

esDBpedia:

núcleo de los datos semánticos del **es**pañol



#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

El español en Internet



Mariano Rico

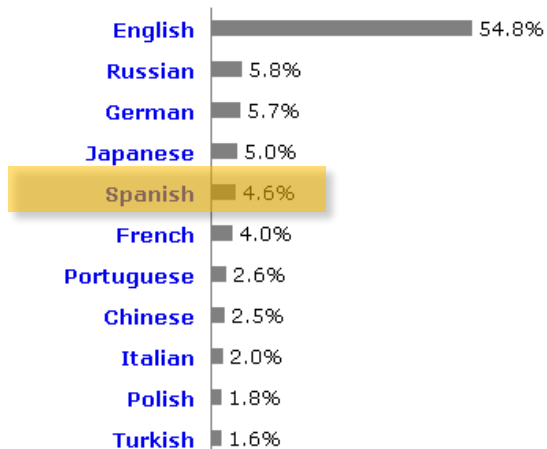
@marianorico



Following

English is used by 54.8% of all the websites. Spanish is fifth with 4.6% after Rus, Ger, Jap.

w3techs.com/technologies/o...



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)

WIKIPEDIA

English

The Free Encyclopedia

4 985 000+ articles

Deutsch

Die freie Enzyklopädie

1 864 000+ Artikel

Русский

Свободная энциклопедия

1 259 000+ статей

中文

自由的百科全书

845 000+ 條目

Português

A enciclopédia livre

891 000+ artigos

Español

La enciclopedia libre

1 206 000+ artículos

日本語

フリー百科事典

986 000+ 記事

Français

L'encyclopédie libre

1 670 000+ articles

Italiano

L'enciclopedia libera

1 228 000+ voci

Polski

Wolna encyklopedia

1 137 000+ hasel

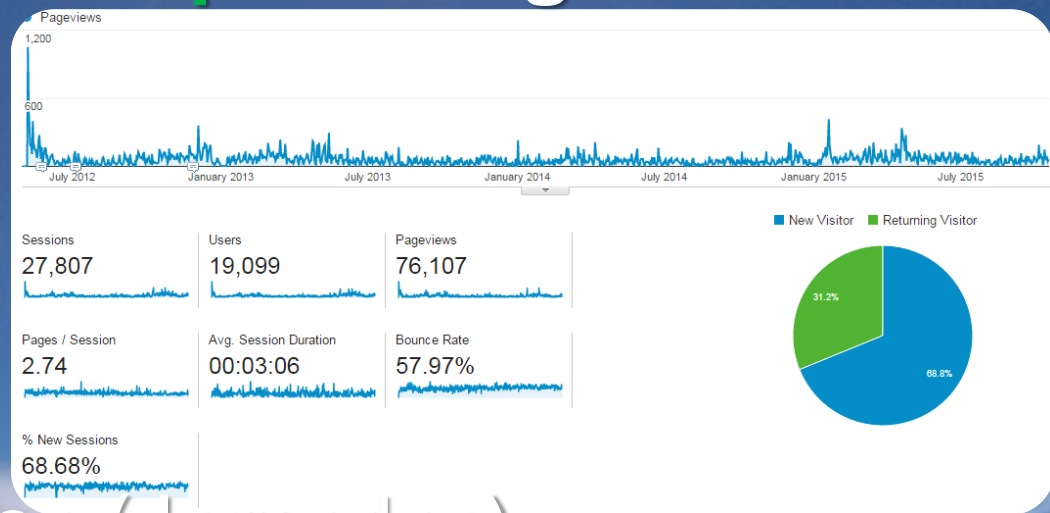


¿Qué es esDBpedia?

- 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**

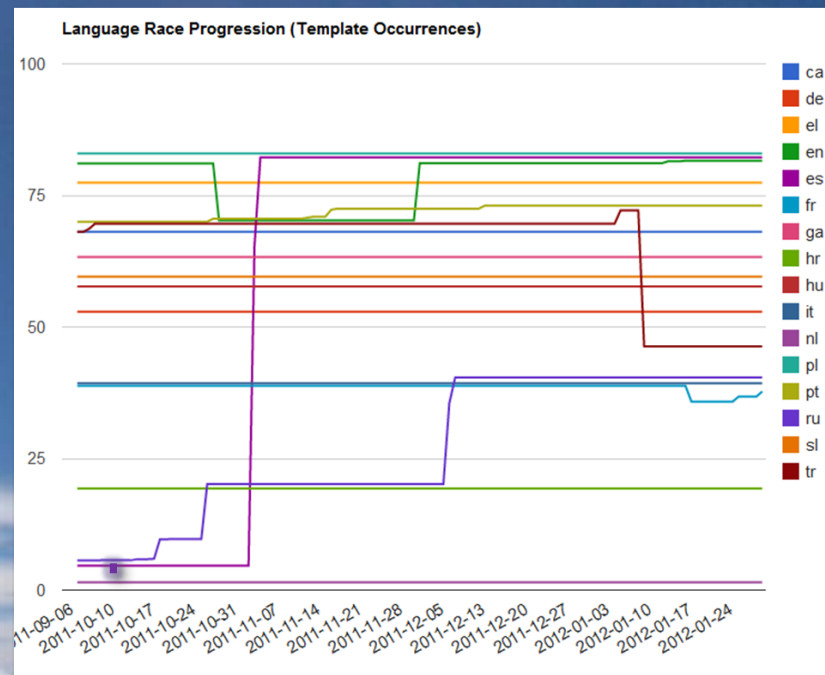
Visibilidad esDBpedia

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



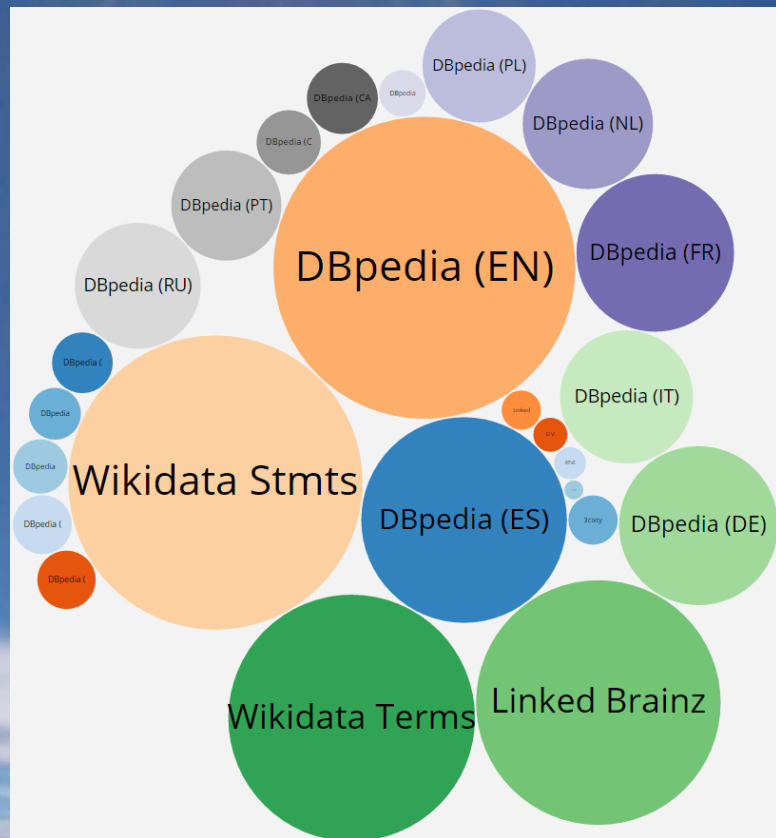
Mapeos, mapeos y mapeos

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



Datos xxDBpedias

- Primicia:
esDBpedia es la primera después de la inglesa



esDBpedia para lingüistas

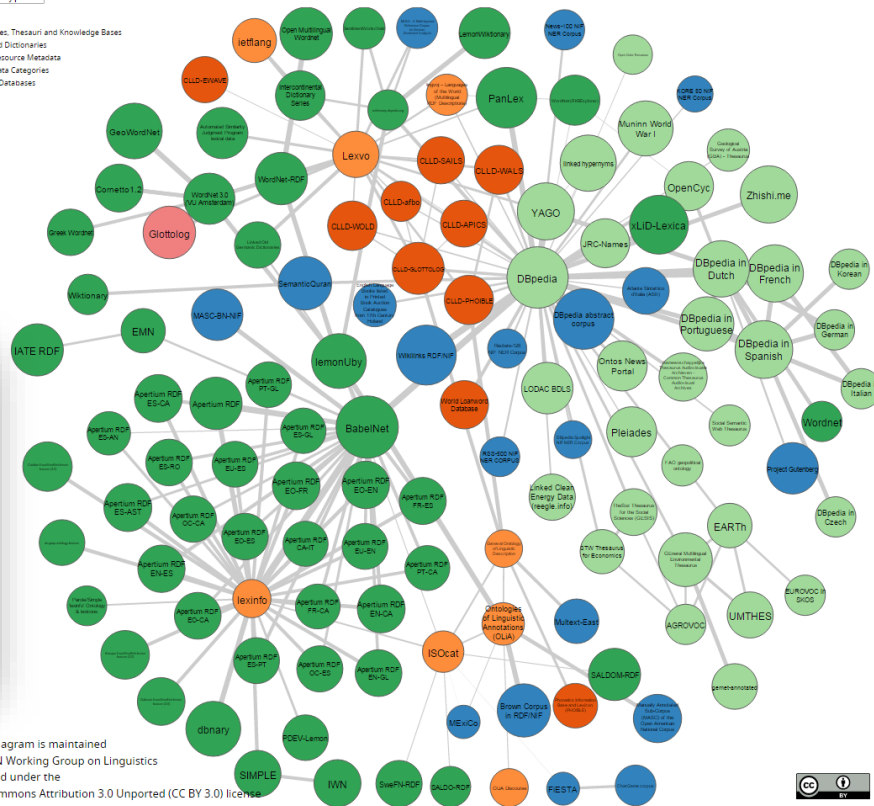
- Linguistic LOD

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

Legend

By resource type Download SVG

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



The LOD diagram is maintained by the OKFN Working Group on Linguistics and provided under the Creative Commons Attribution 3.0 Unported (CC BY 3.0) license





Datos con sentido

ESDBPEDIA COMO FUENTE DE DATOS SEMÁNTICOS

Datos con semántica



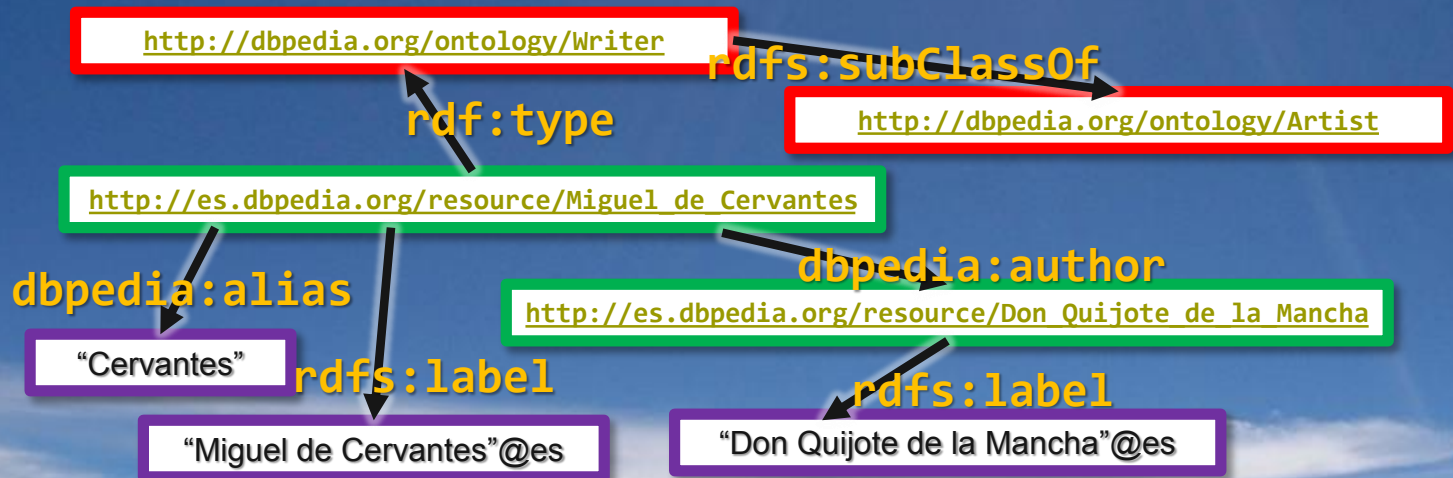
MAKING
SENSE of
DATA

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

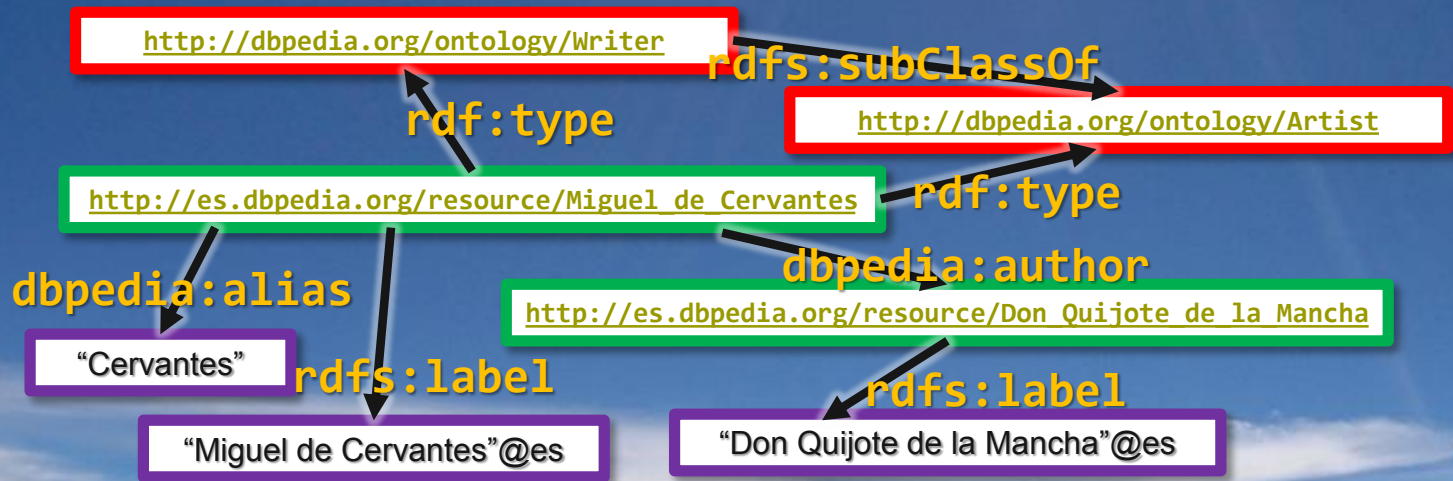
Datos con semántica

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



Datos con semántica

- Razonando: “Cervantes es un artista”



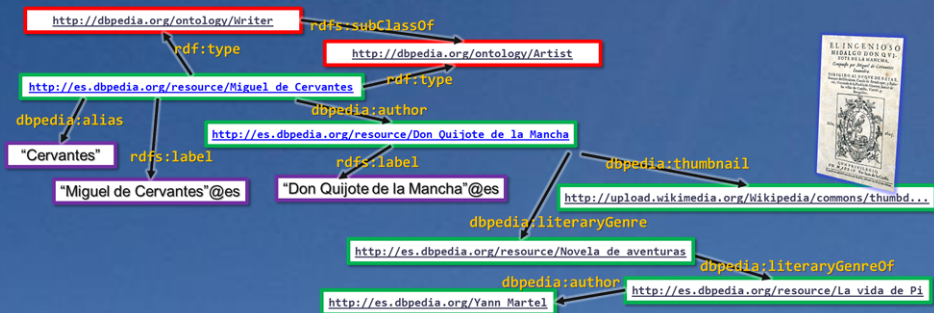
Datos con semántica

- Añade más datos enlazados



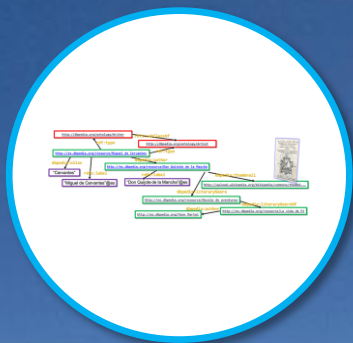
Datos con semántica

- Añade más datos enlazados



Datos con semántica

- Dataset



esDBpedia

- Dataset con datos de la Wikipedia del español

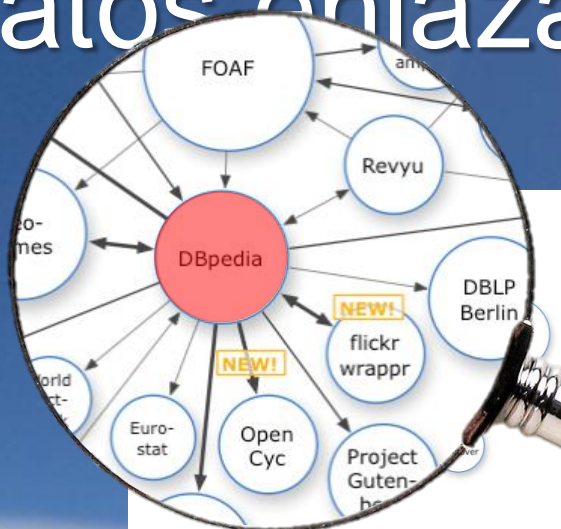


- 160 millones de datos de 1 millón de entradas de Wikipedia del español.
- Enlazados a ontología DBpedia
 - Cientos de conceptos
 - Miles de relaciones entre conceptos

DBpedia, núcleo de los datos enlazados

2007

25 datasets

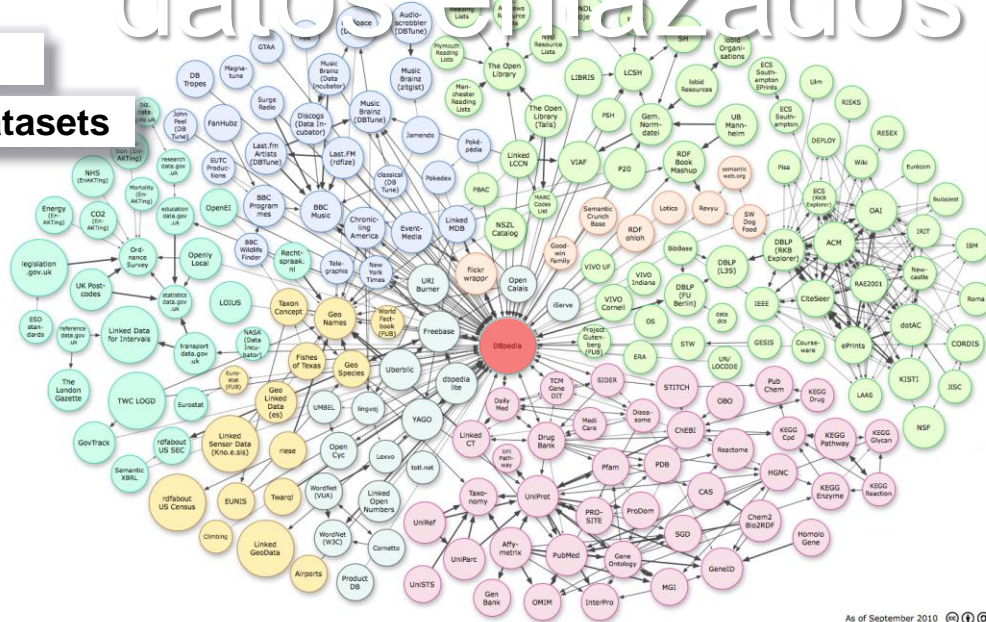


DBpedia, núcleo de los

datos enlazados

2010

203 datasets



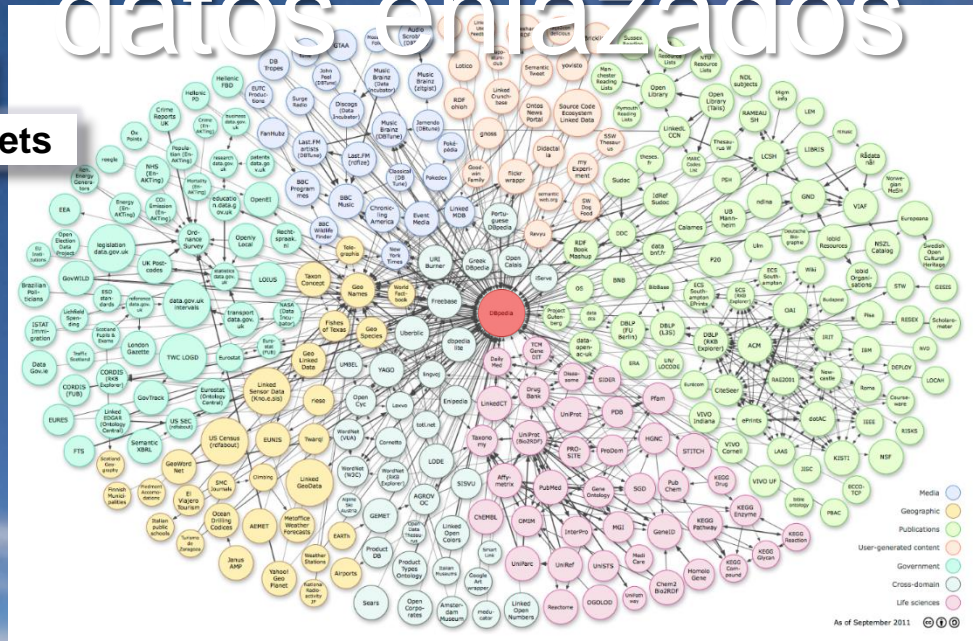
As of September 2010

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

DBpedia, núcleo de los datos enlazados

2011

295 datasets



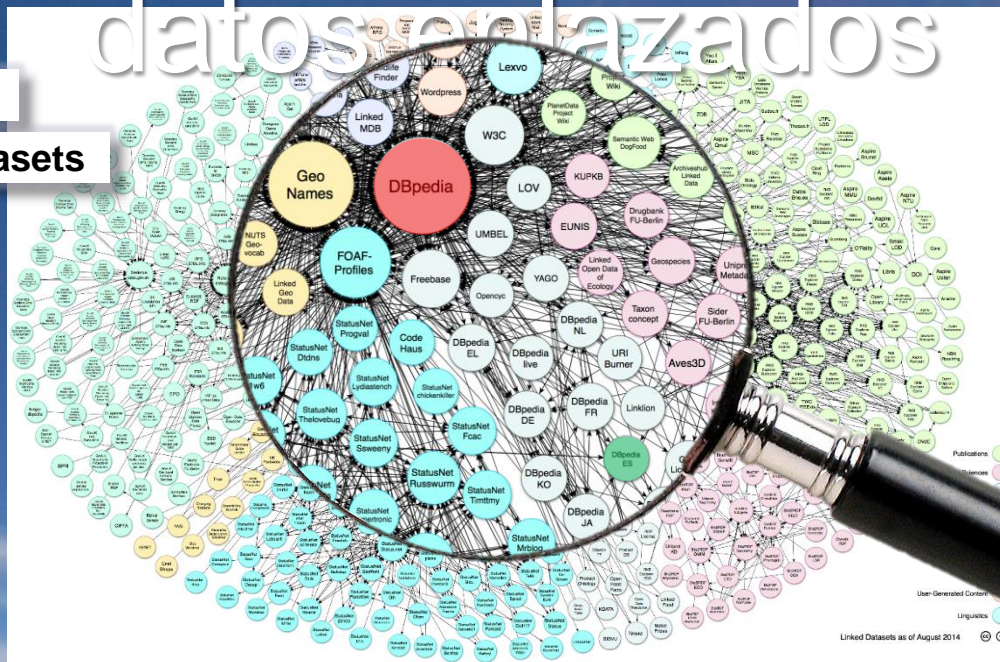
<https://www.kitware.com/2011/09-10/kit-cloud-013>

DBpedia, núcleo de los

datos enlazados

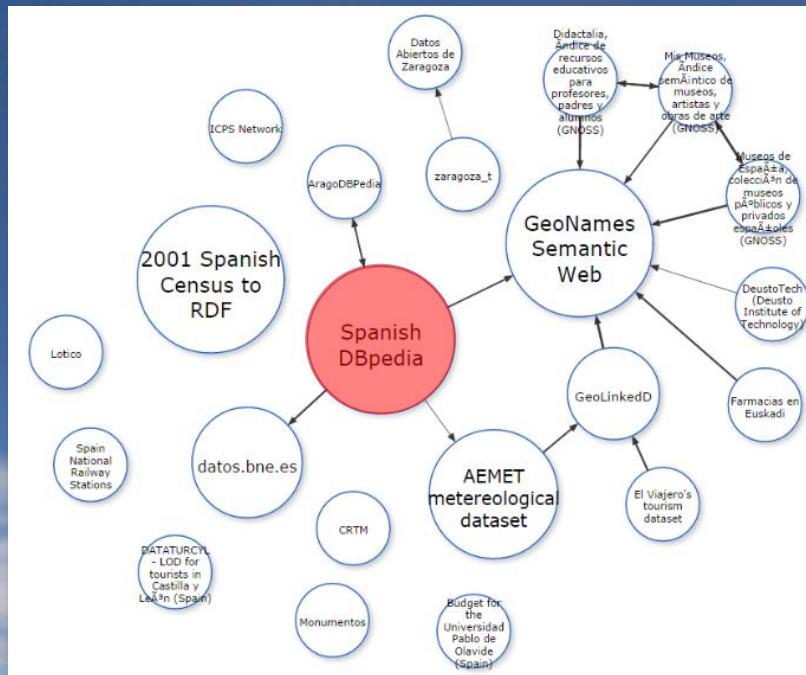
2014

570 datasets



esDBpedia

- Núcleo de los datos enlazados del español



esDBpedia

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



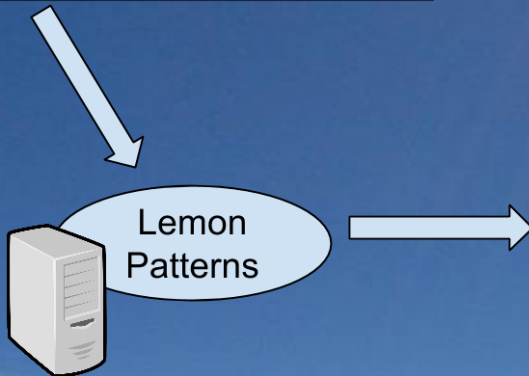
De datos a frases

LEXICALIZACIONES Y MULTILINGÜISMO

lemon y lemon patterns

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

```
6  /// Classes
7
8  ClassNoun("species",dbpedia:Species) with plural "species",
9  ClassNoun("bacteria",dbpedia:Bacteria),
10 ClassNoun("archaea",dbpedia:Archaea),
11 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>

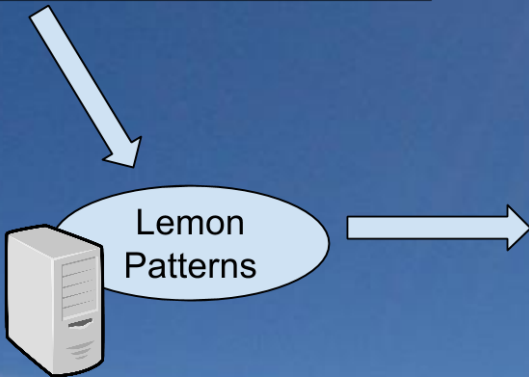
        <lemon:isA>
          <lemon:Argument rdf:about="dbpedia_es_1#especie__noun/subject"></lemon:Argument>
        </lemon:isA>
      </lemon:LexicalSense>
    </lemon:sense><lemon:synBehavior>
    <lemon:Frame rdf:about="dbpedia_es_1#especie__noun/frame">
      <rdf:type rdf:resource="http://www.lexinfo.net/ontology/2.0/lexinfo#NounPredicateFrame"></rdf:type>
      <lexinfo:subject rdf:resource="dbpedia_es_1#especie__noun/subject">
    </lexinfo:subject>
    </lemon:Frame>
```

Lemonizando DBpedia

lemon y lemon patterns

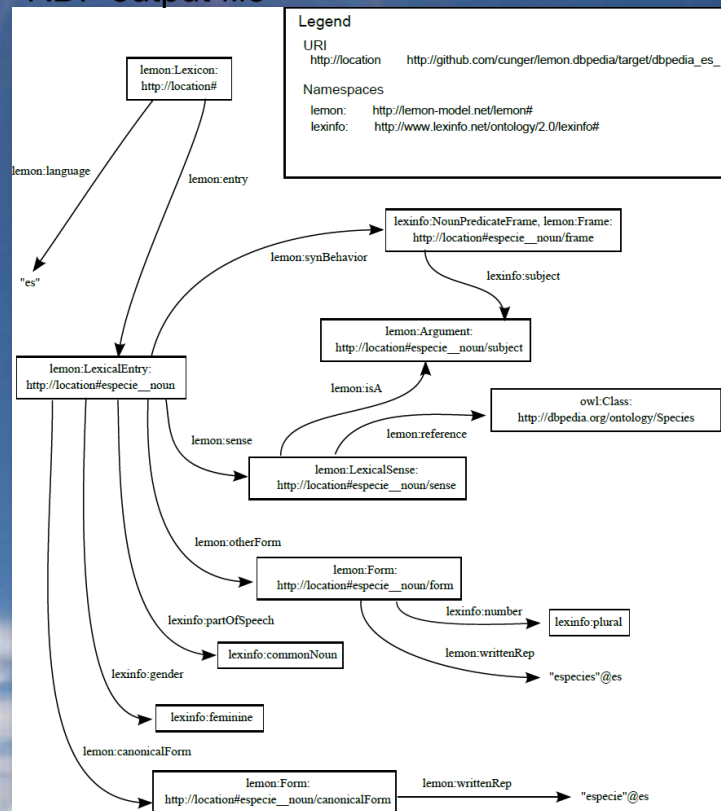
Lemon file (lemon.dbpedia/es/animals_plants.ldp)

```
6  /// Classes
7
8  ClassNoun("species",dbpedia:Species) with plural "species",
9  ClassNoun("bacteria",dbpedia:Bacteria),
10 ClassNoun("archaea",dbpedia:Archaea),
11 ClassNoun("animal",dbpedia:Animal),
```



Lemonizando DBpedia

RDF output file



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

The screenshot shows the 'Lemon Assistant' web interface. The main navigation bar includes 'Lemon Assistant', 'Create', 'Configuration', and 'Info'. A sidebar on the left titled 'Choose a lemon pattern' lists options: 'Class Noun', 'State Verb', 'Relational Noun' (selected), 'Intersective Adjective', and 'Relational Adjective'. The main content area is titled 'Create a relational noun' and contains several input fields and options:

- Singular form:** 'author' (input field)
- Plural form:** 'authors' (input field)
- Uses preposition
- Mapping:** 'Linear' (dropdown menu)
- Subject:** 'Shakespeare' (input field)
- Ontology property:** 'author' (dropdown menu)
- Object:** 'Macbeth' (input field)

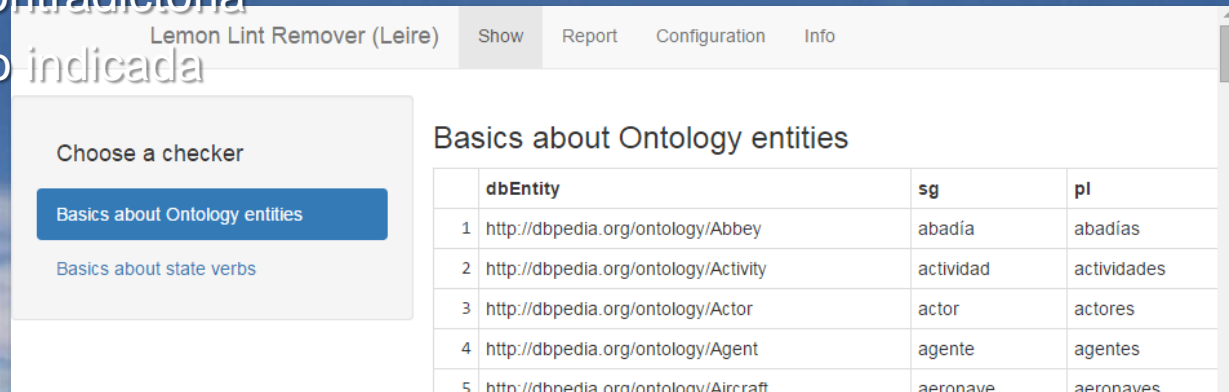
Below these fields, a 'Sentence' section displays 'Shakespeare is the author of Macbeth' in a blue box, with an option Choose Subject and Object from the EP configured.

At the bottom, a 'Lemon pattern code' section shows the following code:

```
RelationalNoun ("author", <http://dbpedia.org/ontology/author>,
propSubj = CopulativeArg,
propObj = PossessiveAdjunct)
```

Leire

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



Lemon Lint Remover (Leire) Show Report Configuration Info

Choose a checker

- Basics about Ontology entities
- Basics about state verbs

Basics about Ontology entities

	dbEntity	sg	pl
1	http://dbpedia.org/ontology/Abbey	abadía	abadías
2	http://dbpedia.org/ontology/Activity	actividad	actividades
3	http://dbpedia.org/ontology/Actor	actor	actores
4	http://dbpedia.org/ontology/Agent	agente	agentes
5	http://dbpedia.org/ontology/Aircraft	aeronave	aeronaves

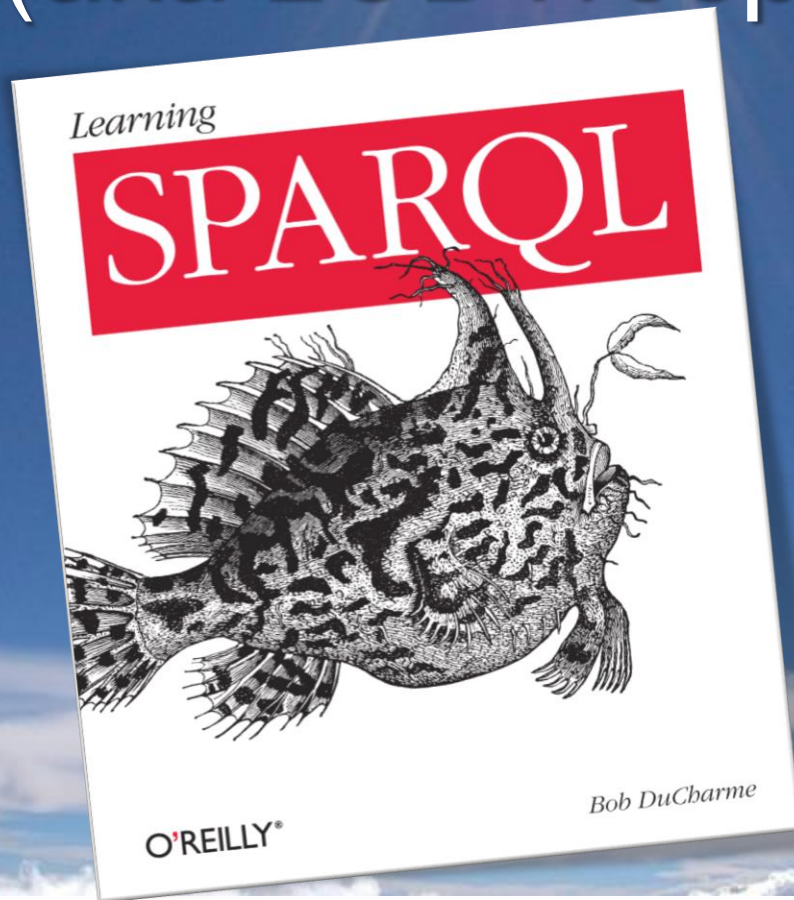


Usando las lexicalizaciones

PROYECTOS FUTUROS

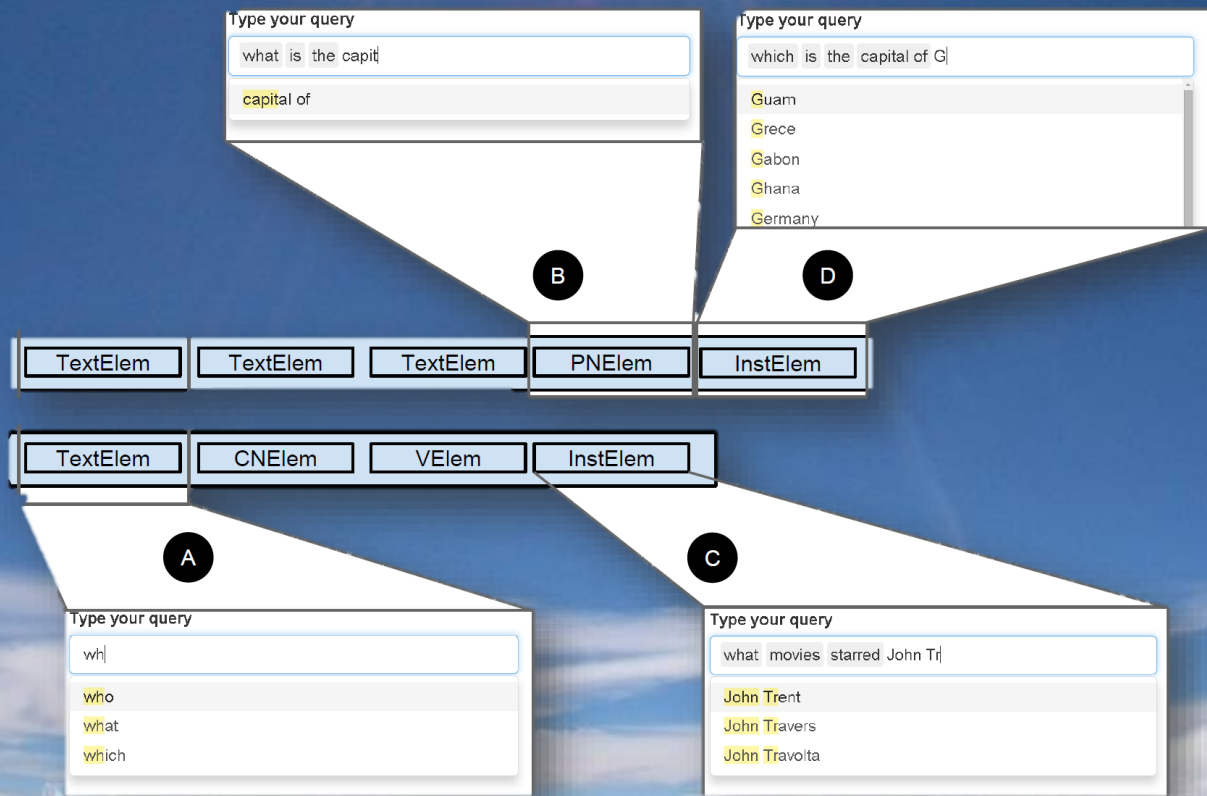
interQA (aka LODTrooper)

- 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



Thanks a LOD for your attention

Mariano.Rico@upm.es

