B.2. Cuando la búsqueda se vuelve semántica: *SWoogle*

Por **Eva Méndez**

Méndez, Eva. "Cuando la búsqueda se vuelve semántica: SWoogle". En: *Anuario ThinkEPI*, 2007, pp. 65-68.

"Hay que romper el círculo vicioso: los creadores de contenidos Web no usan metadatos porque los buscadores no los indizan; los buscadores no indizan metadatos porque los creadores de información no los usan"

"En vez de crear una alternativa semántica a Google, SWoogle crea un Google para la Web semántica"

AHORA QUE TODO EL MUNDO, para bien o para mal, habla de *Google*, voy a hablar de *Swoogle*[1], algo así como un *Google* para la Web semántica.

Cada día se usa *Google*, pero no cada día se ha habla tanto de EL buscador con mayúsculas como en los últimos meses. Desde el *Herald tribune*[2] hasta *El país*[3] toda la prensa internacional se hizo eco de *Quaero*, la pretendida competencia europea a *Google*, noticia que además ha sido "bloggeada" y comentada por doquier[4].

El dominio *quaero.com* pertenece a una compañía de marketing de Charlotte, NC, EUA, que no tiene nada que ver con el proyecto europeo y que no la debió de ver más gorda en su vida por la cantidad de accesos que ha tenido en este tiempo (deseosos de ver cuál es la promesa Europea de la recuperación de información en la Web, muchos hemos tecleado esa dirección, ya que *la buena*[5] se cerró al público a la espera de la estelar aparición); es posible que *Thomson* o los franceses o los alemanes les compren el dominio por una buena cantidad.

Mientras la vieja Europa trata de dar forma a su política de información a través del programa *i2010* (donde las bibliotecas tendrán mucho que decir, por cierto) y de hacer frente al dominio cultural electrónico anglosajón, particularmente al norteamericano, en Estados Unidos se siguen haciendo buscadores en las universidades (que quién sabe si un día cotizarán en bolsa).

En 1994 **Filo** y **Yang** crearon *Yahoo!*; en 1998 **Brin** y **Page** crearon *Google*. En el caso de *SWoogle*[1] el buscador no sale de las aulas de *Stanford*, sino de los laboratorios de investigación de la *Universidad de Maryland*, Baltimore, en el seno del grupo *eBiquity*[6] y de un proyecto financiado desde 2004, con todos los honores, por la *Nacional Science Foundation* (¡el dinero que suele dar tan preciada institución para el fomento de la investigación en USA!). El grupo *eBiquity* está dirigido por **Tim Finin** (otro Tim que añadir a la historia de la Web junto a **Tim Berners Lee**, **Tim O'Reilly** y **Tim Bray**) y trabaja, entre otros temas, todas las tecnologías relacionadas con la Web semántica, sobre todo metadatos, ontologías y redes sociales basadas en *FOAF*[7].

Voy a hablar de *Swoogle* porque en febrero de 2006 se lanzó la nueva versión del proyecto con grandes mejoras que lo hacen muy prometedor y, cuanto menos, interesante.

Web semántica: contexto tecnológico que necesita herramientas (y buscadores)

Si los buscadores son el tema más en boga en los últimos meses, la Web semántica es uno de los temas más boga en los últimos

años y cobra especial interés si correlacionamos ese conjunto de tecnologías y siglas (RDF, XML, OWL, etc. etc.) con los buscadores.

Uno de los problemas de la Web semántica –que he comentado ya en otras ocasiones[8]–, es que las aplicaciones pertenecen aún a la nebulosa del *middleware* (software intermediario que conecta dos aplicaciones), difíciles de hincarles el diente por ajenos a la programación.

De todas maneras el verdadero gran problema es que no existen buscadores de carácter global que permitan búsquedas *all-the-web* basándose en RDF[9] o en metadatos DC, que eternizan la situación *pescadilla que se muerde la cola* destacada en 1997 en los estudios, workshops, y demás eventos reflexivos en torno a la validez o no de los metadatos y a la búsqueda de masa crítica en contexto de información Web-global[10]: los creadores de contenidos Web no usan metadatos porque los buscadores no indizan en función de esos metadatos y los buscadores no indizan en función de los metadatos porque los creadores de información no los usan o lo hacen de una manera errónea y/o capciosa.

Hasta ahora los buscadores de carácter general (ni *Google*, ni *Quaero* supongo) no indizan RDF, ni metadatos, ni basan su recuperación por materias en ontologías, con lo cual la Web semántica queda reducida a:

–Un conjunto de sitios (*islas semánticas*) que utilizan sus ontologías ad-hoc o desarrollan sus motores de búsqueda aplicados a su metainformación en RDF o OWL, dentro de un sitio o dominio informativo particular, como por ejemplo *SWED*[11], un portal semántico para información de medio ambiente.

–Un conjunto de aplicaciones que entienden algunas notaciones semánticas (*RSS*, *FOAF*), basadas en programación relativamente sencilla (*Ruby*, *Ajax*), que permiten agregar y difundir contenidos en un contexto de información dirigido al usuario (a un usuario cooperativo), a las que denominamos de una forma intuitiva, y no menos afortunada, *Web 2.0*.

Entretanto la Web semántica definida como una base de datos enlazada globalmente que permita búsquedas precisas y fiables[12] o el motor de búsqueda inteligente que vaticinaba **Tim Berners Lee**, el del *I want to buy a pair of shoes*[13] es algo completamente parcial todavía.

SWoogle 2006

SWoogle es un sistema de indización y recuperación para documentos de la Web semántica (*Semantic web documents* o *SWD*s), o lo que es lo mismo documentos escritos básicamente en RDF y OWL, aunque

también DAML en algunos casos. Este buscador recupera, procesa, analiza e indiza *documentos SW* que estén disponibles online, pero lo más curioso es que lo hace a través de un sistema de búsqueda y resultados de interfaz Web similar a *Google*.

El grupo de trabajo *eBiquity*[6] que desarrolla este proyecto en la *Universidad de Maryland*, parte de la base de que *Google* ha cambiado la forma en que accedemos a la información Web y que se ha convertido en una tecnología clave para la búsqueda de información. Por ello *SWoogle*, en vez de crear una alternativa –semántica– a *Google*, crea un *Google* para la Web semántica, lo cual parece, tanto desde un punto de vista estratégico, como operativo en términos de funcionamiento y aceptación, brillante.

La nueva versión de *SWoogle* tiene un modelo y una base más simple que la anterior, y hasta un diseño mucho más claro. Recoge más de 850.000 documentos web-semánticos recolectados de la Web, bien buscando directamente en ficheros RDF y OWL

o a través de páginas web (html) que pueden contener documentos SW. Más de 10.000 ontologías disponibles en la Web (1.0, y 2.0), almacenadas en una base de datos MySQL en forma de URIs, pero también permite buscar en los términos de cada vocabulario/esquema/ontología. Por ejemplo, podemos buscar todos los esquemas que contienen la propiedad "title".

Ahora mismo *SWoogle* es utilísimo para los implementadores de la Web semántica y de software relacionado, ya que permite:
– estudiar la magnitud y el crecimiento;
– buscar y recopilar clases y propiedades (términos de la Web semántica, SWTs) o las ontologías en que se conforman;
–a poyar productos de carácter semántico.

Hasta hace muy poco tiempo las *herramientas semánticas* (editores de ontologías o sistemas para la creación de esquemas de metadatos) habían sido de acceso libre como *Protégé*[14] o las aplicaciones de *MindSWap* (*Swoop* y *Smore*)[15] y/o pertenecientes a proyectos en desarrollo. Sin embargo, también a principios de este año, la famosa compañía *Altova* (los creadores de *XMLSpy*) han lanzado *SemanticWorks*[16] un editor de vocabularios que trabaja tanto con RDF y XML como con *N-triples*[17].

SWoogle permite medir, controlar y analizar los vocabularios de la Web semántica u ontologías (a efectos de *SWoogle* tenemos que considerar *Dublin core* o *FOAF* como una ontología –cosa que a mí particularmente me parece errónea–). Aún no es un buscador dirigido al usuario final para encontrar recursos Web, sino que es más bien un "parabuscador" para buscar, clasificar e incluso validar documentos y vocabularios de la Web semántica.

Otros buscadores de la Web semántica y futuro de la búsqueda semántica

SWoogle no es la única iniciativa de un buscador para la Web semántica existen otros parabuscadores (me he inventado la palabra, pero me gusta, refleja la misión de estas herramientas de buscar ontologías y vocabularios dirigidos a los agentes de software para que éstos mejoren las búsquedas):
–*SemanticWeb Search*[18], que tiene buen aspecto y vocación comercial, que busca no sólo vocabularios sino recursos basados en ellos según el parámetro de query: "busca un recurso que sea de un tipo de vocabulario con alguna propiedad que contenga el término x", donde el "vocabulario" puede ser RDF, RSS, FOAF, DOAP[19] o calendarios y otras aplicaciones sencillas RDF, y "x" es una propiedad específica de ese tipo de vocabulario.
–Existen asimismo tentativas más o menos exhaustivas de inventariar las ontologías existentes en la Web semántica, pero en ningún caso con las opciones de búsqueda que ofrece *SWoogle*. Algunas de ellas son *Ontology Library* (de *SemanticWeb Central*)[20] o el propio proyecto del *Consorcio* (*W3C*) en este sentido: *Ontaria*[21], que lleva un tiempo en revisión, esperemos que para mejorar sus resultados, que nunca han sido como los de *SWoogle*.
– También la investigación europea a hechos sus pinitos en buscadores semánticos como el proyecto del *Inria* (*Institut National de Recherche en Informatique et en Automatique*), la primera *casa europea* del W3C, que ha desarrollado *Corese* (*Conceptual resource search engine*)[22], un motor de búsqueda basado en gráficos conceptuales, que se centra en la visualización de la Web semántica.

A pesar de todas estas iniciativas de búsqueda de ontologías y vocabularios para la Web semántica, aún no podemos hablar de una búsqueda global basada en RDF, sino sólo de búsquedas globales de y en RDF.

La clave de la recuperación de información en esa gran base de datos enlazada a la que nos referíamos antes[12], quizás tenga que ver con el desarrollo de un lenguaje de consultas que dote a esa "especie de base de datos" de la Web semántica de la misma consistencia que SQL da a las bases de datos relacionales reales. *Sparql*[23] es ese lenguaje de consulta donde definitivamente se darán la mano SW y Web 2.0 permitiendo un lenguaje de interrogación preciso para RDF y un protocolo de recuperación que posibilite aunar los recursos distribuidos. La existencia del lenguaje y del protocolo *Sparql*, junto a otras tecnologías como los microformatos, hacen cada vez más válida mi idea de que la Web 2.0 es (como todo) una actitud... Pero si quieren de esto hablamos otro día.

Más información sobre SWoogle:

–Blog de SWoogle
http://ebiquity.umbc.edu/blogger/index.php?cat=24

–Swooglers, un grupo de discusión en *Google* sobre *SWoogle* (aunque parezca un trabalenguas)
http://groups.google.com/group/swooglers/

Notas:

1. SWoogle
http://swoogle.umbc.edu/

2. "Europeans weigh plan on google challenge". En: *International herald tribune*, 18 enero 2006.
http://www.iht.com/bin/print_ipub.php?file=/articles/2006/01/17/business/quaero.php

3. "Europa desafía el poder de Google". En: *El país*, 15 enero 2006 *http://www.elpais.es/* (suscriptores)
y "Quaero: los ojos de Europa en la Red: análisis español al buscador franco-alemán que se presenta en febrero". En: *El país*, 27 enero 2006.
http://www.elpais.es/articulo/elpportec/20060127elpepunet_7/Tes/internet/Quaero/ojos/Europa/Red

4. Véase la nota de **Ernest Abadal** y **Lluís Codina** sobre la excepción cultural y la polémica de *Quaero*, publicada en este Anuario.

5. *Thomson*, líder del consorcio europeo de *Quaero*, cerró la página de pruebas del buscador el 12 de enero de 2006. Esta empresa no tiene relación con la norteamericana del mismo nombre propietaria de *Dialog*, *ISI* y muchas otras fuentes de información. Esta *Thomson* se dedica a multimedia, telefonía, cine y TV, y ha engullido marcas como *RCA*, *CSF*, *General Electric*, *Technicolor*, etc. *https://www.thomson.net/*

6. *eBiquity* (*University of Maryland*).
http://ebiquity.umbc.edu/

7. *FOAF* (*Friend of a friend*) es un vocabulario RDF para describir personas y relaciones entre personas que fundamenta algunas de las redes sociales de hoy en día como *LiveJournal* o *MyOpera*.
http://www.foaf-project.org/

8. "La Web semántica: una Web más bibliotecaria". *Boletín CLIP, nº 41*.
http://www.sedic.es/p_boletinclip41.htm

9. RDF (resource description framework)
http://www.w3.org/RDF/

10. Metadata workshop: report of the workshop held in Luxembourg, 1-2 Dec. 1997. [s.l.]: European Commission, Directorate General XIII – E/4, February 1998, pp. 9-10, 13.
ftp://ftp.cordis.lu/pub/libraries/docs/metadata1.pdf

11. SWED (Semantic web enviromental directory).
http://www.swed.org.uk/

12. Esta es la definición de Web semántica que encabeza los trabajos del *ILRT* (*Institute for Learning & Research Technologies* de la *Universidad de Bristol*) en este sentido.
http://www.ilrt.bris.ac.uk/projects/web_futures

13. **Lee, Tim B.** *Weaving the Web: the original design and ultimate destiny of the World Wide Web by its inventor*. San Francisco: Harper-Collins, 1999, p. 133.

14. Protégé. National resource for biomedical ontologies and knowledge bases.
http://protege.stanford.edu/

15. *Smore* y *Swoop* son aplicaciones opensource desarrolladas también en la *Universidad de Maryland*, en este caso en el laboratorio *MindSWap* para editar ontologías.
http://www.mindswap.org/2005/SMORE/
http://www.mindswap.org/2004/SWOOP/

16. SemanticWorks
http://www.altova.com/download/semanticworks/semantic_web_rdf_owl_editor.html

17. http://www.w3.org/2001/sw/RDFCore/ntriples/

18. SemanticWeb search
http://www.semanticwebsearch.com/

19. DOAP: *Description of a project* es un vocabulario RDF, similar a FOAF en cuanto al tipo de propiedades que maneja, para describir proyectos de software opensource.
http://usefulinc.com/doap

20. Proyecto de inventario o repositorio de ontologías de SemWebCentral.
http://www.semwebcentral.org/index.jsp?page=ontologies

21. Ontaria (W3C)
http://www.w3.org/2004/ontaria/

22. Corese
http://www-sop.inria.fr/acacia/corese/

23. Sparql: Query Language and Data Access Protocol for the Semantic Web. El lenguaje de consultas es un borrador del W3C y el último borrador del Protocolo de acceso SPARQL es del 25 de enero.
http://www.w3.org/TR/rdf-sparql-query/
http://www.w3.org/TR/rdf-sparql-protocol/

Eva Mª Méndez Rodríguez
Departamento de Biblioteconomía y Documentación. Universidad Carlos III de Madrid.
emendez@bib.uc3m.es