IV Jornada "Temas Actuales en Bibliotecología"
Centro Médico de Mar del Plata – Universidad Nacional de Mar del Plata
1 de Noviembre de 2013, Mar del Plata

Potenciando la indización web de los contenidos de los repositorios digitales a través de la creación de *sitemaps*

Cristian Merlino-Santesteban
Centro de Documentación. Facultad de Ciencias Económicas y Sociales. Universidad Nacional de Mar del Plata
csantest@mdp.edu.ar

Los repositorios digitales son una de las estrategias definidas por el movimiento acceso abierto para difundir de manera abierta la producción académica y científica generada por una institución o una comunidad científica en particular. Para poder ganar presencia y visibilidad y, de este modo, poder maximizar el acceso en la World Wide Web a los contenidos dispuestos en abierto, esa creación intelectual debe ser distribuida y diseminada a través de proveedores de servicios. Los agregadores de contenido y los recolectores OAI-PMH (*Open Archives Initiative-Protocol Metadata Harvesting*) son herramientas valiosas para llevar adelante esa tarea, pero sin duda son los motores de búsqueda web quienes cumplen un papel preponderante en ese sentido.

Los buscadores web se caracterizan por ser sistemas de uso intensivo y extensivo, lo cual les confiere un lugar destacado en el plano de las herramientas de recuperación de información en la Web. Dado que el mercado de búsquedas en Internet es muy concentrado (comScore 2011; 2013), cuando hacemos alusión a ellos en realidad nos estamos refiriendo básicamente a tres buscadores: Google, Bing y Yahoo!. Y si somos más específicos, a tan sólo uno de éstos, Google, que concentra la mayor parte de las búsquedas realizadas en Internet (comScore 2011; 2013).

Para indizar los contenidos de las sedes web, los motores de búsqueda cuentan con programas autónomos, denominados robots o *crawlers*, que recorren la estructura conectiva de la Red a fin de encontrar nuevos sitios, y de mantener actualizada la base de datos de los recursos ya indexados. Considerando, por un lado, que el funcionamiento de estos programas es muy demandante en recursos de cómputo y de red, y por otro, el constante y acelerado crecimiento de la Web, se planteó la necesidad de proponer otro mecanismo auxiliar que mejorara e hiciera más eficiente el desempeño de los robots de indexación. Dicho mecanismo fue el protocolo Sitemap, al cual adhirieron los buscadores mencionados en 2006 (Schonfeld y Shivakumar, 2009).

Los archivos *sitemap* son archivos XML (*Extensible Markup Language*) que listan los URL (*Uniform Resource Locator*) de un sitio junto a otros metadatos adicionales: fecha de la última actualización, frecuencia de modificación y prioridad (Sitemaps, 2013). Estos archivos ayudan a informar en detalle a los buscadores qué páginas de un sitio están disponibles para su indización, propiciando así una tarea de rastreo más inteligente.

Estructura de un archivo *sitemap*:

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
   <url>
      <loc>http://www.repositorio.edu.ar/doc1/</loc>
      <lastmod>2013-09-29</lastmod>
      <changefreq>monthly</changefreq>
      <priority>0.7</priority>
   </url>
   <url>
      <loc>http://www.repositorio.edu.ar/doc2/</loc>
      <lastmod>2013-09-29</lastmod>
      <changefreq>monthly</changefreq>
      <priority>0.7</priority>
   </url>
</urlset>
```

Donde cada entrada `<url>` está conformada por:
`loc`, campo obligatorio, representa el URL de un recurso web.
`lastmod`, campo optativo, en formato fecha y hora W3C, representa la última fecha de modificación.
`changefreq`, campo optativo, representa la frecuencia con la que puede cambiar la página. Los valores aceptados incluyen *always*, *hourly*, *daily*, *weekly*, *monthly*, *never*.
`priority`, campo optativo, representa la importancia relativa del URL respecto al resto del sitio web (conjunto de páginas).

Con el objeto de alivianar la labor de los servidores web, los archivos *sitemap* no pueden superar los 10 megabytes y los 50 mil URL. En caso de sedes web muy grandes, un archivo *sitemap* se puede emplear como un índice de distintos archivos *sitemap* (SitemapIndex), que a su vez pueden estar comprimidos en formato gzip (GNU ZIP).

Estructura de un archivo índice de *sitemap*:

```
<?xml version="1.0" encoding="UTF-8"?>
<sitemapindex
xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
   <sitemap>
<loc>http://www.repositorio.edu.ar/sitemap1.xml.gz</loc>
      <lastmod>2013-09-29</lastmod>
   </sitemap>
   <sitemap>
<loc>http://www.repositorio.edu.ar/sitemap2.xml.gz</loc>
      <lastmod>2013-09-29</lastmod>
   </sitemap>
</sitemapindex>
```

Donde cada entrada <sitemap> está constituida por:
`loc`, campo obligatorio, representa el URL de un archivo *sitemap*.
`lastmod`, campo optativo, en formato fecha y hora W3C, representa la última fecha de modificación.

Si como punto de partida presuponemos que todo repositorio digital de acceso abierto tiene por objetivo implícito propiciar la indización de los metadatos y los objetos digitales que alberga, la creación de *sitemaps* que identifiquen de manera precisa y detalla cada uno de los ítems documentales que componen su colección es, en ese sentido, un paso vital. De este modo se podrá ayudar a los robots de los buscadores web a descubrir, recolectar y procesar dichos contenidos de forma más eficiente y, consecuentemente, a potenciar notablemente su acceso.

Pese a que el uso del protocolo Sitemap no garantiza que todos los URL serán rastreados (Sitemaps, 2013), éste permite dar visibilidad a ciertos contenidos que pueden ser invisibles o no accesibles para las políticas tradicionales de funcionamiento de los rastreadores.

En definitiva, la presencia y disponibilidad efectiva de los contenidos de los repositorios digitales en los motores de búsqueda requiere no sólo adherir a protocolos de interoperabilidad sino también a protocolos que faciliten la tarea de los *crawlers*, en este caso el protocolo Sitemap.

Bibliografía

comScore. (2013). *comScore Releases August 2013 U.S. Search Engine Rankings*. http://www.comscore.com/Insights/Press_Releases/2013/9/comScore_Releases_August_2013_U.S._Search_Engine_Rankings

comScore. (2011). *Google Sites Accounts for 9 of 10 Searches Conducted in Latin America*. http://www.comscore.com/Insights/Press_Releases/2011/5/Google_Sites_Accounts_for_9_of_10_Searches_Conducted_in_Latin_America

Schonfeld, U., & Shivakumar, N. (2009). Sitemaps: above and beyond the crawl of duty. Presentado en *Proceedings of the 18th International Conference on World Wide Web* (pp. 991-1000). ACM.

Sitemaps. (2013). *sitemaps.org*. http://www.sitemaps.org