Datos enlazados para bibliotecarios

José Manuel Morales del Castillo
Biblioteca Daniel Cosío Villegas
El Colegio de México

En 2001 Berners-Lee, Hendler y Lassila (Berners-Lee, Hendler y Lassila, 2001) publicaron un trabajo en el que nos describen una web en la que al momento de realizar una consulta, dispondríamos de uno o varios agentes software inteligentes que se interconectarían con otros agentes para agregar información obtenida de diferentes fuentes dispersas y proporcionar una respuesta relevante, precisa y confiable (lo que haría que la tarea de ojear listados con docenas de resultados irrelevantes pasara a la historia). En esta evolución de la web, que ellos denominan Web semántica, la información se estructura en un intrincado entramado de conceptos interrelacionados entre sí que son interoperables, reutilizables, comprensibles para los humanos y procesables por máquina.

Son muchos los se preguntan en qué ha quedado el proyecto, de qué se trata realmente y dónde quedaron todas esas buenas intenciones descritas en el trabajo de Berners-Lee. Pues, aunque no lo parezca, la idea de desarrollar este modelo de web ha ido permeando poco a poco y, si bien es verdad que estamos lejos de poder decir que la web actual es semántica, sí que se han estado dando pasos firmes hacia su progresiva "semantización".

En esta web, el elemento esencial ya no es el documento, sino el dato y la red de relaciones semánticas que se establece con otros datos, todos ellos interoperables y reutilizables. Por lo tanto, el proceso de semantización es inherente a esta web de datos. Los profesionales de la información también juegan un papel destacado en este nuevo escenario donde los recursos web tienen que ser "catalogados" para poder ser semánticamente accesibles. La consolidación de este modelo podría servir para reivindicar el papel profesional de los bibliotecarios y su reinserción a la web como actores destacados (García Marco, 2013).

No siempre tenemos claro a qué nos estamos enfrentando o qué se espera realmente de nosotros. La idea de este trabajo es tratar de definir de una manera accesible qué es la web semántica o web de datos, qué elementos esenciales la conforman, en qué momento de desarrollo se encuentra el proyecto y cómo encajan en él las bibliotecas y los bibliotecarios. Comencemos.

El origen

Lo que movió a Berners-Lee, Hendler y Lassila a idear el modelo de web semántica fue la necesidad de resolver el problema de la sobrecarga de información. Todos hemos utilizado un motor de búsqueda (como Google) y hemos sufrido la situación de recibir como respuesta un listado interminable de resultados, de los cuales un gran porcentaje poco o nada tienen que ver con lo que estábamos buscando.

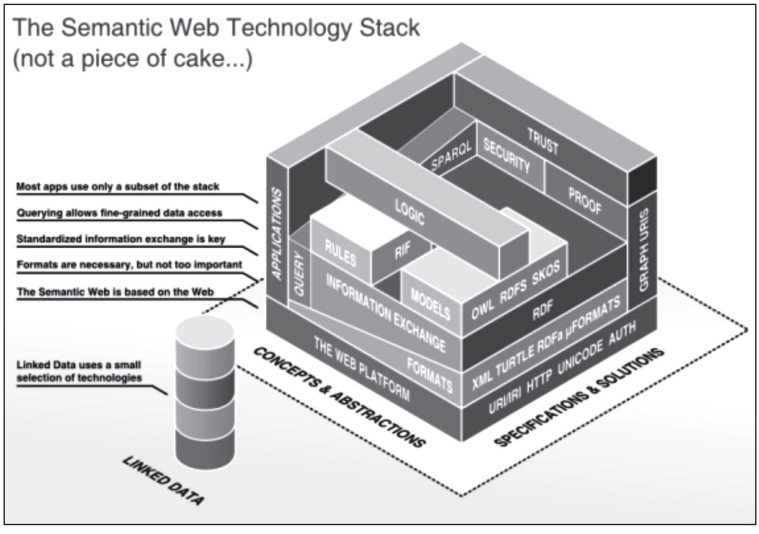
Esto ocurre porque HTML, el lenguaje de etiquetado en el que están descritos la mayoría de los documentos web, no está diseñado para caracterizar contenidos, sino que sirve principalmente para estructurar cómo se va a presentar la información en pantalla. Los motores de búsqueda explotan esta circunstancia realizando correspondencias entre las palabras clave definidas por el usuario en su consulta y cadenas de caracteres definidas dentro de etiquetas con un valor semántico potencial (como las etiquetas *title* o *meta*). Esto explica por qué obtenemos una mezcla de resultados

relevantes (no podemos obviar esa circunstancia) con otros completamente descontextualizados. La solución, pues, implicaba conseguir que la web pudiera determinar exactamente qué es lo que el usuario está buscando y ofrecerle una respuesta precisa y fiable al tomar en cuenta el contexto en el que la consulta se ha realizado. De esta manera, se hará posible una recuperación de información que se genera de manera dinámica a partir de fuentes dispersas (aunque conectadas semánticamente) y que agrega diferentes tipos de datos y formatos (texto, imagen, mapas, gráficas, sonido, video) a través de un interfaz único.

El modelo

La construcción de la web semántica se estructura en diferentes capas de desarrollo superpuestas unas sobre otras e interoperables entre sí (es decir, los elementos desarrollados en las capas inferiores pueden ser utilizados en las superiores y viceversa).

Figura 1. Modelo de capas tridimensional (Nowak, 2009)



En la base se ubica la denominada capa sintáctica y es donde se definen elementos como las URI (identificadores uniformes de recursos que permiten referir de manera unívoca a todos los elementos que intervienen en la web semántica como conceptos, personas o cosas), el protocolo de transferencia de información *http*, el sistema de codificación universal UNICODE, y las convenciones sintácticas definidas en metalenguajes como XML (Extensible Markup Language) (W3C, 2003) para garantizar la interoperabilidad, la reutilización y el intercambio de información.

Sobre ellos, en la capa semántica, se establece el modelo de datos que permite describir la información de manera que pueda ser interpretada semánticamente por máquinas. Este modelo de datos, basado en la lógica descriptiva (o lógica basada en la semántica), viene definido en el estándar RDF (Resource Description Framework) (W3C, 2014a), donde se especifica que la información se tiene que definir en forma de tripletas o sentencias Sujeto/Propiedad/Objeto para poder ser comprensible para las máquinas.

Figura 2. Ejemplo de tripleta RDF

Las tripletas RDF se pueden representar mediante grafos orientados y se interpretan de la siguiente manera: las propiedades que caracterizan a los sujetos (entendidos como cualquier tipo de entidad, evento o cosa) y les permiten relacionarse con otras entidades, eventos o cosas, o bien presentan un valor específico (como una cadena de caracteres). De acuerdo con esta convención, vemos que hay dos clases de tripletas: las que tienen como valor una entidad ("la casa está pintada de azul") y las que tienen como valor una cadena de caracteres ("el coche cuesta $10 000").

No obstante, estructurar la información en forma de tripletas no basta para definir una semántica. Es necesario definir claramente el contexto en el que la información deber ser interpretada y para ello hay que recurrir a herramientas más complejas, como

las ontologías, que se definen como esquemas de conocimiento, como lo puedan ser los tesauros, pero que no se limitan a definir una estructura jerárquica. Van más allá, ya que permiten categorizar los conceptos o entidades, definir el tipo de atributos que los caracterizan y establecer cómo se relacionan unos con otros en un determinado contexto para que de esta manera los agentes software sean capaces de obtener inferencias, o lo que es lo mismo, nuevo conocimiento no explícito, a partir de una serie de asertos y un conjunto de reglas (que se definen en la capa de reglas con el estándar RIF (Rule Interchange Format) (W3C, 2013).

A su vez son varios los estándares que sirven para generar esquemas de conocimiento. Dependiendo del que elijamos, podemos definir diferentes niveles de granularidad o profundidad de acuerdo con nuestras necesidades. Por ejemplo, si queremos desarrollar ontologías con el máximo de expresividad tenemos el estándar OWL (Ontology Web Language) (W3C, 2012a), aunque esto implica un mayor costo computacional que merma la eficiencia operativa (dicho de otro modo, a mayor complejidad, mayor dificultad para obtener inferencias de una manera eficiente). Como alternativas, tenemos RDFS (Resource Description Framework Schema) (W3C, 2014b) y SKOS (Simple Knowledge Organization System) (W3C, 2012b) que ofrecen, respectivamente, soluciones para el desarrollo de ontologías ligeras y tesauros (menos potentes que una ontología definida en OWL, pero con los que es más fácil operar). Por lo tanto, la elección de unos u otros estándares depende exclusivamente de la naturaleza del problema que queramos resolver.

Por último, las capas superiores del modelo (la capa lógica, de seguridad de prueba y de confianza) son las que se encargarían de garantizar la fiabilidad y seguridad de la información que circula en la web, procurando que los datos proporcionados por los agentes fueran no sólo precisos, sino también verídicos. Al día de hoy no se han desarrollado estas capas ya que algunos de los niveles inferiores aún se encuentran en fase de consolidación.

La aplicación práctica del modelo

Como acabamos de ver, el modelo teórico desplegado para caracterizar la web semántica está bien definido, cuenta con una multitud de estándares y vocabularios, pero ¿hasta qué punto hay un desarrollo práctico de éste? Y, si nos centramos en el ámbito bibliotecario, ¿de qué manera puede afectar al desarrollo de la profesión? Tratemos de responder a estas preguntas.

La naturaleza y complejidad del modelo de web semántica le han dado un ritmo de crecimiento y una cadencia propias, alejada de otros problemas que han encontrado una solución tecnológica casi inmediata. En este caso, se requiere de la intervención de múltiples factores y de una conjunción de sinergias que no podrían relizarse en los cortos plazos que los entornos tecnológicos imponen. No obstante, —buenas noticias—, sí que se están dando pasos hacia la web semántica; quizá no en la forma original en que se concibió el proyecto, pero sí de una manera firme hacia algo tangible. Con respecto a los esquemas de conocimiento y la necesidad de decidir entre sacrificar expresividad o facilidad de procesamiento, los desarrolladores han optado por el pragmatismo y la implementación de una web semántica *lightweight* o *ligera* que permite aprovechar algunas de sus principales funcionalidades pero a bajo costo computacional. En otras palabras, nos movemos hacia la versión beta de la Web de datos.

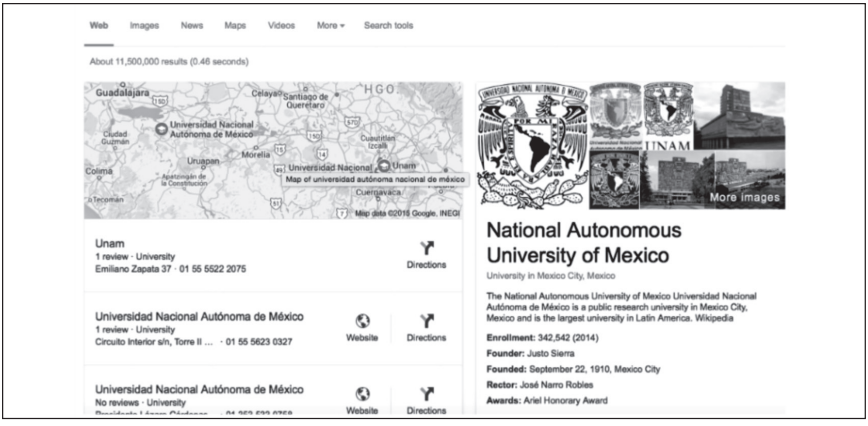
Así, frente al desarrollo de recursos descritos de manera nativa usando RDF (lo cual requiere un esfuerzo considerable), se ha optado por embeber metadatos RDF en el código fuente de otros vocabularios anfitriones. De esa manera poco invasiva, se aprovechan tecnologías que ya están plenamente operativas para enriquecer la descripción de los recursos y, consecuentemente, abrir la puerta a la obtención de resultados más precisos.

Éste es el caso de RDFa (Resource Description Framework in attributes) (W3C, 2015), una sintaxis que permite especificar metadatos de acuerdo al modelo RDF (es decir, atributos con el formato de tripletas sujeto-propiedad-objeto), pero embebidos en el código HTML5 (dentro de sus propias etiquetas). Basados en esta sintaxis,

se han desarrollado varios vocabularios específicos que pretenden facilitar el proceso de semantizar de origen los recursos de la web (democratizando el proceso de catalogación).

El ejemplo más claro es el vocabulario de microdatos Schema.org, desarrollado en 2012 por Google, Yahoo! y Microsoft, cuyo objetivo es que sus buscadores puedan proporcionar respuestas enriquecidas formadas a partir de la agregación de datos de diferentes fuentes (como Wikipedia y otras fuentes de datos estructuradas). Por ejemplo, si buscamos una universidad en Google, además del consabido listado de recursos, el buscador genera en el momento un agregado de datos que incluye mapas, logos o enlaces a biografías de egresados de esa universidad, lo que enriquece la experiencia del usuario con información adicional potencialmente relevante.

Figura 3. Ejemplo de resultados enriquecidos con datos enlazados.



Datos enlazados para bibliotecarios

El ejemplo anterior es un ejemplo claro del uso de datos enlazados en la web. La cuestión que puede surgir ahora es cómo podemos aprovechar esto en las bibliotecas. En principio, para poder trabajar con datos enlazados es necesario disponer de un conjunto de

datos, y eso es precisamente algo en lo que las bibliotecas son especialmente ricas. Sin ir más lejos, tenemos el catálogo bibliográfico como una invaluable fuente de datos semiestructurada que es posible compartir sin incurrir en ninguna violación de derechos de autor (ya que estaríamos compartiendo exclusivamente metadatos, no el contenido de las obras).

Para publicar conjuntos de datos enlazados (y enlazables), debemos tener claras algunas cuestiones esenciales. En primer lugar, debemos de aplicar el test de las 5 estrellas definido por Tim Berners-Lee para datos enlazados (Berners-Lee, 2010), una serie de criterios que nos dan una idea de la calidad de los datos que vamos a publicar:

1. Los datos deben ser legibles por máquina.
2. Estar definidos en formatos no propietarios.
3. Expresados de acuerdo al modelo de datos de RDF.
4. Deben definir vínculos entre sí y con otros conjuntos de datos.
5. Deben estar publicados bajo licencias abiertas, como dominio público o bajo una licencia Creative Commons Zero.

De acuerdo al número de criterios que cumplamos, será el número de estrellas que calificará nuestro conjunto (5 es la situación ideal).

El siguiente paso consistiría en enlazar nuestros datos tanto entre sí como con otros conjuntos de datos. Para ello, y dependiendo del tamaño del conjunto, podemos hacerlo de manera manual (cuando trabajamos con conjuntos pequeños) o de manera semiautomatizada (para conjuntos mayores) utilizando un software especializado como SILK (University of Manheim, 2016)).

Por último, lo que hay que hacer es publicar el conjunto de datos para que estén disponibles para todos aquellos que quieran reutilizarlo. Se puede hacer, por ejemplo, mediante un simple archivo estático RDF/XML que contenga todos los elementos del conjunto; embebiendo el conjunto en un archivo HTML mediante RDFa, o utilizando bases de datos específicas que permiten el almacenamiento de tripletas RDF (Heath & Bizer, 2011).

Es importante resaltar en este punto que la publicación de un conjunto de datos a su vez implica la adquisición de un compromiso claro por parte de la institución para garantizar su veracidad, mantenimiento y actualización continua. De otro modo, no tendría sentido realizar este esfuerzo.

En el ámbito de las bibliotecas, los archivos y los museos, existe un activo movimiento de profesionales que promueven la publicación de este tipo de conjuntos de datos abiertos y enlazables (LODLAM, 2010), donde realmente la única exigencia que se hace a las bibliotecas interesadas en participar es que contribuyan poniendo a disposición del público su conjunto de datos de una manera legible por máquina. De este modo, se estará posibilitando que cualquier persona pueda trabajar con estos datos en bruto y estructurarlos de manera adecuada para su explotación y aprovechamiento por parte de toda la comunidad.

La web de datos y las bibliotecas

La aparición de los datos abiertos enlazados tiene más implicaciones de las que en principio podríamos imaginar dentro del ámbito de las bibliotecas. De entrada, la necesidad de estructurar los datos usando la estructura que define RD hace necesario que se superen modelos tradicionales que se han quedado obsoletos ante este nuevo panorama, como es el caso de las reglas de catalogación angloamericanas AACR2 y el formato MARC. En el caso de las reglas de catalogación se utiliza RDA (Resource Description and Access) (Picco & Ortiz Repiso, 2012), un estándar que proporciona un nuevo marco para la descripción de recursos de información de acuerdo a los modelos conceptuales definidos en FRBR (Functional Requirements for Bibliographic Records) (Tillett, 2005) y FRAD (Functional Requirements for Authority Data) (Sardo, 2004) para la descripción de registros bibliográficos y registros de autoridades, respectivamente.

En esencia, lo que se define en ambos modelos es un conjunto de entidades y las relaciones que pueden establecer entre sí (lo cual

encaja perfectamente con la filosofía de representación de la realidad de RDF). En concreto, se definen tres grandes grupos de entidades (con sus correspondientes atributos) que engloban las diferentes formas en las que los recursos de información se pueden presentar (distinguiendo entre obra, expresión, manifestación e ítem), las personas o entidades corporativas que de alguna manera se relacionan con estos recursos, y las materias que representan el contenido de dichos recursos (conceptos, objetos, eventos y lugares).

Como vemos, la integración entre los modelos de datos definidos por RDF y FRBR se puede hacer de manera natural, por lo que la futura inclusión de los registros catalogados con RDA en la web semántica no será traumática y permitirá contar de inmediato con una masa crítica de recursos semánticamente accesibles para el desarrollo de aplicaciones y servicios de diversa índole.

El otro gran reto pendiente es la superación del formato MARC por un modelo de codificación más flexible. Con este espíritu surgió Bibframe (Bibliographic Framework) (Library of Congress, 2015), una iniciativa de la Library of Congress que pretende proporcionar un esquema para describir y conectar datos bibliográficos entre sí. De entrada, MARC y Bibframe (http://bibframe.org/tools/) presentan notables diferencias que van más allá de lo meramente estético. Bibframe no sólo rompe con la obsoleta codificación alfanumérica de campos y subcampos al utilizar el lenguaje natural para hacerlos fácilmente comprensibles e interpretables de un sólo vistazo, sino que también proporciona nuevas maneras de diferenciar entre los contenidos y sus manifestaciones, de caracterizar entidades bibliográficas (obra, instancia, autoridad y anotaciones), de identificarlas de manera unívoca y de explicitar las relaciones que se definen entre ellas (como las relaciones obra-obra, obra-instancia, obra-autoridad).

No obstante, esta idea de ruptura con los esquemas tradicionales de la catalogación no debe hacernos pensar que todo el trabajo que durante décadas han realizado las bibliotecas caerá en saco roto, ya que gracias a herramientas de mapeo será posible hacer la correspondencia entre campos MARC y Bibframe, y así aprovechar la ingente cantidad de registros que ya existen al usar RDA como un puente entre ambos.

A MODO DE CONCLUSIÓN

La web semántica se perfila cada vez más como una realidad palpable. Si bien los resultados son algo modestos si los comparamos con las expectativas creadas en un principio, no se puede negar que son estas tecnologías las que están marcando el camino hacia donde se dirige la web.

Se ha optado por sacrificar la expresividad (es decir, la capacidad de expresar semántica) en aras de conseguir descripciones más sencillas procesables por máquina a un bajo coste computacional, lo cual me parece un buen trato si con ello conseguimos implantar una web semántica *lightweight* que sirva como base para alcanzar mayores cotas de desarrollo a corto o medio plazo.

Por ello, los profesionales de la información no debemos permanecer ajenos a estos movimientos tecnológicos, sino que nuestra actitud debe ser la de adoptar una posición decidida por la innovación. Sin ir más lejos, la comunidad bibliotecaria se ha convertido en uno de los pilares fundamentales del movimiento Link Open Data (datos abiertos enlazados), cuyo objetivo es compartir y hacer semánticamente accesibles piezas de información, datos y conocimiento en la web semántica. En el caso de las bibliotecas, con lo que se está contribuyendo es con los metadatos de los registros almacenados en los catálogos de las bibliotecas, lo que genera una masa crítica de recursos semánticamente accesibles a disposición de cualquiera que los necesite.

Por lo tanto, no podemos esperar a que un día nos comuniquen que la web semántica ya está plenamente implantada para empezar a considerarla. Debemos contribuir desde ya a la adopción e implantación del modelo no sólo para reivindicar nuestra figura profesional, sino para convertirnos en elementos clave en el nuevo escenario tecnológico que se avecina. No se trata tanto de estar preparados para cuando llegue el cambio, sino de actuar de manera proactiva para que el cambio llegue.

Referencias bibliográficas

Berners-Lee, T. (2010). 5 Stars Open Data. *5 Star Data*. Disponible el 2 de julio de 2016 en http://5stardata.info/.

Berners-Lee, T., J. Hendler y O. Lassila. (2001). The Semantic Web. *Scientific American*.

García Marco, F. J. (2013). Schema.org: la catalogación revisitada. *Anuario ThinkEPI*, (1), 169–172.

Heath, T., y C. Bizer. (2011). Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, *1*(1), 1–136. doi:10.2200/S00334ED1V01Y201102WBE001.

Library of Congress. (2015). Overview (Bibliographic Framework Initiative Technical Site - BIBFRAME.ORG). Disponible el 18 de mayo de 2015 en http://bibframe.org/.

LODLAM. (2010). LODLAM Linked Open Data in Libraries, Archives & Museums. Disponible el 15 de noviembre de 2015 en http://lodlam.net/.

Nowak, B. (2009). The Semantic Web. Not a Piece of Cake. Entrada de blog. Disponible el 27 de septiembre de 2015 en http://bnode.org/blog/2009/07/08/the-semantic-web-not-a-piece-of-cake.

Picco, P. y V. Ortiz Repiso. (2012). RDA, el nuevo código de catalogación: cambios y desafíos para su aplicación. *Revista Española de Documentación Científica*, *35*(1), 145–173. doi:10.3989/redc.2012.1.848

Sardo, L. (2004). Functional Requirements for Authority Records. *Bollettino AIB*, *44*(4), 455–470.

Tillett, B. (2005). What is FRBR? A Conceptual Model for the Bibliographic Universe. *Australian Library Journal*, *54* (marzo 2015), 24–30. doi:Article.

Universidad de Manheim. (2016). Silk: the Linked Data Integration Framework. Disponible el 3 de julio de 2016 en http://silkframework.org.

W3C (2015). RDFa Core 1.1. Disponible el 18 de mayo de 2015 en http://www.w3.org/TR/rdfa-syntax/.

———— . (2014a). RDF - Semantic Web Standards. Disponible el 18 de mayo de 2015 en http://www.w3.org/RDF/.

———— . (2014b). RDF Schema 1.1. Disponible el 18 de mayo de 2015 en http://www.w3.org/TR/rdf-schema/.

———— . (2013). RIF Overview. Disponible el 28 de septiembre de 2015 en http://www.w3.org/TR/rif-overview/.

———— . (2012a). OWL 2 Web Ontology Language Document Overview. Disponible el 18 de mayo de 2015 en http://www.w3.org/TR/owl2-overview/.

———— . (2012b). SKOS Simple Knowledge Organization System-home page. Disponible el 29 de sepiembre de 2015 en http://www.w3.org/2004/02/skos/.

———— . (2003). Extensible Markup Language (XML). Disponible el 17 de mayo de 2015 en http://www.w3.org/XML/.