SIW PL9. Extracción de información semántica.

Hugo Fonseca Díaz uo258318@uniovi.es

Escuela de Ingeniería Informática. Universidad de Oviedo.

19 de noviembre de 2020

1. Información estructurada

En esta primera sección se obtendrá información estructurada de tres textos mediante el uso de tipos definidos en schema.org [1]. Dicha información se define a continuación:

- Miles Davis: es una entidad de tipo Person. Sus propiedades son:
 - familyName: Davis
 - givenName: Miles
 - hasOccupation: Occupation(name: jazz musician)
 - nationality: Country(name: United States of America)
 - url: https://www.wikidata.org/wiki/Q93341
- Barack Obama: es una entidad de tipo Person. Sus propiedades son:
 - familyName: Obama
 - givenName: Barack
 - hasOccupation: Occupation(name: President)
 - url: https://www.wikidata.org/wiki/Q76
- European Union: es una entidad de tipo Organization. Sus propiedades son:
 - name: European Union
 - legalName: European Union
 - url: https://www.wikidata.org/wiki/Q458
- Washington: es una entidad de tipo City. Sus propiedades son:
 - name: Washington DC

- url: https://www.wikidata.org/wiki/Q61
- Euro a Dólar: es una entidad de tipo ExchangeRateSpecification. Sus propiedades son:
 - currency: EUR
 - currentExchangeRate: UnitPriceSpecification(priceCurrency: USD, price: 1.3)
 - url: https://www.ecb.europa.eu/stats/policy_and_exchange_rates/euro_reference_exchange_rates/html/index.en.html
- The New York Times: es una entidad de tipo Newspaper. Sus propiedades son:
 - name: The New York Times
 - url: https://www.wikidata.org/wiki/Q9684
- John McCarthy: es una entidad de tipo Person. Sus propiedades son:
 - familyName: McCarthy
 - givenName: John
 - hasOccupation: Occupation(name: computer scientist)
 - url: https://www.wikidata.org/wiki/Q92739
- LISP: es una entidad de tipo ComputerLanguage. Sus propiedades son:
 - name: LISP
 - url: https://www.wikidata.org/wiki/Q132874

2. Modelado RDF

En esta sección se muestra el modelado RDF en formato Turtle con el que se ha representado la información listada previamente. Dicho modelado es el siguiente:

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

# Primer texto
wikidata:Q93341 rdf:type schema:Person ;
    schema:familyName "Davis" ;
    schema:givenName "Miles" ;
    schema:hasOccupation [
        rdf:type schema:Occupation ;
        schema:name "jazz musician"
    ] ;
    schema:nationality wikidata:Q30 .
```

```
# Segundo texto
wikidata:Q76 rdf:type schema:Person ;
   schema:familyName "Obama" ;
   schema:givenName "Barack" ;
   schema:hasOccupation [
       rdf:type schema:Occupation ;
       schema:name "president"
   ] .
wikidata:Q458 rdf:type schema:Organization;
   schema:name "European Union" ;
   schema:legalName "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/index.en.htm</pre>
   rdf:type schema:ExchangeRateSpecification ;
   schema:currency "EUR" ;
   schema:currentExchangeRate [
       rdf:type schema:UnitPriceSpecification ;
       schema:priceCurrency "USD" ;
       schema:price 1.3
   ] .
# Tercer texto
wikidata: Q9684 rdf: type schema: Newspaper;
   schema:name "The New York Times" .
wikidata:Q92739 rdf:type schema:Person;
   schema:familyName "McCarthy" ;
   schema:givenName "John" ;
   schema:hasOccupation [
       rdf:type schema:Occupation ;
       schema:name "computer scientist"
   ] .
wikidata:Q132874 rdf:type schema:ComputerLanguage ;
   schema:name "LISP" .
```

En la **Figura 1** se observa el output de la herramienta *RDFShape* al introducir el fichero en formato *Turtle*. Se observa que el fichero está bien formado. Una vez hecho eso, convertimos el fichero a *JSON-LD* mediante el uso de la herramienta *RDF Translator* [2]. En la **Figura 2** se puede ver la validación de la *Google Structured Data Testing Tool* [3] sobre el *JSON-LD* generado previamente.

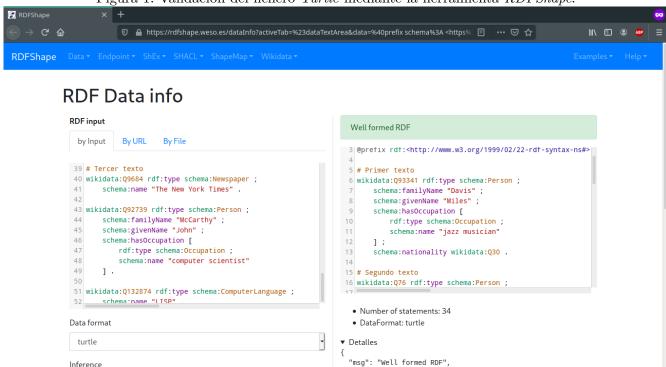


Figura 1: Validación del fichero Turtle mediante la herramienta RDFShape.

3. Obtención automática de información estructurada

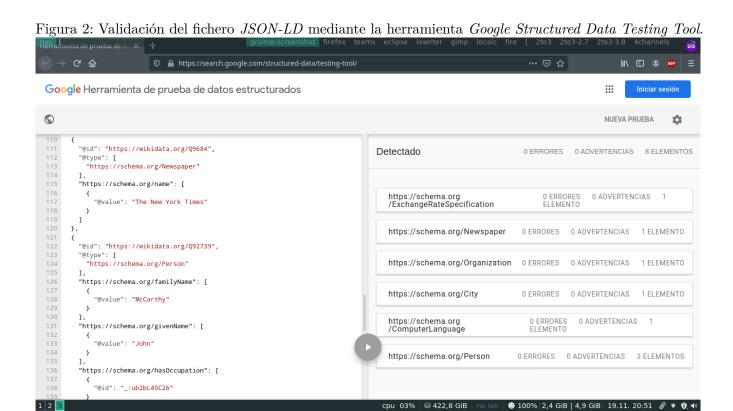
En esta parte del documento se procederá a obtener información estructurada mediante el uso de varias herramientas. Estas son *Intelligent Tagging* de *PermID* [4], *DBpedia Spotlight* [5] y *FRED* [6].

pu 05% | ⊜ 422,8 GiB

A continuación se muestra una serie de ejemplos en forma de capturas de pantalla. Estos ejemplos ilustrarían el proceso realizado para obtener la información estructurada por medio de la herramienta FRED para el texto "Miles Davis was an american jazz musician.". En la Figura 3 se muestra la herramienta FRED con el texto en lenguaje natural. En la Figura 4 se muestra la conversión a JSON-LD del fichero RDF-XML obtenido en el previo paso (también se transforma a N3 para una inspección manual, aunque no se muestra en la captura). Por último, en la Figura 5 se muestra la validación del fichero JSON-LD. Estos pasos se realizan para todos los textos y los ficheros obtenidos pueden observarse en las carpetas adjuntadas a este documento.

Referencias

- [1] Página de Schema.org, https://schema.org/. Última vez accedido 19 de noviembre de 2020.
- [2] Página del RDF Translator, https://rdf-translator.appspot.com/. Última vez accedido 19 de noviembre de 2020.



- [3] Página de la Google Structured Data Testing Tool, https://search.google.com/structured-data/testing-tool/. Última vez accedido 19 de noviembre de 2020.
- [4] Página de la herramienta Intelligent Tagging de Refinitiv, https://permid.org/onecalaisViewer. Última vez accedido 19 de noviembre de 2020.
- [5] Página de la herramienta *DBpedia Spotlight*, https://www.dbpedia-spotlight.org/demo/. Última vez accedido 19 de noviembre de 2020.
- [6] Página de la herramienta FRED, http://wit.istc.cnr.it/stlab-tools/fred/demo/. Última vez accedido 19 de noviembre de 2020.

Figura 3: Obtención del fichero en formato RDF-XML mediante la herramienta FRED. ← → ♂ ☆ STEAB - FRED FRED is a tool for automatically producing RDF/OWL ontologies and linked data from natural language sentences. The method is based on Combinatory Categorial Commun. Discourse Representation Theory, Linguistic Frames, and Ontology Design Patterns. Results are enriched with Named Entity Resolution (NEP) and Word-Series Disambiguation (WSO). A paper describing FRED has been published in the Semantic Web Journal by IOS Press. Please refer to it in scientific publications (BibTex). The documentation about how to use FRED as as a HTTP REST service can be found here. To use FRED from via its API you need an API key. You can request a key by filling this form. Output format: RDF/XML - Read it! CAMQOE Result roff:nos

whos: rdf="http://www.wd.org/1999/82/22-rdf-syntas-nss"

walns: j.d="http://www.notologydesignpatterns.org/ont/fred/quantifiers.owls"

walns: j.l="http://www.notologydesignpatterns.org/ont/seniotics.owls"

walns: y.l="http://www.wd.org/2002/07/orgl"

xalns: y.l="http://www.wd.org/2002/07/orgl"

xalns: j.2="http://www.wd.org/2002/07/orgl"

xalns: j.2="http://www.wd.org/2002/07/orgl"

xalns: j.4="http://www.notologydesignpatterns.org/ont/fred/domain.owls"

xalns: j.5="http://www.notologydesignpatterns.org/ont/fred/soc.ouls"

xalns: j.5="http://www.notologydesignpatterns.org/ont/fred/soc.ouls"

xalns: j.7="http://www.notologydesignpatterns.org/ont/fred/soc.ouls"

xalns: j.7="http://www.notologydesign cpu 06% | © 422,8 GiB | no lan | © 100% | 2,6 GiB | 4,7 GiB | 19.11. 21:55 Ø ▼ 🐧 🕩

RDF Translator 70% ⋯ ☑ ☆ RDF Translator is a multi-format conversion tool for structured markup, it provides translations between data formats ranging from RDFXMUL to RDFa or Microdata. The service allows for conversions triggered either by URI or by direct text input. Furthermore it comes with a straightforward REST API for developers. s; 2.16sists rdf-datatype="http://mm.ml.org/2001/NMLSchema/monlegativeInteger">16-/j.2.hegins>
5; 2.1efersTo rdf-resource="http://mm.monlogydesignpatterns.org/ont/fred/domain.oml@docuverse">
5; 2.1efersTo rdf-resource="http://mm.monlogydesignpatterns.org/ont/fred/domain.oml@docuverse">
5; 1.1efenctser off-resource="http://mm.monlogydesignpatterns.org/ont/fred/domain.oml@docuverse">
6-crdf-label.rdf-datatype="http://mm.monlogydesignpatterns.org/ont/fred/domain.oml@docuverse">
6-crdf-label.rdf-datatype="http://mm.monlogydesignpatterns.org/ont/fred/domain.oml@docuverse">
6-crdf-label.rdf-docuverse">
6-crdf-label.rdf-docuverse</->
6-crdf-label.rdf-docuverse</->
6-crdf-label.rdf-docuverse</->
6-crdf-label.rdf-docuverse</->
6-crdf-label.rdf-docuverse</->
6-crdf-label.rdf-docuverse</->
6-crdf-label.rdf-docuverse</->
6-crdf-label.rdf-docuverse</->
6-crd Spennpos rdf.resource="http://www.ontologydesignpatterns.org/ont/fred/pos.onl@pennpos">
Description
Description rdf:about="http://www.ontologydesignpatterns.org/ont/fred/pos.onl@pennpos">
Fitype rdf.resource="http://www.ns.org/2002/07/onl@DjectProperty">
Titype rdf.resource="http://www.ns.org/2002/07/on In the outreasures in the provided in the control of the control o df:Description rdf:about="http://www.ndlogydesignpatterns.org/ont/dul/DUL.on#associated# -df:typerdf:resourc="http://www.ad.org/2002/07/on#80bjectPoperty"/>
rdf:Description#
df:Description rdf:about="http://ontologydesignpatterns.org/cp/on#/semiotics.ow##denotes">
rdf:Description rdf:about="http://ontologydesignpatterns.org/cp/on#/semiotics.ow##denotes">
rdf:typerdf:resource="http://onw.wä.org/2002/07/on##0bjectProperty"/>
rdf:typerdf:resource="http://oww.wä.org/2002/07/on##0bjectProperty"/> Input RDF/XML V Output JSON-LD V cpu 08% | ⊜ 422,8 GiB

Figura 4: Obtención del fichero en formato JSON-LD mediante la herramienta RDF Translator.

Figura 5: Validación del fichero JSON-LD mediante la herramienta Google Structured Data Testing Tool. ... ☑ ☆ → C û Google Herramienta de prueba de datos estructurados • NUEVA PRUEBA \$ Detectado 0 ERRORES 0 ADVERTENCIAS 18 ELEMENTOS "@id": "http://www.ontologydesignpatterns.org/ont/fred/domain.owl#offset_33 "@type": [
 "http://www.essepuntato.it/2008/12/earmark#PointerRange" l, "http://ontologydesignpatterns.org/cp/owl/semiotics.owl#denotes": [http://www.w3.org/2002/07 /owl#ObjectProperty 0 ERRORES ELEMENTOS 0 ADVERTENCIAS 12 "@id": "http://www.ontologydesignpatterns.org/ont/fred/domain.owl#music http://www.essepuntato.it/2008/12 /earmark#PointerRange 0 ERRORES ELEMENTOS 0 ADVERTENCIAS "@id": "http://www.ontologydesignpatterns.org/ont/fred/domain.owl#Miles],
"http://ontologydesignpatterns.org/cp/owl/semiotics.owl#hasInterpretant": ["@id": "http://www.ontologydesignpatterns.org/ont/fred/domain.owl#Music],
"http://www.essepuntato.it/2008/12/earmark#begins": ["@type": "http://www.w3.org/2001/XMLSchema#nonNegativeInteger",
"@value": "33"],
"http://www.essepuntato.it/2008/12/earmark#ends": ["@type": "http://www.w3.org/2001/XMLSchema#nonNegativeInteger", "@value": "41"

8