Laboratorio 12 - SPARQLing Alumni

WIKIDATA (`http://www.wikidata.org`) es una versión semi estructurada de WIKIPEDIA en la cual los usuarios editan directamente datos semiestructurados sobre entidades en lugar de editar artículos de texto enriquecido. Por ejemplo, podemos encontrar información (semi-)estructurada sobre la Universidad de Chile en WIKIDATA: `https://www.wikidata.org/wiki/Q232141`. Esta URL se puede encontrar ingresando un texto (por ejemplo, `"universidad de chile"`) en el cuadro de búsqueda en la esquina superior derecha. Podemos ver información en múltiples idiomas; puedes seleccionar el idioma principal en el menú superior.

WIKIDATA es una fuente diversa de información, con datos incompletos y muchas relaciones sobre diferentes tipos de entidades. Entonces, en lugar de utilizar una base de datos relacional con un esquema estricto (o una estructura de árbol que no soporta ciclos), WIKIDATA está modelado en RDF: un formato con estructura de grafos que fue estandarizado por la W3C. Por ejemplo podemos ver el RDF de la Universidad de Chile en la sintaxis de Turtle[1]: `https://www.wikidata.org/wiki/Special:EntityData/Q232141.ttl`.

WIKIDATA provee un servicio de consulta en SPARQL donde los usuarios (incluso usted) pueden escribir sus consultas sobre el dataset completo: `https://query.wikidata.org/`.

Dado que los datos son realmente diversos y que no existe un esquema a seguir, hacer consultas sobre los datos puede ser más complejo. Primero que todo, para ayudarle, en el servicio de consulta indicado arriba hay una larga lista de ejemplos que puede explorar para entender cómo funciona el sistema. Por otro lado, los ejemplos de SPARQL vistos en la Clase 13 pueden ser de ayuda. Además, está la documentación de SPARQL en `https://www.w3.org/TR/sparql11-query/`.

Finalmente, veremos un ejemplo simple para comenzar. También puede encontrar estos ejemplos en `sparql-ejemplos.txt` en UCursos. Digamos que queremos encontrar los nombres (en español) de todas las universidades de Chile. ¿Cómo empezar? Mirando un ejemplo, podemos ver que la propiedad `instancia de` (`P31`) se usa para decir que la entidad es una `universidad` (`Q3918`).[2] Entonces, podemos hacer la siguiente consulta[3]

```
SELECT *
WHERE {
 ?uni wdt:P31 wd:Q3918 .
}
LIMIT 10
```

La instrucción `LIMIT 10` evitará que el servicio nos tenga que retornar **todas** las universidades del mundo (o al menos las que están registradas en WIKIDATA). Pruebe la consulta en el servicio.

[1] `https://www.w3.org/TR/turtle/`

[2] Note que, dado que WIKIDATA no es exclusivo de un lenguaje en particular, los IDs usados para los nodos (e.g., `Q3918`, "*universidad*") y propiedades (e.g., `P31`, "*instancia de*") son numéricos.

[3] Los prefijos `wdt:` y `wd:` están definidos por defecto por el servicio. Puede usar el botón `Prefixes` para revisarlos.

Si hace click en cualquier resultado, podrá revisar que efectivamente se trata de una universidad. Sin embargo, estamos retornando códigos, no los nombres. Para eso podemos ejecutar:

```
SELECT ?nom
WHERE {
 ?uni wdt:P31 wd:Q3918 .
 ?uni rdfs:label ?nom .
}
LIMIT 10
```

En esta consulta, usamos `SELECT` para retornar solo el nombre (y no el ID). Entonces, si ejecutamos esta consulta ¡tendremos los nombres! Pero hay nombres en una gran variedad de idiomas y nosotros los queremos solo en español.

```
SELECT ?nom
WHERE {
 ?uni wdt:P31 wd:Q3918 .
 ?uni rdfs:label ?nom .
 FILTER(lang(?nom)="es")
}
LIMIT 10
```

Bien, ¡vamos progresando! Ahora queremos encontrar las universidades que están en Chile. Mirando un ejemplo podemos ver que la relación `país` (`P17`) está definida con la entidad `Chile` (`Q298`). Así que basta con agregar lo siguiente a nuestra consulta:

```
SELECT ?nom
WHERE {
 ?uni wdt:P31 wd:Q3918 .
 ?uni rdfs:label ?nom .
 ?uni wdt:P17 wd:Q298 .
 FILTER(lang(?nom)="es")
}
```

Acá también quitamos el `LIMIT 10` porque estaremos felices de ver todas las respuestas (no hay tantas). Ahora deberíamos ver nuestra Universidad de Chile en los resulta... *pero ¡no está!* ¿Cómo es que tenemos casi todas las universidades de Chile pero no la Universidad de Chile? Lo que pasa es que WIKIDATA dice que la Universidad de Chile (`https://www.wikidata.org/wiki/Q232141`) es una `instancia de` una `universidad pública` (`wd:Q875538`), no de una `universidad` (`wd:Q298`), pero la clase `universidad pública` es una `subclase de` (`wdt:P279`) la clase `universidad`. Pues, para capturar no solo las instancias directas de las universidades, sino también las instancias de sus subclases, tenemos que cambiar la consulta a:

```
SELECT ?nom
WHERE {
 ?uni wdt:P31/wdt:P279* wd:Q3918 .
 ?uni rdfs:label ?nom .
 ?uni wdt:P17 wd:Q298 .
 FILTER(lang(?nom)="es")
}
```

Ahora está la Universidad de Chile. La expresión `wdt:P31/wdt:P279*` busca caminos que empiezan con `wdt:P31` (*instancia de*) seguido por una cadena de cero-o-más `wdt:P279` (*subclase de*).

Lo que sigue no es estrictamente SPARQL, pero es interesante: el servicio soporta visualizaciones con los resultados de las consultas, así que mostraremos las universidades sobre un mapa.

```
#defaultView:Map
SELECT ?nom ?coord
WHERE {
 ?uni wdt:P31/wdt:P279* wd:Q3918 .
 ?uni rdfs:label ?nom .
 ?uni wdt:P17 wd:Q298 .
 ?uni wdt:P625 ?coord .
 FILTER(lang(?nom)="es")
}
```

El símbolo `#` da un comentario que el sistema interpreta para cargar una visualización en particular.

Usted debe entregar las consultas en SPARQL para encontrar los siguientes resultados en Wikidata. Tenga cuidado de proyectar solo las variables pedidas y de quitar las soluciones duplicadas. En el caso de las universidades, debe incluir las instancias de subclases de universidades. Siempre devuelva los nombres en español. No es necesario usar ninguna visualización. (Tenga en cuenta que los datos *pueden* estar incompletos: la base de conocimiento depende de sus voluntarios.) Tendrá que entregar un archivo `.txt` con la consulta SPARQL para cada pregunta.

Tip: El servicio de consulta de Wikidata tiene varias formas de ayudarnos a generar una consulta en SPARQL. Una parte difícil es saber qué identificador usar. Si escribe `wdt:` o `wd:` en una consulta (p.ej., en la clausula `WHERE`), puede presionar Ctrl + Espacio para abrir un diálogo de búsqueda. Luego puede buscar por el nombre de la entidad o propiedad que desee incluir.

P1. 10 PUNTOS La lista de las personas educadas en una universidad chilena. Retorne el nombre de la persona y de la universidad.

P2. 10 PUNTOS La lista de las personas educadas en universidades chilenas que son políticos (tienen la ocupación de ser político) y miembros de un partido político. Retorne el nombre de la persona, de la universidad, y de su(s) partido(s).

P3. [10 PUNTOS] La lista de obras literarias (instancias de `Q7725634` y sus subclases) escritas por personas educadas en o empleados por una universidad chilena. Retorne el nombre de la obra, del autor y de la universidad.

P4. [10 PUNTOS] La lista de las mujeres educadas en una universidad chilena que son músicas o cantantes (tienen la ocupación de ser músico o cantante). Retorne el nombre de la persona y de la universidad.

P5. [10 PUNTOS] La lista de personas educadas en universidades chilenas que han ganado algún Premio Nobel o algún Premio Óscar. Retorne el nombre de la persona, de la universidad, y del premio.[4]

P6. [10 PUNTOS] La lista de las películas dirigidas por personas educadas en universidades chilenas. Retorne el nombre de la película, del director, de la universidad; la fecha de estreno de la película; y, si está disponible, el identificador de IMDb de la película (si no está disponible, debe devolver el resultado igual con el identificador de IMDb en blanco). Ordene los resultados por fecha (el más reciente primero).[5]

Podemos ver que, si sabemos algo de SPARQL, WIKIDATA nos permite responder consultas muy específicas, las que serían horribles de responder en WIKIPEDIA. El único problema es la incompletitud: mientras un modelo basado en grafos nos permite modelar datos diversos e incompletos de forma fácil, es difícil saber cuándo estamos obteniendo todos los resultados deseados. Sin embargo, podemos obtener rápidamente al menos algunos resultados y, a medida que WIKIDATA esté siendo editada cada vez más por más y más voluntarios, podemos esperar que la completitud aumente (incluso podemos ayudar a completarlo).

[4] Los premios específicos –como el Premio Nobel de la paz, el Óscar a la mejor película– se relacionan con su tipo de premio –como un Premio Nobel o un Óscar– a veces con `parte de` (`P361`) y a veces con `instancia de` (`P31`).

[5] Puede haber duplicados si una película tiene varias fechas en países diferentes o varios directores relevantes; no nos importa acá.