Gestión de Big Data: Datos y Usos - PEC4

Jose Maria Mengibar Fornieles

Master en Business Intelligence

1. Datos Abiertos

Para describir la tipología de los datos disponibles en Datos.gob.es podemos diferenciar en:
Tipo de estructura y formato: estructurados en diferentes formatos JSON, XML, ODS, CSV, XLS
Aunque los recursos disponibles presentan diferentes extensiones de archivos, todos son formatos de contenido estructurado, sean distribuidos por columnas (CSV, SCEL, ODS) u objetos con atributos clave valor anidados (JSON, XML).

La web nos permite diferentes modos de acceso a los recursos. A través del catálogo, a través de la API o a través de Sparql:

Catálogo: El contenido está dividido en diferentes categorías. Pudiendo filtrar el contenido a través del menú por diferentes características como: autor, formato, categoría, ect ..

API:
Enlace http://datos.gob.es/es/apidata#/
La API nos permite acceder al contenido a través de servicios-web. Se proporciona una tabla de todos los request get disponibles (ya que únicamente se nos permite „leer“ datos) y las posibles respuestas procedentes del servidor.
Además se ofrece una interfaz en la que podemos probar los diferentes servicios, variando sus parámetros y observando la respuesta.

SPARQL:
La web ofrece la posibilidad de acceder a los datos mediante un formulario en el que introducir las consultas Sparql.

La web ofrece un catalogo de aplicaciones que utilizan los recursos de datos de la propia web, permitiendo a cualquier usuario añadir su propia web. Las aplicaciones están agrupadas por diferentes categorías y deben aportar un mínimo de campos e información para su publicación.

2. Datos Abiertos

Identificación :

La herramienta permite acceder una dirección única con la que diferenciar y acceder al contenido.

En caso contrario, es decir, si la dirección proporcionada no es la exacta al recurso al que se quiere acceder, por ejemplo: http://dblp.uni-trier.de/pers/hd/b/Berne
el sistema es lo bastante rápido e inteligente como para buscar los recursos relacionados y generar un listado de sugerencias que podrían coincidir con el objeto de la búsqueda.

Descripción
Al obtener el sitio resultado de nuestra búsqueda se presentan una serie de artículos en donde también aparece el autor buscado.

En las entradas relacionadas con el objeto buscado, se nos muestra un menú en el que nos permite seleccionar el formato de descarga de la entrada o record.

Entre los formatos disponibles encontramos :

Bibtext (http://www.bibtex.org/) :

```
@inproceedings{DBLP:conf/www/RegaliaJM17,
  author    = {Blake Regalia and
               Krzysztof Janowicz and
               Grant McKenzie},
  title     = {Revisiting the Representation of and Need for Raw Geometries on the
               Linked Data Web},
  booktitle = {Workshop on Linked Data on the Web co-located with 26th International
               World Wide Web Conference {(WWW} 2017)},
  year      = {2017},
  crossref  = {DBLP:conf/www/2017ldow},
  url       = {http://ceur-ws.org/Vol-1809/article-04.pdf},
  timestamp = {Wed, 03 May 2017 15:13:12 +0200},
  biburl    = {http://dblp.org/rec/bib/conf/www/RegaliaJM17},
  bibsource = {dblp computer science bibliography, http://dblp.org}
```

Ris:

```
Provider: Schloss Dagstuhl - Leibniz Center for Informatics
Provider: University of Trier
Database: dblp computer science bibliography
Content:text/plain; charset="utf-8"

TY  - CPAPER
ID  - DBLP:conf/www/Berners-Lee05
AU  - Berners-Lee, Tim
TI  - WWW at 15 years: looking forward.
BT  - Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005
SP  - 1
PY  - 2005//
DO  - 10.1145/1060745.1060746
UR  - http://doi.acm.org/10.1145/1060745.1060746
ER  -
```

También se permite exportar el enlace al recurso en RDF, XML, en los que del mismo modo se proporciona información de otros contenido relacionados con el objeto de nuestra búsqueda.
Si observamos el contenido del elemento descargado de RDF N-Triples vemos que consiste en un cuerpo de relaciones mediante sujeto predica y objeto:
Por ejemplo:

<http://dblp.uni-trier.de/rec/conf/sigopsE/Berners-Lee88>
<http://www.w3.org/2002/07/owl#sameAs> <http://dblp.org/rec/conf/sigopsE/Berners-Lee88> .

Pudiendo interpretar el recurso 1 es igual que el recursos 2

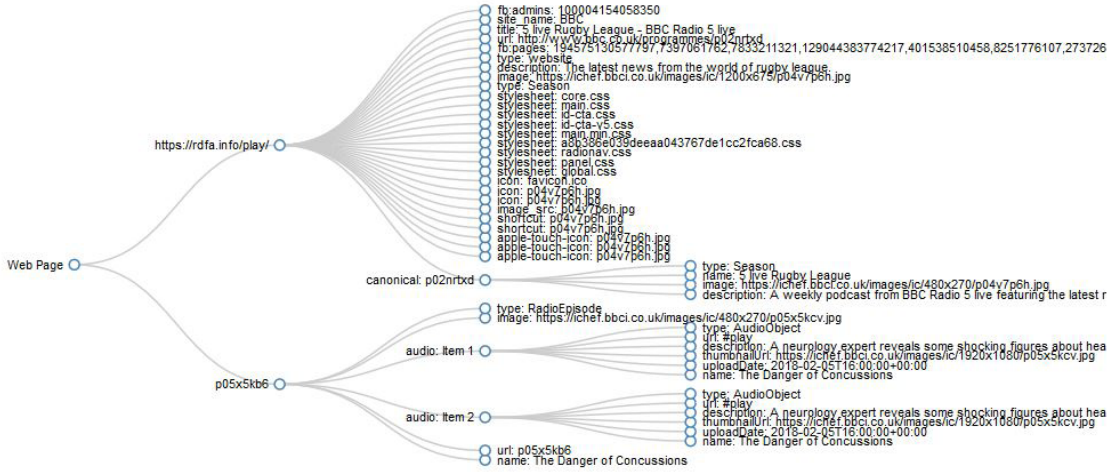
Enlace:
La página del recurso esta repleta de enlaces sobre otros autores, artículos y eventos relacionados con nuestra búsqueda.

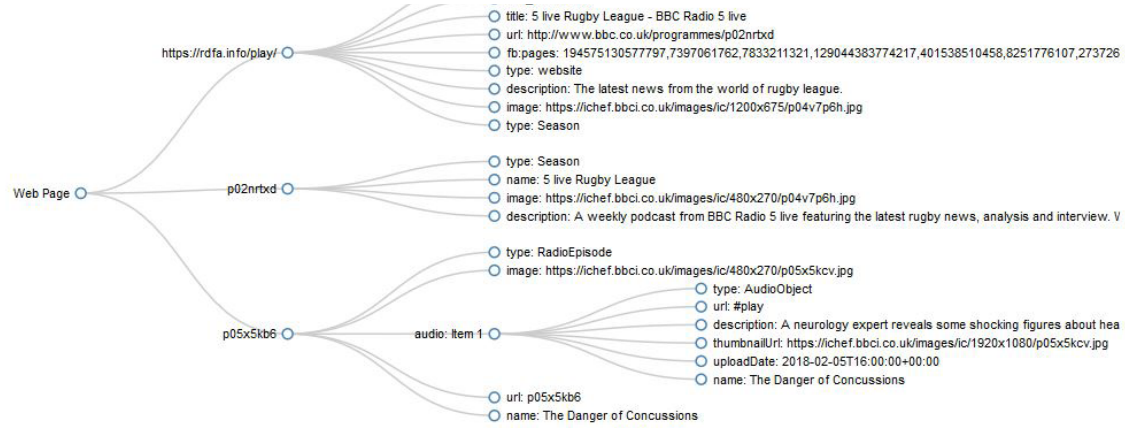
En cada bloque de enlaces se resalta el objeto buscado inicialmente, en nuestro caso Tim Berners-Lee: dentro de un conjunto de de elementos relacionados con el objeto, en nuestro caso otros autores que han colaborado en publicaciones. Cada entrada del listado posee un menú de opciones que permite exportar o acceder a otros formatos del recurso. Entre las opciones se encuentran DOI (Identificador de objeto digital) de función similar a la URI, texto crudo sin formato, enlace a la versión en pdf.

En este sentido se proporciona un sistema enlazado de contenido. Pero con la aparición de las redes sociales es posible además que sean los usuarios los que creen nuevos enlaces entre contenidos que quizás aparentemente no tienen relación.
En el menú de cada entrada se proporciona un listado de plataformas en las que compartir la entrada del articulo. Ya que se tratan de texto científicos, posiblemente una de las opciones mas interesantes de Mendeley (https://www.mendeley.com/), que es un gestor de archivos, principalmente pdf.
En Mendeley los usuarios poseen colecciones que pueden compartir con otros usuarios si lo desean. De este modo aun se expande más la posibilidad de compartir elementos en los cuales además es posible ordenar, categorizar y etiquetar el contenido, expandiendo los principios de identificación, descripción y enlace.

3. RDF

Usando la herramienta https://rdfa.info/play/ se ha obtenido el grafo de los elementos encontrados y la versión en texto de las extensiones RDFa del sitio http://www.bbc.co.uk/programmes/p02nrtxd





Elementos encontrados en la visualización en grafo.

Con la herramienta „Quick and Dirty RDF browser“ obtenemos en el enlace del ejercicio 3 bloques con 38 tripletas.

El primer bloque que encontramos esta relacionado con los meta-datos propios del sitio web y otros relacionados con la red social facebook.
Donde el propio sitio es el sujeto, los atributos en los metadatos son el predicado y el valor mostrado el objeto.
Mediante los meta-datos pertenecientes al protocolo Open Graph proporcionamos información de nuestro sitio en caso por ejemplo de que sea compartido en las redes sociales.

Extracto del RDF browser.

En la web de http://ogp.me/ encontramos toda la información necesaria acerca las propiedades y meta-datos open graph y su implementación en el código web, en este caso en la parte *head* de la estructura html.
http://ogp.me/#metadata

El prefijo definido mediante: *@prefix og: <http://ogp.me/ns#>* es utilizado en los para definir los metadatos: nombre del sitio, título, url, tipo, descripción e imagen.
Ejemplo y aplicación en el Html:
og:title <meta property=“og:title“ content=“The Rock“ />

„<fb:pages>“ proporciona un listado de sitios web relacionados con la BBC.
El elemento de la lista *7833211321* sería interpretado construyendo la dirección https://www.facebook.com/7833211321 en https://www.facebook.com/bbc1xtra/

El sujeto http://www.bbc.co.uk/programmes/p02nrtxd posee variatributos en los tags anidados:

```
<div class="map__intro island" vocab="http://schema.org/" typeof="Season" resource="http://www.bbc.co.uk/programmes/p02nrtxd">
  <h1 class=" visually-hidden" property="name">5 live Rugby League</h1>
    ....
```

Mediante una serie de atributos en el tag encontramos:

- vocab="http://schema.org/" con el que especificamos el vocabulario,
- el tipo typeof="Season"
- el predicado „name" mediante „property" y el objeto valor „5 live Rugby League"

Encontramos la información relacionada en http://schema.org/docs/datamodel.html en el que se especifica donde añadir la fuente de schema
y en http://schema.org/Season se especifica que el tipo Season es una clase especifica para contenido multimedia.

Otro bloque identificado corresponde al nodo en el grafo *p05x5kb6.*
Si observamos en primer lugar el código fuente en el Navegador, vemos como se encuentran los diferentes elementos anidados. Ciertos elementos poseen el atributo „property".
Como se indica en https://www.w3.org/TR/html-rdfa/#property-copying,
se incorporan mecanismos en la estructura que permiten definir atributos para los elementos anidados.

Visualización del codigo Html en panel de Firefox.

Así el bloque principal contiene un elemento del tipo AudioObject. En el objeto siguiente se referencia la url del recurso. (http://schema.org/AudioObject)

```
http://www.bbc.co.uk/programmes/p05x5kb6                                    a Radio Episode
    ⟶ rdf:type ⟶ http://www.bbc.co.uk/programmes/RadioEpisode
    ⟶ http://www.bbc.co.uk/programmes/image ⟶ "https://ichef.bbci.co.uk/images/ic/480x270/p05x5kcv.jpg"
    ⟶ http://www.bbc.co.uk/programmes/url ⟶ "<span xmlns="http://www.w3.org/1999/xhtml" class="programme__title
        gamma"> <span xmlns="http://www.w3.org/1999/xhtml" property="name">The Danger of Concussions</span>
        </span>"^^rdf:XMLLiteral
```

Extracto del RDF browser.

4. SPARQL

4.1 Información relacionada con el recurso de la dbpedia
etiquetado en inglés con el nombre de “Manchester United F.C.”.

```
SELECT DISTINCT ?resource
WHERE {
  ?resource rdfs:label ?name
  FILTER ( regex (?name , „Manchester United F.C.“, „i“) && lang(?name) = „en“)
}
```

4.2 Nombres de los clubs en los que ha entrenado anteriormente el
actual manager del equipo de fútbol representado por el recurso de la dbpedia
etiquetado en inglés con el nombre de “Manchester United F.C.”

```
SELECT (SAMPLE(?clubs) AS ?clubResource ) (SAMPLE(?name) AS ?clubName)
WHERE {
 <http://dbpedia.org/resource/Manchester_United_F.C.> dbo:manager ?manager.
 ?manager dbo:managerClub ?clubs.
 ?clubs foaf:name ?name
 FILTER (lang(?name) = „en“ )
}
GROUP BY ?clubs
```

http://dbpedia.org/page/Manchester_United_F.C.
http://dbpedia.org/page/Jos%C3%A9_Mourinho

4.3 Nombre de las jugadoras de tenis cuyo país es España junto con la clasificación actual que ocupan en el ranking de competiciones individuales.

```
SELECT  str(?name)as ?nombre ( xsd:integer( REPLACE( ?ranking, „\\D“, „“, „i“)) AS  ?posicion  )
WHERE {
 ?resource dct:description ?desc.
 ?resource foaf:gender ?gender.
 ?resource foaf:name ?name.
 ?resource dbp:currentsinglesranking ?ranking

 FILTER ( regex (?desc, „Spanish tennis player“, „i“) && (?gender= „female“@en ) )

}
ORDER BY ASC(?posicion)
```

http://dbpedia.org/page/Garbi%C3%B1e_Muguruza

4.4 Carreras en las que el piloto de nombre “Fernando”@en y apellido “Alonso”@en ha sido considerado el piloto más rápido, mostar el año de su consecución y el nombre de la carrera. Los registros obtenidos se ordenarán, en modo ascendente, por el año y posteriormente por el nombre de la carrera.

```
SELECT  xsd:integer(?raceYear) as ?fecha  str(?raceName) as ?carrera
WHERE {
 ?resource dct:description ?desc.
 ?resource foaf:givenName „Fernando“@en.
 ?resource foaf:surname „Alonso“@en.
 ?race dbo:fastestDriver  ?resource.
 ?race dbp:yearOfRace ?raceYear.
 ?race dbp:nameOfRace ?raceNameRes.
 ?raceNameRes dbp:name ?raceName

 FILTER ( regex (?desc, „racing driver“, „i“) )
}
ORDER BY ASC(?fecha) ?carrera
```

http://dbpedia.org/page/Fernando_Alonso
http://dbpedia.org/page/Singapore_Grand_Prix