Práctica 9 de Sistemas de Información para la Web

by

Ángel García Menéndez

Universidad de Oviedo

Práctica 9 de Sistemas de Información para la Web

by
Ángel García Menéndez

Obtención de información estructurada

A continuación se ofrecen tanto el listado de entidades identificadas en los textos como sus tipos y propiedades dentro de los ofrecidos por *schema.org*.

Miles Davis

Como entidad escogeremos `Person` que si bien resulta generalista, posee las suficientes propiedades para poder ajustarse a nuestras necesidades. El conjunto de propiedades escogidas serán:

• familyName: Davis
• hasOcupation: Occupation: (name: jazz musician)
• nationality: Country(name: United States of America)
• name: Miles
• URL: https://www.wikidata.org/wiki/Q93341

Barack Obama

De la misma forma, recurriremos a `Person` para modelar al ex-presidente de EEUU. En este caso las propiedades sería:

• familyName: Obama
• hasOcupation: Ocupation (name: politician)
• jobTitle: president
• name: Barack
• URL: https://www.wikidata.org/wiki/Q76

European Union

El caso de la Unión Europea es particular, puesto que (discutiblemente), podría considerarse que entra dentro de la categoría de estado. Sin embargo, tomando una aproximación conservadora, la consideraremos como `Organization`, a falta de un concepto más específico:

• name: European Union
• URL: https://www.wikidata.org/wiki/Q458

Whashington

Aunque no se especifica si se trata de la ciudad de Washington o del estado del mismo nombre, asumiremos que se trata de la capital de EEUU, y por ende haremos uso de `City` . Un posible listado de propiedades sería:

• name: Washington DC

• URL: https://www.wikidata.org/wiki/Q61

cambio euro/dolar

No existe una entidad que sea *moneda*, aunque sí un tipo de cambio, `ExchangeRateSpecification`, que es lo que discute en el texto, y por ende se ajusta adecuadamente. Las propiedades sería:
• currency: EUR
• currentExchange: UnitPriceSpecification(price: 1.3, priceCurrecy: USD)
• URL: https://www.ecb.europa.eu/stats/policy_and_exchange_rates/euro_reference_exchange_rates/html/eurofxref-graph-usd.en.html

New York Times

En el caso del NYT, existe la entidad `Newspaper` para describir periódicos. Como conjunto de propiedades:
• name: New York Times
• URL: https://www.wikidata.org/wiki/Q9684

John McCarthy

Nuevamente, ante la falta tanto de entidades como de información en sí, se recurre a la entidad `Person`:
• familyName: McCarthy
• name: John
• URL: https://www.wikidata.org/wiki/Q92739

LISP

Schema.org ofrece una entidad denominada `ComputerLanguage`, que además se ejemplifica con el propio LISP.
• name: LISP
• URL: https://www.wikidata.org/wiki/Q132874

Modelado RDF

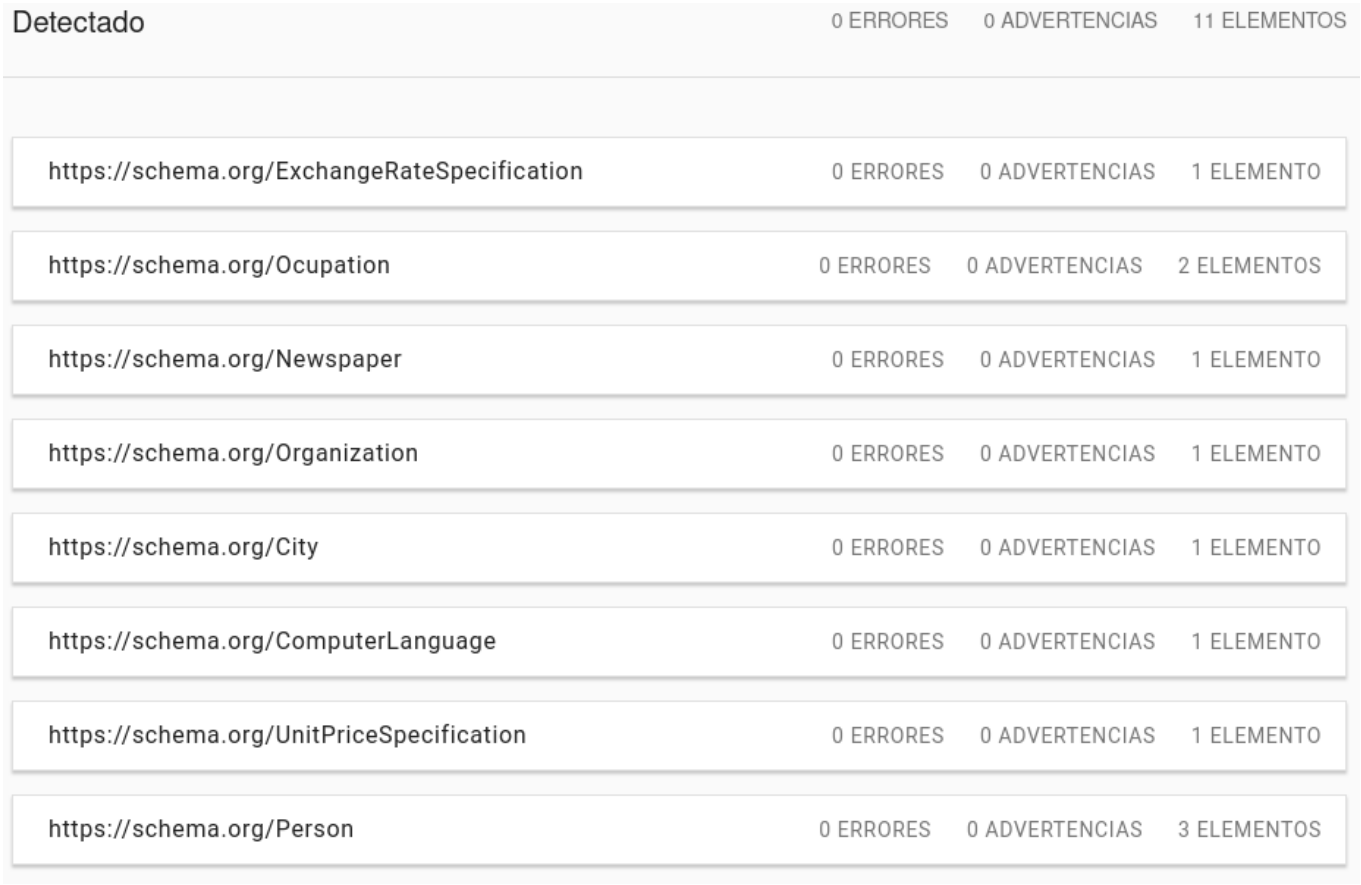
A continuación se muestra el modelado de la información previamente estructurada a RDF, empelando para ello el lenguaje Turtle.

```
@prefix schema: <https://schema.org/> .

@prefix wikidata: <https://wikidata.org/> .

@prefix rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
```

```
# First text wikidata:Q93341 rdf:type schema:Person ;
    schema:familyName "Davis";
    schema:hasOcupation [
        rdf:type schema:Ocupation;
        schema:name "Jazz Musician"
    ];
    schema:name "Miles";
    schema:nationality wikidata:Q30.

# Second text wikidata:Q76 rdf:type schema:Person;
    schema:familyName "Obama";
    schema:name "Barack";
    schema:jobTitle "President";
    schema:hasOcupation [
        rdf:type schema:Ocupation;
        schema:name "Politician"
    ].

wikidata:Q458 rdf:type schema:Organization;
    schema:name "European Union".

wikidata:Q61 rdf:type schema:City;
    schema:name "Washington DC".

<https://www.ecb.europa.eu/stats/policy_and_exchange_rates/
euro_reference_exchange_rates/html/eurofxref-graph-
usd.en.html> rdf:type schema:ExchangeRateSpecification;
    schema:currency "EUR";
    schema:currentExchange [
        rdf:type schema:UnitPriceSpecification;
        schema:price 1.3;
        schema:priceCurrency: "USD"
    ].

# Third text wikidata:Q9684 rdf:type schema:Newspaper;
    schema:name "New York Times".

wikidata:Q92739 rdf:type schema:Person;
    schema:name "John";
    schema:familyName "McCarthy".
```

```
wikidata:Q132864 rdf:type schema:ComputerLanguage;
    schema:name "LISP".
```

Comprobamos con la herramienta *RDFShape* que la sintaxis es correcta, y con su conversión a JSON-LD lo sometemos a la herramienta datos estructurados de Google para confirmar que, efectivamente, la información puede extraerse correctamente.



Obtención automática de información estructurada

Los tres textos han sido sometidos a análisis automatizado con las tres herramientas propuestas. Dichos análisis han sido convertidos a RDF-XML, Turtle y JSON-LD. Con objeto de amenizar el presente escrito, estos se proporcionan como ficheros separados, clasificados por texto. Los nombres son autoexplicativos.

Señalar lo tedioso de la tarea, pues ninguna de las herramientas permitían descargar el RDF en forma de fichero, o al menos un botón de *copiar al portapapeles*. Además, el traductor de sintaxis RDF requería de el borrado manual del contenido previo para realizar una nueva conversión.

Sobre el uso de ontologías

Cada sistema emplea su propio conjunto de ontologías para el tipado.

Por un lado, **OpenCalais** hace uso casi de forma exclusiva de su propia ontología. Esto resulta de especial notoriedad cuando se examina los datos extra que es capaz de proporcionar en realción al texto, así como para representar la estrucutra del propio texto.

Por su parte, **FRED** emplea, aparte de la suya propia, un abanico más amplio de ontologías. Por ejemplo, están presentes *The DOLCE+DnS Ultralite ontology (dul)* o *Boxer*, amén de otros recursos de diversa índole para poder ofrecer sus servicio.

Finalmente, **DBpedia Spotlight** hace uso de su propia ontología, la cual está basada en la información semántica extraída de la Wikipedia.

Concordancias entre ontologías

Como es de esperar, hay entidades y propiedades comunes entre las onotologías. A continuación se enumeran y explican las encontradas en este caso:

- `Person`: entidad presente en las 3, pues es prácticamente fundamental en el lenguaje humano.
- `Position`: parte de la ontología de OpenCalais, puede considerarse equivalente a `Occupation`, pues ambas denotan profesiones.
- `Organization`: similar, por no decir exactamente igual, a la de schema.org.
- `City`: de nuevo, casi indistinguible de la entidad de schema.org.
- `Leader`: propia de la ontología de FRED, pudiendo ser equivalente a la propiedad `jobTitle` de `Person` con el valor “Leader”.
- `Huddle`: propia de FRED, sin equivalencia en schema.
- `Fear`: propia de FRED, sin equivalencia en schema.
- `Grow`: propia de FRED, sin equivalencia en schema.
- `Future`: propia de FRED, sin equivalencia en schema.
- `Euro`/`Dollar`: no existen en schema, pero puede realizarse como se vio previamente en el ejercicio anterior.
- `Published Medium`: propia de OpenCalais, aunque existen varias equivalentes en schema.
- `Programming Language`: casi igual a `ComputerLanguage`
- `Report`: presente de forma exactamente igual en schema.
- `Invent`: propia de FRED, sin equivalente en schema.
- `topical concept`: propia de DBpedia
- `populated place`: propia de DBpedia, aunque equivalente a `Place` de schema con algo más de especificidad.
- `scientist`: propia de DBpedia, equivalente a `Occupation` de schema con el nombre acorde.
- `Newspaper`: presente con el mismo nombre y características que en schema.

Empleando *sameAs.org* se pueden confirmar la mayoría de equivalencias que se han

señalado, con la salvedad de las propias de FRED, las cuales parecen ser muy específicas.

Todas las herramientas han cumplido su propósito de forma correcta. OpenCalais se sienta como la más madura, tanto en términos de interfaz como en lo acertado de su etiquetado, aunque la representación en RDF resulta excesivamente verbosa. FRED se queda algo atrás, aunque la presentación en diagrama resulta útil como alternativa a manejar diferentes sintaxis RDF. El uso que hace de sus ontologías es, al menos aparentemente, el más complejo, aunque la necesidad de que lo sea puede estar en entredicho. Finalmente, DBpedia Spotligth, si bien cumple, se siente bastante menos refinada que las otras dos. No sólo en el apartado de la usabilidad, sino incluso en la forma en que etiqueta, y el hecho de que no permite exportar la información a ningún tipo de formato estándar es una gran lacra para el servicio.

Sobre el uso de ontologías propias

De forma instintiva, se podría pensar que el mejor curso de acción para con las ontologías sería el tratar de universalizarlas en la medida de lo posible. Pudiera concebirse incluso la creación de una suerte de *ontología universal*, que se pudiese emplear en cualquier rama de conocimiento. En el momento en que se aborda desde un punto de vista realista los problemas empiezan a surgir.

Como ocurriese con otros aspectos de la Recuperación de Información cuando se aplicaron a Internet, el uso de ontologías se ve lastrado por la naturaleza heterogénea de la web. Existen innumerables cantidad de textos, de incontables temáticas y campos, y por tanto las ontologías generalistas palidecen por sus propias limitaciones. Esto unido a la creencia general de los grupos humanos de ser portadores de la verdad universal, acaba derivando en que cada herramienta y servicio opten por crear su propio conjunto ontológico que consideren que verdaderamente se ajusta a sus necesidades concretas.

Existe sin embargo otra cara de la moneda, pues en un mundo interconectado, más en la informática, y más en la web, la universalización, aunque esta sólo sea hasta un cierto punto, acaba volviéndose una necesidad para garantizar la longevidad de los proyectos. Más aún en un campo como la Web Semántica, cuyo principal atractivo era, y supuestamente sigue siendo después de cuantiosas reinterpretaciones, el enlace de conceptos en pos de la creación de un posible grafo de conocimiento.

EL proceso de alineación de ontologías no está exento de dificultades, pues en general involucra relacionar términos y conceptos a niveles diferentes, y atendiendo a criterios muy diversos. En cualquier caso, se trata en una labor necesaria, e incluso fundamental, cuando se trata con ontologías, especialmente en los entornos digitales, y debe ser tratado en consecuencia.