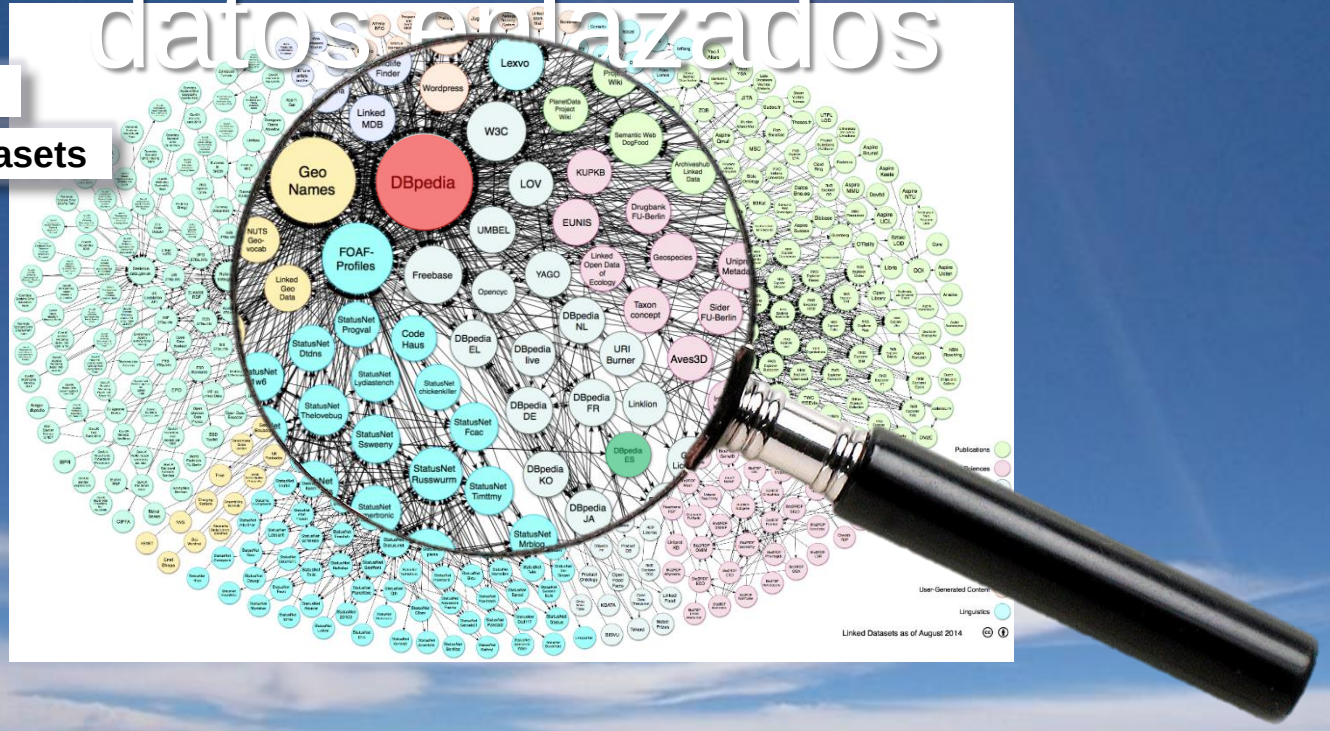
http://lod-cloud.net/versions/2011-09-19/lod-cloud.png

#DigHumesDBpedia

DBpedia, núcleo de los datos enlazados

2014

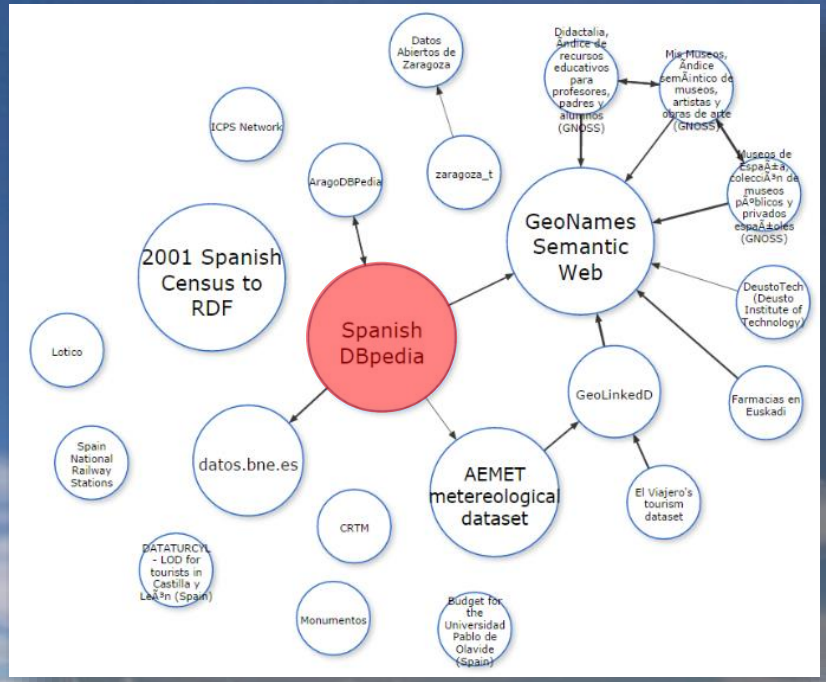
570 datasets



#DigHumesDBpedia

esDBpedia

- Núcleo de los datos enlazados del español



esDBpedia

- Crea tu dataset y enlázalo con esDBpedia
- Nosotros ponemos los enlaces inversos ☺

De datos a frases

LEXICALIZACIONES Y MULTILINGÜISMO

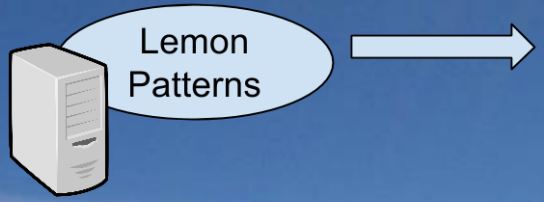
#DigHumesDBpedia

lemon y lemon patterns

Lemon file (lemon.dbpedia/es/animals_plants.ldp)

```
//// Classes

ClassNoun("species",dbpedia:Species) with plural "species",
ClassNoun("bacteria",dbpedia:Bacteria),
ClassNoun("archaea",dbpedia:Archaea),
ClassNoun("animal",dbpedia:Animal),
```



RDF output file

```
...
<lemon:entry>
  <lemon:LexicalEntry rdf:about="dbpedia_es_1#especie__noun">
<lemon:canonicalForm>
  <lemon:Form rdf:about="dbpedia_es_1#especie__noun/canonicalForm">
    <lemon:writtenRep xml:lang="es">especie</lemon:writtenRep>
  </lemon:Form>
</lemon:canonicalForm>
<lexinfo:gender rdf:resource="http://www.lexinfo.net/ontology/2.
0/lexinfo#feminine"></lexinfo:gender>
<lexinfo:partOfSpeech rdf:resource="http://www.lexinfo.net/ontology/2.
0/lexinfo#commonNoun"></lexinfo:partOfSpeech>
<lemon:otherForm>
<lemon:Form rdf:about="dbpedia_es_1#especie__noun/form">
  <lemon:writtenRep xml:lang="es">especies</lemon:writtenRep>
  <lexinfo:number rdf:resource="http://www.lexinfo.net/ontology/2.
0/lexinfo#plural" xmlns:lexinfo="http://www.lexinfo.net/ontology/2.
0/lexinfo#"></lexinfo:number>
</lemon:Form>
    </lemon:otherForm>
<lemon:sense>
<lemon:LexicalSense rdf:about="dbpedia_es_1#especie__noun/sense">
   <lemon:reference>
     <owl:Class rdf:about="http://dbpedia.org/ontology/Species"></owl:Class>
   </lemon:reference>

   <lemon:isA>
<lemon:Argument rdf:about="dbpedia_es_1#especie__noun/subject"></lemon:
Argument>
   </lemon:isA>
 </lemon:LexicalSense>
    </lemon:sense><lemon:synBehavior>
<lemon:Frame rdf:about="dbpedia_es_1#especie__noun/frame">
  <rdf:type rdf:resource="http://www.lexinfo.net/ontology/2.
0/lexinfo#NounPredicateFrame"></rdf:type>
  <lexinfo:subject rdf:resource="dbpedia_es_1#especie__noun/subject"
></lexinfo:subject>
</lemon:Frame>
```

Lemonizando DBpedia

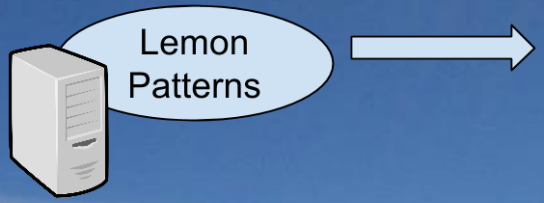
#DigHumesDBpedia

lemon y lemon patterns

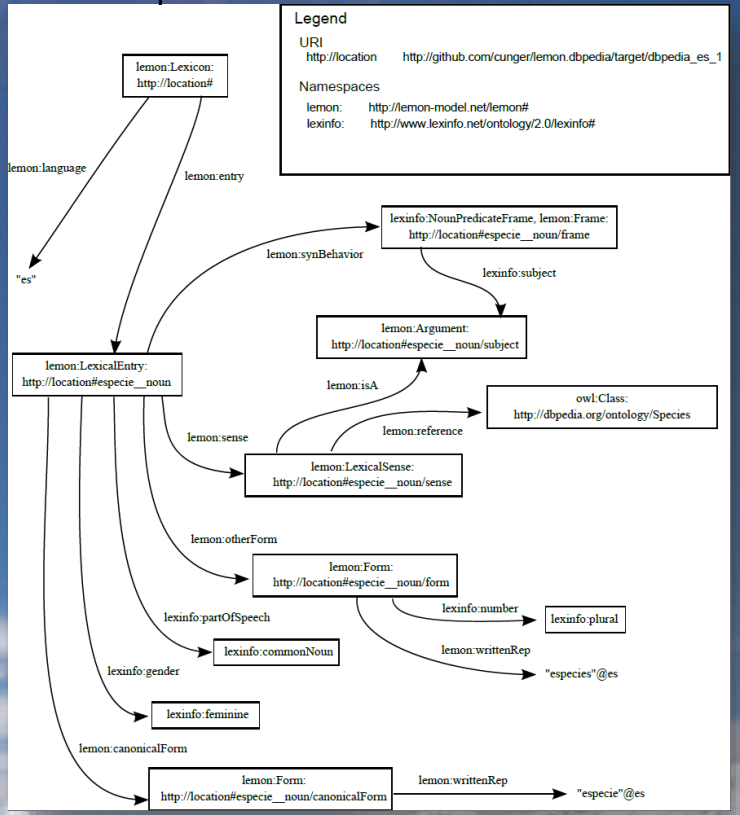
Lemon file (lemon.dbpedia/es/animals_plants.ldp)

```
//// Classes

ClassNoun("species",dbpedia:Species) with plural "species",
ClassNoun("bacteria",dbpedia:Bacteria),
ClassNoun("archaea",dbpedia:Archaea),
ClassNoun("animal",dbpedia:Animal),
```



RDF output file



Lemonizando DBpedia

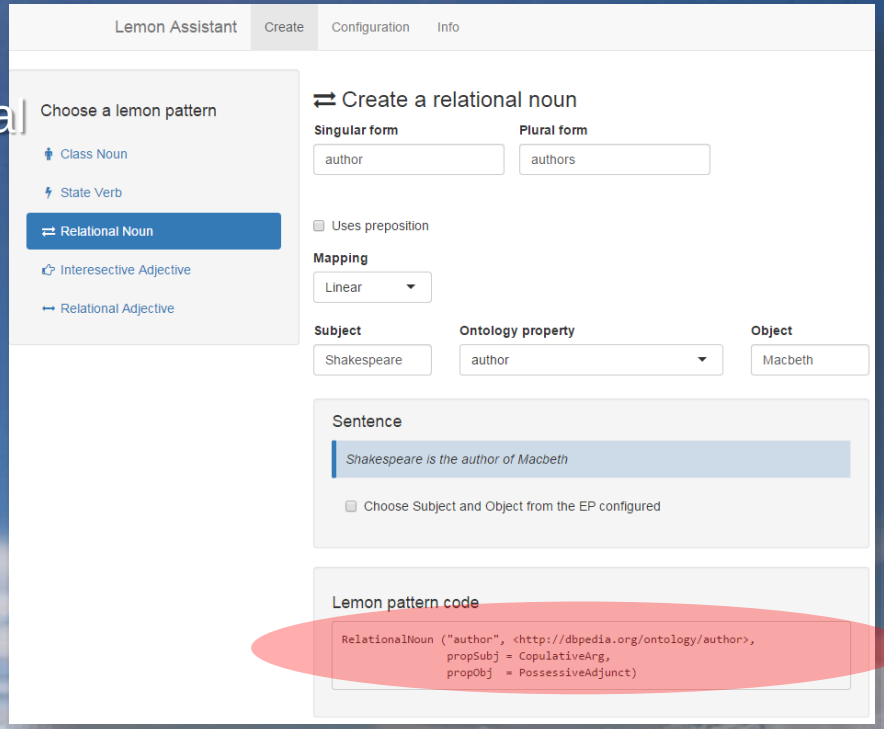
Lemonade: endulzando *lemon*

- Lemonade es
  - Una librería R
    - R es un entorno de desarrollo sencillo de instalar
      - Multiplataforma (iOS, Linux, Windows)
      - Orientado a análisis de datos (métodos estadísticos)
      - Orientado a datos masivos
      - Rápido
        - » Ejecución en paralelo de código C/C++/Java/PHP
        - » Distribuible (entre máquinas)
  - Un servicio (integración sencilla en apps Java/Javascript)
    - Diseñado para BIG data

Lemon Assistant

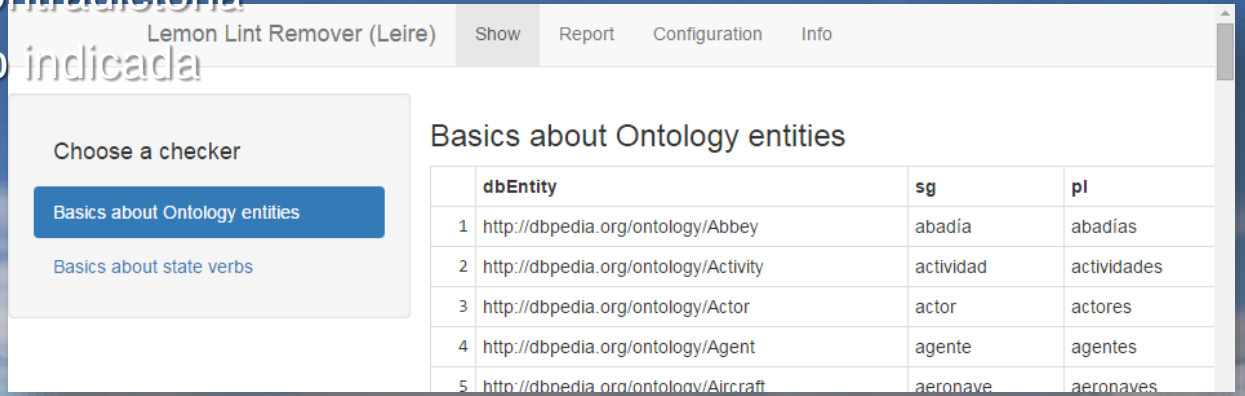
http://lider2.dia.fi.upm.es:3838/lemonAssistant/

- Creación intuitiva de *lemon patterns*
  - Guiado por frases en lenguaje natural
- Linked data → Lenguaje natural
- Algunos lemon patterns (más, pronto)
- Multilingual (inglés, español, alemán)
  - Gramáticas de GF (Grammatical Framework)
    - Gramáticas de muchos idiomas



Leire

- Apunta a un triple store (Fuseki)
- Bang!
  - Listas con información sobre nombres, verbos, preposiciones…
  - Informe de inconsistencias
    - Información contradictoria
    - Información no indicada



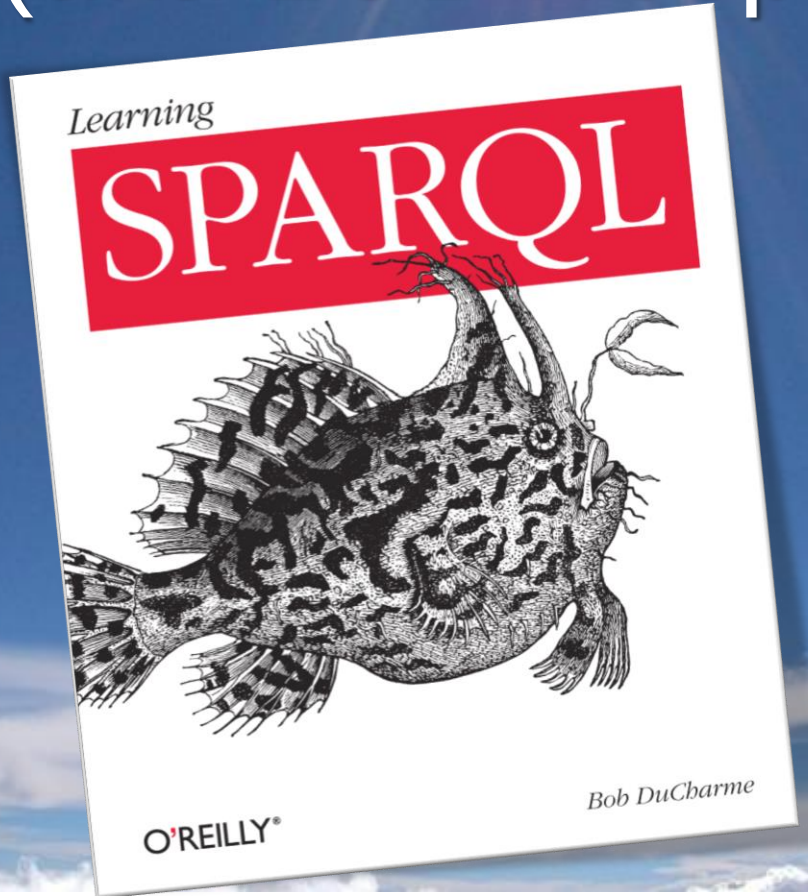
Usando las lexicalizaciones

PROYECTOS FUTUROS

#DigHumesDBpedia

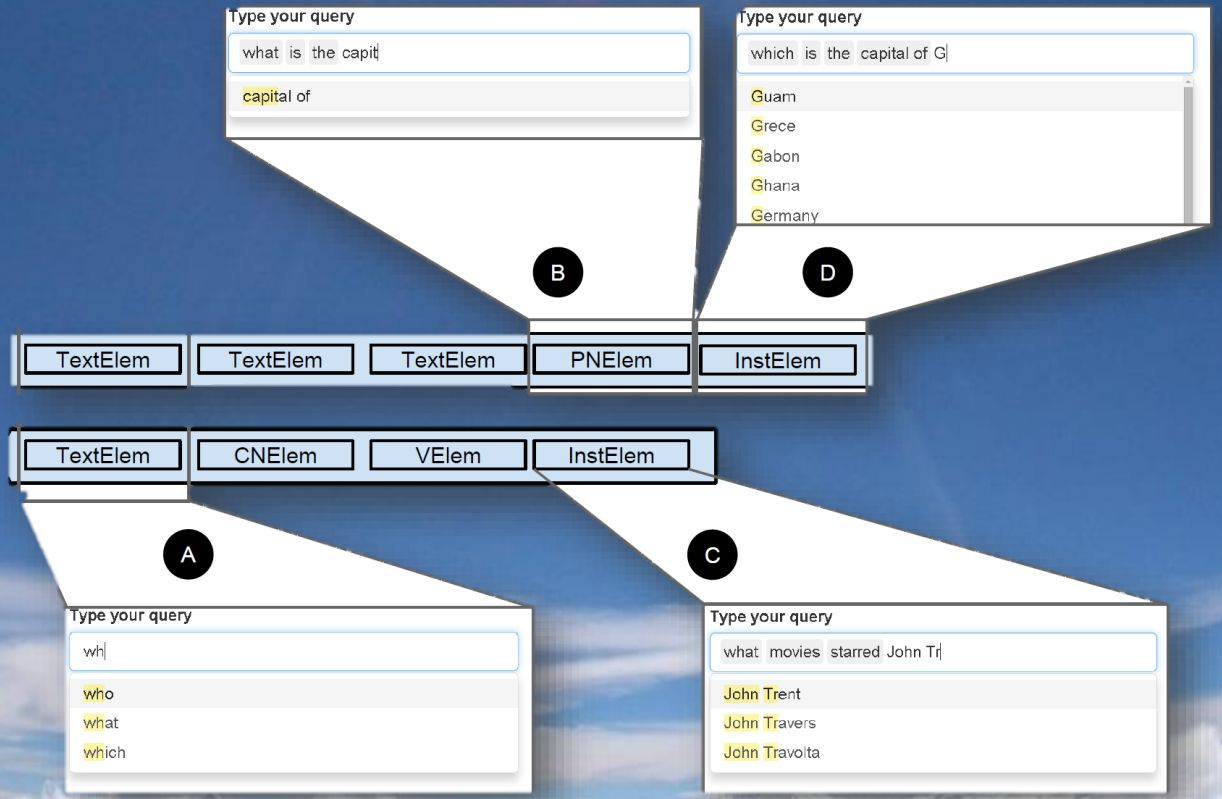
interQA (aka LODTrooper)

- Bye, bye, SPARQL



interQA (aka LODTrooper)

- Uso de las lexicalizaciones
- Consultas en NL guiado
  - Para cada tipo de consulta SPARQL
    - Varias formas de preguntar en LN
- Multilingüe
- Cualquier dataset



Agradecimientos

- Mapeadores
- Organizaciones
  - MINECO
    - Contrato Juan de la Cierva
    - Proyecto INFRA
  - MECD
    - Programa José Castillejo
  - Unión europea
    - Proyecto LIDER
  - OEG-UPM

#DigHumesDBpedia

Thanks a LOD for your attention

Mariano.Rico@upm.es

Ontology Engineering Group

mylider