F. SISTEMAS Y TECNOLOGÍAS DE INFORMACION

Quince años de web semántica: de las tecnologías a las buenas prácticas

Fifteen years of the semantic web: from technologies to good practice

Juan-Antonio Pastor-Sánchez

Pastor-Sánchez, Juan-Antonio (2016). "Quince años de web semántica: de las tecnologías a las buenas prácticas". *Anuario ThinkEPI*, v. 10, pp. 264-268.

http://dx.doi.org/10.3145/thinkepi.2016.58

Publicado en *IweTel* el 29 de enero de 2016

Resumen: El despliegue eficaz de la web semántica depende no solamente de los desarrollos tecnológicos, también es necesario aplicar pautas para el proceso de publicación de datos estructurados en la Web. Parte del trabajo del grupo de trabajo del *W3C* sobre datos en la Web se centra en la definición de buenas prácticas y de vocabularios para representar los conjuntos de datos. La madurez de la web semántica se refleja en la importancia cada vez mayor de los aspectos relacionados con el consumo de los datos publicados, tipos de licencia, confianza, dinámicas de publicación y de reutilizacion de los datos, entre otros procesos.

Palabras clave: Web semántica; Datos; Web; Vocabularios; Buenas prácticas; Conjuntos de datos.

Abstract: The effective deployment of the semantic web depends not just on technological developments, but also on applying guidelines for the process of publishing structured data on the Web. Part of the work of the *W3C* working group about data on the Web is focused on the identification of best practices and defining vocabularies to represent datasets. The maturity of the semantic web is reflected in the increasing importance of aspects related to the use of published data, types of licenses, trust, publication dynamics, and data reusing, among other processes.

Keywords: Semantic web; Data; Web; Best practices; Datasets.

Introducción

En mayo de 2016 se cumplirán quince años desde la publicación por parte de **Berners-Lee**, **Hendler** y **Lassila** (2001) del trabajo donde se introducían los principales objetivos y componentes de la web semántica. En dicho trabajo se mencionaron, sin entrar en detalle, algunas de las tecnologías que han ido definiendo la arquitectura global de la web semántica, así como los principios en torno a los cuales se ha desarrollado.

El paso de la publicación de documentos a la expresión de su significado ha sido el hilo conductor para la elaboración de estándares y tecnologías cuyo objetivo es la representación del conocimiento en la Web. La interoperabilidad sintáctica mediante XML y la interoperabilidad semántica mediante RDF y OWL conforman las principales tecnologías en torno a las cuales se ha construido la web semántica.

En estos quince años, la práctica ha demostrado que no basta con contar con tecnologías maduras para la publicación de datos. Podría afirmarse que el despliegue de la web semántica depende de las posibilidades de aplicación/explotación de los datos así como de la aplicación de una serie de pautas, claras y precisas para el proceso de su publicación.

El grupo de trabajo del *W3C* sobre buenas prácticas de datos en la web (*Data on the web best practices*, *DWBP*) tiene como objetivo la elaboración de un ecosistema de datos abiertos facilitando el entendimiento entre los desarrolladores y los editores de los datos. Esto se consigue mejorando la coherencia de los datos con la oportuna orientación a los editores. Para ello el grupo de trabajo enfoca sus tareas en dos líneas:

- definición de buenas prácticas para aplicar en la implementación tecnológica durante la publicación de los datos;
- definición de vocabularios para representar indicadores de calidad y uso de los conjuntos de datos.

https://www.w3.org/2013/dwbp/wiki/Main_Page

Entender los conjuntos de datos

El concepto de conjunto de datos o *dataset* ha evolucionado con la práctica según se han identificado nuevas necesidades para su publicación. Un aspecto que está marcando la madurez de las tecnologías de la web semántica es que ya no se piensa tanto en las soluciones tecnológicas de edición y publicación como en el consumo de los datos publicados, tipos de licencia, fiabilidad, dinámicas de publicación o reutilización, entre otros procesos. Por lo tanto, es preciso que la implementación y mantenimiento de los conjuntos de datos tengan un enfoque independiente de su posible explotación a posteriori.

Resulta evidente que un conjunto de datos que se publique en una web está rodeado de una serie de elementos tecnológicos. Fundamentalmente la propia arquitectura de la web, los estándares (formatos, RDF, OWL) y los vocabularios utilizados para representar los datos y los mecanismos para acceder a los mismos.

Además de los datos propiamente dichos o valores de datos (*data values*) un conjunto de datos debe contener los metadatos con información sobre su contenido y estructura, procedencia, indicadores de calidad o medios para acceder y utilizar los datos. A su vez los valores de datos han de organizarse en diferentes distribuciones que responderían a diferentes necesidades de explotación o consumo. Cada distribución debe incluir información sobre su formato, estructura, licencia de uso, etc.

"El despliegue de la web semántica depende en gran medida de la aplicación de pautas para la publicación de los datos"

Sin duda son aspectos que van más allá de los relacionados con el modelado o la representación (que se mantienen esenciales). De igual importancia es considerar dentro del concepto de conjunto de datos su ciclo de vida en el entorno *linked open data* (*LOD*). Ya no basta con considerar el modelado, la publicación y la actualización periódica de un *dataset*. También hay que considerar otras operaciones tales como (**Auer** *et al.*, 2012):

- interconexión con otros conjuntos de datos;
- elaboración de vocabularios controlados para la clasificación de los datos;
- enriquecimiento;
- análisis de la calidad;
- prever su explotación mediante mecanismos para el descubrimiento o extracción de datos;
- su consulta, búsqueda y navegación.

Retos y buenas prácticas

La publicación de datos en la Web se fundamen-

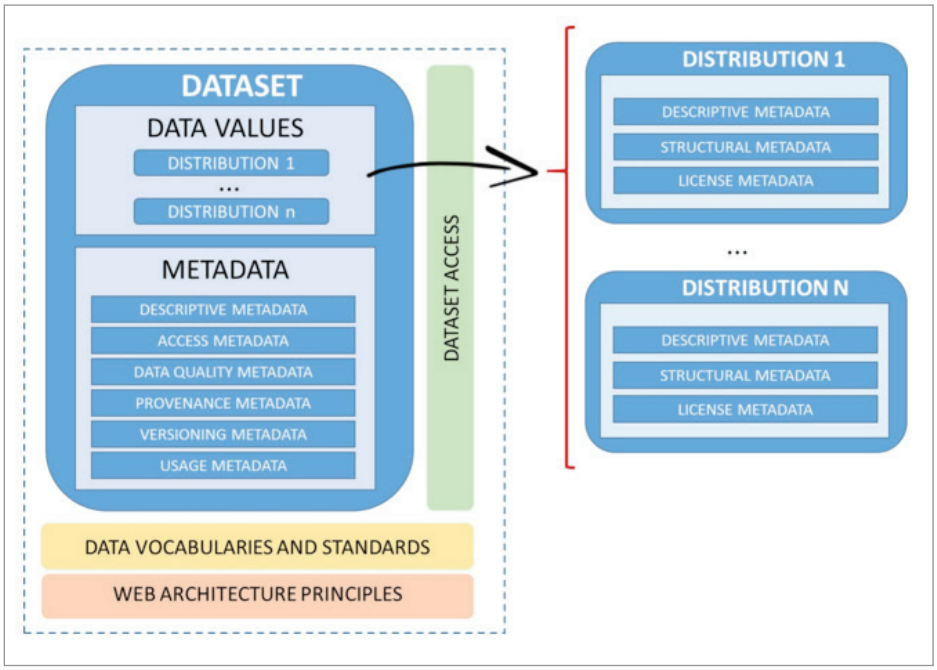

Figura 1. Composición de los conjuntos de datos y componentes relacionados con su publicación y uso
https://www.w3.org/TR/dwbp

ta en su interoperabilidad mediante el uso de estándares. Desde este punto de vista, tanto los formatos, como los mecanismos de identificación y acceso, la reutilización de vocabularios y la aplicación de metadatos descriptivos son aspectos esenciales para alcanzar niveles adecuados de formalización.

Con el tiempo se han planteado nuevos retos relacionados con la privacidad y seguridad de los datos publicados, así como la verificación de su calidad y procedencia. La definición de licencias que indiquen las posibilidades de explotación también deben formar parte de las políticas de publicación de los conjuntos de datos. Hay que destacar que la disponibilidad de diferentes versiones de los *datasets*, su enriquecimiento y su preservación/mantenimiento son características que ya se consideran inherentes a la calidad de los mismos.

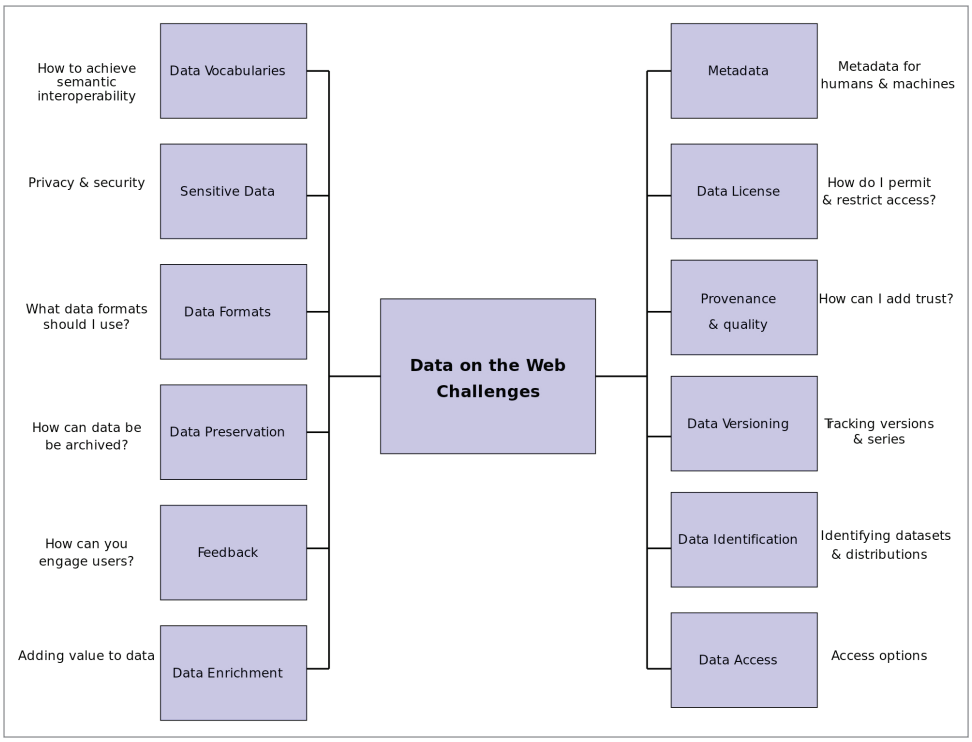

Figura 2. Requisitos y retos de la publicación de datos en la Web
https://www.w3.org/TR/dwbp

"El concepto de conjunto de datos ha evolucionado con la práctica según se han identificado nuevas necesidades para su publicación"

El grupo de trabajo *DWBP* trabaja en la elaboración de un documento de buenas prácticas que identifica de un modo muy concreto los retos anteriormente mencionados (*W3C*, 2016a). Dichos retos se han identificado a través de una serie de requisitos de casos de uso (*W3C*, 2015a):

- Uso de vocabularios de datos para incrementar la interoperabilidad semántica: mediante la reutilización de vocabularios y la elección de niveles adecuados de modelado y formalización.
- Asegurar la privacidad y seguridad de datos sensibles: identificando las partes de los conjuntos de datos no accesibles de forma pública e informando a los usuarios del motivo de dicha restricción.
- Utilizar múltiples estándares para que los datos estén disponibles en diferentes formatos que hagan más versátil su reutilización.
- Atraer a los usuarios y consumidores de los datos indicando los mecanismos para contactar con los editores de los mismos y publicando como datos abiertos los resultados de dicha retroalimentación.
- Enriquecer los datos para aportar valor añadido mediante otros conjuntos de datos, metadatos, bases de datos u otros recursos electrónicos.
- Proporcionar metadatos del conjunto de datos legibles por máquina y por personas: estos metadatos, ya sean descriptivos, estructurales o de aspectos más concretos, deben especificarse mediante términos de vocabularios estándar.
- Proporcionar información sobre la licencia de uso del *dataset* de manera que los consumidores sean conscientes de los límites y posibilidades de reutilización de los datos.
- Incluir información sobre la procedencia y calidad de los datos, de manera que el conjunto de datos pueda ser evaluado para medir la fiabilidad y confianza de los datos que contiene.
- Gestionar adecuadamente las distintas versiones de un mismo conjunto de datos, al tiempo que se evita realizar cambios en las APIs o servicios de acceso a los datos que obliguen a realizar cambios a los clientes que ya estén haciendo uso de los mismos.

- Usar URIs persistentes es algo indispensable, tanto para identificar a los conjuntos de datos en su totalidad, como a los diferentes recursos que contiene. Esto también es aplicable a las distintas versiones del *dataset*.
- Proporcionar múltiples mecanismos para acceder en tiempo real a los datos actualizados en la medida de lo posible, no sólo mediante la descarga de todo el conjunto en algún formato estándar, sino también mediante servicios web y con APIs adecuadamente documentadas.

"Con el tiempo se han planteado nuevos retos sobre la privacidad, seguridad, calidad y procedencia de los datos publicados"

A partir de estos retos se propone un total de 32 buenas prácticas que se agrupan en torno a una serie de beneficios:

- Comprensión: las personas pueden tener una mejor comprensión sobre la estructura, naturaleza y significado de los datos, así como del conjunto de datos en su globalidad y los metadatos que lo acompañan.
- Proceso: las máquinas pueden ser capaces de procesar automáticamente y manipular los datos dentro del *dataset*.
- Descubrimiento: las máquinas pueden descubrir de forma automática conjuntos de datos o datos específicos dentro del conjunto.
- Reutilización: se incrementan las posibilidades de reutilización de datos por parte de diferentes grupos de consumidores.
- Confianza: mejora la confianza de los consumidores en el conjunto de datos.
- Conectividad (*linkability*): es posible crear vínculos entre los recursos, tanto a nivel de conjunto como de datos individuales.
- Acceso: tanto las personas como las máquinas pueden acceder a datos actualizados en diferentes formatos y niveles de detalle.
- Interoperabilidad: facilita el acuerdo entre editores y consumidores de datos.

Cada una de estas prácticas se describen detalladamente, mediante la oportuna justificación, objetivos perseguidos, propuestas prácticas y ejemplificadas de implementación y mecanismos de prueba, así como las correspondientes evidencias en el documento de requisitos y casos de uso.

Sin duda alguna se trata de un enfoque mucho más evolucionado que la propuesta que en su momento se hizo en el ámbito contexto de *LOD* (*linked open data*)(*W3C*, 2014). Por otro lado, estas buenas prácticas tienen una aplicación directa en proyectos de publicación de datos abiertos en contextos concretos, como el de las administraciones públicas (**Pastor-Sánchez**, 2014).

Cabe destacar que el mismo grupo de trabajo *DWBP* está trabajando en dos vocabularios:

- el primero de ellos tiene como finalidad la representación de aspectos, indicadores y métricas sobre la calidad de los conjuntos de datos (*W3C*, 2015b), ampliando la funcionalidad del vocabulario *DCAT* desarrollado en su momento (*W3C*, 2014b);
- el segundo se encuentra en fase de borrador de trabajo muy inicial y su objetivo es representar los usos que se realiza de un *dataset* por parte de los consumidores de datos (*W3C*, 2016b).

Ampliando la visión de la web semántica

La web semántica, como toda tecnología emergente, ha tenido un ciclo de evolución en el que las expectativas se han acotado en torno a desarrollos concretos. A lo largo de los últimos cinco o seis años algunos autores (**Heath**; **Bizer**, 2011, capítulo 6; **Herman**, 2011; **Saorín**; **Peset**; **Ferrer-Sapena**, 2013) ya apuntaron que los aspectos técnicos no son suficientes para asegurar el éxito y correspondiente supervivencia de una tecnología.

Figura 3. Beneficios de las buenas prácticas para la publicación de datos en la Web. Elaboración propia a partir de: *https://www.w3.org/TR/dwbp*

Nos encontramos por tanto ante propuestas que van más allá de los mencionados aspectos y se centran en las dinámicas de publicación, acceso y reutilización de los datos. Durante la puesta en práctica de los procesos de publicación de datos se han identificado dificultades concretas. Por este motivo ha sido preciso elaborar una serie de pautas (buenas prácticas) que orienten y mejoren las distintas tareas asociadas a los procesos de publicación de datos en la Web.

Pese a todo no olvidemos que la Web es un ecosistema tecnológico complejo en constante evolución. Si bien ahora los trabajos del *DWBP* resultan de una gran utilidad, en el futuro se identificarán nuevos problemas, se crearán nuevas tecnologías y se llevarán a cabo nuevas aplicaciones prácticas. Esto implica que siempre será preciso realizar un trabajo de continua revisión de dichas pautas. En consecuencia, la web semántica, cuyo dominio de las tecnologías resultaba relativamente sencillo hace unos años, va a precisar, para su correcta comprensión y aplicación, ir más allá de los aspectos técnicos y conocer en profundidad otros relacionados con el ciclo de vida de los datos en la Web.

Bibliografía

Auer, Sören; **Bühmann, Lorenz**; **Dirschl, Christian**; **Erling, Orri**; **Hausenblas, Michael**; **Isele, Robert**; **Lehmann, Jens**; **Martin, Michael**; **Mendes, Pablo N.**; **Van-Nuffelen, Bert**; **Stadler, Claus**; **Tramp, Sebastian**; **Williams, Hugh** (2012). "Managing the lifecycle of linked data with the LOD2 stack". En: *11th Intl semantic web conf*, Boston, MA, USA, November 11-15, Proceedings, Part II.
http://dx.doi.org/10.1007/978-3-642-35173-0_1

Berners-Lee, Tim; **Hendler, James**; **Lassila, Ora** (2001). "The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities". *Scientific American*, v. 284, n. 5, pp. 28-37.
http://goo.gl/aQH1S0

Heath, Tom; **Bizer, Christian** (2011). *Linked data: evolving the Web into a global data space*. San Raphael, CA: Morgan & Claypool Publishers. ISBN: 978 1608454310
http://linkeddatabook.com/book

Herman, Ivan (2011). *Semantic web adoption and applications*.
http://www.w3.org/People/Ivan/CorePresentations/Applications/Applications.pdf

Pastor-Sánchez, Juan-Antonio (2014). "Aspectos prácticos para proyectos de datos abiertos en las administraciones públicas". *Anuario ThinkEPI*, v. 8, pp. 313-317.

Saorín, Tomás; **Peset, Fernanda**; **Ferrer-Sapena, Antonia** (2013). "Factores para la adopción de *linked data* e implantación de la web semántica en bibliotecas, archivos y museos". *Information research*, v. 18, n. 1, paper 570.
http://InformationR.net/ir/18-1/paper570.html

W3C (2014a). *Best practices for publishing linked data*. W3C Working Group, Note 09 January.
https://www.w3.org/TR/2014/NOTE-ld-bp-20140109

W3C (2014b). *Data catalog vocabulary (DCAT)*. W3C Recommendation, Note 16 January.
http://www.w3.org/TR/vocab-dcat

W3C (2015a). *Data on the web best practices, use cases, & requirements*. W3C Working Group, Note 24 February.
http://www.w3.org/TR/2015/NOTE-dwbp-ucr-20150224

W3C (2015b). *Data on the web best practices: Data quality vocabulary*. W3C Working Draft, 17 December.
https://www.w3.org/TR/2015/WD-vocab-dqv-20151217

W3C (2016a). *Data on the web best practices*. W3C Working Draft 12 January.
http://www.w3.org/TR/2016/WD-dwbp-20160112

W3C (2016b). *Data on the web best practices: Dataset usage vocabulary*. W3C Editor's Draft 26 January 2016.
http://w3c.github.io/dwbp/vocab-du.html

Juan-Antonio Pastor-Sánchez
Universidad de Murcia
pastor@um.es