MARTÍN ÁLVAREZ ESPINAR

Web Semántica y datos enlazados

La Web de los documentos

La *World Wide Web* o, simplemente Web, nació a finales de los años 80 en el Centro Europeo de Investigación Nuclear (CERN) situado en Suiza. La principal motivación de su creador, Tim Berners-Lee, fue la ineficiente gestión de la información de este instituto en ese momento, donde sus miles de investigadores multidisciplinares generaban ingente cantidad de datos procedente de sus estudios y experimentos. Berners-Lee se planteó la necesidad de disponer de una plataforma universal que permitiese la gestión de la información de esa comunidad investigadora mundial. Estos matemáticos y físicos almacenaban la información en diversos formatos sin seguir patrones homogéneos y utilizando tecnologías y entornos tecnológicos distintos, lo que propiciaba que dicha información estuviese aislada a merced de la voluntad de los propios investigadores.

En 1989, Berners-Lee elaboró una propuesta para homogeneizar la representación de la información y permitir que estos documentos de carácter científico estuvieran representados de forma coherente y mediante formatos comunes, representados de forma natural a través de las redes (Berners-Lee, 1990). Pocos

Martín Álvarez Espinar (email: martin.alvarez@fundacionctic.org)
Oficina W España/CTIC

Recibido: 10-07-2015. Aceptado: 13-09-2015
Citación: Álvarez Espinar, Martín (2015). "Web semántica y datos enlazados". *Tábula,* n. 18, pp. 21-43

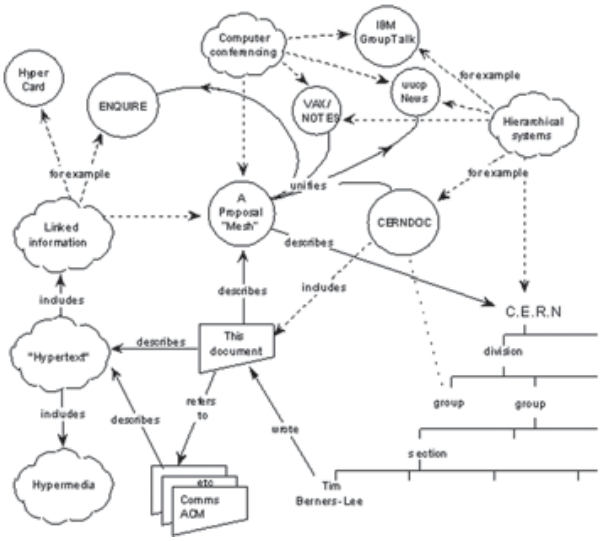

Fig. 1. Propuesta de gestión de la información por Tim Berners-Lee (1989)

meses después de su primer boceto (Fig. 1), creó el primer prototipo funcional de un editor/navegador de documentos Web, denominado *Worldwideweb* (Connolly, 2000), lo que supuso la concepción de la Web de hoy en día, que funcionaba sobre la infraestructura de Internet ya establecida por aquel entonces.

Las primeras páginas Web ya se basaban en una versión preliminar del Lenguaje de Etiquetado de Hipertexto o HTML, un estándar que supuso una revolución para la puesta en común de lenguajes universales que permitían representar información de forma uniforme e independiente del sistema (plataforma informática, navegadores u otro software). La simplicidad y utilidad del HTML, basado en el Lenguaje estándar de etiquetado general o SGML (*ISO 8879:1986 - Standard Generalized Markup Language (SGML)*, 1986), un versátil método, internacionalmente aceptado para representar lenguaje natural en unidades estructurales básicas, como párrafos, encabezados o listas de elementos, fue la que fomentó la confianza de la comunidad científica e hizo rápida su adopción.

Los colectivos académicos primero y las corporaciones privadas después, comenzaron a interesarse por esta nueva forma de comunicación cuyas tecnologías básicas iban siendo mejoradas progresivamente y se acabaría convirtiendo en algo universal. Cinco años después, en 1994, Berners-Lee fundó el World Wide Web Consortium (W3C)[1], un consorcio neutral que pretendía aunar los esfuerzos de la industria y la comunidad investigadora en la generación de las piezas necesarias para llevar a la Web a su máximo potencial, desarrollando estándares libres, abiertos y gratuitos.

La adopción de las tecnologías de la Web, estandarizadas por consenso en el W3C, se extendió por todo el mundo de forma exponencial. Indicativo de esto

es que a finales de los años noventa ya se habían creado más de 30 millones de páginas Web, superando el billón de documentos una década después (Alpert & Hajaj, 2008). Por ese entonces, la Web era considerada como la solución universal para la gestión del conocimiento que estaba revolucionando todos los sectores de la sociedad.

En la actualidad, el Consorcio W3C está formado por más de 400 organizaciones miembro distribuidas por todo el mundo y, desde su creación ha publicado más de 300 estándares para la Web.

La Web de los datos y servicios

La Web original, por el tipo de contenido que albergaba, estaba concebida como un grafo de documentos enlazados que, aunque eran procesables automáticamente (podían ser visualizados y analizados), se dirigían exclusivamente a las personas, que eran quienes finalmente iban a interpretar los textos, gráficos o datos presentes en los mismos.

La arquitectura esencial de la Web ofrece flexibilidad y escalabilidad ilimitada, lo que ha motivado la investigación y desarrollo de nuevas tecnologías para conseguir que la Web no sólo fuese dirigida a los humanos, sino que también sea una plataforma para conectar máquinas, con el objetivo de que esta Web de documentos también gestione servicios y datos.

La creación del Lenguaje de Etiquetado Extensible o XML (Bray et al., 2006), primero, y los servicios web (Booth et al., 2004) para la comunicación después, marcaron dos hitos relevantes en la interoperabilidad de sistemas, ofreciendo un mecanismo para llevar a cabo una comunicación normalizada entre plataformas heterogéneas distribuidas. Esto ha permitido solucionar retos que la sociedad y la industria planteaba —por ejemplo, el comercio electrónico o las agencias de viajes *online*, emergentes por aquel entonces— para dotar de interoperabilidad a las máquinas.

La creación del XML supuso una revolución en el intercambio de la información, desde la primera especificación publicada en 1996, permitiendo normalizar y estructurar la información a través de un sencillo mecanismo de intercambio y procesamiento. Desde su creación, la mayoría de los sistemas de intercambio de información de todo el mundo se han basado en este lenguaje. Sólo por citar algunos: SVG, para representar gráficos vectoriales; SOAP, para servicios Web; RSS o ATOM, para canales de suscripción; XBRL, para informes financieros; o MARCXML, para recursos bibliográficos.

Semántica en la Web

La representación de los datos en XML y su distribución mediante los servicios web ha servido para solucionar problemas acotados y en ámbitos concretos, pero la falta de flexibilidad de los modelos —difícilmente escalables— hace necesario evolucionar el concepto de Web, en busca de esa gran base de datos distribuida, flexible y universal, que sea comprensible y procesable por cualquier plataforma tecnológica.

El desarrollo tecnológico liderado por el W3C mantiene la universalidad del acceso a los documentos HTML para cualquier persona, pero ahora incluyendo mecanismos de representación de conceptos y cosas de la vida real que comprenden las máquinas. Así la Web se puede comportar como una infinita base de conocimiento, permitiendo la interoperabilidad entre cualquier sistema, independientemente de las tecnologías utilizadas para el procesamiento de la información. El entendimiento entre máquinas que antes se conseguía a nivel léxico y sintáctico, ahora se lleva a un plano superior, la semántica.

De esta forma, aparte de la orientación universal y accesible de la Web, sistemas automáticos pueden discernir los tipos de recursos por el concepto que representan. Esto es lo que se conoce como la Web Semántica (Berners-Lee, Hendler, & Lassila, 2010).

La ambigüedad del lenguaje natural

Uno de los grandes retos de los desarrolladores de aplicaciones y servicios sobre la Web, es la ambigüedad de las representaciones iconográficas y el lenguaje natural. Esto es algo que aún está presente desde la creación de la Web y, prueba de ello, es la frustración experimentada diariamente por millones de usuarios que buscan contenido en la Web a través de sus buscadores favoritos y no acceden fácilmente a los resultados adecuados.

Por ilustrar esto, si un usuario interesado en información básica sobre circuitos electrónicos buscase la palabra 'chips' en cualquiera de los buscadores conocidos, éstos devolverían resultados confusos y, en la mayoría de los casos, nada relacionados con el objetivo (Fig. 2): vídeos una famosa serie de televisión norteamericana (CHIPs[2]), imágenes de patatas fritas o fichas de casino (*chips* en inglés)[3], pero ni rastro de componentes electrónicos.

Con la definición y manejo de conceptos semánticos en la Web, se ofrece una abstracción de la representación de la información subyacente en los documentos HTML de la "Web clásica". Esta abstracción da lugar a la eliminación de posibles ambigüedades del lenguaje natural en las aplicaciones o servicios en la Web.

La esencia de la Web semántica ofrece soluciones naturales a éste y a otros problemas similares, ya que cualquier cosa de la vida real se podrá describir de

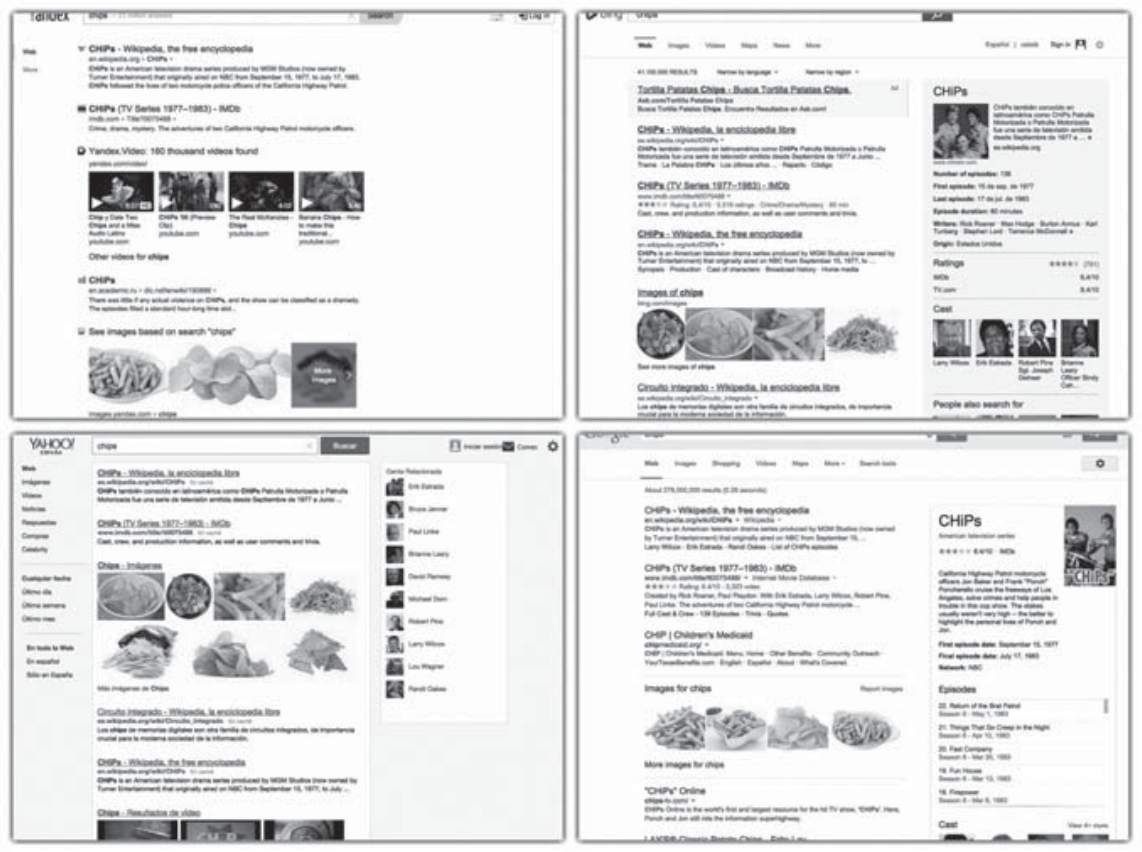

Fig. 2. Resultados ante la búsqueda 'chips' en Yandex, Bing, Yahoo! y Google

forma conceptual. Uno de los grandes avances de esta forma de descripción de metadatos, es que se permite añadir etiquetas e información en varios idiomas simultáneamente y, gracias a que todos los objetos o documentos están identificados por direcciones Web, cualquier persona o máquina puede hacer referencia, acceder y procesar estas descripciones, comprendiendo el concepto que representa.

Dotando de semántica a la web

Con las tecnologías de la Web Semántica, cualquier objeto real o ficticio es susceptible de ser descrito semánticamente a través de los metadatos que lo caracterizan —por ejemplo, una persona podría ser descrita por su nombre y apellidos, características físicas, lugar donde vive, alias que utiliza en las redes sociales, etc.—.

La finalidad de estos metadatos es permitir que tanto personas como máquinas comprendan el concepto de lo que se describe. Los metadatos dirigidos a las personas se plasmarán en documentos HTML, y la versión para las máquinas seguirá un modelo conocido como **Infraestructura de Descripción de Recursos o RDF** (Manola, Miller, & McBride, 2014).

El RDF es una sintaxis genérica estándar que permite asociar propiedades a los objetos o cosas que se describen, siguiendo una estructura basada en tripletas lógicas: **sujetoàpredicadoàobjeto** (Fig. 3).

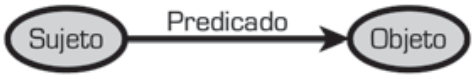

Fig. 3. Tripleta RDF básica

En este esquema, **'predicado' es la propiedad que caracteriza al 'sujeto' con un valor u 'objeto'**. Por ejemplo, para describir una persona de nombre John, se le podrían añadir aquellas propiedades y valores —predicados y objetos— que definan al individuo con sus valores correspondientes: "John es una persona" y "John se apellida Smith". Esto se puede representar gráficamente como se muestra a continuación (Fig. 4).

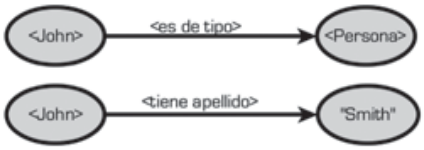

Fig. 4. Tripletas RDF que describen a una persona

Cualquier recurso puede ser definido por infinitas tripletas que generan grafos dirigidos. En estos grafos, los nodos son recursos de información (sujetos u objetos) y sus vértices las propiedades que permiten caracterizar a las cosas descritas.

Al ejemplo anterior se le pueden añadir las descripciones: "John vive en Boston" y "John estudió en el MIT". Ya que ambas tripletas se refieren al mismo sujeto, se pueden fusionar ambos nodos dando lugar al siguiente grafo (Fig. 5).

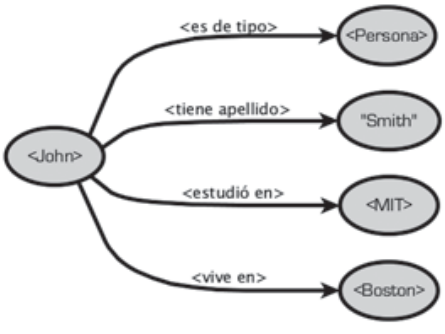

Fig. 5. Grafo RDF que describe a una persona

Los objetos de las tripletas pueden ser recursos que están descritos de la misma forma. Por ejemplo, se puede describir con detalles la entidad "MIT", que se hace referencia en el ejemplo anterior: MIT es una organización que está situada en Boston y tiene como sitio web "http:/mit.edu". Esto generaría el siguiente grafo (Fig. 6).

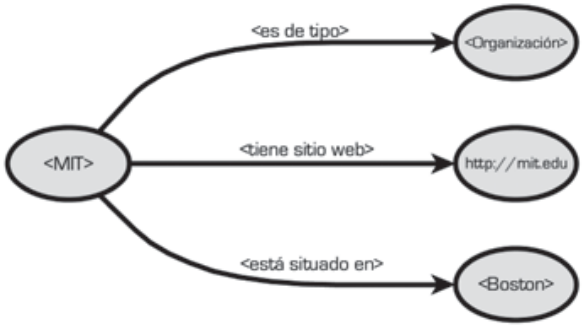

Fig. 6. Grafo RDF que describe una organización

Se observa que ambos grafos (Fig. 5 y Fig. 6) comparten los nodos etiquetados como <*MIT*> y <*Boston*>, por lo que se podrían fusionar en una única representación que combine los nodos coincidentes (Fig. 7).

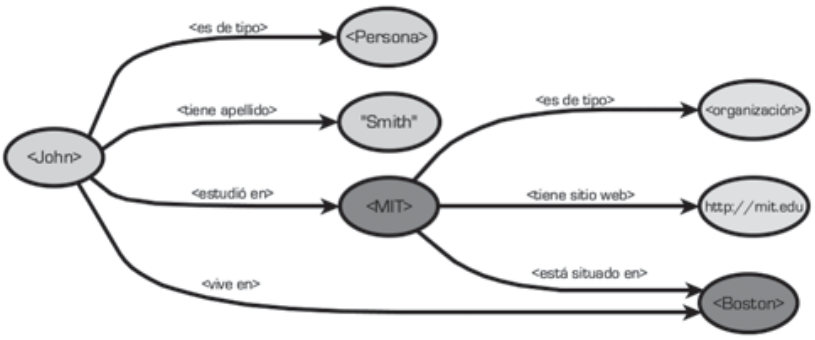

Fig. 7. Grafo RDF que combina dos sub-grafos

La flexibilidad de este modelo es ilimitada, lo que permite añadir nuevas propiedades en cualquier momento, incluso enlazando a otros grafos en la Web si existe una relación semántica entre los recursos de ambos.

Identificación unívoca de recursos

De la misma forma que cualquier documento Web está identificado con una dirección Web única, cualquier recurso descrito en la Web mediante RDF debería ser accesible en la Web. Por ello, cualquier recurso con entidad propia tendrá asociado

una referencia unívoca basada en Identificadores de Recursos Uniformes o URI[4] (Cyganiak, Wood, & Lanthaler, 2014), esto es, una dirección Web única y persistente en el tiempo. Gracias a estos identificadores universales, cualquier máquina o persona puede hacer referencia conceptos, elementos físicos o virtuales, de forma inequívoca en la Web. Esto es importante en el modelo RDF ya que, salvo excepciones puntuales, los sujetos y predicados de las tripletas siempre tienen un URI asociado[5].

En el caso de los valores de las propiedades, éstos pueden ser recursos identificados mediante URI (serán sujetos de otra tripleta) o valores literales que contienen descripciones textuales, números, fechas u otros datos estáticos.

El ejemplo del grafo anterior incluye dos objetos que son literales —los correspondientes al apellido de la persona ("Smith") y al sitio web del organismo ("MIT")—. En la figura siguiente (Fig. 8), se ilustra a los recursos representados por elipses e identificados por URI, y dos literales como rectángulos con valores textuales[6].

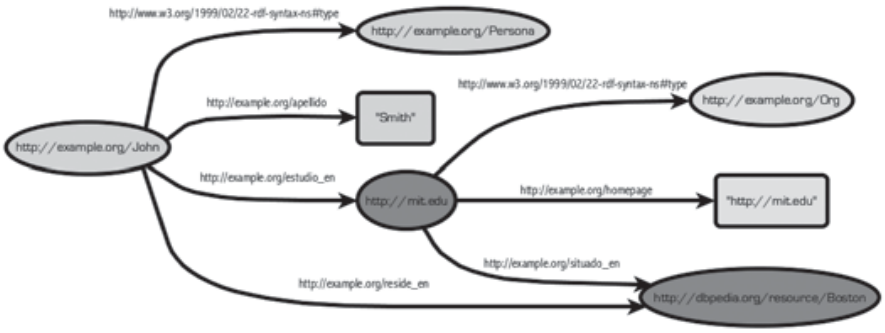

Fig. 8. Grafo RDF con recursos identificados mediante URI

Cabe destacar que los recursos pueden formar parte del grafo RDF en cualquier posición y los literales siempre serán nodos finales.

La información descrita en los grafos RDF será procesable automáticamente y, por lo tanto, cualquier programa podrá recorrer y procesar la información presente en el grafo de una forma automatizada, permitiendo inferir relaciones lógicas entre recursos.

Datos abiertos enlazados

En el ejemplo anterior se aprecia cómo los identificadores de los recursos —nodos y propiedades— pueden pertenecer a distintos dominios de Internet (*example.org*, *www.w3.org*, *mit.edu* y *dbpedia.org*). Esto posibilita que los grafos de tripletas RDF hagan referencia mediante URI públicos a recursos descritos y mantenidos por

entidades externas. Puede observarse como dos propiedades hacen referencia a la ciudad de Boston, identificado como *http://dbpedia.org/resource/Boston*, y descrito semánticamente en base a información procedente de la Wikipedia[7].

Para describir las propiedades en las tripletas se pueden reutilizar términos de vocabularios ya existentes, de forma similar a la reutilización de datos procedentes de otros sistemas. En el ejemplo se puede observar una de las propiedades básicas de RDF que denota el tipo de recurso que se describe: '*type*', identificada por *http://www.w3.org/1999/02/22-rdf-syntax-ns#type* (la propiedad ha sido definida por el W3C y está en el dominio www.w3.org).

Este nuevo paradigma de enlace entre recursos semánticos en cualquier dominio de la Web da lugar al término de **datos abiertos enlazados** o *linked open data* (Bizer, Cyganiak, & Heath, 2008). Los conjuntos de datos específicos por dominios temáticos conforman la nube de los datos abiertos enlazados (Fig. 8) (Cyganiak, 2014), donde cada nodo representado es una colección de datos en RDF que se enlaza con otras.

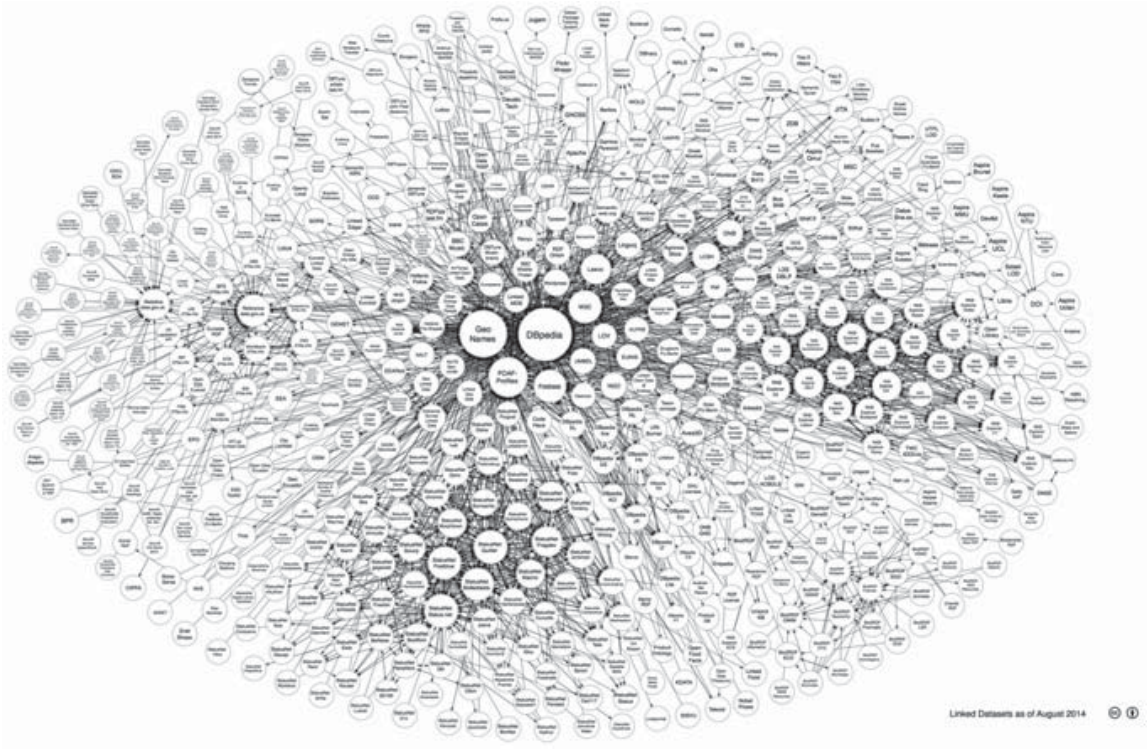

Fig. 9. La nube de los datos abiertos enlazados

La versatilidad de las tecnologías de la Web semántica permiten que las descripciones que dotan de significado a los recursos sean enriquecidas por expertos reconocidos en materias concretas y que cualquiera pueda reutilizar estas descripciones libremente.

En el ejemplo mencionado anteriormente donde se hacía referencia a la ciudad de Boston a través de la dirección *http://dbpedia.org/resource/Boston*, el sistema podría acceder a la descripción de este recurso, interpretar su tipo (lugar poblado), y conocer todos aquellos valores relacionados con este lugar (estadísticas socio-económicas, características geográficas o administrativas, personajes asociados al lugar, etc.) que previamente se han declarado. Todo esto sin necesidad de duplicar esfuerzos y permitiendo a los sistemas inferir los valores de este lugar y procesarlos automáticamente, incluso enriquecer con datos propios o externos. Por ejemplo, a través de una propiedad que denota equivalencia, se puede indicar que este recurso es el mismo que otro descrito por la base de datos geográfica Geonames[8], identificado como *http://sws.geonames.org/4930956/*. Esta nueva relación semántica permitiría a cualquier programa acceder a información más detallada sobre la toponimia oficial de este lugar en más de 80 idiomas.

¿Máquinas inteligentes?

Se ha comentado que, gracias a estas técnicas, las máquinas que procesen la información podrán comprender los conceptos y significados de las cosas, ¿pero cómo?

De la misma forma que se describen las características de individuos concretos (mediante propiedades y los valores que toman), se describen los conceptos abstractos y los tipos de la información del modelo a representar. Esto se hace mediante vocabularios y ontologías —vocabularios controlados que definen un dominio de información concreto— definidos también bajo el modelo RDF y consensuadas universalmente. Siguiendo los ejemplos anteriores, para representar la información del grafo se necesitaría un vocabulario con las características genéricas de las personas, organizaciones y lugares.

Básicamente, estos vocabularios pueden contener **clases** —el tipo de recurso— y **propiedades** que serán usadas como predicado en las tripletas. En ese mismo esquema se debería indicar el **rango** de valores específicos que pueden tomar dichas propiedades y su **dominio** —o el tipo del 'sujeto' al que pueden ir asociadas—. Estos vocabularios son los que realmente permiten a las máquinas interpretar los conceptos y el tipo de información descrita.

Las tecnologías de la Web Semántica

Notaciones RDF

RDF no es un lenguaje en sí, sino una sintaxis abstracta. Este modelo puede ser representado y distribuido en la Red mediante distintos formatos. A continuación se citarán los principales formatos en los que se puede serializar los grafos RDF, ilustrado con el ejemplo visto hasta el momento. Seleccionar un formato u

otro dependerá del potencial uso que se le desee dar y las herramientas de las que se dispongan para el procesamiento posterior de los datos.

Uno de los primeros formatos, que tuvo gran impacto ya hace más de una década, es el **RDF/XML** (Gandon & Schreiber, 2014), una especialización del lenguaje XML que incorpora las características específicas del RDF.

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
  xmlns:org="http://example.org/vocab/org/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:per="http://example.org/vocab/persona/"
>
  <per:Persona rdf:about="http://example.org/persona/John">
    <per:tiene_apellido>Smith</per:tiene_apellido>
    <per:estudio_en>
      <org:Organizacion rdf:about="http://example.org/org/MIT">
        <org:situado_en rdf:resource="http://dbpedia.org/resource/Boston"/>
        <org:tiene_sitio_web rdf:resource="http://mit.edu"/>
      </org:Organizacion>
    </per:estudio_en>
    <per:vive_en rdf:resource="http://dbpedia.org/resource/Boston"/>
  </per:Persona>
</rdf:RDF>
```

Código 1. Ejemplo de notación RDF/XML

La notación RDF/XML, demasiado enfocada al tratamiento automático, dificulta la escritura e interpretación por parte de las personas, ya que no es demasiado intuitivo. Por eso mismo, se han definido notaciones textuales más compactas y legibles para representar las tripletas y los grafos RDF de una forma clara y más enfocada en el usuario.

Uno de estos lenguajes textuales es **Turtle**, cuya sencillez para la representación de tripletas lo hacen uno de los más intuitivos (Beckett, Berners-Lee, Prud'hommeaux, & Carothers, 2014).

```
@prefix org: <http://example.org/vocab/org/> .
@prefix per: <http://example.org/vocab/persona/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://example.org/persona/John> a per:Persona ;
    per:estudio_en <http://example.org/org/MIT> ;
    per:tiene_apellido "Smith" ;
    per:vive_en <http://dbpedia.org/resource/Boston> .

<http://example.org/org/MIT> a org:Organizacion ;
    org:situado_en <http://dbpedia.org/resource/Boston> ;
    org:tiene_sitio_web <http://mit.edu> .
```

Código 2. Ejemplo de notación Turle

N-Triples es un subconjunto de Turtle, que expresa las tripletas de un grafo por líneas (Beckett, 2014).

```
<http://example.org/org/MIT> <http://example.org/vocab/org/situado_en> <http://dbpedia.org/resource/Boston> .
<http://example.org/persona/John> <http://example.org/vocab/persona/vive_en> <http://dbpedia.org/resource/Boston> .
<http://example.org/persona/John> <http://example.org/vocab/persona/tiene_apellido> "Smith" .
<http://example.org/org/MIT> <http://example.org/vocab/org/tiene_sitio_web> <http://mit.edu> .
<http://example.org/org/MIT> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.org/vocab/org/Organizacion> .
<http://example.org/persona/John> <http://example.org/vocab/persona/estudio_en> <http://example.org/org/MIT> .
<http://example.org/persona/John> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://example.org/vocab/persona/Persona> .
```

Código 3. Ejemplo de notación N-Triples

En la actualidad la sintaxis que está teniendo más acogida es **JSON-LD**, que lleva el paradigma de la Web Semántica y los datos enlazados a la notación JSON para objetos JavaScript (Sporny, Longley, Kellogg, Lanthaler & Lindström, 2014), un formato muy extendido globalmente en los entornos de desarrollo Web.

```
{
  "@context": {
    "dbpedia": "http://dbpedia.org/resource/",
    "org": "http://example.org/vocab/org/",
    "per": "http://example.org/vocab/persona/",
    "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
    "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
    "xsd": "http://www.w3.org/2001/XMLSchema#"
  },
  "@graph": [
    {
      "@id": "http://example.org/org/MIT",
      "@type": "org:Organizacion",
      "org:situado_en": {
        "@id": "dbpedia:Boston"
      },
      "org:tiene_sitio_web": {
        "@id": "http://mit.edu"
      }
    },
    {
      "@id": "http://example.org/persona/John",
      "@type": "per:Persona",
      "per:estudio_en": {
        "@id": "http://example.org/org/MIT"
      },
      "per:tiene_apellido": "Smith",
      "per:vive_en": {
        "@id": "dbpedia:Boston"
      }
    }
  ]
}
```

Código 4. Ejemplo de notación JSON-LD

Existen otras sintaxis alternativas como **TriG**, una extensión de Turtle que permite representar varios grafos y así definir varios conjuntos de datos en el mismo documento (Bizer & Cyganiak, 2014), o **N-Quads**, otra opción basada para representar colecciones de grafos RDF mediante líneas de texto (Carothers, 2014).

También para las personas

Aunque la Web Semántica se centra en la interoperabilidad entre máquinas, cualquier recurso identificado por un URI y descrito en RDF, debería tener una representación adecuada y accesible para las personas —habitualmente en documentos HTML—.

El propio servidor Web que contiene el recurso deberá hacer la distinción si distribuye un documento u otro, dependiendo de quién sea el peticionario. Si una persona escribe la dirección en un navegador Web, ésta debería recibir un documento en formato HTML. En el caso de que sea un programa con capacidad de procesamiento RDF, el servidor interpreta la petición y sirve la versión RDF, en alguno de sus formatos de representación. Esta gestión se realizará a través de reglas internas en el servidor y mediante los códigos específicos del protocolo de la Web, el HTTP (Berrueta & Phipps, 2008).

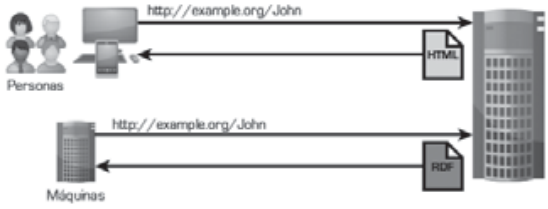

Fig. 10. Gestión del contenido por parte del servidor ante distintas peticiones

RDFa, la solución intermedia

Una de las formas más populares para representar datos en la Web es la integración de la descripción semántica en los sistemas de gestión de contenidos (o CMS) o portales Web. RDFa es una de las tecnologías que permite esto, incluyendo los metadatos —como tripletas RDF— incrustadas en el código fuente de los documentos HTML (Herman, Adida, Sporny, & Birbeck, 2015). Esto se hace de forma transparente para los usuarios y también permite el procesamiento automático por aquellas máquinas que sepan hacerlo.

La forma de representar esta información mediante la notación RDFa es similar a la notación RDF/XML pero a través de atributos especiales incluidos en el HTML, como se muestra a continuación.

```
<div xmlns="http://www.w3.org/1999/xhtml"
  prefix="
    org: http://example.org/vocab/org/
    rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
    per: http://example.org/vocab/persona/
    rdfs: http://www.w3.org/2000/01/rdf-schema#"
  >
  <div typeof="per:Persona" about="http://example.org/persona/John">
    <div property="per:tiene_apellido" content="Smith"></div>
    <div rel="per:estudio_en">
      <div typeof="org:Organizacion" about="http://example.org/org/MIT">
        <div rel="org:tiene_sitio_web" resource="http://mit.edu"></div>
        <div rel="org:situado_en" resource="http://dbpedia.org/resource/Boston"></div>
      </div>
    </div>
    <div rel="per:vive_en" resource="http://dbpedia.org/resource/Boston"></div>
  </div>
</div>
```

Código 5. Ejemplo de notación RDFa en el código HTML

Los principales motores de búsqueda en la Web impulsan y soportan esta tecnología, lo que les confiere la posibilidad de interpretar los contenidos que sus robots indexan. De ahí, que estos sistemas permitan hacer búsquedas sobre tipos de contenido concretos y devuelvan resultados adecuados ante las búsquedas de los usuarios, lo que lo convierte en una solución efectiva para el problema planteado ante la ambigüedad de los documentos únicamente basados en texto y multimedia.

Para asegurar que tanto productores como consumidores de información puedan gestionar información semántica y compartan vocabularios, los fabricantes

de buscadores Web, Bing, Google, Yahoo! y Yandex han creado **schema.org**[9]**,** un conjunto de vocabularios normalizados para describir entidades en la Web: eventos, personas, organizaciones, lugares, productos de venta, entre otros. El propósito de schema.org es ofrecer un conjunto de términos consensuados para que los desarrolladores y editores de contenido en la Web puedan describir y representar datos concretos e indiquen el tipo de información asociado a los documentos HTML. Esto permitirá un subsecuente procesamiento automatizado por parte de estos buscadores o por parte de terceros.

La evolución en la eficiencia de estos buscadores u otras aplicaciones en la Web ya es una realidad con la que se puede experimentar. Por ejemplo, al introducir la cadena "World Wide Web Consortium" en uno de estos buscadores[10], entre los resultados obtenidos, el sistema muestra una representación del resultado adecuándola a su naturaleza específica: denominación, tipo de organismo, logo, descripción, fecha de creación, etc. (Fig. 11).

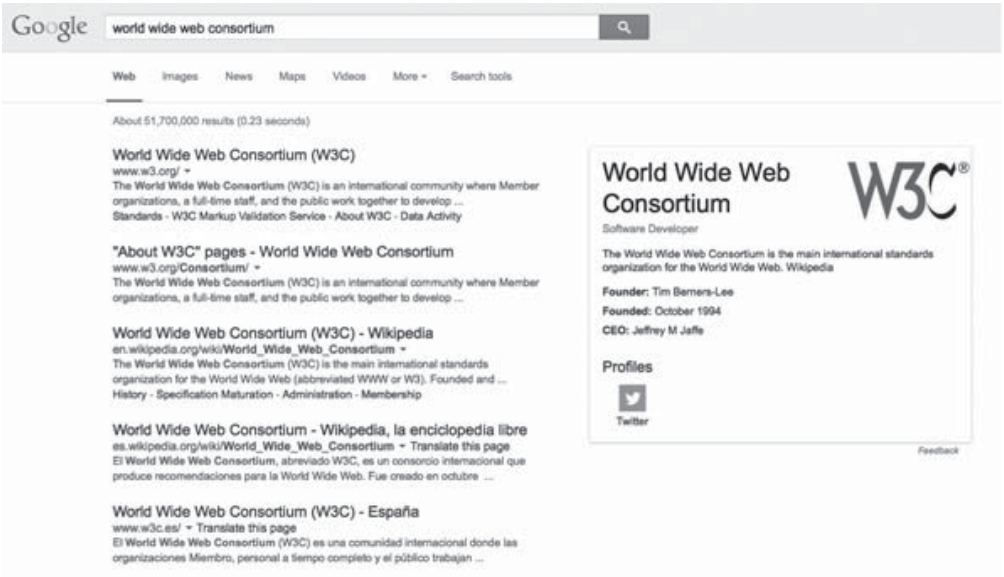

Fig. 11. Resultado de la búsqueda de "World Wide Web Consortium" en Google

Otros estándares

La Web Semántica y los datos enlazados van más allá del modelo de RDF y los vocabularios, ya que existe un compendio de tecnologías y mecanismos adicionales que hacen posible mantener esta gran base de datos global. Como se ha visto hasta el momento, estas tecnologías no son siempre novedosas, ya que se utilizan muchos de los fundamentos de la Web tradicional —como URI, HTML o XML—.

A continuación se muestra un resumen con algunas de las tecnologías básicas que componen la Web Semántica (Fig. 12). Entre ellas se pueden encontrar mecanismos para: identificación de recursos basada en **URI**; codificación mediante juegos de caracteres estándares (**UTF**); sintaxis basada en **XML** (o las notaciones vistas anteriormente), junto a los espacios de nombres (**NS**) para simplificar la notación; definición de tipos de datos básicos mediante **Esquema XML** (enteros, fechas, horas, texto, booleanos, etc.); descripción de metadatos mediante el modelo **RDF**; descripción de vocabularios con Esquema RDF (**RDF Schema**) y **OWL**; o consultas sobre los grafos RDF con **SPARQL**.

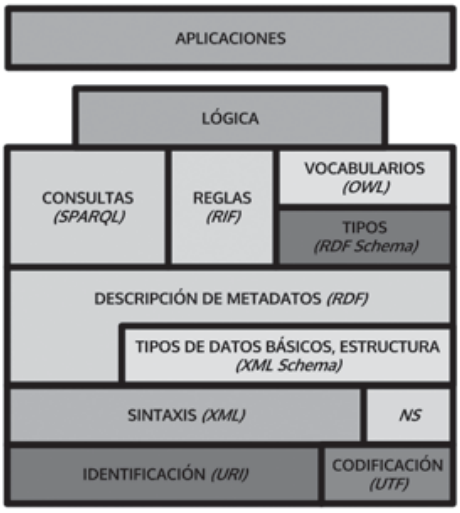

Fig. 12. Pila de tecnologías de la Web Semántica

Sobre estas tecnologías se implementa la capa lógica que permitirá el desarrollo de aplicaciones que procesen y usen los datos de forma segura, eficiente y eficaz.

Vocabularios comunes

Como se ha citado anteriormente en el caso de schema.org, existen vocabularios estándar o de uso frecuente para determinados dominios que reflejan el ámbito semántico en el que se trabaja. Uno de los ejemplos más extendidos es el conjunto de Términos de **Dublin Core**[11] para la descripción de metadatos bibliográficos.

Desde el punto de vista técnico, estos vocabularios se definen utilizando el lenguaje para esquemas RDF —**RDF Schema**— (Brickley & Guha, 2014) , el Lenguaje de Ontologías Web —**OWL**— o, más frecuentemente, una combinación de ambos. Por su menor expresividad, RDF Schema se utiliza sobre todo en vocabularios básicos orientados a la compartición de datos entre distintas fuentes de información, mientras que OWL permite formular expresiones semánticas

mucho más ricas, aunque técnicamente resulta más complejo y puede tener un soporte limitado por las aplicaciones de destino ("OWL 2 Web Ontology Language Document Overview (Second Edition)," 2012).

No siempre se conocen vocabularios adecuados para la representación de la información con la que se trabaje, por lo que, habitualmente será necesario crear el vocabulario desde cero, o adaptar alguno de los ya existentes. Las tecnologías mencionadas posibilitan la modificación o actualización en cualquier momento, sin la necesidad de impactos de relevancia en las implementaciones existentes.

De forma general, cuando exista un vocabulario aplicable al conjunto de datos resulta preferible reutilizarlo o extenderlo en lugar de crear uno nuevo desde cero. Al aprovechar un vocabulario ya existente se fomenta la interoperabilidad entre los agentes que usen esos esquemas ya creados.

Hay varios criterios que deberemos tener en cuenta a la hora de seleccionar un vocabulario externo para representar los datos:

- **Debería ser de uso frecuente y reconocido**. Aunque existen pocos vocabularios reconocidos como estándar, hay una serie de ellos que son internacionalmente reconocidos y ampliamente usados, como: Dublin Core, FOAF para personas y sus relaciones (Brickley & Miller, 2014), SKOS para vocabularios controlados o vCard para información de contacto de personas y organizaciones (Iannella & McKinney, 2014).
- **Adecuada cobertura y expresividad**. Debe valorarse si el grado de expresividad del vocabulario es suficiente para el nivel de desagregación de la información a representar y si éste permite definir todos los conceptos y relaciones con el detalle deseado.
- **Mantenimiento y actualización**. Debe tenerse en cuenta si el vocabulario a reutilizar está siendo mantenido activamente, y si existe una política clara de actualización, y mecanismos de recepción de incidencias, ya que podría evolucionar en el tiempo y requerir cambios.
- **Exposición pública**. El vocabulario a reutilizar debe estar publicado abiertamente de la misma forma que cualquier recurso semántico, es decir, empleando URI resolubles para identificar las clases y propiedades con suficiente documentación, para las personas y para las máquinas.

Sistemas de organización del conocimiento

Otro de los retos cruciales que soluciona la Web Semántica es la gestión de los sistemas de representación del conocimiento (glosarios, taxonomías, tesauros, etc.). El Sistema Simple de Organización del Conocimiento, o **SKOS** (Isaac&Summers, 2009), es un vocabulario estándar que permite definir estas clasificaciones de

conocimiento de una forma sencilla, describiendo los términos en múltiples idiomas y relacionando unos con otros por sus relaciones semánticas de jerarquía, dependencia, abstracción, concreción o similitud. A través de SKOS se pueden definir modelos de la realidad mediante esquemas de conceptos sobre dominios específicos.

En la actualidad ya son muchos los organismos de referencia que han creado y expuesto los esquemas de conceptos y vocabularios que gestionan. Entre ellos destacan los vocabularios y taxonomías publicadas por la Biblioteca del Congreso de los EE.UU.[12], la Biblioteca nacional alemana[13], la húngara[14], la española o la francesa[15], la NASA (Ashish, 2005), la FAO, el New York Times[16], la Oficina de Publicación de la Unión Europea[17] o varios gobiernos en todo el mundo.

Como referencia mundial, la Biblioteca del Congreso de los EE.UU. ha puesto a disposición de la comunidad mundial varias colecciones de taxonomías y tesauros, útiles para la descripción de propiedades de recursos bibliográficos u otros ámbitos. Mantiene conjuntos de datos descritos en RDF y SKOS en varios idiomas, que se han convertido en referencia mundial: una completa taxonomía de temas[18], las familias de idiomas con sus dialectos[19], todas las áreas geográficas[20], o una forma estándar para representar el formato de las fechas[21].

Web Semántica en bibliotecas y archivos

La gestión del conocimiento de bibliotecas y archivos es una de las principales aplicaciones de la Web Semántica. El uso de estas tecnologías permiten una gestión eficiente de las colecciones de documentos que custodian estas instituciones, así como una mayor efectividad en la exposición de los recursos para consulta y reutilización por parte de agentes externos.

De hecho, los profesionales de este campo son grandes expertos en clasificación y catalogación, algo necesario a la hora de aplicar los vocabularios de metadatos y los esquemas de clasificación adecuados. Un claro ejemplo de esta convergencia de sinergias es la Iniciativa de Metadatos de Dublin Core, quien estandarizó una serie de propiedades que describen recursos bibliográficos que pueden ser representadas en RDF (Nilsson, Powell, Johnston, & Naeve, 2008). Este es el conjunto de términos más usado en todo el mundo para la descripción de recursos semánticos[22].

Una implementación del concepto de lo que podría ser el futuro de las bibliotecas y archivos puede verse en la iniciativa Europeana[23], una plataforma impulsada por la fundación que lleva el mismo nombre, y que ha puesto a disposición del público una biblioteca virtual que aglutina más de treinta millones de recursos culturales y científicos procedentes de miles de museos, bibliotecas, archivos y galerías de arte en Europa[24].

La mayor parte de los objetos que se listan en Europeana[25] son imágenes y textos, aunque también se almacenan otros materiales audiovisuales. Esta colección virtual tiene un valor incalculable para toda la sociedad, pero donde realmente está el valor es en la accesibilidad de la información. Todos los objetos presentes en esta gran base de datos, contienen metadatos con información descriptiva basada en los datos enlazados (Isaac & Haslhofer, 2012).

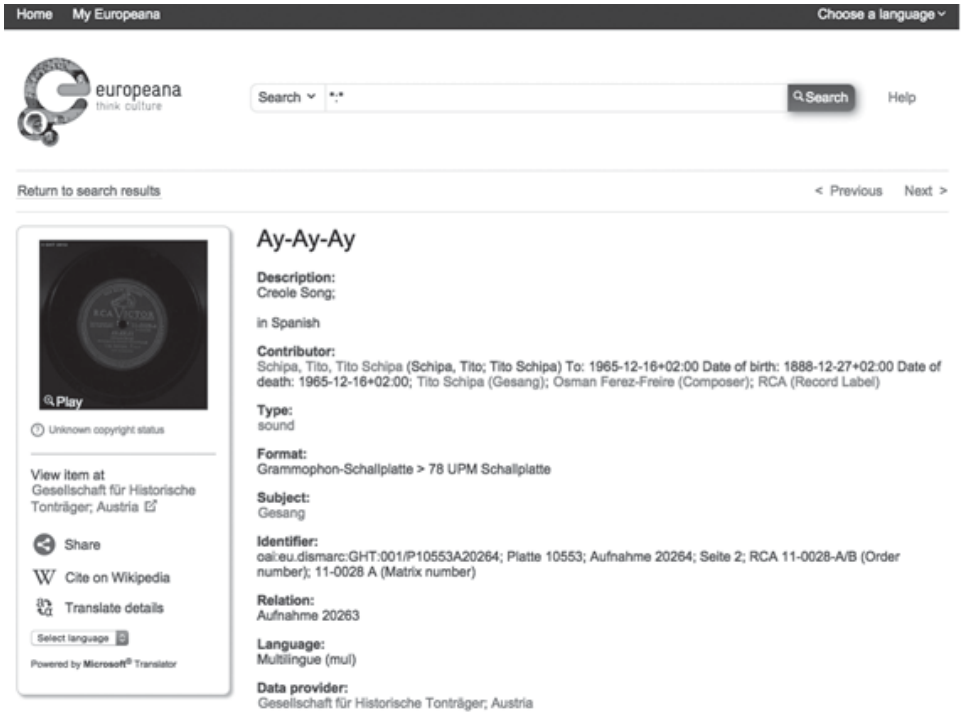

Fig. 13. Registro de audio catalogado en Europeana

Europeana, para homogeneizar los metadatos de los objetos han creado un vocabulario llamado **Modelo de Datos de Europeana** o **EDM**[26] basado en los estándares comúnmente usados en museos y bibliotecas —como LIDO[27], EAD[28] o METS[29]— y alineado con los principios de la Web Semántica.

EDM usa componentes de otros vocabularios como Dublin Core[30], **DCAT** —para describir catálogos de datos (Maali, Erickson, & Archer, 2014)—, Creative Commons[31] —descripción de licencias—, u ORE—para intercambio y reutilización de objetos de OAI—[32]. Además de esto, para representar la información relativa a la cobertura geográfica, Europeana utiliza los datos procedentes del repositorio de Geonames. De esta forma se aprecia cómo los distintos dominios de información son relacionados desde un vocabulario común y reutilizados para que formen parte del propio modelo de la aplicación, permitiendo la interoperabilidad plena con otros sistemas que sigan este esquema.

Como se ha mencionado, la **Biblioteca Nacional de España** (BNE) también gestiona diversos recursos en la nube de los datos enlazados. A través de su portal de datos abiertos[33], la BNE ofrece un tesauro con los principales temas bibliográficos (por ejemplo, *http://datos.bne.es/tema/XX525252* es el URI que identifica el tema 'bioquímica'), que a su vez están relacionados con la lista de temáticas ofrecida por la Biblioteca del Congreso de los EE.UU.

Aparte de la lista de temáticas, la BNE también gestiona obras y autores con las relaciones con otras fuentes de datos externas, como la Biblioteca Nacional Alemana, DBpedia o Libris[34]. Esto les permite ofrecer un portal de consulta con datos enriquecidos —por ejemplo, con información sobre los autores procedente de la DBpedia, como se puede apreciar en la descripción de un autor representado en la ilustración siguiente (Fig. 14)—.

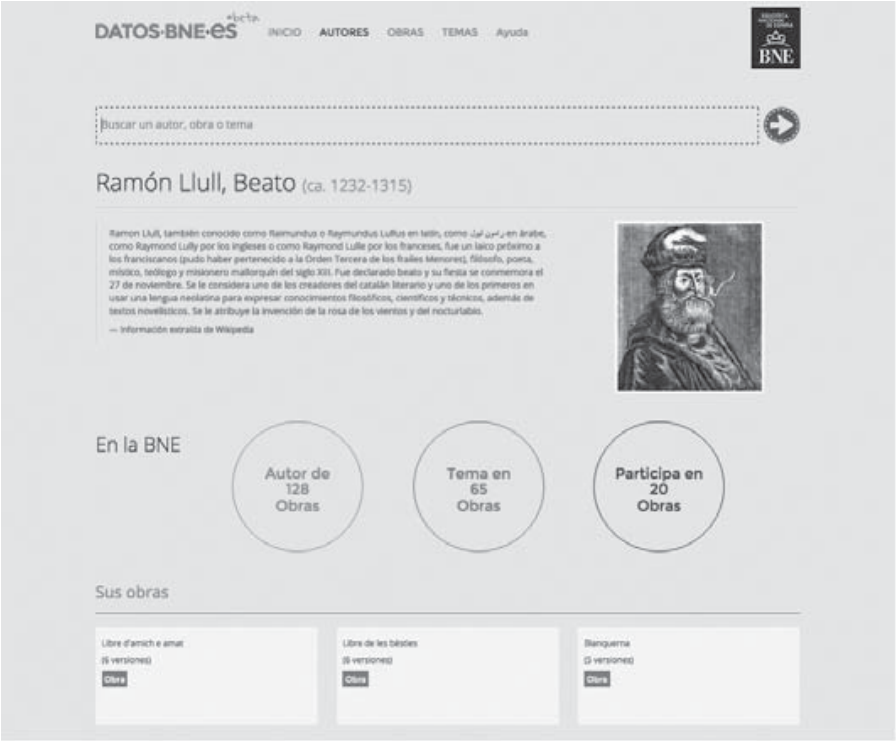

Fig. 14. Información sobre Ramón Lull y sus obras en la BNE

De la misma forma que lo hace Europeana, los recursos de la BNE y sus metadatos descriptivos están disponibles para su reutilización. Su conjunto de datos está enlazado con otros conjuntos de datos distribuidos en la Web, como se puede apreciar en el detalle de la nube de datos enlazados (Fig. 15).

En definitiva, la evolución de las bibliotecas y archivos digitales pasa por la aplicación de los principios de la Web Semántica y los datos enlazados, de forma

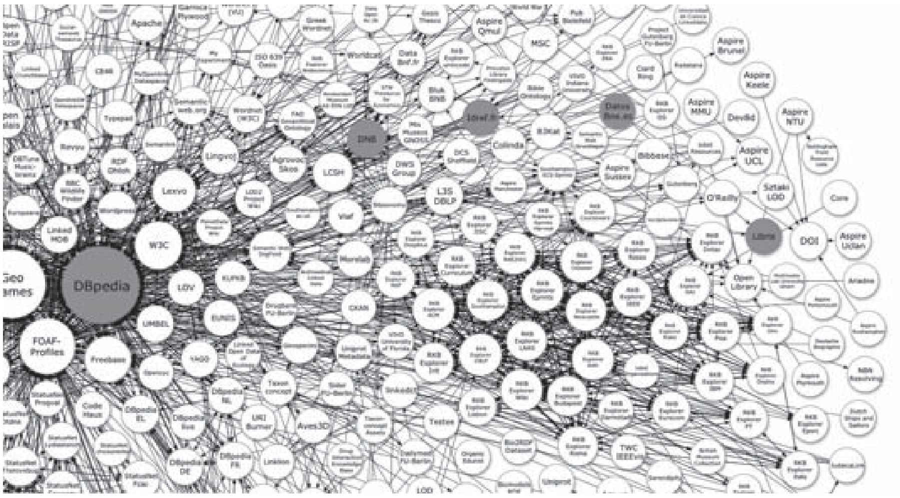

Fig. 15. Detalle de la nube de datos enlazados con las relaciones de BNE resaltadas

que se asegure una mejora en la eficiencia de la gestión de los documentos y metadatos, así como una interoperabilidad universal efectiva, con independencia del idioma, cultura o cualquier aspecto tecnológico.

Gracias a estas tecnologías descritas, lideradas por el modelo RDF y el uso de esquemas y vocabularios consensuados, se asegurará la interoperabilidad semántica plena, permitiendo el crecimiento de una Web robusta y eficiente.

Bibliografía

Alpert, J., & Hajaj, N. (2008). Official Google Blog: We knew the web was big... Retrieved March 23, 2015, from http://googleblog.blogspot.com.es/2008/07/we-knew-web-was-big.html

Ashish, N. (2005). Semantic-Web Technology : Applications at NASA 2 NASA Semantic Web based Systems. *Integration The Vlsi Journal*, 1–7.

Beckett, D. (2014). *RDF 1.1 N-Triples*. *W3C*. Retrieved from http://www.w3.org/TR/n-triples/

Beckett, D., Berners-Lee, T., Prud'hommeaux, E., & Carothers, G. (2014). *RDF 1.1 Turtle*. Retrieved from http://www.w3.org/TR/turtle/

Berners-Lee, T. Information Management: A Proposal, W3C (1990). Retrieved from http://www.w3.org/History/1989/proposal.html

Berners-Lee, T., Hendler, J., & Lassila, O. (2010). The Semantic Web. *Scientific American*. Retrieved from http://www.cs.umd.edu/~golbeck/LBSC690/SemanticWeb.html

Berrueta, D., & Phipps, J. (2008). *Best Practice Recipes for Publishing RDF Vocabularies*. Retrieved from http://www.w3.org/TR/swbp-vocab-pub/

Bizer, C., & Cyganiak, R. (2014). *RDF 1.1 TriG*. Retrieved from http://www.w3.org/TR/trig/

Bizer, C., Cyganiak, R., & Heath, T. (2008). *How to publish Linked Data on the Web*. Retrieved from http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/

Booth, D., Haas, H., McCabe, F., Newcomer, E., Champion, M., Ferris, C., & Orchard, D. (2004). *Web Services Architecture*. Retrieved from http://www.w3.org/TR/ws-arch/

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., Yergeau, F., & Cowan, J. (2006). *Extensible Markup Language (XML) 1.1 (Second Edition)*. Retrieved from http://www.w3.org/TR/xml11

Brickley, D., & Guha, R. V. (2014). *RDF Schema 1.1*. Retrieved from http://www.w3.org/TR/rdf-schema/

Brickley, D., & Miller, L. (2014). *FOAF Vocabulary Specification*. Retrieved from http://xmlns.com/foaf/spec/

Carothers, G. (2014). *RDF 1.1 N-Quads*. Retrieved from http://www.w3.org/TR/n-quads/

Connolly, D. W. (2000). *A Little History of the World Wide Web*. *W3C*. Retrieved from http://www.w3.org/History.html

Cyganiak, R. (2014). *The Linking Open Data cloud diagram*. Retrieved from http://lod-cloud.net/

Cyganiak, R., Wood, D., & Lanthaler, M. (2014). *RDF 1.1 Concepts and Abstract Syntax*. *W3C*. Retrieved from http://www.w3.org/TR/rdf11-concepts/#dfn-iri

Gandon, F., & Schreiber, G. (2014). *RDF 1.1 XML Syntax*. Retrieved from http://www.w3.org/TR/rdf-syntax-grammar/

Herman, I., Adida, B., Sporny, M., & Birbeck, M. (2015). *RDFa 1.1 Primer - Third Edition*. Retrieved from http://www.w3.org/TR/xhtml-rdfa-primer/

Iannella, R., & McKinney, J. (2014). vCard Ontology - for describing People and Organizations. Retrieved June 2, 2015, from http://www.w3.org/TR/vcard-rdf/

Isaac, A., & Haslhofer, B. (2012). Europeana linked open data–data. europeana. eu. *Semantic Web*, *0*(0). Retrieved from http://iospress.metapress.com/index/YWK6780L76800887.pdf

Isaac, A., & Summers, E. (2009). SKOS Simple Knowledge Organization System Primer. Retrieved June 2, 2015, from http://www.w3.org/TR/skos-primer/

ISO 8879:1986 - Standard Generalized Markup Language (SGML). (1986). Retrieved from http://www.iso.org/iso/catalogue_detail.htm?csnumber=16387

Maali, F., Erickson, J., & Archer, P. (2014). *Data Catalog Vocabulary (DCAT)*. Retrieved from http://www.w3.org/TR/vocab-dcat/

Manola, F., Miller, E., & McBride, B. (2014). *RDF 1.1 Primer*. Retrieved from http://www.w3.org/TR/rdf11-primer/

Nilsson, M., Powell, A., Johnston, P., & Naeve, A. (2008). *Expressing Dublin Core metadata using the Resource Description Framework (RDF)*. Retrieved from http://dublincore.org/documents/dc-rdf/

OWL 2 Web Ontology Language Document Overview (Second Edition). (2012). Retrieved June 2, 2015, from http://www.w3.org/TR/owl2-overview/

Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., & Lindström, N. (2014). *JSON-LD 1.0*. Retrieved from http://www.w3.org/TR/json-ld/

Notas

[1] World Wide Web Consortium: http://www.w3c.es

[2] CHiPs (IMDb): http://www.imdb.com/title/tt0075488/

[3] Gracias a la incorporación de tecnologías semánticas, los navegadores hoy en día ya distinguen entre los distintos conceptos de los resultados que ofrecen, por lo que habría que matizar el tipo de contenido buscado.

[4] Estrictamente los identificadores pueden contener caracteres especiales adecuados a los distintos juegos de caracteres para cualquier idioma, considerándose IRI, o Identificadores de Recursos Internacionalizados.

[5] Puede darse el caso en que nodos que no tengan relevancia se representen como nodos en blanco, recursos que no tienen URI asociado.

[6] Aunque parezca confuso, los literales pueden contener también direcciones Web, y estas pueden coincidir con los URI que identifican a otros recursos.

[7] Boston en la Wikipedia: http://en.wikipedia.org/wiki/Boston

[8] Geonames: http://www.geonames.org

[9] schema.org: http://schema.org

[10] Búsqueda de la cadena "World Wide Web Consortium" en Google: https://www.google.com/#q=World+Wide+Web+Consortium

[11] DCMI Terms: http://dublincore.org/documents/dcmi-terms/

[12] Library of Congress - Technical Center: http://id.loc.gov/techcenter/metadata.html

[13] German National Library Linked Data Service: http://www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkeddata_node.html

[14] Biblioteca Nacional de Hungría: http://nektar2.oszk.hu

[15] Open Data de la Bibliothèque nationale de France: http://data.bnf.fr

[16] Linked Open Data. The New York Times: http://data.nytimes.com

[17] EU Law and Publications: https://publications.europa.eu

[18] Library of Congress Subject Headings: http://id.loc.gov/authorities/subjects.html

[19] ISO 639-5 Codes for the Representation of Names of Languages: http://id.loc.gov/vocabulary/iso639-5.html

[20] MARC List for Geographic Areas: http://id.loc.gov/vocabulary/geographicAreas.html

[21] Extended Date/Time Format Datatypes Scheme: http://id.loc.gov/datatypes/edtf.html

[22] Linked Open Vocabularies - DCMI Metadata Terms (dcterms): http://lov.okfn.org/dataset/lov/vocabs/dcterms

[23] Europeana: http://www.europeana.eu

[24] Facts & Figures Europeana: http://pro.europeana.eu/about-us/factsfigures

[25] http://statistics.europeana.eu/page/content/2014/content-2014

[26] Europeana Data Model Documentation: http://pro.europeana.eu/share-your-data/data-guidelines/edm-documentation

[27] LIDO - Lightweight Information Describing Objects Version 1.0: http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf

[28] EAD; http://www.loc.gov/ead/eadabout.html

[29] METS schema: http://www.loc.gov/standards/mets/

[30] DCMI Metadata Terms: http://dublincore.org/documents/2012/06/14/dcmi-terms/

[31] Creative Commons: https://creativecommons.org

[32] The OAI Object Reuse and Exchange (ORE): http://www.openarchives.org/ore/terms/

[33] Portal de datos de la Biblioteca Nacional de España: http://datos.bne.es

[34] Libris: http://libris.kb.se