Dades enllaçades

PID_00271438

Blas Torregrosa García

Temps mínim de dedicació recomanat: 2 hores

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

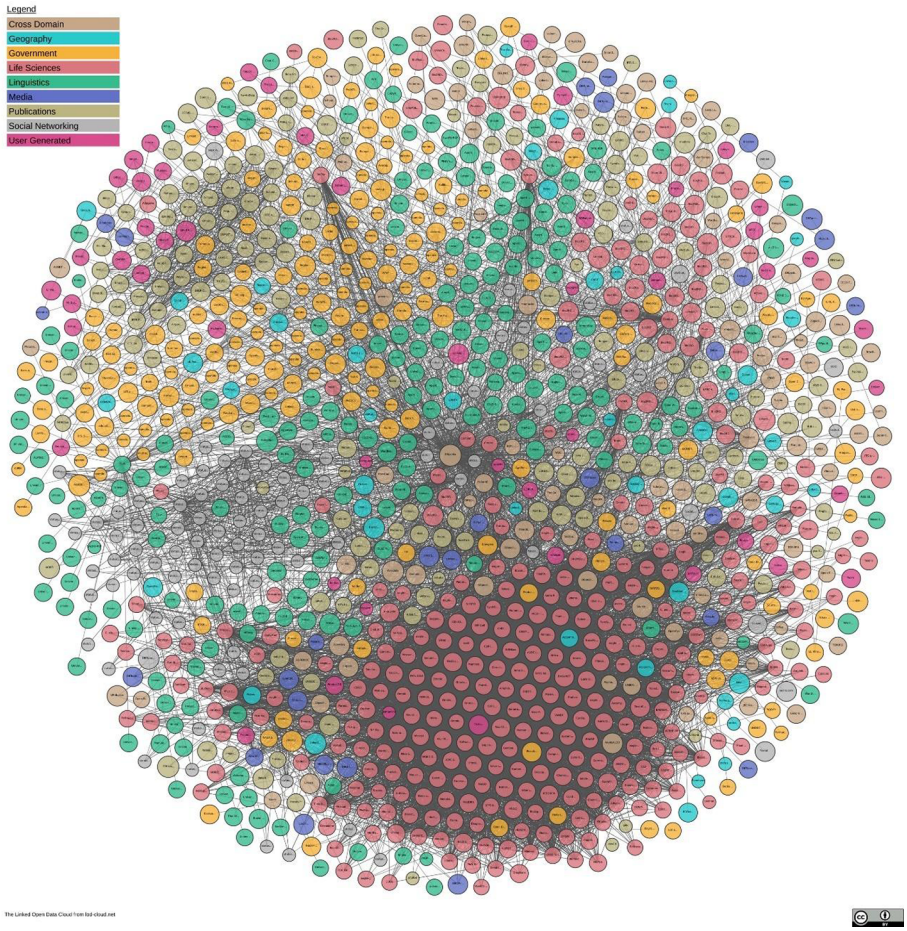
En aquest mòdul proporcionarem una visió general sobre què són les dades enllaçades (*linked data*), tractarem els cinc nivells de dades enllaçades segons Tim Berners-Lee i els beneficis que aporten, com publicar dades enllaçades a partir de bases de dades relacionals i presentarem exemples de conjunts de dades enllaçades.

1. Dades enllaçades

La web està evolucionant des d'una web de documents enllaçats cap a una web de dades mitjançant l'ús de tecnologies que afegeixen semàntica, de manera que es facilita el tractament de les dades per part de les aplicacions. L'horitzó que es pretén aconseguir és el d'una base de dades global, amb una gran quantitat d'aplicacions que puguin accedir a un conjunt creixent de dades.

Les dades es publiquen a la web per mitjà de diferents organitzacions i estan emmagatzemades en diverses localitzacions i en diferents formats. Per facilitar la construcció de la web de dades, cal establir una manera estàndard de connexió entre aquestes. En els apartats següents es descriu el que es coneix com a **dades enllaçades** (*linked -open- data o LOP*) i l'estat actual es pot visualitzar en la figura 1.

Figura 1. Núvol de dades obertes i enllaçades



Font: lod-cloud.net/

1.1. Què són les dades enllaçades?

Les **dades enllaçades** proporcionen un mètode per interconnectar les dades de diferents fonts de dades.

Les dades enllaçades es basen en tecnologies web estàndard, com ara HTTP, RDF i URI. Aquestes tecnologies, a excepció d'RDF, s'han utilitzat a la web de documents. En aquest cas es pretén utilitzar aquestes tecnologies perquè també els programes informàtics puguin accedir, enllaçar i interpretar les dades. Això permet que dades de diferents fonts puguin ser connectades, consultades i analitzades.

Per a això, cal possibilitar la **interoperabilitat** entre els sistemes que gestionen les dades. Es diu que dos sistemes són interoperables quan estan en condicions d'intercanviar amb èxit informació entre ells. Hi ha diferents enfocaments per aconseguir la interoperabilitat:

1) Mapatge (*mapping*) entre els conceptes de cada font. Hi ha d'haver una descripció d'informació global que tradueixi els conceptes d'una font a una altra.

2) Intermediació, que insereix entre cada font de dades una capa intermèdia que tradueix les dades. Aquesta capa pot ser un programari addicional, un conjunt de regles, una ontologia, un agent de programari, etc.

3) Basada en consultes, estratègia en què es plantegen consultes que s'avaluaran en cada font de dades.

Aquests enfocaments no són mútuament excloents. Per exemple, un sistema pot incloure intermediaris mentre que també té una descripció global d'informació.

La integració de la informació és un terme que sovint es confon amb interoperabilitat. Per aconseguir la integració no és suficient amb el compliment dels estàndards que permetin la comunicació.

La **integració** és el procés segons el qual la informació que s'origina en diverses fonts i sistemes es combina per permetre el seu processament conjunt.

Les dificultats d'integració de fonts heterogènies és conseqüència de les nombroses formes en què les dades es poden emmagatzemar, organitzar o comunicar. En particular, el problema d'heterogeneïtat pot incloure algunes de les qüestions següents:

- **Diferents models de dades**. Per exemple, hi pot haver dades en bases de dades relacionals, arxius XML o bases de dades NoSQL.

- **Diferències de vocabulari**. En un sistema la propietat «temps» pot aparèixer en un altre com a «durada».

- **Desajust sintàctic**. Les mateixes dades poden aparèixer en un arxiu XML com a /dades/descripció i en un altre com a /dades/@descripció, o com a /descripció/dades en un tercer.

- **Desajust semàntic**. En el model RDF hi ha el concepte de classe/subclasse que no existeix en el model relacional.

1.2. Els quatre principis

Les dades enllaçades es basen en quatre **principis bàsics** enunciats per Tim Berners-Lee:

1) Identificació: utilitzar URI per identificar les coses (recursos). Segons aquest principi, els elements que es desitja compartir hauran de tenir una adreça web (URI) que els identifiqui.

> Per exemple, per identificar l'astronauta Neil Armstrong es pot utilitzar l'URI http://dbpedia.org/resource/neil_armstrong, que l'identifica en el conjunt de dades de la DBpedia.

2) Consulta: utilitzar HTTP URI, de manera que possibiliti buscar a la web i saber més de la semàntica d'aquests recursos.

> Per fer-ho, en el cas de l'exemple anterior, simplement cal navegar pel recurs http://dbpedia.org/resource/neil_armstrong que ens redirigeix a la pàgina web del recurs: http://dbpedia.org/page/neil_armstrong.

3) Descripció: proporcionar informació útil sobre l'URI utilitzant estàndards web (RDF i SPARQL).

> En el cas de Neil Armstrong, per exemple, s'inclou informació sobre la seva vida i el fet que va ser el primer ésser humà a trepitjar la Lluna. Aquesta informació haurà d'estar representada mitjançant RDF.

4) Enllaç: incloure enllaços a altres URI perquè es pugui navegar per les dades i descobrir informació relacionada.

> En el cas de l'exemple, entre d'altres, s'enllaçaria el recurs amb els seus companys de missió Buzz Aldrin (http://dbpedia.org/resource/buzz_aldrin) i Michael Collins (http://dbpedia.org/resource/michael_collins_(astronaut)), amb la missió Apol·lo 11 (http://

dbpedia.org/resource/apollo_11) i amb la resta de les missions del programa Apol·lo (http://dbpedia.org/resource/apollo_program).

1.3. El model de cinc estrelles

La web de dades és un espai heterogeni, en què es publiquen diferents tipus de dades en els més diversos formats i estructures. Com a manera d'orientar la publicació de dades obertes a la web i d'acord amb aquesta nova visió semàntica, Tim Berners-Lee va suggerir un esquema de desenvolupament de cinc estrelles. Segons aquest esquema, les dades obertes es poden convertir en dades enllaçades si s'interrelacionen entre si. Les dades enllaçades són la base tècnica per crear una web de dades en què les dades estarien connectades d'acord amb els quatre principis anteriors.

Figura 2. Model de cinc estrelles

Font: www.w3.org/designissues/linkeddata.html

A continuació veurem aquesta escala composta per cinc nivells:

1) Nivell «1 estrella». **Publicar les dades a la web** (en qualsevol format).

Aconseguir una estrella significa que els usuaris poden:

- accedir a les dades,
- consumir les dades,
- emmagatzemar-les localment,
- manipular les dades,
- compartir les dades.

Mentre que per a l'editor resulta fàcil i còmode publicar les dades.

2) Nivell «2 estrelles». **Publicar les dades com a dades estructurades**.

Quan s'aconsegueixen dues estrelles, els usuaris poden:

- processar les dades,
- agregar (resumir) les dades,
- realitzar càlculs,
- visualitzar les dades,
- exportar-les a un altre format (estructurat).

Mentre que per a l'editor de dades encara resulta senzill publicar les dades.

3) Nivell «3 estrelles». **Utilitzar formats no propietaris**.

Aconseguir tres estrelles ajuda els usuaris de dades a fer tot el que es pot fer amb el nivell anterior i, a més, es podrien manipular les dades sense cap programari propietari. De la mateixa manera, com a editor de dades, pot ser necessari un convertidor per exportar les dades des del format propietari.

4) Nivell «4 estrelles». **Usar estàndards oberts** (URI, RDF i SPARQL) per identificar qualsevol cosa a la web.

Aconseguir quatre estrelles permet als usuaris de dades realitzar tot l'anterior i a més:

- les dades es poden vincular usant URI,
- es pot accedir parcialment a les dades,
- es poden reutilitzar les eines i llibreries existents.

D'altra banda, per a l'editor de dades el model RDF pot ser més exigent que altres models de dades (tabular en Excel o CSV, o arbres en XML o JSON), encara que les dades també es poden combinar de forma segura amb altres dades.

5) Nivell «5 estrelles». **Enllaçar les dades a altres dades per proporcionar context**.

Aconseguir cinc estrelles permet que els usuaris facin tot l'anterior i a més:

- descobrir noves dades mentre se'n consumeixen altres,
- tractar amb els URI no trobats,
- vincular dades amb temes relacionats és una qüestió de confiança.

D'altra banda, com a editor de dades cal fer que sigui possible descobrir les dades, la qual cosa augmenta el seu valor. Cal invertir recursos per vincular les dades a altres existents a la web.

Des de la primera estrella fins a l'última, l'obertura de les dades enllaçades està recolzada per una millor estructuració, una millor interoperabilitat i, per tant, una millor reutilització com un recurs de dades a la web.

1.4. API de dades enllaçades

Un dels quatre principis de les dades enllaçades estableix que s'ha d'usar un URI per accedir a un recurs, i que s'han d'obtenir dades rellevants sobre el recurs identificat.

La idea de l'API de dades enllaçades (LD API) és oferir una manera senzilla d'accedir a les dades enllaçades via web, permetent que els conjunts de recursos estiguin exposats com a URI i facilitant la seva consulta.

A més, també es possibilita, mitjançant paràmetres de consulta, filtrar, paginar i ordenar els resultats. L'API admet diversos formats de resultats, incloent JSON, XML, RDF/XML i Turtle.

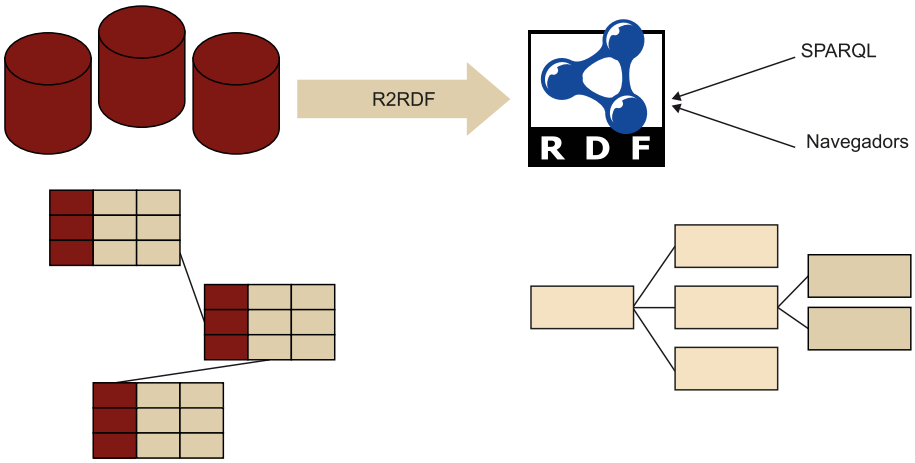
Per desenvolupar programari amb dades enllaçades cal comprendre el model de dades RDF (amb les serialitzacions associades) i el llenguatge de consulta SPARQL que han demostrat ser una barrera per a l'adopció de dades enllaçades.

Una API de dades enllaçades és una especificació de codi obert i hi ha algunes solucions programari que la implementen.

1.5. Publicació de bases de dades relacionals com a dades enllaçades

La majoria del contingut dinàmic dels llocs web prové de bases de dades relacionals, com ara MS SQL, mySQL, Oracle o PostgreSQL. Publicar les dades com a RDF fa d'aquestes dades més accessibles a la web.

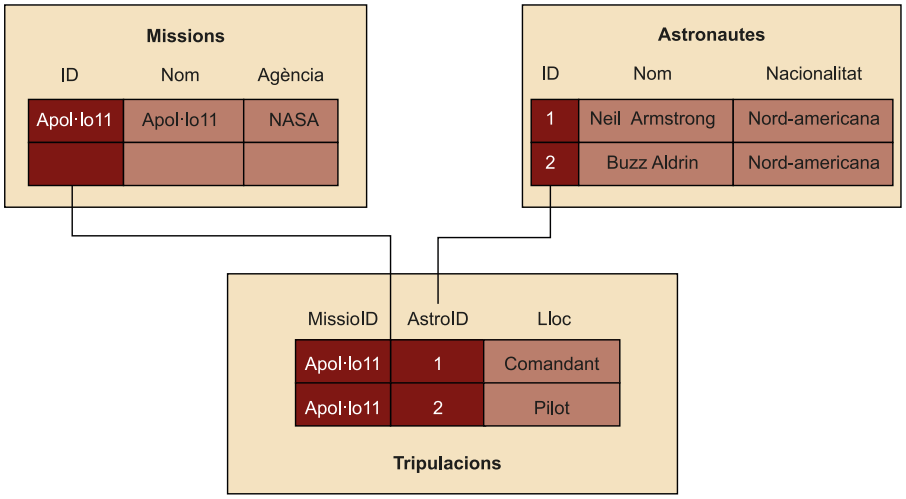
Figura 3. Publicació de bases de dades relacionals amb RDF



Suposant que la informació que la base de dades emmagatzema estigui en taules, per realitzar la traducció des de la base de dades relacional cal generar un recurs RDF per cadascuna de les files de les taules.

Suposem que tenim tres taules en una base de dades relacional, una per recopilar missions espacials, una altra per als astronautes i una tercera que les enllaça com a tripulació. En totes les taules hi ha un identificador únic que és la clau primària de cada taula.

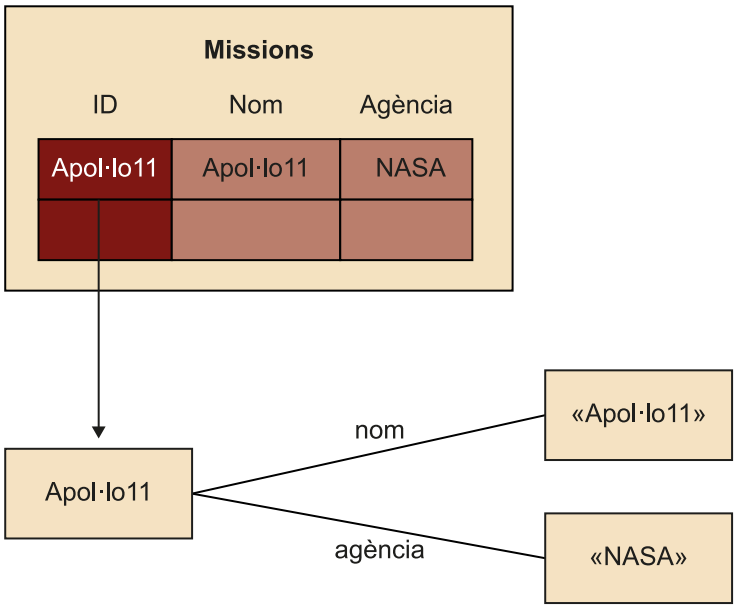
Figura 4. Taules d'exemple



Suposem que es comença per la taula «Missions». En primer lloc, es genera un nou recurs de dades RDF a partir de cada fila de la taula, utilitzant la columna ID per generar un nou concepte amb RDF, l'identificador del qual és un URI associat.

A continuació, es consideren els elements restants de la taula. D'aquesta forma la columna «Nom» genera un nou literal i així successivament amb totes les columnes. I una vegada tinguem totes les columnes, cal generar les relacions entre els recursos.

Figura 5. Transformació de la fila en un recurs RDF



El **mapatge** (*mapping*) és una relació entre una entitat d'una base de dades relacional i un concepte en un graf RDF.

Hi ha dues formes de generar mapatges:

1) Mapatge directe (*direct mapping*). Proporciona una traducció automàtica de la base de dades relacional a RDF. Per fer la traducció cal generar un URI per a cada taula, per a cada atribut de la taula i per a cada clau primària de cada taula. Per indicar a quina taula pertany cada fila es genera una tripleta. I per cada atribut de cada taula es genera una tripleta.

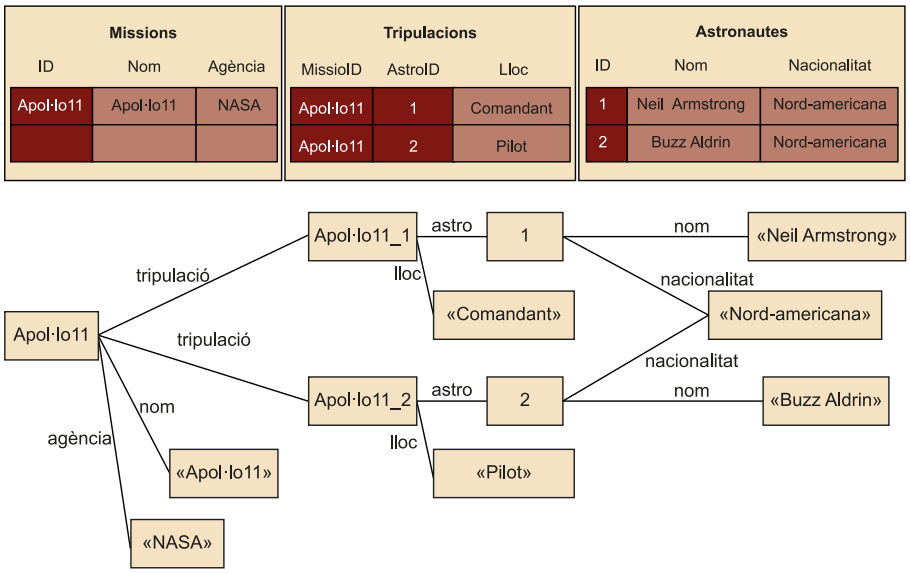
2) R2RML (RDB to RDF *Mapping Language*). Proporciona un llenguatge de mapatge per descriure una traducció de la base de dades relacional a RDF. Mitjançant l'ús d'R2RML és possible descriure com es realitzarà la traducció mitjançant regles que detallen la representació de les dades i inclouen vocabulari RDFS i OWL.

El sistema de traducció que executa els mapatges i que genera les noves dades en RDF es denomina *ETL*.[1] El procés d'ETL consta de tres fases:

[(1)]Acrònim de l'anglès, *Extract Transform Load*.

1) Fase d'extracció (*extract*): es llegeixen les dades de la base de dades relacional.
2) Fase de transformació (*transform*): es transformen a RDF utilitzant el sistema de traducció.
3) Fase de càrrega (*load*): s'emmagatzemen les tripletes RDF.

Figura 6. Exemple de transformació de taules en graf RDF



2. Exemples de dades enllaçades

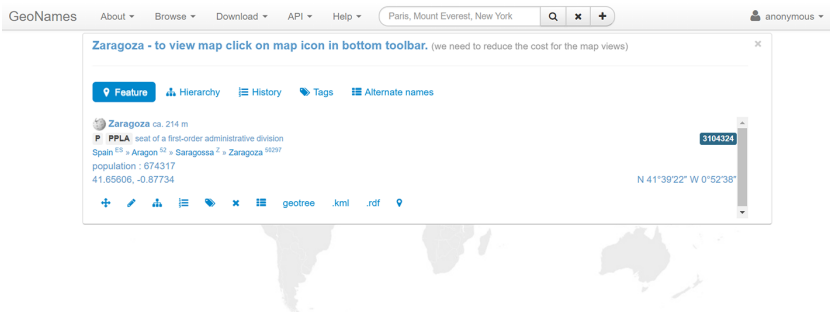
2.1. GeoNames

GeoNames és una base de dades que conté informació geogràfica sobre més de 10 milions de llocs de tot el món. Per a cada lloc es descriu el seu nom (en diferents idiomes), la seva ubicació (latitud, longitud i elevació), la seva població, el seu codi postal i la seva categoria. La categoria permet indicar el tipus d'element que s'està descrivint (zona, carretera, límit administratiu, etc.). La ubicació es representa mitjançant el vocabulari Basic Geo (WGS84), que permet representar la geoposició dels elements definits utilitzant RDF.

Els URI de GeoNames tenen la forma següent «https://sws.geonames.org/codigo», en què el codi identifica l'element que descriurà.

Per exemple, en el cas de la ciutat de Saragossa, el seu URI és http://sws.geonames.org/3104324, que redirigeix a la versió HTML del recurs https://www.geonames.org/3104324/zaragoza.html. Amb aquest URI es poden obtenir les dades RDF o presentar el mapa.

Figura 7. Dades de Saragossa



Font: www.geonames.org/3104324/zaragoza.html

2.2. La Biblioteca Nacional d'Espanya

El projecte de la Biblioteca Nacional d'Espanya i del Grup d'Enginyeria Ontològica de la Universitat Politècnica de Madrid té com a objectiu l'exploració de dades bibliogràfiques de manera diferent als catàlegs tradicionals, oferint una altra forma de navegació pels diferents recursos de la biblioteca i enriquint les seves pròpies dades amb altres dades externes.

Es pot accedir a les dades a partir del propi portal o des d'una interfície SPARQL. A més, també es poden obtenir abocaments complets de les dades. La biblioteca utilitza la seva pròpia ontologia en què defineix les seves entitats, que són equivalents a les entitats descrites en l'ontologia FRBR.

Ontologia FRBR

El model FRBR va ser publicat l'any 1998 i descriu els registres bibliogràfics per mitjà de les entitats «Obra», «Expressió», «Manifestació» i «Ítem», a més d'autors i matèries. L'any 2010 es va publicar l'ontologia FRBR, que és la que s'ha utilitzat a les biblioteques d'Espanya.

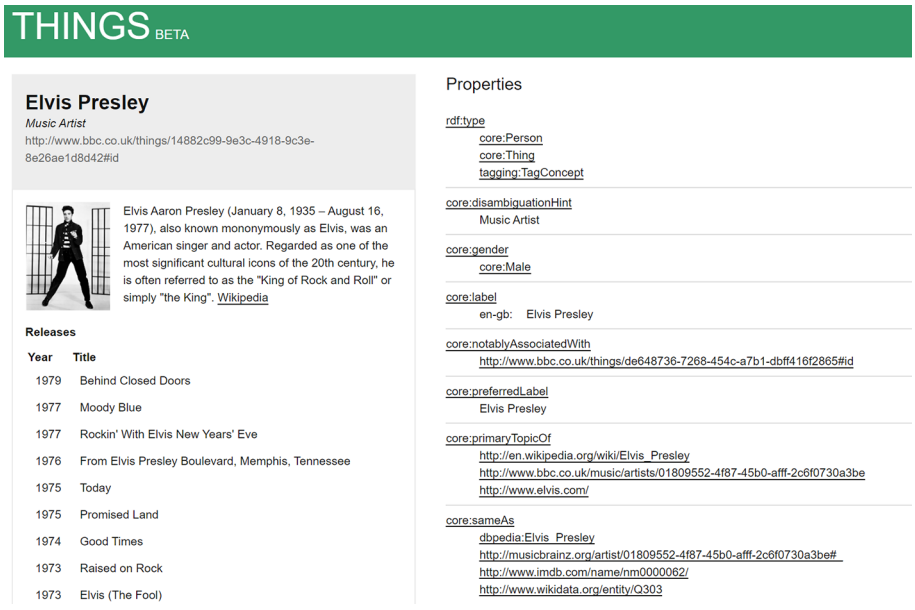
Figura 8. Portal de dades de la Biblioteca Nacional d'Espanya

Font: datos.bne.es/persona/xx1718747.html

2.3. BBC Things

BBC Things utilitza tecnologies de dades enllaçades que permeten accedir a temes d'interès per al públic del grup BBC (British Broadcasting Corporation), com ara persones, llocs, organitzacions, competicions esportives, etc. La BBC posseeix un conjunt propi d'ontologies que defineixen els diferents temes que es poden consultar a BBC Things.

Figura 9. Dades sobre Elvis Presley a BBC Things (HTML)



Font: www.bbc.co.uk/things/14882c99-9e3c-4918-9c3e-8e26ae1d8d42

La figura 9 mostra la pàgina HTML resultant d'una consulta sobre el músic i actor Elvis Presley que inclou les dades en tripletes associades al recurs.

Bibliografia

Allemang, D.; Hendler, J. (2011). *Semantic Web for the working ontologist* (2a. ed.). Morgan Kaufmann.

Antoniou, G.; Van Harmelen, F. (2004). *A Semantic Web Primer*. Cambridge: MIT Press.

Heath, T.; Bizer, C. (2011). «Linked Data: Evolving the Web into a Global Data Space (1a. ed.)». *Synthesis Lectures on the Semantic Web: Theory and Technology* (vol. 1, núm. 1, pàg. 1-136). Morgan & Claypool.