El potencial de la publicación semántica en el contexto de difusión y citación

Eduardo Alvarez[1]
Layla Michán[2]

Resumen

Se investigó la implementación de la publicación semántica en revistas académicas. Los principales estándares sobre los que se basa la tecnología semántica son el Marco de Descripción de Recursos (*Resource Description Framework*, RDF), el sistema de consulta (SPARQL) y la estructura ontológica (*Web Ontology Language,* OWL). Dichos estándares son utilizados en las distintas instancias encargadas de generar contenidos durante el proceso de publicación científica: las revistas, editoriales, bases de datos, así como en las aplicaciones y servicios de información asociados a éstas. Existen iniciativas locales, regionales y mundiales encargadas de promover y fomentar la implementación de la Web Semántica en la publicación académica.

Palabras clave: trabajo editorial; publicación semántica; futuro editorial; revistas; artículos.

Abstract

The implementation of semantic publication in academic journals was investigated. The main criteria on which semantic technology is based are the Resource Description Framework (RDF), the query system (SPARQL) and the ontological structure (OWL, Web Ontology Language). These specific are used in the diffe-

1 https://orcid.org/0000-0002-1572-164X. Meta-datos.
2 https://orcid.org/0000-0002-5798-662X. Facultad de Ciencias, UNAM, México.

rent instances responsible for generating content during the scientific publication process: magazines, editorials, databases, as well as in the applications and information services associated with them. There are local, regional and global initiatives in charge of promoting and encouraging the implementation of the Semantic Web in academic publishing.

Keywords: editorial work; semantic publication; future publishing; journals; papers.

Introducción

Tim Berners-Lee fue uno de los artífices de la Web en 1989, es el personaje reconocido como su fundador; doce años después vaticinó su transformación: "[...] probablemente cambiará profundamente la naturaleza misma de cómo se produce y comparte el conocimiento científico, de una manera que ahora apenas podemos imaginar". A tal modficación la denominó Web Semántica (Berners-Lee y Hendler 2001).

Desde ese momento que acuñó el término hasta ahora se han sucedido un gran conjunto de avances tecnológicos en esa dirección: se fundó el World Wide Web Consortium (W3C), la comunidad internacional que desarrolla los estándares abiertos para asegurar el crecimiento a largo plazo de la Web. Ésta establece los recursos semánticos utilizados para la integración de datos, modelado, interpretación y explotación de información para la Web Semántica, manteniendo los estándares y buenas prácticas.

La Web Semántica se refiere a la publicación de documentos electrónicos acompañados de etiquetado y marcado semántico que expresan el significado de los elementos. En la Web Semántica, la información publicada se acompaña de metadatos que describen la información, proporcionando así un contexto "semántico" (British Medical Journal [BMJ] 2010).

El objetivo es mejorar el Internet ampliando la interoperabilidad entre los sistemas informáticos y usando programas de agentes inteligentes (algoritmos) cuya funcionalidad es buscar información sin operadores humanos. La estructuración de datos en la Web Semántica se basa en la creación de lenguajes semánticos para representar la información, entre los que se destacan el lenguaje XML, RDF y el OWL (Berners-Lee y Hendler 2001).

El procedimiento semántico de los documentos digitales en la Web consiste en asignar etiquetas que provienen de un vocabulario controlado abierto que debe estar estructurado en RDF o OWL; hasta el momento se encuentran registrados 767 vocabularios ligados abiertos (Linked Open Data). Los datos vinculados son esenciales para conectar realmente la Web Semántica. Con un poco de pensamiento, resulta algo bastante fácil de hacer, y se convierte en una segunda naturaleza (Berners-Lee 2006).

Los vocabularios más sofisticados son las ontologías diseñadas en OWL que proporcionan mejores opciones de búsqueda y también ofrecen una mejor búsqueda textual (DeLeon y Dumontier 2008). Una ontología es una especificación explícita de una conceptualización. El término es prestado de la filosofía, donde una ontología es una cuenta sistemática de la existencia. Podemos describir la ontología de un programa mediante la definición de un conjunto de términos representativos. En una ontología las definiciones asocian los nombres de las entidades en el universo del discurso (por ejemplo, clases, relaciones, funciones u otros objetos) con texto legible que describe lo que esos nombres deben significar y axiomas formales que restringen la interpretación y el buen uso de estos términos (Shotton 2009).

Un ejemplo del uso de las ontologías en la Web es el que hace el portal de noticias inglés BBC con su ontología empleada para describir y durar sus contenidos. Otro ejemplo corresponde con la ontología más usada en biología, que es, sin duda, Gene Ontology (Figura 1).

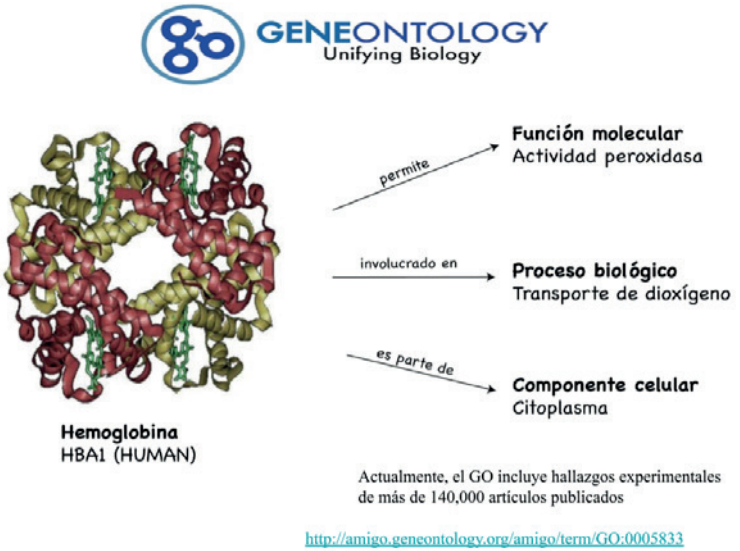

Figura 1. Anotación semántica de información publicada en literatura biomédica, representada en Gene Ontology. Fuente: elaboración propia con información de Gene Ontology Consortium (2020).

La semántica en las publicaciones académicas

La publicación académica ha sufrido una profunda transformación en la era digital, pues el incremento acelerado del número de revistas, la inmensa cantidad de aplicaciones diseñadas específicamente para indexar, recuperar, procesar y visualizar la información sobre éstas, los artículos o los autores y la implementación de métodos innovadores como la minería de textos, la Web Semántica, los datos ligados, los grandes datos y la ciencia de datos han llevado a las revistas científicas a otro nivel (Berners-Lee y Hendler 2001).

La literatura de investigación semántica incluye todo lo que mejore el significado, facilite su descubrimiento automatizado, permita su vinculación a otros artículos semánticamente relacionados, proporcione acceso a los datos en forma procesable o facilite la integración de datos entre documentos. Shotton (2009) propone seis reglas para que los editores de revistas implementen la tecnología semántica:

1. Comenzar de manera simple y mejorar la funcionalidad de forma incremental.
2. Esperar grandes cosas de los autores.
3. Explorar por completo las habilidades internas existentes.
4. Usar los estándares establecidos siempre que sea posible.
5. Publicar conjuntos de datos en bruto en la Web.
6. Publicar metadatos del artículo, particularmente listas de referencias, en forma legible por máquina.

Los objetivos de la publicación semántica son mejorar los artículos de revistas académicas, ayudando a la publicación de datos y metadatos y proporcionando un acceso interactivo "animado" al contenido. La vinculación de la literatura académica con datos de investigación y la anotación automática de documentos académicos a ontologías (Peroni 2017), además de dar mayor impacto en el contenido, harán que la revista se consulte más y, en última instancia, cuente con una mayor citación.

Desde la aparición e implementación del JATS-XML, se han agregado otras formas de preservación digital, otra estructura de información como SPARQL, JSON-XML, RDF y OWL. Las ontologías diseñadas precisamente en OWL proporcionan el marco conceptual en el cual los procesos científicos y los flujos de trabajo pueden ser estructurados y compartidos para ser interoperables, proporcionando el contenido y el contexto de diálogos en línea dentro de comunidades virtuales.

Nuestro objetivo es investigar cuál es el valor de la publicación semántica en el ámbito académico; para ello nos proponemos: 1) definir las condiciones necesarias para que se implemente la publicación semántica en las revistas académicas e 2) identificar las herramientas y tecnología que permitan implementar esta práctica.

Metodología

Se realizó una investigación en la Web a través de la exploración y búsqueda de información. Se definieron tres dimensiones utilizadas para el estudio:

1. Los objetos de este estudio que corresponden a todas las instancias que participan en el ciclo de publicación

académica de artículos científicos, y que son las encargadas de generar, diseñar, difundir, procesar y preservar los contenidos, son: las editoriales y las revistas académicas, las bases de datos de literatura académica, los servicios y aplicaciones asociadas a cualquiera de ellos, y las iniciativas locales, regionales o mundiales encargadas de promover e incentivar la tecnología semántica en la publicación académica.

2. Las condiciones indispensables para que la publicación se pueda realizar como Acceso Abierto: especificaciones como XML, RDF y OWL.
3. Los diferentes roles en el proceso de publicación académica: los autores, editores, evaluadores y lectores.

Estas tres dimensiones fueron analizadas con el fin de investigar y mapear el estado general que tiene la Web Semántica en la publicación académica (Figura 2).

Figura 2. Tres dimensiones de la Web Semántica de la publicación académica utilizadas en la investigación: objetos de estudio, tecnología semántica y roles. Fuente: elaboración propia.

Resultados y discusión

En cuanto a la tecnología, identificamos doce especificaciones que deben seguir las instancias encargadas de procesar artículos científicos para hacer publicación semántica (Figura 3).

La implementación del Acceso Abierto y Ciencia Abierta en el sector editorial, junto con la suma de las tecnologías computacionales, podrán desarrollar nuevos métodos de publicación, así como la implementación de procedimientos efectivos en el registro y difusión de resultados de investigación, verificación y replicación del estudio, del mismo modo que la reutilización de datos.

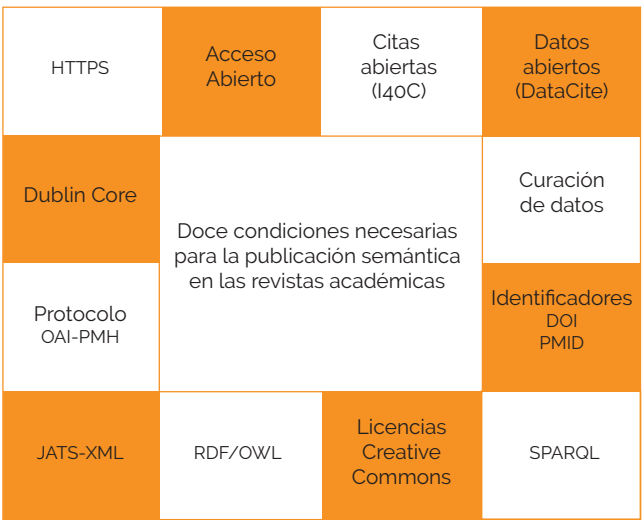

Figura 3. Resultado de la investigación en revistas y editoriales, que muestra doce condiciones para realizar publicación semántica. Fuente: elaboración propia.

Actualmente existen casos de revistas, editores académicos y casas editoriales que fomentan el enriquecimiento de los artículos publicados y por publicar, sumando la infraestructura semántica en cada caso. La propuesta es identificar, mediante una consulta web –especializada y en recursos de Acceso Abierto–, aquellas promociones a la publicación semántica, el objetivo y alcance de este tipo de publicación, su impacto en la revista y/o artículo, así como en otros que han sido publicados con reutilización de datos semánticos.

La estructura de un artículo semántico comienza con el etiquetado del documento en XML (trabajo que se hace en la editorial o instancia que publica); posteriormente, viene el trabajo de curación y etiquetado, generalmente se realiza en RDF y por un curador especializado (trabajo hecho en la editorial o institución de investigación); y finalmente la representación del conocimiento en ontología OWL (trabajo que realizan las colecciones semánticas) (Abad-Navarro *et al.,* 2020) (Figura 4).

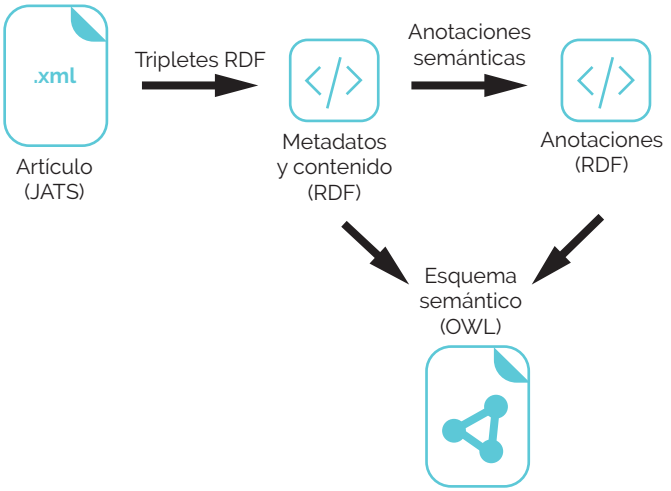

Figura 4. *Workflow* en la formación de documentos semánticos.
Fuente: modificado de Abad-Navarro *et al.* (2020).

Dentro de las colecciones web, son muy pocas las que contienen documentos estructurados para la generación de publicaciones semánticas. Un ejemplo es Biotea, una representación semántica con anotaciones sobre un subconjunto (incluye artículos de 7,407 revistas) de Acceso Abierto de PubMed Central (PMC). Esta colección cuenta con identificadores y un documento estructurado en XML, lo cual permitió el desarrollo de RDF de los artículos analizados (Figura 5). Tener entidades caracterizadas semánticamente hace posible que los agentes de *software* las procesen de varias maneras, por ejemplo, usando la asociación enfermedades-poblaciones-intervenciones para vincular los registros de salud o mediante el uso de la asociación gen-proteína-enfermedad para vincular las vías metabólicas. En

Biotea, las anotaciones se representan utilizando un formalismo interpretable por máquina. Como se ilustra en el prototipo, las notas se utilizan para clasificar, vincular, interconectar, buscar y filtrar (Garcia *et al*. 2018).

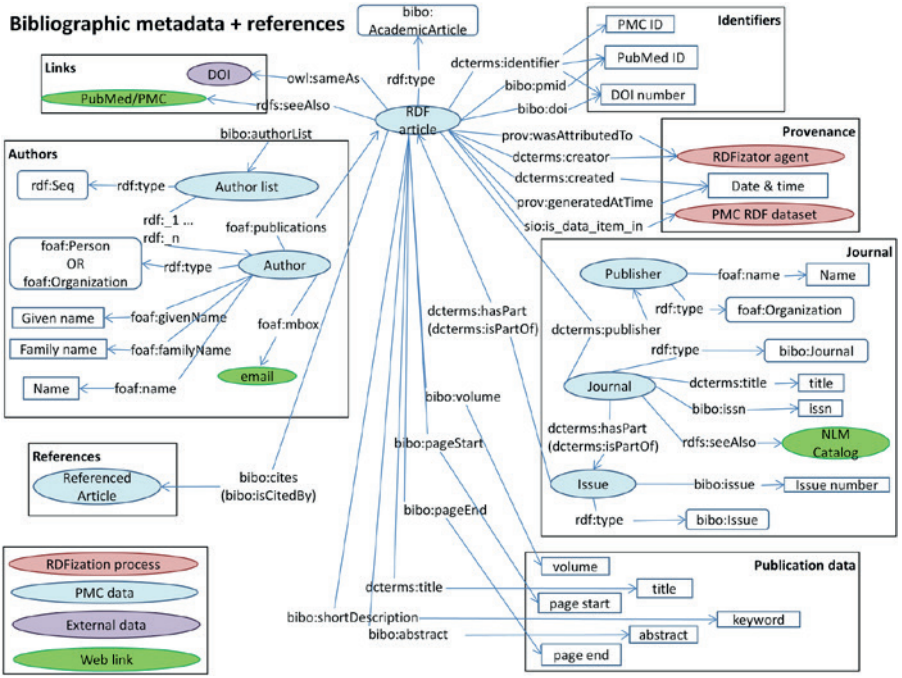

Figura 5. Estructura de metadatos de un artículo en RDF para hacer anotaciones semánticas en una colección especializada (PMC). Fuente: Garcia *et al*. (2018).

Otra aplicación de datos semánticos son las evaluaciones de hipótesis. La clave del éxito de *e-Science* es la capacidad de evaluar computacionalmente la validez de las hipótesis compuestas por expertos contra los datos experimentales publicados en artículos. HyQue, es una herramienta de Web Semántica para consultar bases de conocimiento científico y evaluar hipótesis biológicas. La base de conocimientos se consulta utilizando SPARQL y las consultas pueden incluir referencias a instancias o tipos. Los resultados de las mismas se evalúan en referencia a la estructura lógica de una hipótesis para calcular una puntuación que indica el nivel de soporte que los datos prestan a

tal hipótesis. Éstas, así como los datos de respaldo o refutación, están representadas en RDF y directamente vinculadas entre sí, lo que permite a los científicos navegar de datos a hipótesis y viceversa (Callahan, Dumontier y Shah 2011).

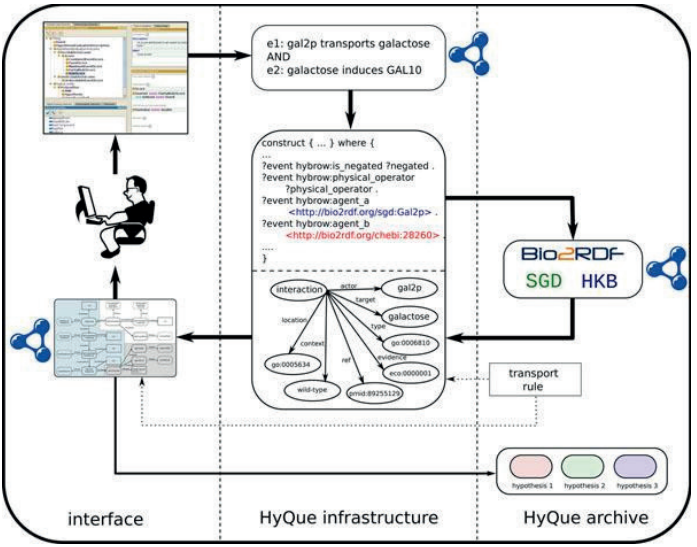

Figura 6. Esquema de la plataforma HyQue. Un usuario formula una hipótesis utilizando términos de la ontología de hipótesis (arriba a la izquierda), que se convierte en una consulta SPARQL correspondiente (centro superior). Las reglas de evaluación se aplican a los datos recuperados por la consulta SPARQL (centro inferior) para generar puntajes de soporte y contradicción. Se presenta al usuario una descripción general de los datos utilizados para evaluar la hipótesis junto con puntajes de apoyo/contradicción (abajo a la izquierda). Fuente: tomado de Callahan *et al*. (2011).

Hay cuatro formas de generar artículos semánticos: las anotaciones semánticas, la interconexión semántica, la integración semántica y el descubrimiento semántico. Pero actualmente las herramientas y los algoritmos computacionales permiten que los investigadores capturen desde un inicio sus resultados de investigación con una tecnología semántica formal; ésta sería la estrategia de publicación semántica genuina y deseable, pero la menos utilizada (Kuhn y Dumontier 2017).

Uno de los impedimentos más importantes, sin duda, es el cultural. Se conoce poco de las nuevas tecnologías informáticas en los ámbitos académicos. Éste definitivamente es un asunto de habilidades digitales y competencias informacionales de los interesados en el tema. Es necesario promover habilidades digitales entre los distintos roles relacionados con la publicación académica, esto es:

- Los autores, para que generen contenidos con características semánticas (ligados y estructurados).
- Los editores, de modo que estén familiarizados con esta tecnología y cuenten con la infraestructura en sus plataformas (como OJS, Crossref, Dimensions, DataCite, entre otros), así como capacitados en la generación de XML y la vinculación con ID's únicos (DOI, ORCID, CRediT, entre otros).
- Los lectores, para que conozcan la ventaja de estas innovaciones tecnológicas, estén familiarizados con ellas y las utilicen. En muchos casos existe la tecnología y es funcional, pero son muy pocos los involucrados que la aprovechan.
- Los curadores especialistas en un tema, de manera que permitan estructurar el conocimiento en RDF y OWL.
- Especialistas en cómputo, para que generen nuevas herramientas de fácil aprendizaje y menor tiempo invertido en programación orientada a objetos.

Por esto sería importante generar materiales de difusión que permitieran a todos los autores entender esta tecnología de una manera sencilla, amigable y visual, por medio de presentaciones, videos cortos, tutoriales e infografías.

Dado que el uso de los artículos científicos, la consulta de las revistas académicas, la recuperación de bases de datos y la implementación de servicios y aplicaciones de publicaciones científicas inicia en la formación universitaria, sería importante que se realizaran talleres, diplomados, sesiones de actualización y clases en línea que permitieran a los alumnos aprender a utilizar estas herramientas para la realización de reportes y lecturas. Una vez que se familiaricen con esta tecnología, comprenderán sus ventajas y la utilizarán de manera potencial.

En cuanto a los autores, revisores y editores, no sólo sería ideal que conocieran las ventajas de esta tecnología, sino que, como usuarios de los artículos, es necesario. La desventaja está en la capacitación técnica del uso de estas herramientas, así como en la vinculación con otras colecciones digitales y la conservación de la interoperabilidad informática.

Referencias

Abad-Navarro, Francisco, José Antonio Bernabé-Díaz, Alexander García-Castro y Jesualdo Tomás Fernandez-Breis. 2020. "Semantic Publication of Agricultural Scientific Literature Using Property Graphs". *Applied Sciences* 10, núm. 3: 861. https://doi.org/10.3390/app10030861

Berners-Lee, Tim y James Hendler. 2001. "Scientific publishing on the 'semantic web'". *Nature,* núm. 410: 1023–1024. https://doi.org/10.1038/nature28055

Berners-Lee, Tim. 2006. "Linked Data". *W3 Organization*. http://www.w3.org/DesignIssues/LinkedData.html

British Medical Journal-BMJ. 2010. "Semantic publishing: how to create richer metadata". *BMJ Group blogs*. https://blogs.bmj.com/bmj-journals-development-blog/tag/semantic-publishing/

Callahan, Alison, Michel Dumontier y Nigam H. Shah. 2011. "HyQue: evaluating hypotheses using Semantic Web technologies". *Journal of Biomed Semantics* 2, suppl. 2: S3. https://doi.org/10.1186/2041-1480-2-S2-S3

DeLeon, Alexander y Michel Dumontier. 2008. "Publishing OWL ontologies with Presto". *CEUR Workshop Proceedings,* 496, artículo 13. http://ceur-ws.org/Vol-496/owled2008dc_paper_19.pdf

Garcia, Alexander, Federico Lopez, Leyla Garcia, Olga Giraldo, Victor Bucheli y Michel Dumontier. 2018. "Biotea: semantics for PubMed Central". *PeerJ* 6: e4201. https://doi.org/10.7717/peerj.4201

Gene Ontology Consortium. 2004. "The Gene Ontology (GO) database and informatics resource". *Nucleic acids research* 32, suppl. 1: D258–D261. https://doi.org/10.1093/nar/gkh036

Kuhn, Tobias y Michel Dumontier. 2017. "Genuine semantic publishing". *Data Science* 1, núm. 1-2: 139-154. https://doi.org/10.3233/DS-170010

Peroni, Silvio. 2017. "Automating semantic publishing". *Data Science* 1, núm. 1-2: 155-173. https://doi.org/10.3233/DS-170012

Shotton, David. 2009. "Semantic publishing: the coming revolution in scientific journal publishing". *Learned Publishing* 22, núm. 2: 85-94. https://doi.org/10.1087/2009202

Shotton, David, Katie Portwin, Graham G. Klyne y Alistair Miles. 2009. "Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article". *PLOS Computational Biology* 5, núm. 4: e1000361. https://doi.org/10.1371/journal.pcbi.1000361

Reseñas curriculares

Eduardo Alvarez. Maestro en Ciencias egresado del Departamento de Innovación Biomédica de CICESE. Actualmente colabora como editor asociado de la revista arbitrada *Economía Creativa* y como *Sponsor* de Crossref en México a través de Meta-datos.

Layla Michán. Doctora e investigadora egresada de la Facultad de Ciencias de la UNAM. Profesora de Licenciatura y Posgrado, especialista en las teorías, métodos, conceptos, herramientas y aplicaciones para el manejo de información, interesada en procesos como estructuración, recuperación, sistematización, curación, análisis y publicación de información científica digital. Actualmente dirige el laboratorio de Bioinformación de la Facultad de Ciencias, UNAM, especializado en indagar la dinámica, aplicaciones y problemas de la literatura, la información y el conocimiento biológico.