Web semántica y datos entrelazados

Trabajo final

David Domínguez Barbero
david.dominguez@posgrado.uimp.es
3 de junio de 2017

Máster en Investigación en Inteligencia Artificial
Universidad Internacional Menéndez Pelayo

Índice

Índice de figuras

1. Introducción

El enlazamiento de datos en la web es un conjunto de técnicas muy útiles para la extracción de información. Este conjunto de técnicas se conocen más comúnmente como *linked data*. Según la página `linkeddata.org`, administrada por *Tom Heath* a favor de la comunidad de *linked data*, indica que el enlazamiento de datos se utiliza para enlazar datos que no lo estaban antes o para facilitar utilizando la web con los que sí que estaban enlazados y así disminuir las barreras entre distintas fuentes de información. En `wikipedia` se indica que el enlazamiento de datos describe una práctica recomendada para la publicación y compartimiento de información u otros conjuntos de datos e incluso conocimiento en la web semántica utilizando URIs y RDF.

En la figura 1, extraída de `lod-cloud.net` se observa el grado de conectividad entre fuentes de información a fecha de 22-09-2010, con 203 datasets, más del doble que en 2009. La primera imagen existe desde el 01-05-2007 con 12 datasets y la última actualización contiene 1139 datasets convirtiendo la misma imagen casi ilegible.

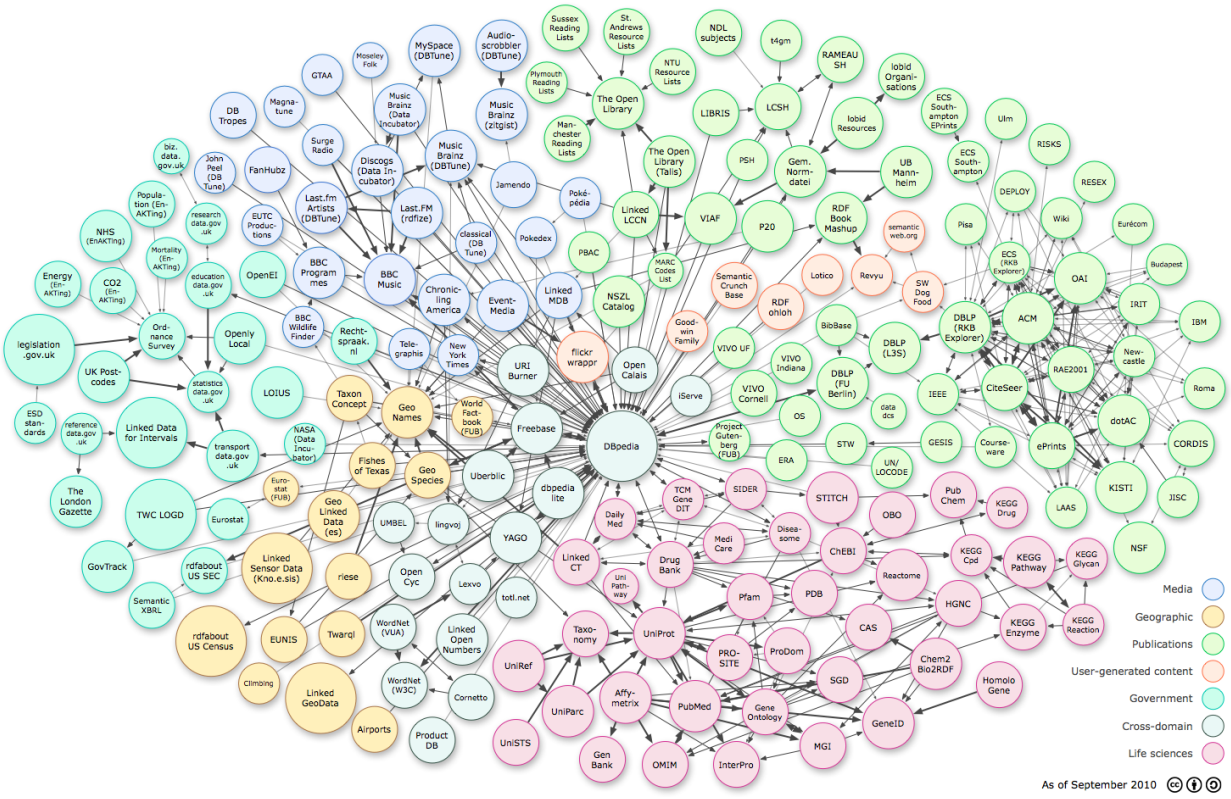

Figura 1: Conectividad en la web a finales de 2010. `lod-cloud.net`

En este trabajo se extrae un conjunto de datos de la web y se realiza un análisis para su posterior transformación de los mismos. Además se desarrolla una aplicación de ejemplo que permita la utilización de estos datos y su enlazamiento en la web.

2. Proceso de transformación

En esta sección se detalla el proceso que han seguido los datos desde su extracción cruda de la red hasta su estado final para su completo uso.

2.1. Selección de los datos

La web contiene un exceso de información. Si se buscan conjuntos de datos se puede encontrar una infinidad de ellos. Aún así su acceso para su tratamiento es más complejo ya que hay distintos factores a tener en cuenta como el formato de los datos, la licencia de los mismos, etc.

En nuestro caso buscamos un conjunto de datos que permita su uso para consumo propio y que además tengan un formato estandar para su transformación. También necesita tener una licencia que permita su publicación por lo que buscamos conjuntos de datos que cumplan todas estas características.

Una de las webs que hemos encontrado y que es bastante idónea para lo que se buscaba es `catalog.data.gov` de la que hemos extraído uno de los miles de conjuntos de datos que contiene esta página. Todos los conjuntos observados de la página contienen una licencia que permite su publicación. El conjunto extraído es *city-of-hartford-birth-information-2002-2012* en su formato csv. La elección del conjunto no tiene ninguna justificación relevante, la página permite buscar conjuntos por etiquetas y se eligieron algunas etiquetas que parecieron más atractivas para tratar los datos, sobre todo con el fin de que la etiquetación de los datos fuese entendible. El conjunto también contiene estos mismos datos en el formato que deseamos de destino, el formato rdf, pero nosotros realizamos la transformación con las técnicas aprendidas en la asignatura.

2.2. Análisis de los datos

Los datos extraídos contienen información de nacimiento de la ciudad de hartford en el estado de Connecticut desde el año 2002 hasta el año 2012.

Los campos que contiene el dataset son:

- Año de nacimiento
- Género
- Etnia
- Raza
- Vecindario
- Distrito

El formato es csv (*Comma-separated values*) por lo que el fichero puede ser abierto por la aplicación de excel. En ella podemos observar los distintos valores del conjunto en cada columna donde, para el año de nacimiento tenemos valores de entre 2002 hasta 2012, el género entre masculino y femenino, distintos tipos de etnias, razas, vecindarios y cuatro zonas para la columna de distritos. Además, el conjunto contiene 22496 instancias.

Los datos tienen relación con la salud pública y están protegidos por la licencia creative-commons `creativecommons.org/publicdomain/zero/1.0/legalcode`

2.3. Extrategia de nombrado

Cada registro se relaciona con todos sus atributos, pero no tenemos ningún identificador específico fuera del mismo identificador de registro así que éste será el identificador. Para cada atributo se busca en distintos vocabularios para comprobar si coincide el nombre del atributo con los atributos de ontologías ya definidas en la web.

2.4. Desarrollo del vocabulario

En este conjunto de datos no se desarrolla ningún vocabulario ya que no es necesario. Cada atributo utiliza el vocabulario de dbpedia ya que para cada atributo existe un atributo en su ontología del vocabulario dbpedia-owl. La ontología es `dbpedia.org/ontology`

Así, tenemos:

- **Birth_Year**. `dbpedia.org/ontology/birthYear`
- **Gender**. `dbpedia.org/ontology/gender`
- **Ethnicity**. `dbpedia.org/ontology/ethnicity`
- **Race**. `vocab.getty.edu/ontology#nationalityNonPreferred`
- **Neighborhood**. `dbpedia.org/ontology/neighborhood`
- **High_School_District**. `dbpedia.org/ontology/district`

2.5. Transformación de los datos

Para realizar la transformación de los datos se ha utilizado la herramienta OpenRefine. Esta herramienta nos permite realizar una limpieza de los datos. Analizamos los datos con la herramienta y comprobamos que existen algunos valores que pueden presentar problemas. Por ejemplo, en la etnia existen varios valores con nombres distintos que representan la misma etnia, como el valor "Brazilian" y "Brasilian". Para solucionar este problema se escoge una de las dos y la otra se cambia

por la escogida. En este ejemplo se escogió Brazilian porque es el nombre que se utiliza en inglés, igual se procede con el resto de casos. También existen otros nombres que están escritos mal, estos nombres se corrigen cambiándolo por el nombre que está bien escrito. También los nombres varían en número donde se escoge el que está en singular.

Normalmente cuando aparecen distintos nombres para especificar la misma cosa existe un número muy grande de casos con el nombre correcto o más común y un número muy reducido de registros con el nombre erróneo.

Para el resto de columnas se aplica el mismo proceso. Sobre las razas existe bastante duda sobre cómo reducirlas ya que no hay un convenio específico con su denominación. La solución aplicada es dejar cada raza como viene y las que indiquen la misma raza elegir la del nombre en inglés.

Además, para las etnias, existen varios registros que contienen varios valores para el mismo campo. En este caso se ha decidido separar para que tengan varios valores en el mismo campo. Por lo tanto aparecerán varias filas con sólo el valor de la etnia pero constituirán el mismo registro.

Una vez limpiado los datos se pasa a la creación del fichero *turtle* con RDF. Para ello es necesario instalar la extensión de RDF. Se crean los prefijos necesarios, el prefijo `db` con la URI `dbpedia.org/ontology/` y el prefijo `gvp` con la URI `http://vocab.getty.edu/ontology#` y luego se asignan los nombres de los atributos a cada columna.

- **Birth_Year**. `db:birthYear`
- **Gender**. `db:gender`
- **Ethnicity**. `db:ethnicity`
- **Race**. `gvp:nationalityNonPreferred`
- **Neighborhood**. `db:neighborhood`
- **High_School_District**. `db:district`

Para la raíz se especifica la clase Person y se le asigna el ID del registro con la expresión `"ID:"+(row.record.index+1)`. Esto es porque tenemos filas con valores que pertenecen a otro registro. Si lo hiciésemos por fila entonces existirían dos IDs distintos para valores que pertenecen al mismo registro. De esta manera, cuando se desee buscar por etnias se puede buscar una sola, encontrar el ID, y con el ID encontrar todas las etnias.

Para la base URI se pone exercise.org como ejemplo ya que no se utiliza.

Se exportan los datos en formato *turtle*. El nombre de este fichero es mi_proyecto.ttl

3. Aplicación y explotación

Una aplicación de los datos sería la de usos estadísticos. Analizar la cantidad de población nacida cada año y también por distritos. En código R habría que instalar la aplicación, además hay que instalar las librerías de xml y curl en linux y luego instalar los paquetes xml, curl y sparql en Rstudio.

Para utilizar Rstudio con sparql se necesita un endpoint de los datos que habría que crear con openrefine pero en este caso no se ha realizado.

En cuanto al código en Rstudio, a continuación aparecen unos ejemplos de códigos que deberían funcionar y extraer una serie de resultados desde el fichero ttl (fichero mi_proyecto.ttl) extraído con openrefine.

```
library(SPARQL)
endpoint <- #endpoint creado
query <-
'
PREFIX db: <http://dbpedia.org/ontology/>
SELECT ?race WHERE{
?race db:race "WHITE"
}'
reslist <- SPARQL(endpoint,query)
df <- reslist$results #Aqui se extrae el resultado $
df
```

En este ejemplo se piden los registros que sean de raza blanca.

4. Conclusiones

Existen diversas fuentes de datos para extraer información de todo tipo como por ejemplo el conjunto de datos extraído para la práctica. Además se puede comprobar que RDF es un formato bastante utilizado ya que algunas páginas ofrecen sus conjuntos de datos también en este formato.

La herramienta OpenRefine es muy útil para analizar y limpiar los datos con los que se va a trabajar. A parte permite la creación de *endpoints* y relaciones entre las columnas. Además es muy fácil editar los datos como añadir o eliminar columnas.

R es una herramienta muy potente en el análisis de datos por lo que la agregación de sparql para el análisis de conjuntos de datos en formato rdf permite que esta herramienta también sea muy útil.

La conexión entre datos de distintas páginas es un concepto que tendrá mucha fuerza ya que la web posee mucha información de todo tipo y, además, el usuario puede utilizarlas ya que muchas

de ellas tienen licencias que lo permiten.

En un futuro la disposición de la información en herramientas de trabajo permitirá a los ordenadores extraer de manera más sencilla y concisa la información clave que necesite el usuario.

5. Bibliografía

Toda la información ha sido extraída del temario y de algunas páginas web como:

- `dbpedia.org`
- `vocab.getty.edu`
- `lov.okfn.org`