Marcado semántico: tecnologías y aplicación para la representación de sistemas de organización del conocimiento en el contexto Linked Open Data

Semantic Markup: technologies and application for the representation of knowledge organization systems in the context of Linked Open Data

Juan Antonio PASTOR SÁNCHEZ

Universidad de Murcia, España, pastor@um.es

Resumen

Se aborda la aplicación del marcado semántico de contenidos web para sistemas de organización del conocimiento. Para ello, se analiza el concepto se marcado semántico web, así como los fundamentos de los principales formatos de marcado semántico. También ofrece una visión general de la integración del marcado semántico en los sistemas de gestión de contenidos. Se contextualiza el marcado semántico a partir de la Arquitectura de la Web Semántica y los principios de Linked Open Data, para exponer un caso de uso de marcado semántico web en una implementación del Tesuro de la UNESCO, modelado con SKOS. Se concluye indicando la necesidad de la aplicación del marcado semántico en servicios web de consulta de vocabularios controlados, como un medio para mejorar la interoperabilidad y reutilización de dichos instrumentos.

Palabras clave: Linked Open Data. Marcado semántico web. Microdatos. Microformatos. RDFa. SKOS. Tesauro de la UNESCO. Web Semántica.

Abstract

This article focuses on the application of web content semantic markup for knowledge organization systems. The Web Semantic Markup concept and the fundamentals of the main semantic markup formats are analized. It also provides an overview of the integration of Semantic Markup in Content Management Systems. The Web Semantic Markup is contextualised into the principles of the Semantic Web Architecture and Linked Open Data, and its use is tested for the publication of the UNESCO Thesaurus as Linked Open Data, modelled with SKOS. As a conclusion, the need for the implementation of Web Semantic Markup in controlled-vocabularies-based on-line services is emphasised as a means to improve interoperability and reuse of these knowledge organization systems.

Keywords: Linked Open Data. Web Semántic Markup. Microdata. Microformats. RDFa. SKOS. UNESCO Thesaurus. Semantic Web.

1. Introducción

Actualmente la Web se enfrenta a dos retos: definir el significado semántico de los contenidos y facilitar la reutilización de la información de los mismos. La Web Semántica ha abordado desde el principio ambas problemáticas desarrollando tecnologías abiertas para incrementar la interoperabilidad de los datos y describiendo recursos de información mediante metadatos y ontologías (Obrst, 2003). En cualquier caso, se parte de un enfoque que aborda la formalización de estructuras de representación del conocimiento. Como consecuencia de todo ello se redefinió en cierto sentido la Web, planteando una arquitectura multinivel en la que se establecen diferentes capas de abstracción.

El modelo de datos RDF (Manola y Miller, 2004) conforma la base para representar recursos en la Web Semántica, aplicando simultáneamente diferentes modelos descriptivos. A través del establecimiento de estructuras atómicas de datos —que conforman tripletas sujeto-predicado-objeto— es posible plantear una separación entre el modelo de datos (en este caso RDF), el modelo descriptivo y el formato para su representación o almacenamiento en entorno digital. Por ejemplo: Dublin Core (modelo descriptivo) puede representarse mediante RDF (modelo de datos) y serializarse utilizando RDF/XML, Notation3 o Turtle (Klyne, Carroll y McBride, 2004) entre otros formatos. Muchos modelos descriptivos (Dublin Core, FOAF, SIOC) disponen de su correspondiente vocabulario RDF que permite su aplicación en el entorno de la Web Semántica.

Las ontologías web plantean descripciones más avanzadas desde el punto de vista formal y se sustancian aplicando alguno de los diferentes

sabores de OWL (Cuenca Grau et al., 2008). Con las ontologías web es posible crear representaciones muy elaboradas de relaciones entre recursos y con un alto nivel de formalización lógica.

Por su parte, la explotación de los conjuntos de datos RDF generalmente se realiza a través de la recuperación selectiva de datos mediante el protocolo/lenguaje de consulta SPARQL. El uso combinado de SPARQL (Langegger, Blochl y Woss, 2007), motores de inferencia y almacenes de datos RDF *(triplestore)* conforman una infraestructura tecnológica con un enorme potencial para la creación de nuevos sistemas de recuperación y reutilización de información semántica.

Este nuevo panorama plantea un escenario en el que los contenidos, generalmente diseñados para su consulta por personas, conviven en la Web con conjuntos de datos que utilizan formatos abiertos para su procesamiento por aplicaciones informáticas. Bien podría plantearse la metáfora de la Web como una gran esfera de todo tipo de datos. Los contenidos diseñados para las personas se localizan en la superficie; y el interior de dicha esfera alberga datos procesables directamente por máquina. Este símil es más interesante si se piensa que —según aumenta el volumen de la esfera, gracias a la proliferación de conjuntos de datos— mayores son las oportunidades para definir y desarrollar nuevos servicios y sistemas en la superficie dirigidos a los usuarios. Pero incluso resulta más sugerente si se plantea el mismo razonamiento a la inversa: el aumento de la superficie de las esfera podría incidir directamente en una mayor proliferación de conjuntos de datos.

Precisamente el marcado semántico Web responde a esta última idea y a una necesidad: la reutilización de la información de los contenidos web (Tomberg y Laanpere, 2009). La Web se desarrolló como un sistema para la comunicación de información cuya estructura se orientaba a su consulta por parte de personas. La aplicación de estilos de visualización ha sido el centro de gravedad en torno al cual han girado los procesos de diseño de contenidos. Las prácticas más extendidas en la aplicación de las tecnologías de la Web Semántica muestran una Web para las personas (la de los contenidos) y una Web para las máquinas (la de los datos). Esta separación también puede conducir a una desconexión entre ambas realidades.

El marcado semántico propone un puente en dos sentidos. El primero de ellos conduce a la estructuración y el marcado de los contenidos web de forma que la información sea semánticamente interoperable y, en consecuencia reutilizable. El segundo recorre el mismo camino pero sentido inverso ya que tiene como destino la reutilización de conjuntos de datos y la aplicación de servicios para enriquecer semánticamente los contenidos web. El primero de ellos nos muestra la tierra prometida de los contenidos web totalmente reutilizables y procesables de forma automática, para extraer su significado exacto utilizando servicios y aplicaciones como, por ejemplo, los motores de búsqueda. El segundo rompe con el tradicional aislamiento de los conjuntos de datos y por ende con la separación entre la Web de las personas y la Web de las máquinas.

El objetivo de este trabajo es ofrecer una visión general sobre las tecnologías aplicadas en el marcado semántico de contenidos web y su aplicación en la representación de sistemas de organización del conocimiento en la Web. Se realiza una introducción a las bases conceptuales y principales formatos del marcado semántico. Se exponen las tecnologías de la Web Semántica implicadas en la publicación Linked Open Data de conjuntos de datos en SKOS y finalmente se muestra un caso de uso de marcado semántico web basado en el Tesauro de la UNESCO.

2. El marcado semántico de contenidos web: concepto y formatos

El marcado de contenidos web generalmente ha utilizado lenguajes como HTML o XHTML (1), cuyo objetivo es la creación de documentos a partir de una serie de elementos estructurales básicos. Dichos elementos suelen abarcar funcionalidades relacionadas con su posterior representación visual para su lectura.

Tomemos el siguiente ejemplo:

```
<h1>Paul Dirac</h1>
<p>Paul Adrien Maurice Dirac (8 de agosto de
1902 - 20 de octubre de 1984) fue un físico
teórico británico nacido en Bristol
(Inglaterra) que contribuyó al desarrollo de
la mecánica y electrodinámica cuánticas y
compartió el premio Nobel de física de 1933
con Erwin Schrödinger.</p>
```

La etiqueta <h1> tiene una función clara: expresar el título de un documento. Mientras que por su parte la marca para la definición de párrafos <p> se utiliza para identificar unidades de contenido dentro de un documento (X)HTML. Sin embargo, la semántica de dichas etiquetas es aportada por el lector de dicho código, ya sea un usuario a través de un navegador web o una aplicación informática.

De hecho, el uso que se hace de la etiqueta <h1> puede variar: algunos CMS la utilizan para representar el nombre de un sitio web, reservando la etiqueta <h2> para el título de las páginas del sitio. El uso de párrafos vacíos fue una costumbre muy extendida para obtener separaciones adicionales entre los mismos. Las pautas de accesibilidad WCAG (Caldwell et al., 2008) pusieron orden y concierto en el uso de elementos estructurales de (X)HTML.

De esa forma, y poco a poco, han desaparecido ciertas prácticas que hacían uso de elementos estructurales con fines visuales o de maquetación. Algunos elementos de HTML5 (como <nav>, <footer>, <header>, <article>, <section>, <summary>) únicamente permiten definir la semántica funcional de un bloque de una página web (Berjon et al., 2013). Es una primera capa muy básica que no permite llegar al nivel de detalle necesario para la reutilización de la totalidad del contenido informativo de una página.

Volviendo de nuevo al párrafo del ejemplo anterior, puede observarse que sería posible extraer los siguientes datos estructurados:

- El nombre completo de una persona: Paul Adrien Maurice Dirac.
- Su fecha de nacimiento: 8 de agosto de 1902.
- Su fecha de fallecimiento: 20 de octubre de 1984.
- Su profesión: físico teórico.
- El lugar de su nacimiento: Bristol (Inglaterra).
- Recibió el premio Nobel de física en 1933.
- Dicho galardón fue compartido con Erwin Schrödinger.

Como puede verse, el “texto plano” del párrafo contiene gran cantidad de información para cuya identificación y estructuración no es suficiente con los elementos de (X)HTML. Se precisan de otros elementos estructurales definir sin ambigüedades el significado de determinados fragmentos de dicho “texto plano”. Desde un punto de vista técnico, el marcado semántico consiste en añadir determinados atributos a las etiquetas (X)HTML que estructuran un documento. A partir de entonces, las mismas etiquetas utilizadas para organizar la información para su lectura por personas, también podrían ser procesadas por aplicaciones informáticas para extraer el significado de los contenidos de la página.

Por otro lado, el marcado semántico de contenidos web no solamente conlleva la aplicación de una serie de tecnologías o formatos. También implica la estructuración adecuada de estos contenidos atendiendo a una serie de criterios de organización lógica, tanto jerárquicos como asociativos, y la aplicación de metadatos y ontologías para su descripción y clasificación. Partiendo de un documento (X)HTML válido, el marcado semántico implica los siguientes pasos:

- Identificación del objeto u objetos sobre los que se van a realizar una o varias declaraciones.
- Definir la taxonomía de dicho objeto desde el punto de vista conceptual: persona, lugar, objeto de arte, tema, sentimiento, documento.
- Identificar las características o propiedades del objeto u objetos susceptibles de ser descritos.
- Identificar posibles relaciones del objeto descrito con otros objetos.
- Seleccionar los esquemas de metadatos u ontologías más adecuadas para representar las declaraciones identificadas previamente.
- Realizar el marcado semántico añadiendo los atributos (X)HTML necesarios para ello.

En principio el formato elegido no debería condicionar el proceso conceptual del marcado semántico de un contenido. Sin embargo, en la práctica la elección de microformatos, microdatos o RDFa imponen una serie de limitaciones que se mostrarán en detalle a continuación.

2.1. Microformatos

Los microformatos tienen como objetivo incluir datos estructurados en documentos (X)HTML. Su objetivo es utilizar estándares existentes para realizar un marcado simple de datos en documentos legibles para las personas, de modo que posteriormente pueda ser reutilizada fácilmente por aplicaciones informáticas (Khare y Çelik, 2006).

Básicamente los microformatos son una forma de aplicar nomenclaturas y convenciones simples en atributos HTML ya existentes para definir la semántica de dichos datos (Méndez, Bravo y López, 2007).

Más concretamente, los microformatos reutilizan los siguientes atributos (X)HTML:

- class: Indica el tipo de objeto sobre el que se realiza la descripción.

- rel: Se utiliza en los hiperenlaces para expresar el tipo de relación que se establece entre el objeto descrito y la URL de destino.
- rev: Se utiliza en los hiperenlaces pero el tipo de relación se define en sentido contrario a "rel", es decir, desde la URL destino hacia el objeto descrito.
- title: Ofrece una alternativa procesable por máquina a un texto marcado legible para personas.

Veamos un ejemplo muy sencillo de uso de microformatos:

```
<div class="vcard">
<h1 class="fn">Paul Dirac</h1>
<p>
<span class="given-name">Paul</span> <span
class="additional-name">Adrien</span>
<span class="additional-name">
Maurice</span>
<span class="family-name">Dirac (<time
class="bday" title="1902-08-08">8 de agosto
de 1902</time> - <time class="dday"
title="1984-10-20">20 de octubre de
1984</time>) fue un
<span class="category">físico</span> teórico
británico nacido en
<span class="birthplace">Bristol
(Inglaterra)</span> que contribuyó al
desarrollo de la mecánica y electrodinámica
cuánticas y compartió el premio Nobel de
física de 1933 con Erwin Schrödinger.</p>
</div>
```

Como puede verse, el atributo clave para el marcado semántico con microformatos es "class", ya que a partir de él se definen los tipos y propiedades del objeto descrito.

Existen distintas especificaciones de microformatos:

- hCalendar: Se utiliza para la descripción de eventos y actividades.
- hCard: Se utiliza para la descripción de personas, organizaciones y contactos. El ejemplo anterior hace uso de esta especificación.
- rel-license: Indica que un enlace apunta a una licencia de uso de un contenido.
- rel-nofollow: Indica que un enlace no debe ser utilizado en procesos de análisis web, como los realizados por los motores de búsqueda.
- rel-tag: Indica que un enlace apunta a una categoría temática bajo la que se ha clasificado el contenido.
- XFN (XHTML Friends Network): Permite definir los enlaces entre personas que conforman una red social.
- XMDP (XHTML Meta Data Profiles): Se utiliza para la definición de perfiles de metadatos legibles por personas y procesable por máquinas.
- XOXO (Extensible Open XHTML Outlines): Es un perfil XMDP para embeber en el código XHTML entradas RSS, Atom o XML en general.

Las anteriores especificaciones se consideran estables y aceptadas por la comunidad de microformatos. Otras muchas se encuentran actualmente en proceso de elaboración (2) para el marcado de recetas, ubicaciones geográficas, entradas RSS, etc. En resumidas cuentas se tratan de modelos de descripción adaptados por convención al marcado (X)HTML.

Los microformatos han alcanzado una gran difusión debido fundamentalmente a su sencillez de aplicación. No obstante, tienen limitaciones en cuanto al modelo de datos utilizado (exclusivamente jerárquico), la imposibilidad de definir recursos referenciables mediante URIs y la dificultad para definir nuevos vocabularios o para combinar otros ya existentes (Sporny, 2011). Desde el punto de vista de la organización del conocimiento, los microformatos carecen de la base sólida de una semántica formal y por tanto resulta difícil mapear los datos estructurados marcados con microformatos a tripletas (declaraciones de sujeto-predicado-objeto) que contengan toda la potencialidad expresiva de RDF.

2.2. RDFa

RDFa (Resource Description Framework in attributes) es un mecanismo genérico para la inclusión de sentencias RDF mediante atributos de elementos de marcado en lenguajes basados en XML (Adida et al., 2012). No obstante, la última especificación de RDFa 1.1 contempla el uso de RDFa en HTML (3).

RDF es el modelo de datos que conforma el núcleo de la Web Semántica. Las descripciones de recursos se realizan mediante tripletas RDF. Estas tripletas son declaraciones formadas por tres componentes: sujeto, predicado y objeto. El recurso sobre el que se realiza la descripción es el sujeto y puede clasificarse mediante taxonomías de clases y subclases. El predicado permite representar propiedades del sujeto o relaciones de éste con otros recursos. El objeto es el valor de la propiedad o el recurso con el que se relaciona el sujeto.

En RDF los recursos descritos se identifican mediante una URI. Dicha URI puede además ser utilizada para acceder al recurso, en cuyo caso se denomina URL. Las propiedades, rela-

ciones y taxonomías de los vocabularios tienen una semántica bien definida y también se identifican mediante una URI. Por ejemplo, la URI <http://purl.org/dc/terms/> identifica el espacio de nombres del esquema de metadatos de Dublin Core. El elemento "title" de dicho esquema, se identifica con la URI <http://purl.org/dc/terms/title>. Es práctica común abreviar las URIs de los espacios de nombres utilizando prefijos. En el caso de Dublin Core cualificado suele utilizarse como prefijo "dcterms" de forma que la referencia al elemento title podría realizarse mediante la expresión "dcterms:title". Así pues, en el ámbito de la Web Semántica, no hay dos recursos, clases, propiedades o relaciones que puedan presentar algún tipo de ambigüedad en cuanto a su identificación puesto que cada una de esas entidades se identifican mediante URIs diferentes.

Actualmente, la especificación de RDFa se denomina RDFa 1.1 Core. Sin embargo su aplicación resulta algo compleja por lo que el W3C (World Wide Web Consortium) ha elaborado un subconjunto denominado RDFa 1.1 Lite (Sporny, 2012) aplicable a la mayor parte de las necesidades del marcado semántico con RDFa.

RDFa define una serie de nuevos atributos:

- about: identifica el recurso al que se refiere la descripción del marcado semántico, "sujeto" en terminología RDF.
- datatype: define el tipo de datos utilizado por un literal.
- inlist: se utiliza para indicar agrupar varios "objetos" a un mismo "sujeto" a través de un único atributo rel o property.
- prefix: permite definir equivalencias entre prefijos y espacios de nombres. De utilidad para abreviar referencias a espacios de nombres.
- property: se usa para expresar relaciones entre el recurso "sujeto" y cualquiera de los recursos "objeto" u "objetos" literales.
- resource: permite expresar el recurso "objeto" de una relación pero sin utilizar enlaces navegables.
- typeof: indica el tipo de clase del recurso "sujeto".
- vocab: permite hacer referencia a vocabularios para utilizar los elementos descriptivos de un modo sencillo para el marcado semántico.

A partir del ejemplo anterior en el que se han utilizado microformatos, el marcado mediante RDFa podría realizarse del siguiente modo:

```
<div prefix="schema: http://schema.org/
foaf: xmlns.com/foaf/0.1/"
typeof="schema:Person foaf:Person"
resource="http://www.ejemplo.org/Paul_Dirac"
>
<h1 property="schema:name foaf:name">Paul
Dirac</h1>
<p><span property="schema:givenName
foaf:givenName">Paul</span>
<span property="schema:additionalName">
Adrien</span>
<span property="schema:additionalName">
Maurice</span>
<span property="schema:familyName
foaf:familyName">Dirac (<span
property="schema:bithDate"
datatype="xsd:date" content="1902-08-08">8
de agosto de 1902</span> - <span
property="schema:deathDate"
datatype="xsd:date" content="1984-10-20">20
de octubre de 1984</span>) fue un <span
property="schema:jobTitle">
físico teórico</span>
<span property="schema:nacionality"
content="UK"> británico</span> nacido en
<span property="schema:birthplace">
Bristol (Inglaterra)</span> que contribuyó
al desarrollo de la mecánica y
electrodinámica cuánticas y compartió el
<span property="schema:award">premio Nobel
de física de 1933</span> con <span
rel="schema:knows foaf:knows"
typeof="schema:Person foaf:Person"
resource="http://www.ejemplo.org/Erwin_Schrö
dinger">
<span property="schema:name foaf:name">Erwin
Schrödinger</span>
</span>.</p>
</div>
```

RDFa es sin duda el formato más verboso. Ello se debe a su mayor potencia y flexibilidad de la capacidad de expresión semántica que la de los microformatos y microdatos.

En RDFa destacan la sencilla combinación de varios vocabularios en un mismo documento (en el ejemplo se han utilizado FOAF y Schema.org), la definición de prefijos y la facilidad para asociar varios tipos o valores de propiedades a un mismo elemento de marcado.

2.3. Microdatos

En 2004 el W3C había apostado por las tecnologías basadas en XML y por tanto parecía que abandonaba HTML en favor de la evolución de XHTML para el desarrollo XHTML 2. Por este motivo se creó el grupo de trabajo Whatwg para actualizar HTML con la consiguiente creación de HTML5. Los microdatos son la propuesta para el marcado semántico de dicho lenguaje ya que todo apuntaba a que RDFa no se adaptaría a HTML5. Finalmente el W3C abandonó XHTML 2 y actualmente coordina las labores de desarrollo

de HTML5 y de la correspondiente adaptación de RDFa.

La especificación de microdatos (Hickson, 2012) establece el marcado semántico a partir de los elementos HTML utilizados en el documento, con el objeto de describir un contenido específico de una página web.

Los microdatos definen cinco nuevos atributos que permiten el marcado semántico en HTML5:

- itemscope: Delimita un ítem de información del contenido web.
- itemtype: Indica el tipo de ítem (identificado mediante itemscope) sobre el que se realizará el marcado semántico. Este atributo contiene una URL válida que identifica unívocamente un tipo o clase de elemento definido mediante un vocabulario.
- itemid: Permite asociar un identificador al item descrito.
- itemprop: Permite especificar el elemento de un vocabulario que se refiere a una propiedad o atributo del contenido web que se está marcando.
- itemref: Permite referirse a ítems definidos en cualquier lugar del documento. Esto permite, por ejemplo, realizar asignaciones de propiedades a dichos ítems desde otros diferentes.

Volviendo al ejemplo anterior, se podría realizar el marcado con microdatos como se muestra a continuación:

```
<div itemscope itemid="#Paul_Dirac"
itemtype="http://schema.org/Person">
<h1 itemprop="name">Paul Dirac</h1>
<p>
 <span itemprop="givenName">Paul</span>
 <span itemprop="additionalName">Adrien
</span>
<span itemprop="additionalName">Maurice
</span>
<span itemprop="familyName">Dirac (<time
itemprop="bithDate" datetime="1902-08-08">8
de agosto de 1902</time> - <time
itemprop="deathDate" datetime="1984-10-
20">20 de octubre de 1984</time>) fue un
<span itemprop="jobTitle">físico
teórico</span>
<span itemprop="nacionality"
content="UK">británico</span> nacido en
<span class="birthplace">Bristol
(Inglaterra)</span> que contribuyó al
desarrollo de la mecánica y electrodinámica
cuánticas y compartió el <span
itemprop="award">premio Nobel de física de
1933</span> con
<span itemprop="knows" itemscope
itemtype="http://schema.org/Person">
<span itemprop="name"
itemid="Erwin_Schrödinger">Erwin
Schrödinger</span></span>.</p>
</div>
```

En el ejemplo se utiliza Schema.org como modelo descriptivo para expresar las diferentes propiedades de los objetos representados (en este caso Paul Dirac y Erwin Schrödinger").

El marcado semántico con metadatos sigue un modelo organizativo jerárquico, si bien el atributo itemref aporta mayor flexibilidad a este respecto a cambio de volver más complejo el código (X)HTML resultante. No permite el uso de tipos de datos y su uso se circunscribe únicamente a HTML5 y XHTML5. No permite incluir más de una sentencia en un mismo elemento y la imposibilidad de utilizar IRIs compactas y definir prefijos de espacios de nombres hace que, en algunas circunstancias, el código sea verboso en exceso. Tampoco existe en todos los casos un método directo para el mapeado del marcado semántico con sentencias RDF (Sporny, 2011; Tomberg y Laanpere, 2009).

No obstante, la aplicación del formato de microdatos en el marcado semántico es relativamente sencilla. Por otro lado, ha tenido una amplia difusión debido al apoyo recibido por parte de los principales motores de búsqueda (Google, Bing, Yahoo! y Yandex) a través del proyecto Schema.org (4), una iniciativa desarrollada conjuntamente por Google, Yahoo!, Bing y Yandex presentada a mediados de 2011.

La inclusión de metadatos estructurados es una de las soluciones que proponen los motores de búsqueda para la solución de la problemática que plantea la enorme proliferación de contenidos en la Web, cuya edición y publicación puede ser realizada por cualquier persona. Sin embargo, Schema.org no se detiene en la mera inclusión de metadatos generales (descripción, autor, fecha de publicación, palabras clave, etc) sino que persigue una descripción más profunda de los aspectos semánticos de los contenidos (García Marco, 2013). Aunque en teoría Schema.org también puede ser utilizado con RDFa (tal y como se ha visto en el ejemplo del epígrafe 2.2), los desarrolladores del proyecto recomiendan el uso de microdatos debido a su mayor facilidad de uso.

3. Integración del marcado semántico en los Sistemas de Gestión de Contenidos

Pese a la experiencia acumulada de las tecnologías y formatos de marcado semántico, éste resulta una tarea compleja, propensa a errores y de difícil mantenimiento. La amplia difusión de la tecnología de Sistemas de Gestión de Contenidos (CMS) sugiere la integración del marcado semántico en estos sistemas para simplificar y hacerlo más eficiente. La inclusión en el código (X)HTML de los atributos de microformatos,

microdatos o RDFa es un proceso que podría realizar un CMS de forma relativamente automática (Pastor, 2012). Generalmente la estructura de almacenamiento de información de estos sistemas utilizan ciertos campos descriptivos estructurados (título, autor, fecha de edición, categorías, resumen) y un campo de texto que contiene el cuerpo informativo de una página web. Algunos CMS, como Drupal, tienen un alto grado de flexibilidad para definir tipos de contenido y estructuras personalizadas de campos descriptivos.

La generación del marcado semántico en un CMS debe partir de un mapeado entre los campos descriptivos y los correspondientes elementos de las ontologías y esquemas de metadatos utilizados como modelos descriptivos. En mayor o menor medida los CMS hacen uso de extensiones, módulos o plugins que amplían sus funcionalidades para realizar el correspondiente marcado semántico en el código (X)HTML. El propio Drupal incorpora en su núcleo un marcado semántico básico (ampliable mediante módulos) en RDFa.

El marcado del cuerpo informativo de una página web es una tarea más compleja. Generalmente los CMS utilizan un editor visual para la creación y modificación de los contenidos. Algunas extensiones asisten al usuario que edita los contenidos permitiendo escoger elementos de varios modelos descriptivos para definir los elementos de marcado semántico, asignar sus correspondientes clases y describir sus propiedades y relaciones. También existen ciertos servicios basados en técnicas procesamiento del lenguaje natural (PLN), que ofrecen APIs para realizar el marcado semántico. Mediante determinadas extensiones de los CMS, es posible conectarse a estos servicios, suministrar el texto del contenido que se desea marcar y obtener el texto debidamente marcado. Algunos de estos servicios son:

- AlchemiAPI: SaaS (5) basado en técnicas de PLN y bases de datos estadísticas y lingüísticas que permite identificar personas, organizaciones, etiquetas, conceptos y relaciones entre distintos tipos de entidades contenidos en un texto e incluso definir equivalencias con conjuntos de datos Linked Open Data.
- Sindice: Es un servicio en línea que indiza fuentes de datos rdf. Permite buscar recursos RDF y utilizarlos para realizar declaraciones sobre los mismos en el marcado semántico (Oren et al, 2008).
- OpenCalais: Es un servicio web de Thompson-Reuters basado en técnicas de PLN capaz de identificar entidades, hechos y eventos de un texto.
- DBpedia Spotlight: ES un servicio que utiliza DBpedia e identifica entidades y su clase o clases correspondientes mediante técnicas de desambiguación semántica (Mendes et al, 2011).

El uso de estos servicios y de sus APIs suele integrarse con los CMS a través de extensiones. En Drupal podemos encontrar módulos como Markup Editor, OpenCalais y Alchemy. En MediaWiki la extensión Semantic MediaWiki permite incorporar marcado semántico a los hiperenlaces. Wordpress dispone del plugin RDFa-CE, que permite el marcado manual y automático utilizando los servicios descritos anteriormente (Khalili, Auer y Hladky, 2012).

En general es posible agrupar los plugins de marcado semántico según las funciones realizadas. En este sentido, la aplicación del marcado semántico en los CMS cubre cinco funciones principales (Martínez, 2013, p. 28):

- Integración de contenidos en el ámbito de las redes sociales mediante la aplicación del protocolo Open Graph (Ko et al., 2010).
- Inclusión de sistemas de navegación del tipo *breadcrumbs* (migas de pan) que indica indica la ruta de navegación recorrida por el usuario en el sitio web y además ubica un contenido en la estructura jerárquica del sitio web.
- Adición de descripciones estructuradas de contenidos para el uso de la tecnología *rich snippets* de google.
- Marcado semántico del cuerpo de una página web.
- Incorporación de metadatos descriptivos para mejorar el Posicionamiento SEO.

La introducción de datos estructurados semánticos como un añadido al cuerpo de los contenidos es una tarea bien resuelta por muchos plugins. Sin embargo, actualmente el marcado semántico del cuerpo de los contenidos se encuentra en un estado inicial de desarrollo (Martínez, 2013, p. 39) y en un futuro podría constituir la principal fuente para el desarrollo de conjuntos de datos RDF.

4. Aplicación del marcado semántico en los sistemas de organización del conocimiento

El marcado semántico podría aplicarse en gran cantidad de tesauros, clasificaciones, listas de

encabezamientos de materia y sistemas de organización del conocimiento en general, disponibles para su consulta a través de la Web. Son numerosos los vocabularios controlados publicados como Linked Open Data que aplican tecnologías de la Web Semántica (Pastor, Martínez y Rodríguez, 2012).

Sin embargo muy pocos de ellos aplican el marcado semántico. En esta sección se realiza una breve visión del contexto de la Web Semántica, Linked Open Data y SKOS como ontología estándar para la representación de sistemas de organización del conocimiento. Finalmente se analiza el caso de marcado semántico del Tesauro de la UNESCO.

4.1. El contexto de la Web Semántica y Linked Open Data

La Web se edificó sobre unos cimientos, conformados principalmente por la localización y el acceso a objetos utilizando direcciones URL, el uso del protocolo HTTP para la petición y transmisión de dichos objetos y la aplicación del lenguaje de marcado HTML para documentos legibles por las personas. La Web Semántica define nuevas capas superpuestas para que la interoperabilidad de la información a nivel sintáctico, estructural y semántico.

La Web Semántica conserva características importantes de la Web actual para identificar, localizar y codificar datos. UNICODE se utiliza para codificar caracteres. Los identificadores uniformes de recursos (URI, Uniform Resource Identifier) se aplican para definir espacios que identifican recursos abstractos o localizables (URL) en Internet. Las URIs son esenciales en la Web Semántica, ya que permiten unificar la localización de recursos con una nomenclatura común y descentralizada.

La familia de especificaciones XML (eXtensible Markup Language) ofrecen un formato para representar una sintaxis jerárquica usada para el intercambio de información por la mayor parte de las aplicaciones web. Los espacios de nombres XML (XML namespaces) se utilizan para referenciar y reutilizar recursos y modelos de descripción. También permiten utilizar varios modelos de descripción en un mismo documento, evitando conflictos debidos a que dos elementos de distintos esquemas tengan el mismo nombre.

Como ya se ha indicado anteriormente al analizar RDFa, el núcleo que hace posible la interoperabilidad semántica de los datos es RDF (Resource Description Framework). Se trata de un modelo de datos muy simple para identificar recursos y describir sus carácterísticas y relaciones con otros recursos. Por su parte, RDFS (RDF Schema) permite definir jerarquías de clases y propiedades de los recursos (Brickley y Guha, 2004). Finalmente, OWL (Web Ontology Language) se utiliza para la creación de ontologías con una semántica mucho más precisa que la de RDFS (Hitzler et al., 2012). Estas tres tecnologías (RDF, RDFS y OWL) se complementan mutuamente, conformando una semántica bien definida para su uso en la ejecución de inferencias mediante reglas lógicas de razonamiento expresadas en RIF (Rule Interchange Format). Los datos están disponibles en almacenes de datos RDF (triplestore) o en ficheros en algún formato que permita la serialización de RDF.

El conjunto de especificaciones SPARQL (SPARQL Protocol and RDF Query Language) define un lenguaje de consulta y gestión de datos RDF, así como un protocolo para realizar dichas operaciones y una serie de formatos para recuperar los datos (W3C SPARQL Working Group, 2012). Todo ello, permite crear servicios web SPARQL Endpoint que permiten la búsqueda en almacenes de datos RDF y la recuperación de las sentencias que cumplan las condiciones de la consulta.

Sin duda, una de aplicaciones más conocidas de las tecnologías de la Web Semántica es la publicación Linked Open Data para la reutilización de conjuntos de datos gestionados de forma distribuida. Los datos pueden ser recuperados a través del protocolo HTTP e interpretados por aplicaciones para su procesamiento o el descubrimiento de conexiones con otros conjuntos de datos a partir de los enlaces establecidos entre los mismos. Esta propuesta se basa en una serie de principios (Bizer, Heath y Berners-Lee, 2009):

- Utilizar URIs para nombrar de manera univoca objetos y recursos.
- Que se pueda acceder a esos datos a través del protocolo HTTP. Es decir, que la URI de un recursos permita también su consulta. Esto se denomina derreferenciación de URIs.
- Suministrar los datos en un formato adecuado en función del cliente que lo solicite. Para las personas es más fácil leer un documento en HTML a través de un navegador, mientras que para una aplicación informática (máquina) es más sencillo procesar datos RDF. Esta técnica se denomina negociación de contenido.
- Usar estándares y tecnologías abiertas para acceder y recuperar los datos. Se evita el

uso de formatos y tecnologías propietarias sujetas a cambios en función de las políticas corporativas de las empresas que las han desarrollado.

- Definir conexiones entre conjuntos de datos, que permitan "descubrir" recursos siguiendo los enlaces entre los mismos. De este modo se evitan duplicidades de datos y consiguientemente su publicación es más eficiente.

No todos los conjuntos de datos Linked Open Data cumplen todos los principios anteriores y generalmente suelen aplicar tecnologías de la Web Semántica.

4.2. SKOS

La ontología más extendida para la representación de sistemas de organización del conocimiento en la Web Semántica es SKOS (Miles y Bechhofer, 2009). Se trata de una ontología OWL y se basa en RDF. Puede aplicarse a cualquier tipo de vocabulario controlado como clasificaciones, tesauros, encabezamientos de materia, taxonomías, tesauros, etc.

En SKOS los elementos de un vocabulario se representan como recursos que pueden ser de tres clases: conceptos (skos:Concept), esquemas de conceptos (skos:ConceptScheme) y colecciones (skos:Collection). Los elementos principales son los conceptos que tienen asignadas distintos tipos de etiquetas en uno o varios idiomas:

- Etiquetas preferentes (skos:prefLabel): Son términos usados en la indización y cuya función es idéntica a la de los términos descriptores de los tesauros.
- Etiquetas alternativas (skos:altLabel): Utilizadas para representar términos sinónimos o cuasi-sinónimos de las etiquetas preferentes. Aportan riqueza léxica y definen diversos puntos de acceso a un concepto que puede representarse con diferentes términos.
- Etiquetas ocultas (skos:hiddenLabel): Su uso no está pensado para su consulta por personas, sino para su interpretación por aplicaciones informáticas para el control y tratamiento de variantes terminológicas, errores ortográficos, diversas formas de acrónimos y abreviaturas, etc.

Existen una serie de propiedades que permiten relacionar conceptos semánticamente:

- Relación jerárquica específica (skos:narrower): Permite conectar un concepto con otro que tiene un significado más específico.
- Relación jerárquica genérica (skos:broader): Es la propiedad inversa de la anterior.
- Relación asociativa (skos:related): Indica que dos conceptos están relacionados semánticamente sin que exista entre ambos un vínculo jerárquico.

Los conceptos pertenecen a esquemas de conceptos y pueden organizarse en colecciones que se utilizan para definir campos semánticos o facetas. Un mismo concepto puede formar parte de varias colecciones o esquemas de conceptos.

Con SKOS pueden definirse relaciones semánticas entre conceptos de diferentes esquemas. Estas relaciones de mapeado se utilizan para indicar que un concepto de un esquema se considera idéntico, o con un significado cercano, genérico, específico o relacionado. Suelen aplicarse para definir relaciones entre elementos de diferentes vocabularios controlados.

La recomendación SKOS define una serie de condiciones de consistencia que debe cumplir el conjunto de entidades y relaciones de un vocabulario, así como un conjunto de reglas que delimitan el ámbito de aplicación de los vocabularios en procesos lógicos de inferencia.

4.3. Un caso de uso del marcado semántico RDFa: el Tesauro de la UNESCO

El tesauro de la UNESCO se publicó en 1977 (Aitchison y Dextre-Clarke, 2004) principalmente para su uso en la consulta a bases de datos de dicha organización. Con el paso del tiempo su aplicación se ha extendido a otros ámbitos: desambigüación terminológica, fuente para elaborar otros tesauros (Dunsire, 2011), docencia, descripción de recursos educativos (García y Jaroszczuk, 2009), etc. La versión web data del año 2000 y actualmente se ofrece en cuatro idiomas: inglés, francés, español y ruso. Es un tesauro multidisciplinar, principalmente monojerárquico (a excepción de los descriptores geográficos) y basado en las normas ISO-2788 e ISO-5964 (Ewketu, 2011). Entre los términos preferentes y no-preferentes, cuyo número varía en función del idioma, se establecen relaciones de equivalencia (sinonimia o cuasi-sinonimia). Las relaciones jerárquicas y asociativas únicamente se establecen entre los términos preferentes. El tesauro consta de siete áreas o temas principales que a su vez albergan un total de 88 micro-tesauros. Actualmente el Tesauro de la UNESCO es conformado por más de 8.600 términos en español y francés, más de 7.100 términos en inglés y casi 7.000 en ruso.

La versión en SKOS del tesauro de la UNESCO se ha desarrollado en el ámbito del proyecto UNESKOS, en el que también se ha modelado la Nomenclatura Internacional de la UNESCO para los campos de Ciencia y Tecnología (8). El tesauro se representa como un esquema de conceptos y las áreas y micro-tesauros como colecciones. Las áreas se asocian al tesauro mediante la propiedad skos:inScheme (<Área> skos:inScheme <Tesauro>). Los micro-tesauros que componen un área de conocimiento se representan con la propiedad skos:member (<Área> skos:member <Micro-tesauro>). Los conceptos se asocian al esquema de conceptos y al micro-tesauro correspondiente con las propiedades skos:inScheme y skos:member respectivamente (<Concepto> skos:inScheme <Tesauro>; <Micro-tesauro> skos:member <Concepto>).

La arquitectura del sistema usa URIs derreferenciables a partir del espacio de nombres http://skos.um.es/unescothes, que se ha utilizado como prefijo para todos los elementos del tesauro.

La negociación de contenido se ha realizado a partir de las recomendaciones existentes (Sauermann y Cyganiak, 2008; Heath y Bizer, 2011). Cada elemento del vocabulario está asociado a su correspondiente URI neutra. A su vez cada formato tiene asignado un determinado sufijo que complementa dicha URI neutra. Por ejemplo, el concepto etiquetado con el término en Español “Información” tiene asignada la URI neutra <http://skos.um.es/unescothes/C01977>. La URL específica para recuperar los datos en formato RDF/XML es <http://skos.um.es/unescothes/C01977/rdfxml>. Si un usuario accede a través del navegador web a una URI neutra el servidor lo redirige automáticamente a la correspondiente URL de la versión HTML. La URL del ejemplo anterior sería <http://skos.um.es/unescothes/C01977/html>.

En el caso de otros clientes web, las solicitudes de objetos en un formato determinado a través de la URI neutra obtienen como respuesta por parte del servidor el código de estado HTTP “303 See others” junto con la URL donde se encuentran los datos en el formato solicitado. Si el cliente se conecta directamente a una URL correspondiente a un formato concreto, recupera los datos en dicho formato.

La implementación de la vista HTML del tesauro de la UNESCO incorpora el marcado semántico utilizando RDFa 1.1. El marcado usa SKOS para expresar las etiquetas y relaciones de los conceptos. Para ilustrar el modo en el que se ha implementado dicho marcado usaremos el mismo ejemplo anterior del término “Información”.

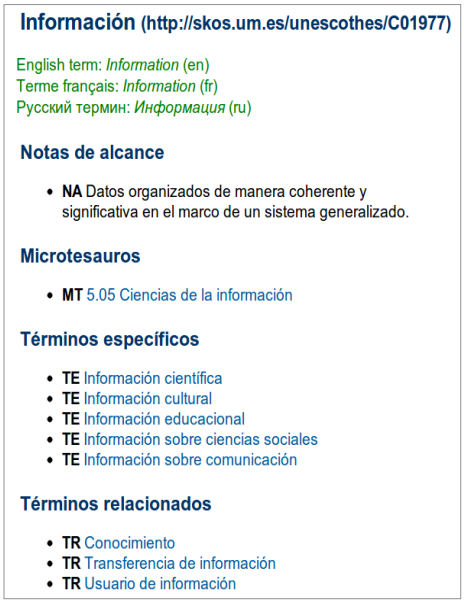

Figura 1. Ejemplo de visualización de la página web del término “Información” del Tesauro de la UNESCO (Fuente: http://skos.um.es/unescothes/C01977)

El código HTML correspondiente al inicio del bloque que muestra los datos del concepto se inicia con una etiqueta en la que se definen los prefijos a utilizar durante el marcado y la clase del objeto sobre el que se va a realizar el marcado (en este caso skos:Concept) y la etiqueta preferente en el idioma actual, el español, puesto que el navegador está consultando la versión del tesauro en dicho idioma.

```
<div id="node" prefix="skos:
http://www.w3.org/2004/02/skos/core#
unescothes: http://skos.um.es/unescothes/"
about="http://skos.um.es/unescothes/C01977"
typeof="skos:Concept">
<h2><span property="skos:prefLabel"
xml:lang="es">Información</span>
<small>(http://skos.um.es/unescothes/C01977)
</small></h2>
```

A continuación se incluyen las etiquetas preferentes del término en el resto de idiomas:

```
<dl>
<dt>English term:
<em property="skos:prefLabel"
xml:lang="en">Information</em> (en)</dt>
<dt>Terme français:
<em property="skos:prefLabel"
xml:lang="fr">Information</em> (fr)</dt>

<dt>Русский термин:
<em property="skos:prefLabel"
xml:lang="ru">Информация</em> (ru)</dt>
</dl>
```

Seguidamente se incluye la nota o notas de alcance del idioma actual:

```
<h3>Notas de alcance</h3>
<ul><li><strong>NA</strong>
<span property="skos:scopeNote"
xml:lang="es"> Datos organizados de manera
coherente y significativa en el marco de un
sistema generalizado. </span></li></ul>
```

Los siguientes elementos marcados semánticamente son las relaciones jerárquicas específicas con otros conceptos, definidas con la propiedad skos:narrower:

```
<h3>Términos específicos</h3>
<ul rel="skos:narrower">
<li resource="unescothes:C03555">
<strong>TE</strong><a
href="http://skos.um.es/unescothes/C03555/ht
ml">Información científica</a></li>
<li resource="unescothes:C00889">
<strong>TE</strong><a
href="http://skos.um.es/unescothes/C00889/ht
ml">Información cultural</a></li>
<li resource="unescothes:C01243">
<strong>TE</strong><a
href="http://skos.um.es/unescothes/C01243/ht
ml">Información educacional</a></li>
<li resource="unescothes:C03716">
<strong>TE</strong><a
href="http://skos.um.es/unescothes/C03716/ht
ml">Información sobre ciencias
sociales</a></li>
<li resource="unescothes:C00689">
<strong>TE</strong><a
href="http://skos.um.es/unescothes/C00689/ht
ml">Información sobre
comunicación</a></li></ul>
```

Finalmente se marcan las relaciones asociativas con la propiedad skos:related:

```
<h3>Términos relacionados</h3>
<ul rel="skos:related">
<li resource="unescothes:C02161">
<strong>TR</strong><a
href="http://skos.um.es/unescothes/C02161/ht
ml">Conocimiento</a></li>
<li resource="unescothes:C01997">
<strong>TR</strong><a
href="http://skos.um.es/unescothes/C01997/ht
ml">Transferencia de información</a></li>
<li resource="unescothes:C02000">
<strong>TR</strong><a
href="http://skos.um.es/unescothes/C02000/ht
ml">Usuario de información</a>
</li></ul>
</div>
```

La implementación del marcado semántico con RDFa no incluye las etiquetas alternativas (términos no-descriptores) en un idioma distinto del idioma actual. Esto puede verse en el código del documento HTML correspondiente al concepto cuyo término descriptor en Español es “Guía de fuentes de información” y cuya URL es <http://skos.um.es/unescothes/C01751/html>. Dicho concepto tiene asociados los términos no-descriptores “Guía de fuentes literarias” y “Guía literaria”, tanto en Español como en el resto de idiomas del tesauro. Sin embargo puede observarse en el código HTML que únicamente se incluye la información relativa a las etiquetas alternativas en Español con su correspondiente marcado RDFa:

```
<h3>Usado por (no descriptores)</h3>
<ul>
<li><strong>UP</strong>
<span property="skos:altLabel"
xml:lang="es">Guía de fuentes
literarias</span></li>
<li><strong>UP</strong>
<span property="skos:altLabel"
xml:lang="es">Guía literaria</span></li>
</ul>
```

Por otro lado cabe resaltar que en los documentos HTML de los conceptos no existe ningún marcado semántico en los enlaces correspondientes a los micro-tesauros. Algo similar sucede en las páginas de los micro-tesauros, ya que el enlace que lleva a las áreas a las que pertenecen tampoco están marcadas con RDFa.

El motivo de dicha ausencia se debe a una limitación del propio SKOS, puesto que de forma nativa no existe una propiedad inversa a skos:member, que permitiría enlazar los conceptos con su correspondiente micro-tesauro y a estos con su área o dominio de conocimiento.

En cualquier caso es posible analizar la URL mediante el servicio del W3C RDFa 1.1 Distiller and Parser (9) siendo posible extraer las correspondientes tripletas RDF:

```
http://skos.um.es/unescothes/C01977
   rdfs:type skos:Concept;
   skos:narrower
    <http://skos.um.es/unescothes/C00689>,
    <http://skos.um.es/unescothes/C00889>,
    <http://skos.um.es/unescothes/C01243>,
    <http://skos.um.es/unescothes/C03555>,
    <http://skos.um.es/unescothes/C03716>;

   skos:prefLabel "Información"@es,
    "Information"@en, "Information"@fr,
    "Информация"@ru;

   skos:related
    <http://skos.um.es/unescothes/C01997>,
    <http://skos.um.es/unescothes/C02000>,
    <http://skos.um.es/unescothes/C02161>;

   skos:scopeNote "Datos organizados de
    manera coherente y significativa en el
    marco de un sistema generalizado."@es
```

Las tripletas extraídas son incompletas en relación a las suministradas en las versiones RDF/XML, N3, Turtle, JSON o JSON-LD del mismo concepto. Dichas versiones incluyen información relativa a todos los idiomas, propiedades skos:inScheme o skos:topConceptOf.

En consecuencia, el marcado semántico debe mejorarse. Para ello debería incorporarse información que pudiera ser extraída por analizadores RDFa, aunque no fuera visible al usuario

cuando consultara la versión (X)HTML a través del navegador.

5. Conclusiones

El marcado semántico permite aplicaciones que van más allá de las que en un principio promueven iniciativas como Schema.org. La mejora de los procesos de recuperación de información en los motores de búsqueda, mediante la aplicación de metadatos, es la punta de lanza del marcado semántico. La inclusión de datos estructurados, a partir del código (X)HTML de los documentos Web, ofrece grandes posibilidades desde el punto de vista de la interoperabilidad semántica y la reutilización de la información.

La actual situación de dispersión en cuanto a los formatos puede resultar problemática. Existe una clara apuesta por parte de los principales motores de búsqueda por los microdatos y Schema.org. Esto plantea ciertas divergencias con respecto al enfoque del W3C y la mayoría de tecnologías de la Web Semántica, así como ciertas incógnitas y dudas que únicamente paso del tiempo permitirá dilucidar: ¿Qué formato es el más adecuado para el marcado semántico? ¿Se impondrá Schema.org como modelo descriptivo estándar para los contenidos web? ¿Acabarán los principales motores de búsqueda incorporando RDFa y otros modelos descriptivos? RDFa es claramente más expresivo desde el punto de vista de formalización, granularidad semántica y potencialidad, pero ¿merece la pena invertir recursos en el marcado semántico con dicho formato en relación a los resultados que se obtienen en el posicionamiento SEO en los motores de búsqueda?

Ciertamente nada impide el marcado semántico utilizando microformatos, microdatos y RDFa simultáneamente. También es posible utilizar conjuntamente otros modelos descriptivos junto con Schema.org para la descripción semántica del contenido web. No obstante, estas alternativas resultan complejas y redundantes.

La solución a estos dilemas pasa ineludiblemente por comprender la naturaleza semántica de los contenidos objeto del marcado semántico, así como el tipo de reutilización que se desee hacer de los mismos. El paradigma actual de los motores de búsqueda es Google y la elección para mejorar el posicionamiento de los contenidos pasa necesariamente por utilizar microdatos y Schema.org. Sin embargo, RDFa es la opción más adecuada en proyectos que precisen un mayor detalle en cuanto a la representación de la semántica de los contenidos, la publicación Linked Open Data y la generación automática de conjuntos de datos RDF a partir de contenidos web.

En cuanto a los modelos descriptivos, cabe mencionar que Schema.org constituye sin duda una opción bastante completa, robusta y extensible para la descripción de gran parte de los contenidos disponibles en la Web. Pese a todo, hay que profundizar en la definición de las relaciones de mapeado entre Schema.org y otras ontologías y esquemas de metadatos de mayor recorrido como FOAF, SIOC, Dublin Core o DBpedia entre otros. En este aspecto es más importante entretejer una red de equivalencias entre modelos descriptivos que adoptar uno u otro. La potencialidad de Linked Open Data no solamente radica en la conexión entre conjuntos de datos, sino también en el mapeado entre modelos descriptivos.

Son numerosos los sistemas de organización del conocimiento cuya consulta está disponible en línea. Sin embargo, su reutilización deja mucho que desear. Es ocasiones, ni siquiera se ha planteado la reutilización de los datos por parte los diseñadores del servicio, tal y como sucede actualmente con el propio servidor web institucional del tesauro de la UNESCO. Por otra parte, a veces se publican conjuntos de datos cuya explotación y reutilización es costosa y compleja. El procesamiento de ficheros RDF de varios cientos de megabytes resulta mucho más ineficiente en comparación con el uso de SPARQL Endpoints.

Si bien las soluciones tecnológicas han alcanzado cierta madurez, existen diferentes planteamientos con respecto a la implementación de las técnicas de derreferenciación de URIs y negociación de contenido. Sería necesaria la creación de algunas pautas en la publicación de vocabularios controlados, extensibles también a las funcionalidades e interfaz de consulta, búsqueda y navegación por parte de los usuarios, y que contemplaran el correspondiente marcado semántico web.

Además, se precisa cierta infraestructura que, a veces, se encuentra con obstáculos y dificultades. Un ejemplo de ello podría ser la ausencia de políticas corporativas en las organizaciones con respecto a la apertura y reutilización de datos. Es preciso que los organismos encargados de la gestión de los vocabularios consideren que la apertura de estos instrumentos constituye una decisión de gran valor para su aplicación y difusión.

El marcado semántico web de servicios de vocabularios controlados hace patente ciertas limitaciones de SKOS: En primer lugar, la ausencia de determinadas relaciones inversas de

las propiedades skos:inScheme o skos:member hace imposible descubrir directamente las colecciones a las que pertenece un concepto, o determinar las colecciones asociadas a un esquema de conceptos. A lo anterior hay que añadir las carencias de SKOS para definir puntos de entrada a la consulta de conjuntos de conceptos de una colección (conceptos principales de una colección). Se trata de algo a tener en cuenta cuando se realiza un modelado de vocabularios organizados en micro-tesauros o facetas. Por lo tanto, es necesario desarrollar extensiones estándares de SKOS que contemplen lo anterior, con el objeto de mejorar la utilidad y eficacia del marcado semántico.

Sería recomendable contar con plataformas que permitan la publicación de vocabularios controlados de un modo sencillo y ágil. Actualmente se están explorando los CMS como herramientas adecuadas en este sentido, con la evidente adaptación mediante extensiones y desarrollos añadidos. Por este motivo los sistemas de organización del conocimiento pueden encontrar en estos sistemas una plataforma para su publicación y gestión, incorporando de forma inmediata el marcado semántico.

Tampoco hay que olvidar que la posibilidad de que los motores de búsqueda procesen el marcado semántico de sitios web de vocabularios controlados abre una nueva puerta a los procesos de indización automatizada en la web, que complementaría los procesos basados en técnicas de recuperación de información.

Por consiguiente, el marcado semántico web de los sistemas de organización del conocimiento se encuadra en una meta más ambiciosa que alberga el marcado semántico de la totalidad de los contenidos web. Posiblemente en ese momento podremos pasar de hablar de “big-data” a “big-knowledge”.

Notas

(1) En adelante cuando sea preciso referirse a ambos lenguajes de marcado de forma conjunta se hará utilizando la expresión (X)HTML.

(2) A este respecto resulta recomendable consultar http://microformats.org/wiki/Main_Page.

(3) El borrador de la recomendación para usar RDFa en HTML4 y HTML5 puede consultarse en http://www.w3.org/TR/rdfa-in-html/.

(4) Más información sobre Schema.org en http://schema.org/docs/documents.html.

(5) SaaS (Software as a Service) es un modelo de negocio software en el que las aplicaciones y los datos que manejan se ubican en servidores web. La acceso a la información para su consulta y mantenimiento se realiza mediante un navegador web y la reutilización de los datos suele realizarse mediante APIs.

(6) Algunos de estos vocabularios son el Tesauro STW de Economía (http://zbw.eu/stw), la Clasificación Decimal de Dewey (http://dewey.info), el Tesauro para materiales gráficos (http://www.t4gm.info), el Tesauro de la UNESCO (http://skos.um.es/unescothes) y la Nomenclatura de Ciencia y Tecnología de la UNESCO (http://skos.um.es/unesco6).

(7) Como prefijo para el espacio de nombres http://www.w3.org/2004/02/skos/core# se ha utiliza “skos”.

(8) Ambos vocabularios modelados por el proyecto UNESKOS están disponibles en http://skos.um.es/unescothes y http://skos.um.es/unesco6.

(9) Dicho servicio está disponible en http://www.w3.org/2012/pyRdfa/.

Referencias

Adida, Ben; Birbeck, Mark; McCarron, Shane; Herman, Ivan (20012). RDFa Core 1.1: Syntax and processing rules for embedding RDF through attributes. W3C Recommendation 07 June 2012. http://www.w3.org/TR/2012/REC-rdfa-core-20120607/.

Aitchison, Jean & Dextre-Clarke, Stella (2004). The thesaurus: a historical viewpoint, with a look to the future. // Cataloging & Classification Quarterly. 37:3-4 (2004) 5-21.

Berjon, Robin; Faulkner, Steve, Leithead, Travis; Doyle, Erika; O'Connor, Edwar; Pfeiffer, Silvia, Hickson, Ian (2013). HTML5: A vocabulary and associated APIs for HTML and XHTML. W3C Candidate Recommendation 6 August 2013. http://www.w3.org/TR/2013/CR-html5-20130806/.

Bizer, Christian; Heath, Tom; Berners-Lee, Tim (2009). Linked data-the story so far. // International Journal on Semantic Web and Information Systems. 5:3 (March 2009) 1-22.

Brickley, Dan; Guha, Ramanathan (2004). RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-rdf-schema-20040210/.

Caldwell, Ben; Cooper, Michael; Guarino, Loretta; Vanderheiden, Gregg (2008). Web Content Accessibility Guidelines (WCAG) 2.0. W3C Recommendation 11 December 2008. http://www.w3.org/TR/2008/REC-WCAG20-20081211/.

Cuenca Grau, Bernardo; Horrocks, Ian; Motik, Boris; Parsia, Bijan; Patel-Schneider, Peter; Sattler, Ulrike (2008). OWL 2: The next step for OWL. // Web Semantics: Science, Services and Agents on the World Wide Web. 6:4 (November 2008) 309-322.

Dunsire, Gordon (2011). Enhancing Information Services Using Machine-to-Machine Terminology Services. // Landry, P.; Bultrini, L.; O'Neill, E.; Roe, S.K. Subject Access: Preparing for the Future. Berlin: IFLA, 2011. 111-126.

Ewketu, Meron (2011). The UNESCO Thesaurus. UN-LINKS Meeting, 28-30 Nov. http://www.unesco.org/library/PDF/The%20UNESCO%20Thesaurus.pdf.

García, Nélida Elba; Jaroszczuk, Susana Eunice (2009). Objetos digitales: una experiencia de representación con metadatos Du-blin Core. // I Encuentro Nacional de Catalogadores: experiencias en la organización y tratamiento de la información en bibliotecas argentinas. Buenos Aires: Biblioteca Nacional, 2009. 193-206.

García Marco, Francisco Javier (2013). Schema.org: la catalogación revisitada. // Anuario ThinkEPI. 7 (2013) 169-172.

Hitzler, Pascal; Krötzsch, Markus; Parsia, Bijan, Patel-Schneider, Peter; Rudolph, Sebastian (2012). OWL 2 Web Ontology Langua-ge Primer (Second Edition). W3C Recommendation 11 December 2012. http://www.w3.org/TR/2012/REC-owl2-primer-20121211/

Khare, Rohit; Çelik, Tantek (2006). Microformats: a pragmatic path to the semantic web. // Proceedings of the 15th international conference on World Wide Web. ACM: New York. 865-866.

Klyne, Graham; Carroll, Jeremy; McBride, Brian (2004). Resource description framework (RDF): Concepts and abstract syntax. W3C recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.

Heath, Tom; Bizer, Christian (2011). Linked Data: Evolving the Web into a Global Data Space (1st edition). // Synthesis Lectures on the Semantic Web: Theory and Technology. 1:1, 1-136. Morgan & Claypool. http://linkeddatabook.com/editions/1.0/.

Hickson, Ian (2012). HTML Microdata. W3C Working Draft 25 October 2012. http://www.w3.org/TR/2012/WD-microdata-20121025/.

Ko, Moo Nam; Cheek, Gorrell; Shehab, Mohamed; Sandhu, Ravi (2010). Social-Networks Connect Services. // Computer archive. 43:8 (August 2010) 37-43.

Langegger, Andreas; Blochl, Martin, Woss, Wolfram (2007). Sharing Data on the Grid using Ontologies and distributed SPARQL Queries. // DEXA '07 Proceedings of the 18th International Conference on Database and Expert Systems Applications. Wa-shington DC: IEEE Computer Society, 2007. 450-454.

Mendes, Pablo; Jakob, Max; García-Silva, Andrés; Bizer, Christian (2011). DBpedia spotlight: shedding light on the web of docu-ments. // Proceedings of the 7th International Conference on Semantic Systems. New York: ACM, 2011. 1-8.

Méndez, Eva; Bravo, Alejandro; López, Leandro (2007). Microformatos: web 2.0 para el Dublin Core. // El profesional de la información. 16:2 (marzo-abril 2007) 107-113.

Manola, Frank; Miller, Eric (2004). RDF Primer. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-rdf-primer-20040210/.

Martínez Martínez, Carolina (2013). El Marcado Semántico en Wordpress. Mucia: Facultad de Comunicación y Documentación (Universidad de Murcia), 2013. Trabajo de Fin de Grado.

Miles, Alistair, Bechhofer, Sean (2009). SKOS Simple Knowledge Organization System Reference. W3C Recommendation 18 August 2009. http://www.w3.org/TR/2009/REC-skos-reference-20090818/.

Obrst, Leo (2003). Ontologies for semantically interoperable systems. // Proceedings of the twelfth international conference on Information and knowledge management. New York: ACM, 2003. 366-369.

Oren, Eyal; Delbru, Renaud; Catasta, Michele; Cyganiak, Richard; Stenzhorn, Holger; Tummarello, Giovanni (2008). Sindice.com: a document-oriented lookup index for open linked data. // International Journal of Metadata, Semantics and Ontologies. 3:1 (Jan-nuary 2008) 37-52.

Pastor-Sánchez, Juan-Antonio (2011). Tecnologías de la Web Semántica. Barcelona: Editorial UOC, 2011.

Pastor-Sánchez, Juan-Antonio (2012). Los "cms" como pieza fundamental en el despliegue de la web semántica. // Anuario Thin-kEPI. 6 (2012) 184-189.

Pastor-Sánchez, Juan-Antonio; Martínez-Méndez, Francisco-Javier; Rodríguez-Muñoz, José-Vicente (2012). Aplicación de SKOS para la interoperabilidad de vocabularios controlados en el entorno de linked open data. // El Profesional de la Informacion. 21:3 (Mayo-Junio 2012) 245-253.

Sauermann, Leo; Cyganiak, Richard. Cool URIs for the Semantic Web. W3C Interest Group Note 03 December 2008. http://www.w3.org/TR/2008/NOTE-cooluris-2008 1203/.

Sporny, Manu (2012). RDFa Lite 1.1. W3C Recommendation 07 June 2012. http://www.w3.org/TR/2012/REC-rdfa-lite-20120607/.

Tennison, Jeni (2011). Microdata and RDFa Living Together in Harmony.

Tomberg, Vladimir; Laanpere, Mart (200) RDFa versus Microformats: Exploring the Potential for Semantic Interoperability of Mash-up Personal Learning Environments. // 2nd workshop on Mash-Up Personal Learning Environments. 102-109. http://ceur-ws.org/Vol-506/tomberg.pdf.

W3C SPARQL Working Group (2012). SPARQL 1.1 Overview. W3C Recommendation 21 March 2013. http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/

Enviado: 2013-08-09.
Aceptado: 2013-09-02.