Laboratorio SPARQL

https://users.dcc.uchile.cl/~hrosales/SPARQL_lab.pdf

WIKIDATA es una versión semi-estructurada de los datos de WIKIPEDIA en la cual los usuarios editan directamente los datos sobre entidades (lugares, cosas, personas, etc.) en lugar de editar artículos de texto enriquecido[1]. Por ejemplo, podemos encontrar información sobre la Universidad Central "Marta Abreu" de Las Villas en `https://www.wikidata.org/wiki/Q6156410`. Esta URL puede ser encontrada ingresando un texto (por ejemplo, `"Universidad Central de las Villas"`) en el cuadro de búsqueda en la esquina superior derecha. Podemos ver información en múltiples idiomas, puedes seleccionar el idioma principal en la parte superior de la página.

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 Central "Marta Abreu" de Las Villas en la URL mencionada en el párrafo anterior[2]. El formato es RDF, cuyos triples se escriben usando la sintaxis de Turtle[3].

Hay a disposición un servicio de consulta en SPARQL donde los usuarios 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 complicado. 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. Además, está la documentación de SPARQL en `https://www.w3.org/TR/sparql11-query/`. Finalmente, veremos un ejemplo simple para comenzar. Digamos que queremos encontrar los nombres (en español) de todas las universidades de Cuba. ¿Cómo empezar?

Tendríamos que mirar un recurso de ejemplo primero, como el de la Universidad Central "Marta Abreu" de Las Villas. Podemos ver en la página que la propiedad `instance of` (`P31`) se usa para decir que el recurso es una `universidad` (`Q3918`)[4]. Entonces, podemos hacer la siguiente consulta[5]

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

[1] `http://www.wikidata.org`

[2] `https://www.wikidata.org/wiki/Special:EntityData/Q6156410.ttl`

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

[4] Note que, dado que WIKIDATA no es específica a un lenguaje en particular, los IDs usados para los nodos (e.g., `Q3918`, "university") y propiedades (e.g., `P31`, "instance of") son numéricos. Encontrar los códigos adecuados puede ser complicado, pero la forma más fácil es mirar ejemplos que los usen.

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

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. Si hace click en cualquier resultado, podrá revisar que efectivamente se trata de universidades. 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 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 Cuba. Entonces, volvamos al recurso de la Universidad Central de Las Villas y veamos cómo se define allí. Podemos ver que existe la relación `país` (`P17`) está definida como `Cuba` (`Q241`). Entonces, solo basta con agregar lo siguiente a nuestra consulta:

```
SELECT ?nom
WHERE {
 ?uni wdt:P31 wd:Q3918 .
 ?uni rdfs:label ?nom .
 ?uni wdt:P17 wd:Q241 .
 FILTER(lang(?nom)="es")
}
```

Acá también quitamos el `LIMIT 10` porque estaremos felices de ver todas las respuestas :). 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 wd:Q3918 .
  ?uni rdfs:label ?nom .
  ?uni wdt:P17 wd:Q241 .
  ?uni wdt:P625 ?coord .
  FILTER(lang(?nom)="es")
}
```

El símbolo `#` denota un comentario que el sistema interpreta para cargar una visualización en particular. ¡Buenísimo! ¿O no?

Desarrolle las siguientes consultas. Siempre retorne los nombres en español. No es necesario añadir ninguna visualización. Tenga en cuenta que los datos *pueden* estar incompletos, la base de conocimiento depende de los voluntarios que agregan información.

P1. La lista de mujeres educadas en universidades cubanas. Retorne el nombre de la mujer y el de la universidad.

P2. La lista de personas educadas en universidades cubanas y cuya fecha de muerte esté disponible. Retorne el nombre de la persona, de la universidad y la fecha de muerte.

P3. La lista de directores de cine educados en universidades cubanas. Retorne el nombre del director o directora y de la universidad.

P4. La lista de las películas de dichos directores. Retorne el nombre de la película, del director y de la universidad.

P5. La lista anterior, pero ordenada por la fecha de estreno de la película, partiendo con la más reciente. Retorne el nombre de la película, del director, de la universidad y la fecha de estreno.

P6. La lista de todos los nombres de las instituciones educacionales en Cuba.

P7. Encuentre la ocupación más común para las personas educadas en universidades cubanas. Retorne el nombre de la causa de la ocupación y la cuenta. Ordene convenientemente.

Podemos ver que, si sabemos algo de SPARQL, Wikidata nos permite responder consultas muy específicas, las que sería horrible responder manualmente. El único problema es la incompletitud: mientras un modelo basado en grafos nos permite modelar datos diversos e incompletos de forma fácil, es dificil saber cuándo estamos obteniendo todos los resultados deseados. Sin embargo, podemos obtener rápidamente al menos algunos resultados y, a medida que Wikidata está siendo editada cada vez más por más y más usuarios, podemos esperar que la completitud aumente.