Datos enlazados

PID_00271447

Blas Torregrosa García

Tiempo mínimo de dedicación recomendado: 2 horas

Universitat Oberta de Catalunya

Blas Torregrosa García

Ingeniero en Informática y máster universitario en Seguridad de las Tecnologías de la Información y de las Comunicaciones (MISTIC) por la Universitat Oberta de Catalunya (UOC). Especializado en ciberseguridad. Profesor colaborador en el máster de Ciencia de Datos de la UOC y profesor asociado en la Universidad de Valladolid (UVA).

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por el profesor: Ferran Prados Carrasco (2020)

Primera edición: febrero 2020

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

Índice

Introducción

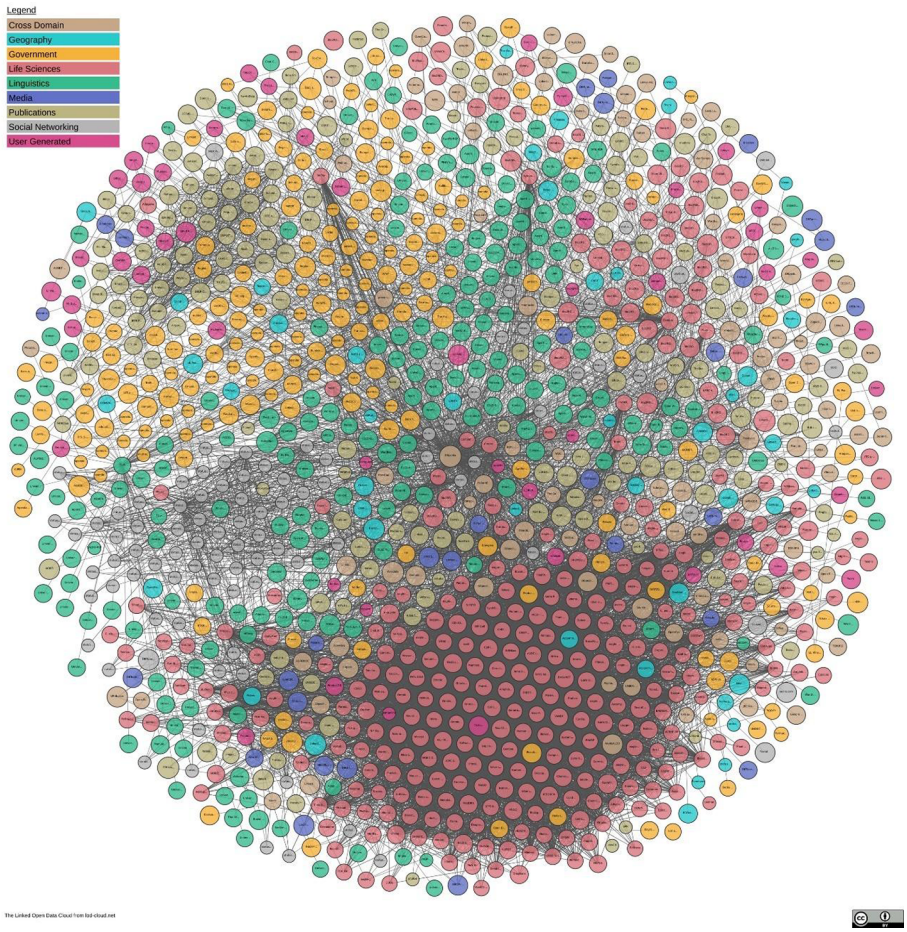
En este módulo proporcionaremos una visión general sobre qué son los datos enlazados (*linked data*), trataremos los cinco niveles de datos enlazados según Tim Berners-Lee, y los beneficios que aportan, cómo publicar datos enlazados a partir de bases de datos relacionales y presentaremos ejemplos de conjuntos de datos enlazados.

1. Datos enlazados

La web está evolucionando desde una web de documentos enlazados hacia una web de datos mediante el uso de tecnologías que añaden semántica, de forma que se facilita el tratamiento de los datos por parte de las aplicaciones. El horizonte que se pretende alcanzar es el de una base de datos global, con una gran cantidad de aplicaciones que puedan acceder a un conjunto creciente de datos.

Los datos se publican en la web por diferentes organizaciones y están almacenados en diversas localizaciones y en diferentes formatos. Para facilitar la construcción de la web de datos es necesario establecer una manera estándar de conexión entre estos. En los apartados siguientes se describe lo que se conoce como **datos enlazados** (*linked -open- data o LOP*) y estado actual se puede visualizar en la figura 1.

Figura 1. Nube de datos abiertos y enlazados



Fuente: lod-cloud.net/

1.1. ¿Qué son datos enlazados?

Los **datos enlazados** proporcionan un método para interconectar datos de distintas fuentes de datos.

Los datos enlazados se basan en tecnologías web estándar, como HTTP, RDF y URI. Estas tecnologías, a excepción de RDF, se han utilizado en la web de documentos. En este caso se pretende utilizar estas tecnologías para que también los programas informáticos puedan acceder, enlazar e interpretar los datos. Esto permite que datos de distintas fuentes puedan ser conectados, consultados y analizados.

Para ello es necesario posibilitar la **interoperabilidad** entre los sistemas que gestionan los datos. Se dice que dos sistemas son interoperables cuando están en condiciones de intercambiar con éxito información entre ellos. Existen diferentes enfoques para lograr interoperabilidad:

1) Mapeo (*mapping*) entre los conceptos de cada fuente. Debe existir una descripción de información global que traduzca los conceptos de una fuente a otra.

2) Intermediación, que inserta entre cada fuente de datos una capa intermedia que traduce los datos. Esta capa puede ser un software adicional, un conjunto de reglas, una ontología, un agente de software, etc.

3) Basado en consultas, estrategia en la que se plantean consultas que se evaluarán en cada fuente de datos.

Estos enfoques no son mutuamente excluyentes. Por ejemplo, un sistema puede incluir intermediarios mientras que también tiene una descripción global de información.

La integración de la información es un término que a menudo se confunde con interoperabilidad. Para lograr la integración no es suficiente con el cumplimiento de los estándares que permitan la comunicación.

La **integración** es el proceso según el cual la información que se origina en varias fuentes y sistemas se combina para permitir su procesamiento conjunto.

Las dificultades de integración de fuentes heterogéneas es consecuencia de las numerosas formas en las que los datos se pueden almacenar, organizar o comunicar. En particular, el problema de heterogeneidad puede incluir algunas de las cuestiones siguientes:

- **Diferentes modelos de datos**. Por ejemplo, puede haber datos en bases de datos relacionales, archivos XML o bases de datos NoSQL.

- **Diferencias de vocabulario**. En un sistema la propiedad «tiempo» puede aparecer en otro como «duración».

- **Desajuste sintáctico**. Los mismos datos pueden aparecer en un archivo XML como /datos/descripción y en otro como /datos/@descripción, o como /descripción/datos en un tercero.

- **Desajuste semántico**. En el modelo RDF existe el concepto de clase/subclase que no existe en el modelo relacional.

1.2. Los cuatro principios

Los datos enlazados se basan en cuatro **principios básicos** enunciados por Tim Berners-Lee:

1) Identificación: utilizar URI para identificar las cosas (recursos). Según este principio, los elementos que se desea compartir deberán tener una dirección web (URI) que los identifique.

> Por ejemplo, para identificar al astronauta Neil Armstrong se puede utilizar el URI http://dbpedia.org/resource/Neil_Armstrong, que lo identifica en el conjunto de datos de la DBpedia.

2) Consulta: utilizar HTTP URI, de forma que posibilite buscar en la web y saber más de la semántica de esos recursos.

> Para hacerlo, en el caso del ejemplo anterior, simplemente hay que navegar al recurso http://dbpedia.org/resource/Neil_Armstrong que nos redirige a la página web del recurso: http://dbpedia.org/page/Neil_Armstrong.

3) Descripción: proporcionar información útil sobre el URI utilizando estándares web (RDF y SPARQL).

> En el caso de Neil Armstrong, por ejemplo, se incluye información sobre su vida y el hecho de que fue el primer ser humano en pisar la Luna. Esta información deberá estar representada mediante RDF.

4) Enlace: incluir enlaces a otros URI para que se pueda navegar por los datos y descubrir información relacionada.

> En el caso del ejemplo, entre otros, se enlazaría el recurso con sus compañeros de misión Buzz Aldrin (http://dbpedia.org/resource/Buzz_Aldrin) y Michael Collins (http://dbpedia.org/resource/Michael_Collins_(astronaut)), con la misión Apollo 11 (http://

dbpedia.org/resource/Apollo_11) y con el resto de las misiones del programa Apollo (http://dbpedia.org/resource/Apollo_program).

1.3. El modelo de cinco estrellas

La web de datos es un espacio heterogéneo, en el cual se publican diferentes tipos de datos en los más diversos formatos y estructuras. Como modo de orientar la publicación de datos abiertos en la web de acuerdo con esta nueva visión semántica, Tim Berners-Lee sugirió un esquema de desarrollo de 5 estrellas. Según este esquema, los datos abiertos pueden convertirse en datos enlazados si se interrelacionan entre sí. Los datos enlazados son la base técnica para crear una web de datos en la que los datos estarían conectados de acuerdo con los cuatro principios anteriores.

Figura 2. Modelo de cinco estrellas

Fuente: www.w3.org/DesignIssues/LinkedData.html

A continuación veremos esta escala compuesta por cinco niveles:

1) Nivel «1 estrella». **Publicar los datos en la web** (en cualquier formato).

Alcanzar una estrella significa que los usuarios pueden:

- acceder a los datos,
- consumir los datos,
- almacenarlos localmente,
- manipular los datos,
- compartir los datos.

Mientras que para el editor resulta fácil y cómodo publicar los datos.

2) Nivel «2 estrellas». **Publicar los datos como datos estructurados**.

Al lograr dos estrellas, los usuarios pueden:

- procesar los datos,
- agregar (resumir) los datos,
- realizar cálculos,
- visualizar los datos,
- exportarlos a otro formato (estructurado).

Mientras que para el editor de datos todavía resulta sencillo publicar los datos.

3) Nivel «3 estrellas». **Utilizar formatos no propietarios**.

Lograr tres estrellas ayuda a los usuarios de datos a hacer todo lo que se puede hacer con el nivel anterior y, además, se podrían manipular los datos sin ningún software propietario. Del mismo modo, como editor de datos, puede ser necesario un conversor para exportar los datos desde el formato propietario.

4) Nivel «4 estrellas». **Usar estándares abiertos** (URI, RDF y SPARQL) para identificar cualquier cosa en la web.

Lograr cuatro estrellas permite a los usuarios de datos realizar todo lo anterior y además:

- los datos se pueden vincular usando URI,
- se puede acceder parcialmente a los datos,
- se pueden reutilizar las herramientas y librerías existentes.

Por otro lado, para el editor de datos el modelo RDF puede ser más exigente que otros modelos de datos (tabular en Excel o CSV, o árboles en XML o JSON), aunque también los datos se pueden combinar de forma segura con otros datos.

5) Nivel «5 estrellas». **Enlazar los datos a otros datos para proporcionar contexto**.

Lograr cinco estrellas permite que los usuario hagan todo lo anterior y además:

- descubrir nuevos datos mientras se consumen otros,
- tratar con los URI no encontradas,
- vincular datos con temas relacionados es una cuestión de confianza.

Por otro lado, como editor de datos hay que hacer que sea posible descubrir los datos, lo que aumenta el valor de los mismos. Es necesario invertir recursos para vincular datos a otros existentes en la web.

Desde la primera estrella hasta la última, la apertura de los datos enlazados está respaldada por una mejor estructuración, una mejor interoperabilidad y, por lo tanto, una mejor reutilización como un recurso de datos en la web.

1.4. API de datos enlazados

Uno de los cuatro principios de los datos enlazados establece que se debe usar un URI para acceder a un recurso, y que se deben obtener datos relevantes acerca del recurso identificado.

La idea de la API de datos enlazados (LD API) es ofrecer una manera sencilla de acceder a datos enlazados vía web, permitiendo que los conjuntos de recursos estén expuestos como URI y facilitando su consulta.

Además, se posibilita también, mediante parámetros de consulta, filtrar, paginar y ordenar los resultados. La API admite diversos formatos de resultados, incluyendo JSON, XML, RDF/XML y Turtle.

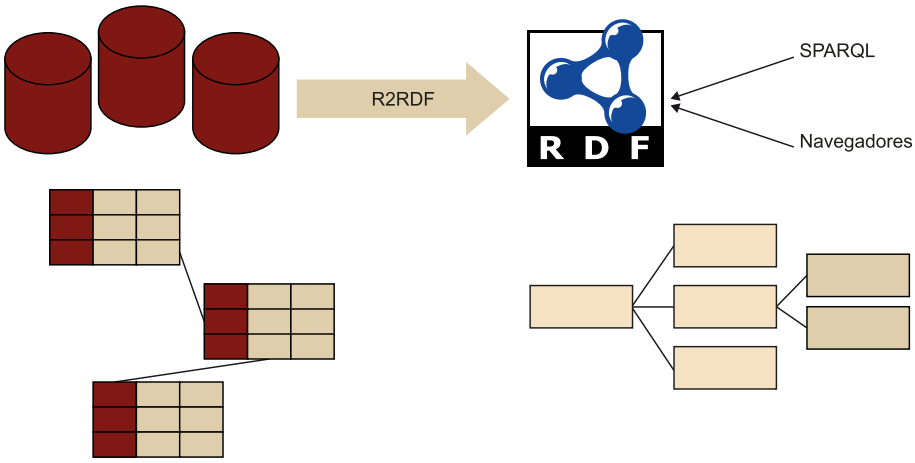
Para desarrollar software con datos enlazados es necesario comprender el modelo de datos RDF (con las serializaciones asociadas) y el lenguaje de consulta SPARQL que han demostrado ser una barrera para la adopción de datos enlazados.

Una API de datos enlazados es una especificación de código abierto, y existen algunas soluciones software que la implementan.

1.5. Publicación de bases de datos relacionales como datos enlazados

La mayoría del contenido dinámico de los sitios web proviene de bases de datos relacionales, como MS SQL, mySQL, Oracle o PostgreSQL. Publicar los datos como RDF hace de estos datos más accesibles en la web.

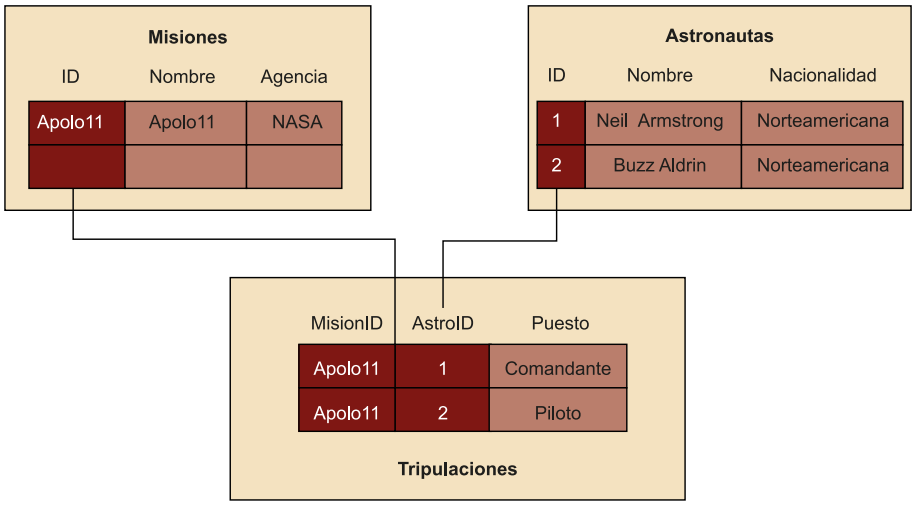
Figura 3. Publicación de bases de datos relacionales en RDF



Suponiendo que la información que la base de datos almacena esté en tablas, para realizar la traducción desde la base de datos relacional hay que generar un recurso RDF por cada una de las filas de las tablas.

Supongamos que tenemos tres tablas en una base de datos relacional, una para recopilar misiones espaciales, otra para los astronautas y una tercera que las enlaza como tripulación. En todas las tablas existe un identificador único que es la clave primaria de cada tabla.

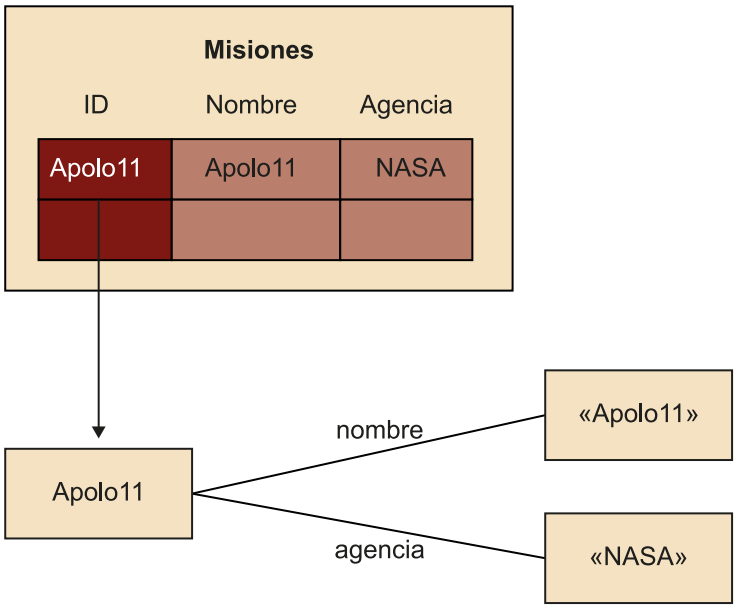
Figura 4. Tablas de ejemplo



Supongamos que se comienza por la tabla «Misiones». En primer lugar se genera un nuevo recurso de datos RDF a partir de cada fila de la tabla, utilizando la columna ID para generar un nuevo concepto en RDF cuyo identificador es un URI asociado.

A continuación se consideran los restantes elementos de la tabla. De esta forma la columna «Nombre» genera un nuevo literal y así sucesivamente con todas las columnas. Y una vez tengamos todas las columnas, hay que generar las relaciones entre los recursos.

Figura 5. Transformación de fila en recurso RDF



El **mapeo** (*mapping*) es una relación entre una entidad de una base de datos relacional y un concepto en un grafo RDF.

Hay dos formas de generar mapeos:

1) Mapeo directo (*direct mapping*). Proporciona una traducción automática de la base de datos relacional a RDF. Para hacer la traducción es necesario generar un URI para cada tabla, para cada atributo de la tabla y para cada clave primaria de cada tabla. Para indicar a qué tabla pertenece cada fila se genera una tripleta. Y por cada atributo de cada tabla se genera una tripleta.

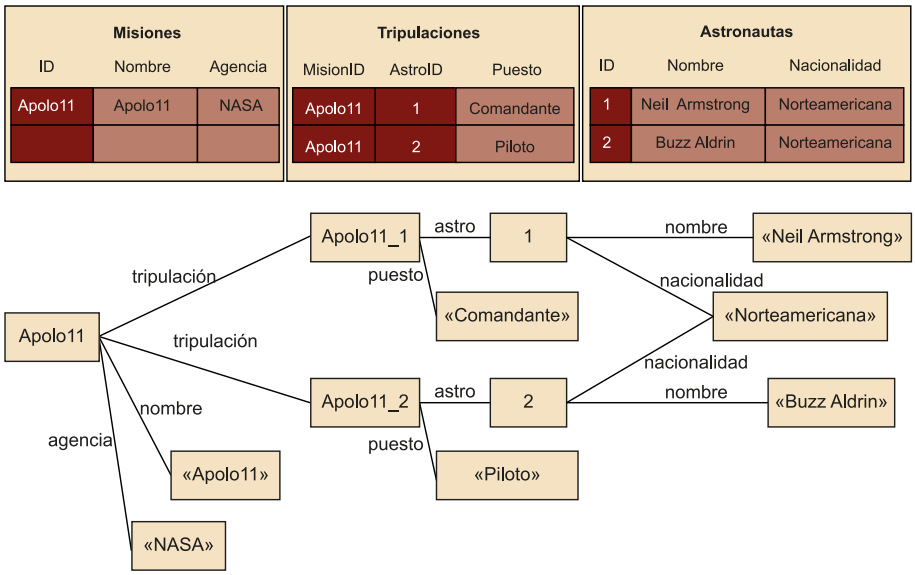
2) R2RML (RDB to RDF *Mapping Language*). Porporciona un lenguaje de mapeo para describir una traducción de la base de datos relacional a RDF. Mediante el uso de R2RML es posible describir cómo se realizará la traducción mediante reglas que detallan la representación de los datos e incluyen vocabulario RDFS y OWL.

El sistema de traducción que ejecuta los mapeos y que genera los nuevos datos en RDF se denomina *ETL*.[1] El proceso de ETL consta de tres fases:

[(1)]Acrónimo del inglés, *Extract Transform Load*.

1) Fase de extracción (*extract*): se leen los datos de la base de datos relacional.
2) Fase de transformación (*transform*): se transforman a RDF utilizando el sistema de traducción.
3) Fase de carga (*load*): se almacenan las tripletas RDF.

Figura 6. Ejemplo de transformación de tablas en grafo RDF



2. Ejemplos de datos enlazados

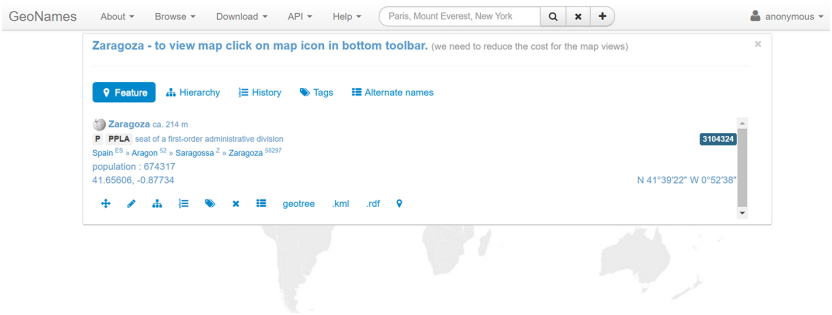
2.1. GeoNames

GeoNames es una base de datos que contiene información geográfica sobre más de 10 millones de lugares de todo el mundo. Para cada lugar se describe su nombre (en distintos idiomas), su ubicación (latitud, longitud y elevación), su población, su código postal y su categoría. La categoría permite indicar el tipo de elemento que se está describiendo (zona, carretera, límite administrativo, etc.). La ubicación se representa mediante el vocabulario Basic Geo (WGS84), que permite representar la geoposición de los elementos definidos utilizando RDF.

Los URI de GeoNames tienen la siguiente forma «https://sws.geonames.org/codigo», donde el código identifica el elemento que se va a describir.

Por ejemplo, en el caso de la ciudad de Zaragoza, su URI es http://sws.geonames.org/3104324, que redirige a la versión HTML del recurso https://www.geonames.org/3104324/zaragoza.html. Con este URI se pueden obtener los datos RDF o presentar el mapa.

Figura 7. Datos de Zaragoza



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

2.2. La Biblioteca Nacional de España

El proyecto de la Biblioteca Nacional de España y del Grupo de Ingeniería Ontológica de la Universidad Politécnica de Madrid tiene como objetivo la exploración de datos bibliográficos de manera diferente a los catálogos tradicionales, ofreciendo otra forma de navegación por los diferentes recursos de la biblioteca y enriqueciendo sus propios datos con otros datos externos.

Puede accederse a los datos a partir del propio portal o desde una interfaz SPARQL. Además también pueden obtenerse volcados completos de los datos. La biblioteca utiliza su propia ontología en la que define sus entidades, que son equivalentes a las entidades descritas en la ontología FRBR.

Ontología FRBR

El modelo FRBR fue publicado en el año 1998 y describe los registros bibliográficos por medio de las entidades «Obra», «Expresión», «Manifestación» e «Ítem», así como autores y materias. En el año 2010 se publicó la ontología FRBR, que es la que se ha utilizado en las bibliotecas de España.

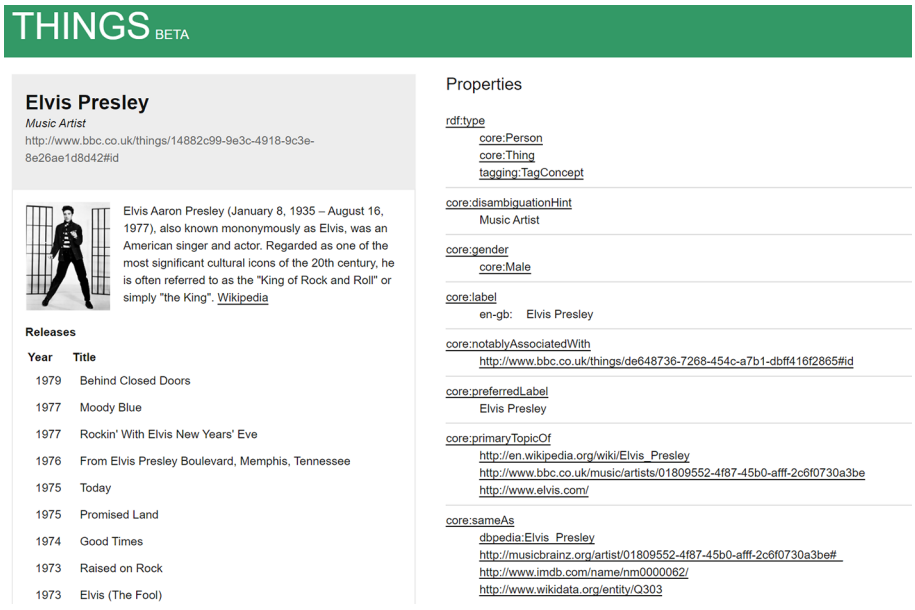
Figura 8. Portal de datos de la Biblioteca Nacional de España

Fuente: datos.bne.es/persona/XX1718747.html

2.3. BBC Things

BBC Things utiliza tecnologías de datos enlazados que permiten acceder a temas de interés para el público del grupo BBC (British Broadcasting Corporation), tales como personas, lugares, organizaciones, competiciones deportivas, etc. La BBC posee un conjunto propio de ontologías que definen los diferentes temas que se pueden consultar en BBC Things.

Figura 9. Datos sobre Elvis Presley en BBC Things (HTML)



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

La figura 9 muestra la página HTML resultante de una consulta sobre el músico y actor Elvis Presley que incluye los datos en tripletas asociadas al recurso.

Bibliografía

Allemang, D.; Hendler, J. (2011). *Semantic Web for the working ontologist* (2.ª 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 (1.ª ed.)». *Synthesis Lectures on the Semantic Web: Theory and Technology* (vol. 1, núm. 1, págs. 1-136). Morgan & Claypool.