

# Google, dsi y la sindicación de contenidos mediante rdf/rss

Por Jorge Serrano Cobos

**LOS PROFESIONALES DE LA INFORMACIÓN** hemos encontrado en *Google*, en los blogs y en los cambios tecnológicos que éstos han producido una herramienta estupenda para obtener información de alta especialización, actualizada y en formato inteligible por los sistemas informáticos, una herramienta que merece la pena conocer y estudiar: la sindicación de contenidos.

La sindicación puede usarse en sistemas de gestión de contenidos (CMS) complejos o simples, en entornos web o en intranets. Está directamente relacionada con la difusión selectiva de información y, a modo de ejemplo práctico de uso como veremos a continuación, incluso a través de una aplicación gratuita que nos permitirá estar actualizados casi al minuto de las últimas noticias que se producen sobre una palabra clave o expresión de búsqueda concreta.

Toda la filosofía de la sindicación de contenidos se basa en el formato rss (*rich site summary*). Estructuraremos la explicación en forma de preguntas y respuestas (faqs), al igual que un taller práctico.

## ¿Para qué sirve rss y la sindicación de información?

Con la proliferación de sitios de noticias, blogs, artículos, etc., generada a diario en la Red, cada vez nos resulta más difícil estar al tanto de toda la información disponible, no es posible ni visitando tus "favoritos" ni buscando en *Google*.

Pero los blogs han puesto de moda la sindicación de noticias: ofrecen su información en formato rss (esto es, sindicación de informa-

### Más información en:

—Grupos de news de Google sobre rdf:

<http://groups.google.com>

(buscar aquí por rdf o por rdf rss)

—Foro Rss-dev de Yahoo groups en el que participa **Aaron Swartz**:

<http://groups.yahoo.com/group/rss-dev/message/722>

—Un truco de *Microdoc news* para buscar documentos rss y rdf que hablen sobre una cadena de búsqueda en *Google*, por ejemplo: `iraq filetype:rss or filetype:rdf or filetype:rss.xml`

—*Microdoc news*:

<http://www.microdocs-news.info/infoSeeker/2003/03/28.html#a447>

ción, permiten que se reutilice). Es legible a través de una serie de aplicaciones (agregadores) que aúnan todas esas informaciones (*feeds*) agregando todos los artículos sindicados (de ahí que a estos programas les llamen de esa forma) y consultando cada cuanto queramos todas esas estupendas fuentes de información. Nosotros sólo tenemos que esperar y leerlas todas desde un único punto de lectura.

## ¿Qué es rss?

Es un formato de intercambio (sindicación) de contenido. Está basado en xml y tiene dos estándares. Uno de ellos toma como referencia a rdf, un formato de metadatos que está íntimamente relacionado con la web semántica, impulsada por **Tim Berners-Lee**. Básicamente permite no tener que na-

vegar web a web para leer cada día lo nuevo que se publica en ellas, pudiendo hacerse sin navegar, con una herramienta de agregación de noticias.

Sobre este punto, más información en:

—¿Cómo lees tus bitácoras favoritas?

<http://fernand0.blogalia.com/?historias/3005>

—Rss para principiantes.

<http://www.matotuonda.com.ar/archives/000119.php>

—La guerra del rss, de **Íñigo Arbildi**.

[http://trucosdeGoogle.blogspot.com/2003\\_02\\_01\\_trucosdeGoogle\\_archive.html#89384546](http://trucosdeGoogle.blogspot.com/2003_02_01_trucosdeGoogle_archive.html#89384546)

—Y para entenderlo con detalle, el tutorial *Rss workshop*.

<http://gils.utah.gov/rss/>

### ¿Qué es un agregador?

Lee y entiende los *feeds*, las fuentes de información sindicadas de cada sitio web elegido, y ofrece los titulares de los contenidos de cada una de ellas. Primero vemos los titulares, así decidimos si interesa leer el resto, pues pinchando sobre él se puede acceder a un resumen, o bien al contenido completo. Se parece mucho a un programa de correo estilo *Outlook*.

Hay muchos sitios donde pueden conseguirse estos agregadores; en este ejemplo usaremos *Feedreader*: simple, cómodo, limpio, en *Windows*.

<http://www.feedreader.com/>

<http://www.hot.ee/isys/feedreader2.4.exe>

Ya tenemos el programa, pero ahora hay que decidir qué fuentes

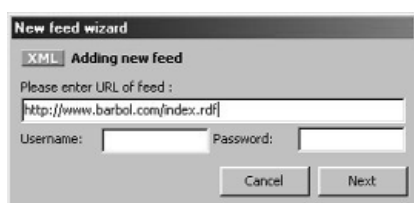


Figura 2

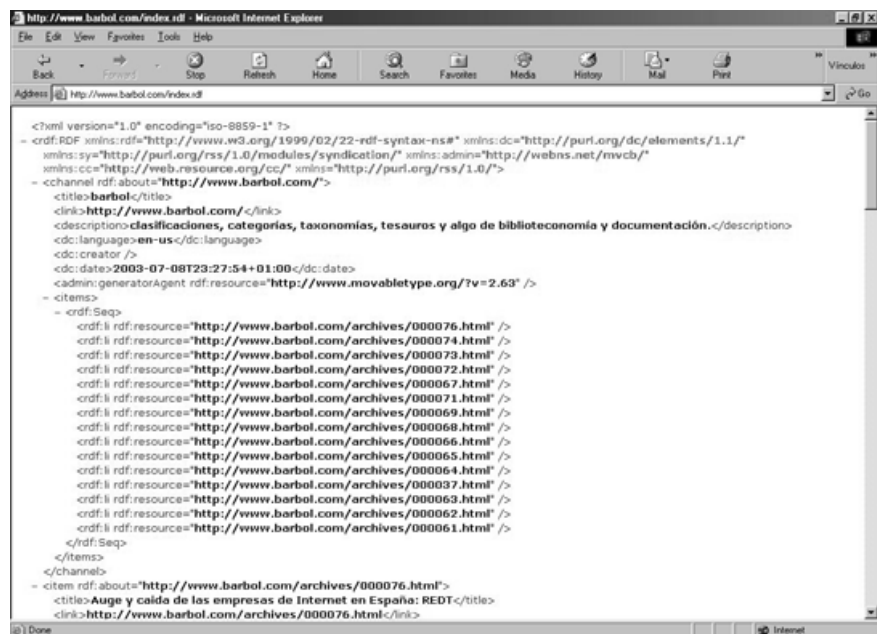


Figura 1

de información se quieren "agregar". Supongamos que hemos localizado una web de la que nos gustaría estar al tanto sin necesidad de apuntarse a un boletín periódico. Un requisito imprescindible es que posea una versión rss del contenido para que se entienda con el agregador *Feedreader*. Para encontrar esta información cogemos como ejemplo el caso de *barbol.com* pues existe un enlace en el que se puede leer: *sindicación (xml)*. Si pinchamos sobre él nos encontraremos con la imagen de la figura 1.

<http://www.bربول.com/index.rdf>

Eso es lo que necesitamos, la url, porque ésa es la fuente (*feed*) con la que vamos a alimentar el agregador. Para agregar dicha versión rss seguimos los siguientes pasos:

—Copiamos la dirección de ese archivo:

<http://www.bربول.com/index.rdf>

—Ahora, en *Feedreader* se abre la opción "new" y allí se pega (figura 2).

Una característica importante de *Feedreader* es que le da igual con qué versión de rss trabaje, así que lo ponemos sin tener en cuenta esta cuestión. Bien, ahora ya po-

demos agregar todas las webs que creamos interesantes y que tengan una versión en rdf/rss.

El problema que se planteará enseguida es que muy pronto tendremos tantas webs agregadas que el volumen de noticias irrelevantes a descartar para encontrar las que realmente interesan será cada vez mayor, incluso aunque agreguemos webs muy especializadas. Veamos dos formas de filtrar más esas fuentes de información.

### ¿Cómo puedo syndicar (y agregar) noticias de Google?

*Voidstar* ha desarrollado una herramienta experimental muy interesante: *Gnews2rss*.

<http://www.voidstar.com/gnews2rss.php>

¿Cómo funciona? introducimos una palabra si existe ya en castellano igual que si se buscara algo en *Google* y se pincha en "create RSS" (figura 3). Es importante definir muy bien la ecuación de búsqueda, por ejemplo: *google OR usability-blogger* (figura 4).

<http://news.google.com/news?hl=en&q=google+OR+usability+blogger&btnG=Search+News>

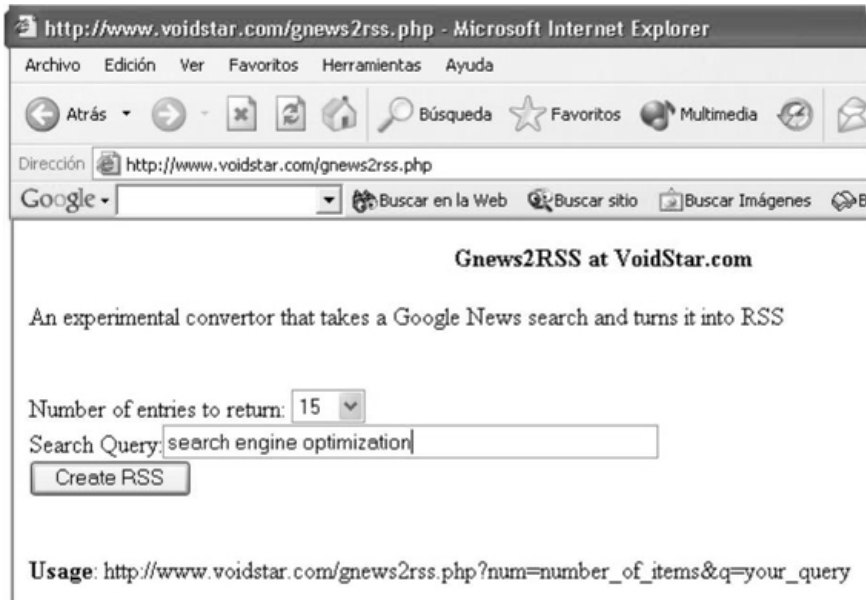


Figura 3

Copiar y pegar la dirección o url resultante en la opción "new" de *Feedreader*. Es posible configurar cada cuánto quiero que el agregador revise si hay novedades, para que haga la búsqueda cada cierto tiempo y de esta forma me dará las noticias del tema que me interesa cada 20 minutos, por ejemplo.

Pero es mejor no utilizar esa frecuencia de consulta porque, si lo hacemos todos, saturaremos el servidor de la web a consultar y porque *Google* prohíbe expresamente que se utilicen consultas automáticas sobre su algoritmo, por lo que podemos abrir y cerrar el programa cuando necesitemos realizar nuestra actualización de noticias.

Como vemos, es una herramienta muy interesante y con muchas posibilidades. Si conocemos *php*, lo mejor sería usar su código fuente, disponible en: <http://www.voidstar.com/gnews2rss.php.txt>

El problema de esta herramienta es que las *News* de *Google* sólo cubren 4.000 servicios de noticias, la mayor parte en inglés. Mientras aumenta el número de fuentes de información, no hay más por ese lado.

Pero lo que nos gustaría es hacer con los blogs o webs de noti-

cias de actualización constante lo mismo que con *Google*: agregar sólo los artículos/posts de los temas que interesen de cada weblog. Todavía no es posible hacer con las búsquedas del *Google* "general" lo mismo que con *Google News*. Todo lo más, **Tara Calishain** ha desarrollado una aplicación que permite transformar mediante *Perl* y una *API* de *Google* (sistema que permite programar una aplicación específica creada a partir del algoritmo de *Google*, de forma que exploremos de diferentes formas las capacidades del motor de búsqueda) las respuestas de *Google* en un

fichero delimitado por comas, el cual se puede exportar a una base de datos de una intranet, por ejemplo.

<http://hacks.oreilly.com/pub/h/164>

La otra forma de conseguir noticias (de blogs) más específicas con respecto a un tema delimitado es más compleja: un pársers rss. Bien: queremos agregar información concreta, pero de blogs o webs que no están en *Google News*. Aquí tenemos que agradecer a **Pedro Palazón** su ayuda, el mayor conocedor de este campo en el ámbito hispano que explica sus descubrimientos en internet. **Pedro** comenta que lo que se podría hacer es agregar sólo aquellos temas de interés mediante un pársers rss.

<http://www.kusor.net/>

### ¿Agregar una web entera o una temática de varias webs?

La cuestión es que quisiéramos no tener que agregar todo lo que aparezca en un blog, sino sólo aquello que hable de lo que a nosotros nos interese. Conforme leamos weblogs, notaremos que muchos (véase los creados con *Movable Type*) tienen categorías para clasificar sus artículos; pues bien, con ellas podemos obtener fuentes de información más específicas. En



Figura 4

este caso no hablamos de agregar una cadena de búsqueda, sino una categoría, una temática concreta.

Por ejemplo, para los que usan como clasificación temática "google" o "usabilidad", debería poder agregar en mi *FeedReader* sólo lo correspondiente a esos temas y así agrupar los *feeds* por temas de webs, no por webs sobre temas, porque muchos blogs hablan de varias cosas diferentes (véase las categorías de mini-d, por ejemplo) y nuestro tiempo de lectura es finito.

<http://www.minid.net/categ.php>

Según **Pedro Palazón**: "en primer lugar crear un páser rss, o emplear uno de los que existen, para buscar la categoría de lo que la gente escribe". Por ejemplo, en el archivo rdf (importante: sólo en formato rdf) de *kusor.net* hay un elemento denominado "*dc:subject*", que contiene la categoría a la que pertenece cada post (artículo). Debemos entonces crear una lista de aquellas sobre las que nos gustaría mantenernos al día, y decirle a ese páser que sólo procesase esas noticias y que descartase las demás.

<http://www.kusor.net/dhtml-weblog/index.rdf>

Hoy por hoy no hay nada fácilmente instalable que permita hacerlo de forma sencilla y agregarlo en *FeedReader* o programa similar sin más. Pero en cualquier caso, un páser muy recomendable es *OnyxRss*, disponible en la web.

<http://readinged.com/onyx/rss/>

### **¿Qué estándar se utilizará en el futuro, rss o rdf?**

En este momento hay muchas discusiones sobre este tema, en concreto alrededor de **Aaron Swartz**, co-creador del estándar rdf, y la guerra entre estándares sigue en pie (*Dublin core* es muy complejo, y más todavía rdf).

<http://google.blogspot.com/>

Todavía no se sabe si va a ser rdf quien gane la partida. Si gana rss, es un formato que sólo syndica titulares y poco más, pero eso incide en su popularidad (lo más sencillo suele ser lo más usado). Si sindicásemos una versión rss de una web y no la rdf, de cada ítem (post, noticia o artículo) sólo leería el título y una descripción, nada de subjects (temáticas).

Rdf está más pensado para, en un futuro ideal, trabajar junto a otros estándares hacia la web semántica (ontologías). Rss es más sencillo de implementar, pero con menos riqueza informativa, menos metadatos.

### **Posibilidades futuras de la sindicación de contenidos**

El éxito de la sindicación y agregación de contenidos es tal hoy día que quizá en unos años cambien hasta los navegadores tipo *Navigator* o *Internet Explorer* para adaptarse a esta forma de navegar. Desde el punto de vista del gestor de información, que filtra datos de muchísimas fuentes distintas para presentar a sus usuarios información relevante para la toma de decisiones, es y será una solución técnica sencilla, ingeniosa y práctica, en constante evolución y mejora.

Vayamos más allá: con un estándar futuro (rss o rdf), podríamos pensar en canales de información multimedia sindicados y agregados, agregación de bases de datos gratuitas o de pago de las que podríamos conocer automáticamente su actualización para cada tema de nuestro interés o del de nuestros usuarios, incluso ofrecer las novedades bibliográficas de la biblioteca de mi barrio.

Todo un mundo de posibilidades.

**Jorge Serrano Cobos**

<http://trucosdeGoogle.blogspot.com>