Vocabularis i taxonomies

PID_00271445

Blas Torregrosa García

Temps mínim de dedicació recomanat: 1 hora

Universitat Oberta de Catalunya

Blas Torregrosa García

Enginyer en Informàtica i màster universitari en Seguretat de les Tecnologies de la Informació i de les Comunicacions (MISTIC) per la Universitat Oberta de Catalunya (UOC). Especialitzat en ciberseguretat. Professor col·laborador del màster de Ciència de Dades de la UOC i professor associat a la Universitat de Valladolid (UVA).

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Ferran Prados Carrasco (2020)

Primera edició: febrer 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Realització editorial: FUOC

Índex

Introducció

Un component fonamental per aconseguir els objectius de la web semàntica és l'ús de vocabularis que posseeixin una semàntica ben definida. Un enfocament que facilita enormement la consecució d'aquest objectiu és el d'utilitzar vocabularis de referència o taxonomies.

1. Vocabularis

Els **vocabularis** defineixen el conjunt de termes utilitzats per una aplicació o per un conjunt d'aplicacions. A més, qualifiquen un conjunt de possibles relacions i defineixen les restriccions sobre els termes que es poden utilitzar en un cert context. Els vocabularis poden ser molt complexos (amb diversos milers de termes) o molt simples (descrivint solament un o dos conceptes).

L'ús principal de vocabularis a la web semàntica és ajudar en la integració de dades en què hi pugui haver ambigüitats entre els diferents conjunts de dades. Suposem que hi ha una biblioteca que vol integrar dades que arriben de diferents editors. Alguns editors poden usar el terme «Autor» per descriure l'escriptor del llibre, mentre que altres editors poden usar el terme «Creador». Per integrar les dades, un vocabulari ha de descriure el fet que l'autor i el creador signifiquin el mateix.

> Els vocabularis controlats recopilen conceptes i termes utilitzats per descriure un camp o àrea d'interès.

Per exemple, per definir una persona en un format entendible per màquines, cal un vocabulari que tingui la definició formal de «Persona». Una opció és utilitzar el vocabulari *Friend of a Friend* (FOAF, literalment 'Amic d'un Amic'), que té una classe «Person» que defineix les propietats típiques d'una persona, entre d'altres el seu nom. Suposem que tenim el fragment XML següent:

Figura 1. Definint la classe i la propietat d'un recurs

```
<Person>
  <name>Neil Armstrong</name>
</Person>
```

Aquest fragment XML proporciona una jerarquia, suposant que hi ha una classe «Person» i que té una propietat «Name». No obstant això, no té context. Cal indicar un vocabulari extern que defineixi aquesta classe i aquesta propietat, utilitzant el mecanisme de l'espai de noms (*Namespace*).

Amb RDF/XML això es pot fer usant l'atribut xmlns. En aquest cas, `xmlns: foaf = "http://xmlns.com/foaf/0.1/"`, que fa referència a l'espai de noms FOAF. Aquest mecanisme permet abreujar-ho mitjançant el prefix foaf. Així `foaf:Person` refereix a http://xmlns.com/foaf/0.1/person.

Figura 2. Descrivint el nom d'una persona usant un vocabulari

```
…
xmlns:foaf="http://xmlns.com/foaf/0.1/"

<foaf:Person>
  <foaf:name>Neil Armstrong</foaf:name>
</foaf:Person>
```

Les definicions dels vocabularis no depenen del format. Així, FOAF es pot usar amb RDF, RDFa, HTML5 Microdata o JSON-LD. El vocabulari sí que depèn de l'àrea d'interès a la qual representa. No obstant això, certs dominis usuals com ara persones o llibres es poden descriure amb classes i propietats de més d'un vocabulari.

1.1. RDF Schema

L'**RDFS** (RDF *Vocabulary Description Language*, originalment RDF *Schema Language*) és un llenguatge basat en RDF per crear ontologies que defineixen dominis de coneixement i les relacions entre ells.

Conjunt de recursos

http://www.w3.org/2000/01/rdfschema# i el prefix associat rdfs.

RDFS és una extensió del vocabulari RDF amb elements bàsics de l'ontologia i reutilitza propietats d'RDF. Les ontologies RDFS es representen com a grafs RDF. Així, RDFS és adequat per descriure diversos tipus de recursos mitjançant propietats específiques. Les **classes** i les **propietats** són el vocabulari RDFS, que inclou un conjunt de recursos predefinits juntament amb el seu significat.

Les **classes** del vocabulari RDFS s'usen per definir:

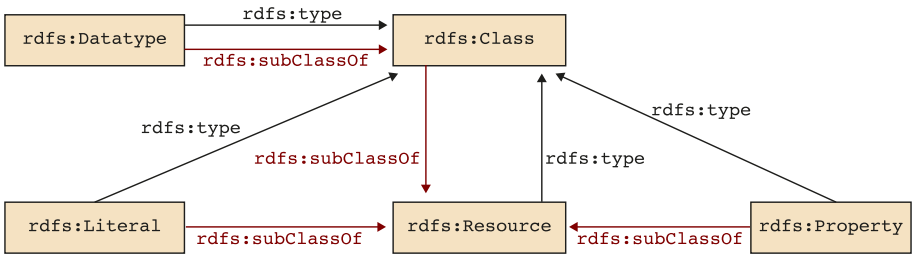
- Recursos (`rdfs:Resource`)
- Literals com ara sencers o cadenes (`rdfs:Literal`)
- Classes (`rdfs:Class`)
- Tipus de dades RDF (`rdfs:Datatype`)
- Contenidors (`rdfs:Container`)
- Propietats de membres d'un contenidor (`rdfs:ContainerMembershipProperty`)

Les **propietats** RDFS poden expressar:

- El subjecte és una subclasse d'una classe (`rdfs:subClassOf`).
- El subjecte és una subpropietat d'una propietat (`rdfs:subPropertyOf`).
- Definir un domini (`rdfs:domain`).
- Rang d'una propietat (`rdfs:range`).
- Afegir un nom llegible per persones (`rdfs:label`).
- Descripció d'un recurs (`rdfs:comment`).
- Identificar un membre d'un recurs (`rdfs:member`).
- Afegir informació relativa al recurs (`rdfs:seeAlso`).

- Proporcionar la definició del recurs (`rdfs:isDefinedBy`).

Figura 3. Relacions entre les classes definides en el llenguatge RDFS



1.1.1. Definint classes i subclasses

Una classe RDFS correspon a un tipus o categoria utilitzada per a una classificació o jerarquia. Amb RDFS, una classe C es defineix mitjançant una tripleta:

Figura 4. Definició de classe

```
C rdf:type rdfs:Class .
```

on `rdfs:Class` és una classe predefinida i `rdf:type` una propietat també predefinida.

De la mateixa manera, les subclasses també es defineixen mitjançant una tripleta:

Figura 5. Definició de subclasse

```
SC rdfs:subClassOf C .
```

La propietat `rdfs:subClassOf` és reflexiva, és a dir, una vegada que una classe RDFS s'ha creat, és subclasse de si mateixa. La propietat `rdfs:subclassOf` s'usa per declarar que una classe és una especialització d'una altra classe més general.

Per definir una instància d'una classe C també s'utilitza una tripleta:

Figura 6. Definició d'instància

```
I rdf:type C .
```

Aquí, la propietat `rdf:type` s'usa per declarar que l'individu I és una instància de la classe C.

1.1.2. Definint propietats, dominis i rangs

Una propietat amb RDF es pot utilitzar d'una de les formes següents:

1) Per indicar el valor d'un atribut per a un recurs.
2) Per establir una relació entre dos recursos.

Per indicar que una propietat és de tipus `rdf:Property` utilitzem `rdf:type`.

Figura 7. Definició d'una propietat

```
:p rdf:type rdf:Property .
```

En moltes aplicacions és útil restringir els recursos als quals una propietat es pot aplicar. RDFS té dos components per gestionar el domini (`rdfs:domain`) i el rang (`rdfs:range`) d'una propietat.

Una restricció de **domini** permet classificar els subjectes d'una relació.

Figura 8. Definició de propietat amb domini

```
:p rdf:type rdf:Property .
:p rdfs:domain :C .
```

Una restricció de **rang** permet classificar els objectes d'una relació.

Figura 9. Definició de propietat amb rang

```
:p rdf:type rdf:Property .
:p rdfs:range :D .
```

D'aquesta forma es poden afegir restriccions sobre el subjecte i sobre l'objecte d'una propietat.

1.1.3. Regles d'inferència amb RDFS

El mecanisme de raonament amb RDFS permet inferir noves tripletes a partir de les existents en un graf RDF. Aquest mecanisme utilitza el significat del vocabulari RDFS: `rdfs:Class`, `rdf:type`, `rdf:Property`, `rdfs:domain`, `rdfs:range` o `rdfs:subClassOf`.

Aquest mecanisme es tradueix en una sèrie de regles d'inferència. A partir de les tripletes següents:

Figura 10. Antecedent

```
:a  rdf:type :C .
:C  rdfs:subClassOf :D.
```

s'infereix la tripleta següent (que es denomina una jerarquia de classes):

Figura 11. Conseqüent

```
:a rdf:type :D .
```

Amb RDFS és possible expressar diferents regles d'inferència que es resumeixen en la taula 1.

Taula 1. Regles d'inferència RDFS

Regla d'inferència	Premissa	Inferència
Jerarquia de classe	`:a rdf:type :C .` `:C rdfs:subClassOf :D.`	`:a rdf:type :D .`
Jerarquia de propietat	`:a :p :b .` `:p rdfs:subPropertyOf :q .`	`:a :q :b .`
Jerarquies de classes	`:A rdfs:subClassOf :B .` `:B rdfs:subClassOf :C .`	`:A rdfs:subClassOf :C .`
Jerarquies de propietats	`:p rdfs:subPropertyOf :q .` `:q rdfs:subPropertyOf :r .`	`:p rdfs:subPropertyOf :r .`
Domini d'una propietat	`:a :p :b .` `:p rdfs:domain :C .`	`:a rdf:type :C .`
Rang d'una propietat	`:a :p :b .` `:p rdfs:range :D .`	`:b rdf:type :D .`

1.2. Exemples de vocabularis

En aquest context cal establir un conjunt de vocabularis de referència, com a manera de facilitar la comunicació de les dades.

1.2.1. El vocabulari schema.org

Amb aproximadament 300 definicions de conceptes, **schema.org** és una de les col·leccions més utilitzades com a esquema d'etiquetatge per a dades estructurades. Schema.org va ser llançat per Google, Yahoo i Bing el 2011. Schema.org conté les definicions dels conceptes més utilitzats, de manera que pot fer anotacions d'accions, treballs creatius, esdeveniments, serveis, conceptes mèdics, organitzacions, persones, llocs o productes.

Classes i propietats

Els tipus (classes) més freqüents d'schema.org i les seves propietats es poden trobar a http://schema.org/docs/gs.html#schemaorg_types, i la llista completa de propietats està a http://schema.org/docs/full.html.

De forma anàloga a l'exemple anterior, si volem descriure un llibre necessitaríem un vocabulari per definir els llibres i les seves propietats típiques. Si volguéssim descriure el títol del llibre, es podria utilitzar la propietat «Name» d'schema.org o utilitzar la propietat «Title» del vocabulari Dublin Core (DC), usant dues declaracions d'espais de noms.

Figura 12. Descripció d'un llibre

```
…
xmlns:schema="http://schema.org/"
xmlns:dc="http://purl.org/dc/terms/"
…
<schema:Book>
  <dc:title>El ingenioso hidalgo Don Quixote de la Mancha</dc:title>
</schema:Book>
```

En aquest cas, `schema:Book` és una abreviatura d'http://schema.org/book, que és una definició de la classe «Book», i `dc:title` refereix a http://purl.org/dc/terms/title, que és una definició de la propietat «títol».

1.2.2. Vocabularis de persones

Les característiques d'una persona i les relacions entre les persones es poden descriure mitjançant una varietat de vocabularis controlats, com es resumeix en la taula 2.

Taula 2. Vocabularis de persona

Vocabulari	Abreviatura	Espai de noms	Ús
Person de schema.org	schema:Person	https://schema.org/person	Nom, cognoms, gènere, nacionalitat, professió, prefix (o sufix) honorífic, etc.
Friend Of A Friend	foaf	http://xmlns.com/foaf/0.1/	Persona, nom, gènere, coneguts, etc.
vcard	vcard	http://www.w3.org/2006/vcard/ns#	Targetes personals electròniques (*electronic business cards*).
Contact	contact	http://www.w3.org/2000/10/swap/pim/contact	Contactes personals, llengua materna, etc.
Bio	bio	http://purl.org/vocab/bio/0.1/	Informació biogràfica.
Relationship	relationship	http://purl.org/vocab/relationship	Relacions entre persones (friendOf, parentOf, etc.)

1.2.3. Vocabularis de llibres i bibliografia

Els llibres es poden descriure amb força precisió utilitzant les propietats del vocabulari http://schema.org/book, que defineix el format, el nombre de pàgines, el titular dels drets d'autor, el gènere literari i altres característiques dels llibres. Dublin Core també s'usa moltes vegades per definir metadades sobre llibres.

L'ISBN[1] és un nombre internacional per dotar a cada llibre d'un codi numèric que l'identifiqui. També és una propietat de la classe «Book» d'schema.org (definida a http://schema.org/isbn).

(1)Acrònim de les seves sigles en anglès, *International Standard Book Number*.

Els llibres que ja s'han llegit o llibres favorits es poden descriure utilitzant l'esquema de llista de lectura per mitjà de l'espai de noms http://purl.org/net/schemas/book/.

PRISM (*Publishing Requirements for Industry Standard Metadata*) descriu molts components sobre contingut imprès, en línia, mòbil o multimèdia, inclosos els següents:

- Creador, col·laborador, propietari dels drets d'autor.
- Llocs, organitzacions, temes, persones, esdeveniments, condicions de reproducció.

- Data de publicació, incloent data d'edició, data de publicació, volum, nombre, etc.

- Restriccions de publicació i reutilització.

PRISM s'usa per descriure l'agregació de continguts, la reutilització de continguts, el descobriment de recursos, la captura d'informació sobre l'ús de drets i anotacions a la web. L'espai de noms per PRISM 2.1 Basic porta el prefix prism i per PRISM 3.0 el prefix és prism-ad.

1.2.4. Vocabularis per a comunitats

SIOC (*Semantically-Interlinked Online Communities*) permet representar informació sobre xarxes socials i comunitats en línia. Aquest vocabulari també ofereix URI per indicar que un recurs és d'un determinat tipus en el context de xarxes socials (un fòrum, un post, un lloc web, etc.). Per exemple, amb el predicat «Name» s'indica un nom d'usuari d'una persona, o `member_of` per indicar si un usuari és membre d'un grup o aplicació.

SIOC

L'espai de noms de SIOC Core és http://rdfs.org/sioc/ns#. La seva especificació està a http://rdfs.org/sioc/spec/.

Facebook utilitza el vocabulari d'OpenGraph (og) que consisteix bàsicament en una plataforma centrada en aspectes socials dels usuaris. El funcionament d'**OpenGraph** permet que les accions específiques realitzades pels usuaris siguin publicades i compartides per mitjà de la seva cronologia, notícies o barra lateral.

OpenGraph

L'espai de noms d'OpenGraph és http://ogp.me/ns#.

2. Taxonomies

Les taxonomies proporcionen els termes o categories per mitjà dels quals es pot descriure una entitat. L'objectiu de la taxonomia és determinar quina classe és la més adequada per a una entitat donada. És similar al treball d'un bibliotecari que ha de classificar els nous llibres segons una taxonomia.

Una **taxonomia** és un vocabulari controlat organitzat en forma jeràrquica.

No obstant això, abans que un bibliotecari pugui realitzar aquesta tasca hi ha d'haver una taxonomia. Melvil Dewey va ser bibliotecari a l'Amherst Library i va desenvolupar el Sistema de Classificació Decimal Dewey en la dècada de 1870. Dewey va idear un sistema de classificació que assignava els temes en un rang entre 0 i 1.000. Una de les principals aspiracions en un sistema de classificació és minimitzar l'ambigüitat. Les jerarquies són eines de classificació útils perquè redueixen l'ambigüitat.

El problema d'aquests sistemes és que una mateixa entitat tingui més d'una categoria i que aquestes categoritzacions estiguin en branques diferents.

Finalment, un **tesaurus** és una taxonomia amb informació addicional sobre cada terme que inclou termes preferits i alternatius. Els termes guarden entre si relacions semàntiques i genèriques: d'equivalència, jeràrquiques i associatives.

2.1. SKOS

SKOS (*Simple Knowledge Organization System*) és una recomanació del W3C per representar esquemes de classificació com a vocabularis controlats, taxonomies i tesaurus. Molts d'aquests sistemes comparteixen una estructura similar i s'utilitzen en aplicacions semblants.

SKOS permet compartir dades entre les aplicacions. Amb SKOS els conceptes s'identifiquen mitjançant URI, amb etiquetes en un o més idiomes i documentats amb diferents tipus de notes. Els conceptes es poden relacionar semànticament entre si, en jerarquies i xarxes d'associació informals, i també se'ls pot agrupar en esquemes conceptuals.

A continuació es presenten els principals elements d'SKOS:

- `skos:Concept`. Constitueix l'element fonamental d'SKOS. És una classe que defineix que un determinat recurs és un concepte.

- `skos:prefLabel`, `skos:altLabel`. Són etiquetes que permeten fer referència a conceptes en llenguatge natural. Aquí `prefLabel` és l'etiqueta per a la definició principal i `altLabel` és una definició alternativa. Per exemple, per a sinònims.

- `skos:broader`, `skos:narrower`. El significat d'un concepte està vinculat amb altres conceptes del vocabulari. Així, `broader` indica que un concepte és més ampli que un altre (un concepte engloba o inclou un altre concepte) i `narrower` indica el contrari, que un concepte és més concret que un altre.

- `skos:related`. Expressa una relació associativa entre dos conceptes.

- `skos:note`. Es refereix a una observació genèrica respecte al concepte. Hi ha la possibilitat de qualificar diferents tipus d'observacions utilitzant `skos:scopeNote`, `skos:historyNote`, `skos:editorialNote` o `skos:changeNote` relatives a l'àmbit, històric, qüestions editorials o canvis realitzats, respectivament.

- `skos:definition`. Indica una definició del concepte.

- `skos:ConceptScheme`. Els conceptes es poden utilitzar com a entitats independents, encara que normalment estan organitzats com a esquemes de classificació o tesaurus. Aquests esquemes es poden representar amb aquesta classe.

El model de dades d'SKOS està basat en RDF i estructura les dades en forma de tripletes que es poden representar en qualsevol format vàlid d'RDF.

Figura 13. Taxonomia amb SKOS

```
@prefix ex: <.> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix skos: <http://www.w3.org/2008/05/skos#> .

ex:Planet   rdf:type        skos:Concept ;
            skos:prefLabel  "Planeta"@ca ;
            skos:related    ex:Satellite .

ex:Dwarf_Planet rdf:type        skos:Concept ;
                skos:prefLabel  "Planeta Nan"@ca ;
                skos:narrower ex:Planet .

ex:Satellite rdf:type       skos:Concept ;
             skos:prefLabel     "Satèl·lit"@ca ;
             skos:altLabel      "Lluna"@ca;
             skos:related       ex:Planet .
```

Bibliografia

DuCharme, R. (2013). *Learning SPARQL* (2a. ed.). O'Reilly Media.

Guarino, N.; Oberle, D.; Staab, S. (2009). *What Is an Ontology?* [en línia]. [Data de consulta: gener 2020]. Disponible a: <https://iaoa.org/isc2012/docs/Guarino2009_What_is_an_Ontology.pdf>

Kumar, A. (2018). *Architecting Data-Intensive Applications*. Packt Pub.

Noy, N. F.; McGuinness, D. L. (2001). *Ontology development 101: A guide to creating your first ontology*. Stanford knowledge systems laboratory technical report (Informe SMI-2001-0880).

Powers, S. (2003). *Practical RDF*. O'Reilly Media.