¿Qué es la "Semantic Publishing"? ¿Cómo puede ayudar a la edición y publicación de revistas científicas?

M. E. Rubio Lucas[1]

[1] *Facultad de Comunicación y Documentación, Universidad de Murcia, mariaester.rubio@um.es*

Gracias al vertiginoso avance de las TICs y a la explosión de información en internet, la información científica ha dejado de estar oculta y reservada a aquellos investigadores que tienen acceso, por medio de costosas suscripciones, a bases de datos comerciales. La forma de consumir información científica, revistas y/o artículos científicos está sufriendo constantes cambios a causa de la penetración de internet en todos los ámbitos académicos-científicos. Aún así, la edición y publicación de las publicaciones científicas no han incorporado totalmente grandes cambios editoriales propios de la revolución tecnológica que posibilita la era digital.

Con la generalización del uso de internet para la difusión de información en formato digital, la publicación de las revistas cambió el formato y el medio, pero no ha llegado aún a aprovecharse del todo de las inherentes ventajas de este medio, más allá de las evidentes ventajas económicas y la inmediatez en el acceso (más si cabe cuando la publicación es de acceso abierto):

> "La edición de revistas científicas en muchos casos aún tiene pendiente abordar la revolución digital con todas sus consecuencias. Hace falta "pensar en digital", abandonar toda dependencia de la versión impresa, que puede mantenerse si es necesario como estrategia de distribución, pero que no debe frenar la explotación y difusión a través de internet". (Rodríguez-Yunta, 2013, p.207)

Mejorar la edición y publicación de las revistas científicas con los medios y herramientas inherentes al nuevo entorno digital, como son los metadatos, facilitaría sin duda alguna mejoras en la recuperación, difusión, posicionamiento, impacto o ubicuidad de aquellos documentos científicos depositados en la web.

En este campo concreto, la propuesta de Web Semántica elaborada por Berners-Lee et al. (2001) y la posibilidad de describir semánticamente la información publicada en este sistema de información (Berners-Lee and Hendler, 2001) abre un mundo lleno de posibilidades para la publicación y recuperación de información en general, que los editores de las revistas científicas pueden y deberían aprovechar. Estamos convencidos de que dotar de significado semántico a la información científica disponible en la web es más necesario que nunca, ya que abre un nuevo espacio donde se facilita la recuperación de información de forma más precisa y relevante, al igual que se favorece la interoperabilidad y la reutilización de información por parte de las máquinas y/o motores de búsqueda. Para ello, el marcado semántico se sirve de propiedades, clases, relaciones o etiquetas de marcado que hacen referencia a un vocabulario concreto y a una sintaxis determinada (como microdatos, microformatos o RDFa, etc.).

La descripción semántica no solo pretende describir los aspectos formales o de contenido de las publicaciones científicas, sino ampliar el alcance de esta descripción a

otros elementos que tradicionalmente han pasado casi desapercibidos: tablas, imágenes, gráficos, figuras, fotografías, o anexos de los artículos; considerando así que "el artículo no es la única unidad de información objeto de una posible recuperación documental". (Rodríguez-Yunta, 2013, p. 208).

Las posibles mejoras que puede aportar la Web Semántica a la edición las revistas científicas podrían ir en la línea de aportar un marcado (o etiquetado) semántico en estas publicaciones, ampliando la idea de publicar digitalmente a un concepto más amplio y con objetivos adicionales. Este innovador campo de trabajo se ha dado en llamar "Publicación Semántica" (del inglés 'Semantic Publishing'), concepto en torno al cual se ha concentrado, en los últimos años, el trabajo del profesor David Shotton, principal investigador internacional en este campo. El autor (2009) define este concepto como todo aquello realizado para mejorar el significado de los artículos de las revistas científicas, facilitando así su descubrimiento, habilitando su enlazado semántico con artículos relacionados, proporcionando acceso a los datos dentro del artículo en forma procesable y/o facilitando la integración de datos entre artículos. Esto implica enriquecer el artículo científico con metadatos que puedan ser procesados y analizados automáticamente para mejorar la verificabilidad de la información publicada y proporcionar la capacidad de descubrimiento de manera automatizada.

La utilización del marcado semántico incrementa el valor intrínseco de los artículos, aumentando implícitamente el valor de la información contenida en ellos, haciéndola más comprensible y fácil de recuperar tanto por las personas como por las máquinas. Otro aspecto significativo de esta idea es que además de potenciar la recuperación de información por parte de las máquinas se aporta a los lectores una mayor profundidad informativa con un acceso más completo y seguro a la información.

Para Shotton (2009), el desarrollo e implementación del enriquecimiento semántico de las publicaciones ayuda a crear nuevas oportunidades de negocio en forma de valor añadido en los servicios de publicación, algo que si bien no parece ser el objetivo primigenio de un editor de revistas científicas, puede convertirse en una ventaja competitiva de gran potencial que merece ser explotada. Además, los lectores y autores se beneficiarían de un acceso más rápido y mejor a los datos relativos a las publicaciones. En esta misma línea Shotton (2009) establece 6 reglas para la publicación semántica:

1. Comenzar de manera sencilla e ir implementando incrementalmente mejoras.
2. Esperar grandes cosas de los autores.
3. Explotar las habilidades existentes de las que se disponen en casa.
4. Utilizar estándares siempre que sea posible.
5. Publicar en la web el conjunto de datos "en crudo".
6. Publicar metadatos de los artículos, particularmente listas de referencias en un lenguaje capaz de ser leído por los sistemas de búsqueda automatizados.

Actualmente existen diferentes tecnologías de la web semántica con la que poder llevar a cabo la descripción semántica de la información en la web. Entre estas tecnologías, merecen ser destacadas, con el objeto de este trabajo, las siguientes:

- Microdatos: forma de etiquetar contenido para describir estructuras de información específicas (personas, eventos, calendarios, etc.) en la web. Utilizan la especificación HTML5 descrita por la W3C para asignar nombres breves y descriptivos a elementos y propiedades. Su objetivo fundamental es posibilitar a los motores de búsqueda, robots y navegadores el procesamiento y la extracción de información semántica de la web para poder responder y mostrar los resultados de manera más precisa a las necesidades de información de los usuarios. Cabe destacar el vocabulario Schema.org promovido por los grandes buscadores Google, Yahoo!, Bing y Yandex con el objetivo de crear, mantener y promover un esquema de datos estructurados en internet.
- Microformatos: tipo de metadatos creados para el marcado semántico de sitios web en HTML. Ayudan a los sistemas de información a identificar estructuras de información relevantes en una página web. No se trata de un nuevo lenguaje de marcado, sino de una forma o método de marcado a través de etiquetas HTML para un tipo concreto de información (opiniones, eventos, productos, personas, etc.).
- RDF ('Resource Description Framework'): modelo de estándar para el intercambio de datos en la web. Gracias a la descripción de la información con este esquema, se le dota a la información de un significado semántico capaz de ser procesado y entendido por las máquinas.
- HTML5: 5º versión del popular lenguaje de marcas HTML desarrollado para describir documentos en la web por parte de la W3C. Permite a los desarrolladores web introducir marcas semánticas en sus etiquetados o marcas.
- XHTML ('eXtensible HyperText Markup Language)'. Se trata de una versión de HTML para mostrar datos definidos en XML.

Son diversos los elementos susceptibles de verse afectados por la Publicación Semántica. Desde nuestro punto de vista se pueden establecer cuatro categorías recogidas a continuación:

Descripción de revistas:

Son aquellos términos o propiedades empleadas para la identificación formal y de contenido de las publicaciones periódicas. Estas propiedades recogen los aspectos identificativos de las revistas tales como título, ISSN, volumen, regularidad, editor y datos de contacto, etc.; es decir, datos importantes en cualquier referencia bibliográfica. Su descripción favorecerá la recuperación de la información contenida en la revista y realizar vínculos de afinidad entre las diferentes revistas del ámbito. Mediante las propiedades de relaciones, se puede establecer el vínculo entre las revistas y los artículos que contiene.

Descripción de artículos:

Se trata de identificadores básicos que permitan localizar y recuperar aquellos artículos referenciados de manera inequívoca como: autores, título, palabras clave, resumen, número de páginas. Así mismo, los artículos se podrán relacionar con las revistas de las que procedan.

Justificación y propósito de las citas:

Esta categoría está compuesta por propiedades que permiten la justificación, argumentación, contextualización, definición de propósitos, etc. de las citas incluidas en el texto del artículo científico. Además, el conocimiento de la ubicación de las citas permitirá evaluar la relevancia o peso de cada una de ellas dentro del artículo (por ejemplo, las menciones en conclusiones o resultados suelen poseer mayor relevancia que las contenidas en la introducción). Las propiedades de esta categoría pueden ser: referencia bibliográfica, citas, propósito de las citas, argumentación, discusión en torno a la cita, etc. Esta categoría es quizá, desde nuestro punto de vista, la más importante y novedosa de las cuatro, ya que contempla elementos que los vocabularios de metadatos no suelen contemplar. Conociendo más información sobre las referencias bibliográficas, el autor y/o lector obtendrá una mayor profundidad informativa sobre la argumentación del texto y otra ventaja adicional es poder establecer relaciones entre los diferentes trabajos científicos a través de sus referencias bibliográficas.

Partes del documento

Se trata de propiedades para la descripción estructural y formal del artículo científico. Pretenden ser una guía de localización para conocer dónde se encuentra la cita, los datos del autor, etc. Entre las más importantes destacan: resumen, palabras clave, paginación, introducción, metodología, resultados, conclusiones, apéndices, notas, gráficos, tablas, imágenes, etc.

Nuestra tarea actualmente consiste en estudiar todos los elementos propuestos para ser descritos semánticamente y elaborar una propuesta de microdatos que permita su descripción en la web. La propuesta deberá ser debatida en foros abiertos de expertos de prestigio como el de la W3C o BiblioGraph.net (http://bibliograph.net/), con cuyo director ya nos hemos puesto en contacto y recibido un primer *'feedback'* de la propuesta. La aceptación de la propuesta de microdatos de los grupos mencionados es de vital importancia para asegurar su expansión, difusión y penetración en la web.

El ejemplo presentado a continuación, es una muestra de cómo realizar el marcado semántico en un artículo científico. El ítem descrito pertenece a la clase "*Periodical*". Con las diferentes etiquetas (como "*journalName*", "*volumeNumber*", "URL", etc.) se identifican algunos de los principales elementos de la revista. Además, permite referenciar aquellos artículos que la componen con la propiedad "*hasArticle*".

```
<div itemscope itemtype="http://schema.org/Periodical">
  <h1 itemprop="journalName">Cuadernos de Gestión de Información</h1>
        <span itemprop='volumeNumber'>2011 </span>
        <span itemprop='hasArticle'>Information Management, today and tomorrow</span>
        <span itemprop='URL'>http://revistas.um.es/gesinfo/issue/view/13051</span>
        <span itemprop="issn">2253-8429</span></span>
        <span itemprop="journalRegularity">Anual</span>
        <span itemprop="editor">Facultad de Comunicación y Documentación, Universidad de Murcia</span>
        <span itemprop="addressCountry">Espanya</span>
        <div itemprop="address" itemscope itemtype="http://schema.org/PostalAddress>
                <span itemprop="addressLocality">Murcia</span>,
                <span itemprop="addressRegion">Murcia</span>
                <span itemprop="postalCode">30100</span>
                <span itemprop="streetAddress">Campus Universitario de Espinardo</span>
        </div>
 </div>
```

A modo de conclusión, la utilización de esquemas de metadatos cobra actualmente una mayor importancia que en otras épocas del desarrollo de la publicación en la web, debido fundamentalmente a la necesidad de descripción de documentos en este sistema de información. La utilización de este tipo de herramientas semánticas, junto al uso de perfiles de aplicación de metadatos y/o ontologías para describir semánticamente la información plasmada en los documentos, es indispensable para favorecer la interoperabilidad y reutilización de la información en la web. Asimismo, se favorece que las máquinas y motores de búsqueda puedan inferir información y ser capaces de establecer relaciones entre documentos debido a su descripción semántica. El vocabulario Schema.org se atisba como una buena apuesta para este tipo de descripción, debido a la facilidad de implementar esquemas de microdatos. Este hecho hace posible su expansión de mano de los desarrolladores web que van a encontrar esta tarea menos tediosa y más fácil de desarrollar.

La Publicación Semántica no sólo persigue facilitar su tarea a los motores de búsqueda para aumentar su efectividad en la búsqueda de información científica en la web, también aporta información adicional a los lectores de los artículos añadiendo una mayor profundidad informativa y un acceso más completo y seguro a la información. Las categorías propuestas intentan cubrir todos los aspectos básicos y una amplia diversidad de opciones en la descripción semántica de las publicaciones científicas en la web. Uno de los aspectos recogidos que se ha considerado más relevante e innovador es incluir la justificación de las citas y de las referencias bibliográficas en el texto (lo que podría llamarse “bibliografía en contexto”). La descripción de las intenciones, argumentaciones, discusiones o justificaciones es un aspecto poco o nada explotado hasta en la edición de publicaciones científicas, por lo que se ha considerado el punto de partida de futuros trabajos relacionados.

BIBLIOGRAFÍA

Berners-Lee, T. & Hendler, J. (2001). Publishing on the semantic web. Nature, 410 (6832), 1023–1024. doi:10.1038/35074206

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The Semantic Web. Scientific American, 284 (5), 28–37. Recuperado de http://www.scientificamerican.com/article/the-semantic-web/

Rodríguez-Yunta, L. (2013). Indización en profundidad y aplicación de metadatos a materiales suplementarios en la edición de revistas. *Anuario ThinkEPI*, 8, 207–210. Recuperado de www.thinkepi.net/

Shotton, D. (2009). Semantic publishing: the coming revolution in scientific journal publishing. *Learned Publishing*, 22(2), 85–94. doi: 10.1087/2009202