Uso de robots.txt y *sitemaps* en la administración pública española

Por **Bonifacio Martín-Galán, Tony Hernández-Pérez, David Rodríguez-Mateos** y **Daniel Peña-Gil**

Resumen: *Se explica la importancia que tienen los ficheros robots.txt y los sitemaps para los sitios web. Se realiza un estudio sobre más de 4.000 webs de la administración pública española para analizar el uso de ficheros robots.txt y sitemaps como medio de optimización para los crawlers o arañas de los motores de búsqueda.*

Palabras clave: *Robots, Crawlers, Sitemaps, Motores de búsqueda, Recuperación de información, Visibilidad, Sitios web.*

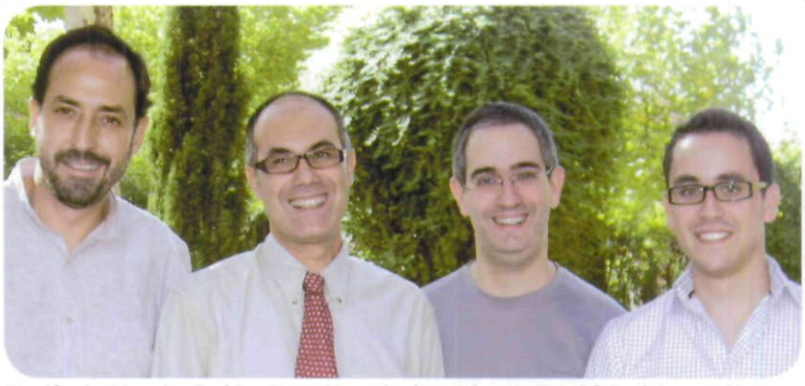
Title: The use of robots.txt and *sitemaps* in the Spanish public administration

Abstract: *Robots.txt and sitemaps files are the main methods to regulate search engine crawler access to its content. This article explain the importance of such files and analyze robots.txt and sitemaps from more than 4,000 web sites belonging to spanish public administration to determine the use of these files as a medium of optimization for crawlers.*

Keywords: *Robots, Crawlers, Sitemaps, Search engines, Information retrieval, Visibility, Web sites.*

Martín-Galán, Bonifacio; **Hernández-Pérez, Tony**; **Rodríguez-Mateos, David**; **Peña-Gil, Daniel**. "Uso de robots.txt y *sitemaps* en la administración pública española". *El profesional de la información*, 2009, noviembre-diciembre, v. 18, n. 6, pp. 625-630.

DOI: 10.3145/epi.2009.nov.05



Bonifacio Martín-Galán, Tony Hernández-Pérez *y* ***David Rodríguez-Mateos*** *son profesores del Departamento de Biblioteconomía y Documentación de la Universidad Carlos III de Madrid y miembros del grupo de investigación Tecnodoc para la aplicación de las tecnologías de la información en bibliotecas, archivos y centros de documentación.* ***Daniel Peña-Gil*** *es becario de investigación de dicho Departamento y colaborador en proyectos de investigación del grupo Tecnodoc.*

Introducción

Los motores de búsqueda en internet construyen sus bases de datos fundamentalmente a partir del trabajo de robots.

Una definición de robot es: "programa que recorre automáticamente la estructura de hipertexto de la Web a través de la recuperación de un documento y, de forma recursiva, va recuperando todos los documentos a los que se hace referencia".

(*tomada de www.robotstxt.org*).

Los robots pueden ser empleados para múltiples propósitos: indexación de páginas de un sitio web, búsqueda de enlaces rotos, validación de códigos html de las páginas, descarga de documentos para la navegación off-line, etc. Es posible por lo tanto diferenciar diversas subcategorías de programas según su función o especialización: *spiders*, *worms*, *web crawlers*, *web ants*, etc. Para referirnos a los que rastrean información en el espacio de la Web lo más preciso es hablar de *crawlers* o *web crawlers*.

Un *crawler* o rastreador es un programa que va descubriendo nuevas páginas y recursos en la Web a partir de un enlace o un conjunto de enlaces (url) a los que se denomina semillas (*seeds*). Generalmente recorre sitios web y extrae cada nuevo enlace que encuentra, permitiendo que otro programa denominado *harvester* (cosechador) descargue las páginas para ser indexadas en la base de datos de un motor de búsqueda. Posteriormente pueden ser buscadas por el usuario a través de un sistema de recuperación de información.

Al territorio o espacio web que un determinado *crawler* recorre en su proceso de rastreo y recolección se le denomina *crawler frontier*. Este espacio puede ser un conjunto finito de páginas dentro de un sitio web, un conjunto de sitios o toda la Web.

En función de las reglas o políticas que siga el rastreador, el contenido de la base de datos sobre la que

Artículo recibido el 01-07-09
Aceptación definitiva: 27-08-09

el usuario puede buscar será una u otra. Por ello, los resultados de una consulta a dos motores de búsqueda son diferentes ya que tienen distintas bases de datos. Y esa es la razón por la que la política de recolección de enlaces y páginas de cada robot es prácticamente un secreto industrial de las empresas de buscadores.

La actividad de los *crawlers* puede ralentizar un sitio web al solicitar un gran número de páginas a un mismo tiempo. A cambio, permite que éste aparezca en uno o varios motores de búsqueda, Por eso, conocer cómo funcionan estos sistemas ayuda a optimizar su actividad en nuestra web, como paso previo al proceso de su posicionamiento en los buscadores.

"Los *crawlers* ralentizan un sitio web al solicitar un gran número de páginas a un mismo tiempo, pero a cambio, permiten que éste aparezca en motores de búsqueda"

Entre los principales problemas con los que se encuentran los *crawlers* (**Schonfeld**, 2009), se incluyen:

– Cobertura: la Web es tan extensa que es prácticamente imposible que se puedan detectar todos los enlaces a páginas o recursos existentes.

– Actualización o refresco: los sitios web cambian rápidamente, con páginas que se añaden, que se modifican o que se borran, por lo que el *crawler* debe establecer una política para que la información que existe en el motor de búsqueda sea lo más actualizada posible.

– Acceso: muchos recursos no son accesibles a los *crawlers* y por tanto no están en disposición de ser buscados. Los problemas de acceso más comunes son:

a. Información dinámica: cada vez más sitios web generan información de forma dinámica, a partir de bases de datos, lo que hace más difícil que el buscador la pueda localizar.

b. Información protegida: directorios o recursos a los que sólo se puede acceder mediante identificación de usuario y palabra clave.

c. Información en formatos no reconocibles para el indexador.

– Duplicidad y versiones: muchos recursos se encuentran a veces duplicados en el mismo o en distintos servidores con distintas urls.

Los rastreadores recorren la Red continuamente en busca de nuevas páginas y, a menudo, su actividad en un sitio web es tan importante en cantidad de visitas como la de los usuarios.

Los administradores de webs deberían tener muy en cuenta la actividad de estos *crawlers* por dos razones:

– Gestión del tráfico web de un sitio. Un administrador puede suministrar información para que el robot evite entrar a recopilar información de directorios y/o ficheros que no se desea que sean recuperables a través de un buscador (porque el sitio o partes de él están en proceso de construcción, por páginas con contenidos anticuados, por información de carácter privado o confidencial, etc.),

– Aumento de confianza en la actualización de la información de un motor de búsqueda respecto a un sitio web. Ya que también puede informar al *crawler* sobre los ficheros que se han suprimido, añadido o modificado y/o sobre la frecuencia en que se suele modificar la información en los directorios de la web.

Esta información tiene un valor muy especial: para el *crawler*, porque le evita tener que descubrir nuevos recursos por sí mismo, dado que el administrador ya le está proporcionando las indicaciones oportunas; y para el administrador, porque reduce sobrecargas de tráfico en su servidor.

Para restringir el acceso a ciertos contenidos, los administradores de webs pueden emplear dos mecanismos: 1) a través de etiquetas meta específicas, para establecer el comportamiento de forma individualizada en cada página; 2) a través del fichero robots.txt, para especificar qué directorios o documentos del sitio no deben ser accedidos por el *crawler*.

Y al contrario, para indicar el contenido que se quiere incluir en un buscador, así como información adicional sobre este proceso, se emplea un protocolo reciente denominado *sitemaps*.

Exclusión de información en la indexación: etiquetas meta y robots

La forma más básica de pasar información a un *crawler* es a través de las etiquetas meta de html, ubicadas en la cabecera de dichos documentos. La recomendación oficial de html establece en su apéndice B una serie de pautas para una etiqueta meta, incluyendo:

– un atributo "name", con un valor que puede ser "robots" (para excluir genéricamente a todos los buscadores), o bien, el nombre de uno concreto.

– uno o varios valores en el atributo "content", a elegir entre "all" (todas las páginas del sitio deben ser indexadas), "index" (indexar la página en la que aparezca), "nofollow" (no seguir los enlaces existentes en la página para el proceso de indexación) y "noindex"

```
<META name="ROBOTS" content="NOINDEX, NOFOLLOW">
<META name="Googlebot" content="noindex">
```

Figura 1. Ejemplo de meta

(no indexar la página en la que aparezca este valor) (figura 1).

http://www.w3.org/TR/html401/appendix/notes.html

Una opción más genérica, al alcance sólo de los administradores de sitios web, es el uso del fichero robots.txt. Funciona mediante un estándar de facto desde 1994, el *Robots Exclusion Protocol* (*REP*), aunque sin aprobar por ninguna organización de normalización.

Consiste en el fichero de texto plano robots.txt, ubicado en el directorio raíz de cada sitio web. Es el primero que busca un *crawler*, y le indica a éste las partes de una web que quedan excluidas del rastreo. El protocolo es muy simple y durante mucho tiempo se le ha relacionado con la ética o buen comportamiento de los robots respecto a un sitio (**Thelwall**, 2005).

No obstante este fichero es público y accesible para todo el mundo: cualquiera puede ver los directorios y/o ficheros que un administrador no quiere que sean analizados por robots. Un mayor nivel de protección se puede conseguir con el fichero oculto *.htaccess*, insertado dentro de la estructura de directorios del sitio web, que aplica directivas similares de acceso para un directorio y sus subdirectorios.

La sintaxis de robots.txt consta de uno o más registros separados por una o más líneas en blanco. Cada registro contiene:

– una o más líneas con la palabra "user-agent" seguida de dos puntos y un valor: un asterisco (si se aplica a todos los *crawlers*) o el nombre del *crawler*.

– una o más líneas con la palabra "disallow" seguida de dos puntos y el nombre de los directorios y ficheros a los que no se quiere que se acceda.

El aumento de los accesos a las webs por parte de los *crawlers* ha hecho necesario indicarles también los contenidos a los que deben acceder y cómo. Los tres grandes motores (*Google*, *Yahoo* y *Microsoft*) comenzaron aumentando las directivas que se podían utilizar en el fichero robots.txt. Entre ellas se incluyen "allow" (documentos y/o directorios que deben ser rastreados), "visit-time" (hora o periodo horario en que los crawler deberían realizar su trabajo) o "crawl-delay" (tiempo en segundos que debe transcurrir entre la solicitud de una página y la siguiente).

http://Googlewebmastercentral.blogspot.com/2008/06/improving-on-robots-exclusion-protocol.html

Inclusión de información en la indexación: *sitemaps*

Posteriormente, en febrero de 2008, estas compañías adoptaron un protocolo común: *Xml Sitemaps Protocol* o *sitemaps*, derivación de un modelo anterior de *Google*, para indicar con más detalles que contenidos debían ser rastreados.

De acuerdo con *Sitemaps.org* "un *sitemap*, en su forma más sencilla, es un archivo xml que enumera las url de un sitio web junto con metadatos adicionales acerca de cada una de ellas: la última actualización, frecuencia de modificación e importancia en relación con las demás url. Los rastreadores web suelen encontrar páginas a partir de vínculos del sitio y a partir de otros sitios".

http://www.sitemaps.org/

El mismo sitio de *Sitemaps.org* advierte que "el uso del protocolo *sitemaps* no garantiza que las páginas web se incluyan en los motores de búsqueda, aunque proporciona sugerencias para mejorar el trabajo de los rastreadores web al rastrear su sitio".

Un fichero *sitemap* puede contener los datos concretos o ser un índice a distintos ficheros *sitemap*.

Las principales reglas de construcción de este archivo xml son:

– Comienza y termina con sendas etiquetas <urlset> y </urlset>. En la primera se especificará el *namespace* de este protocolo.

– Incluye el elemento "url" para dar entrada a cada dirección del sitio web que se desee definir.

– Incluye el elemento "loc" con la url de la dirección.

– Puede incluir elementos opcionales como "lastmod" (última fecha de modificación del fichero), "changefreq" (frecuencia de cambio de los contenidos) y "priority" (importancia en la prioridad de indexación de la página frente a otras del sitio, en un rango de 0.0 al 1.0. La importancia por defecto es 0.5).

– Se llama *sitemap.xml*, y está alojado en el directorio raíz del sitio. No puede tener más de 50.000 direcciones url y no pesa más de 10 MB. Para facilitar la descarga el

Ejemplo 1	Ejemplo 2	Ejemplo 3
User-agent: * Disallow: /cgi-bin/ Disallow: /tmp/	User-agent: * Disallow: /cgi-bin/ User-agent: *Googlebot* Disallow: /videos/	User-agent: *yahooseeker* User-agent: *slurp* Disallow: /links.html

Figura 2. Ejemplos de robots.txt

Ejemplo 5

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
 <url>
 <loc>http://www.prueba.com/</loc>
  <lastmod>2009-01-01</lastmod>
  <changefreq>monthly</changefreq>
  <priority>0.8</priority>
 </url>
</urlset>
```

Figura 3. Ejemplo de sitemap.xml

fichero se puede comprimir en formato gzip (ficheros con extensión .gz).

Una vez creado el fichero *sitemap* se puede incluir la directiva "sitemap" en el fichero robots.txt, con su correspondiente url de ubicación del fichero, y esperar a que el *crawler* vuelva a visitar el sitio; o bien, notificar directamente la ubicación del *sitemap* a los buscadores.

Objetivos y metodología

En este trabajo se ha pretendido: observar el uso de ficheros robots.txt y *sitemaps* por parte de los principales sitios web de la administración española para optimizar el acceso de los *crawlers*; analizar si estas webs favorecen a unos *crawlers* respecto a otros; comprobar si se cumple la tendencia de introducción de *sitemaps* en sitios web (**Kolay**, 2008); y por último, detectar posibles patrones de uso de este tipo de ficheros.

Se utilizó un *crawler* de elaboración propia y específico para que partiendo del sitio web *060.es* localizara los principales sitios de la administración española. En total se obtuvieron datos de 4.108 recursos distintos. Después de su localización un programa se encargaba de descargar los ficheros robots.txt y *sitemaps.xml* del directorio raíz de cada uno. Una vez descargados, los ficheros robots.txt fueron sometidos a un análisis sintáctico a través de la aplicación en línea *Robots.txt syntax checlaker* para descubrir posibles errores. Esta aplicación ya había demostrado su eficacia en estudios anteriores (**Ajay**, 2006), encontrándose disponible libremente para un uso científico.

http://060.es
http://www.sxw.org.uk/computing/robots/check.html

Por otro lado, los ficheros *sitemap.xml* fueron igualmente tratados por un validador xml en línea. La gran cantidad de información suministrada en su respuesta y la alta estructuración de la misma, hizo recomendable su utilización frente a otros servicios similares. Sin embargo, gran parte de los resultados obtenidos en ambos casos resultaban ser ficheros vacíos o documentos html con mensajes de "página no encontrada" o textos similares.

http://schneegans.de/sv/

Se consideraron válidos para el análisis aquellos ficheros que contenían errores sintácticos menos graves. Los dos errores más frecuentes en el caso de robots.txt han sido: no contener "user-agent", y errores de mayúsculas y minúsculas, además de otros errores leves como el uso de espacios en blanco en las rutas que se especifican; el uso de rutas relativas en vez de rutas absolutas; o el uso de comodines o campos (como "allow") no reconocidos en el estándar.

Resultados y discusión

De los 4.108 sitios web analizados, tan solo 715 (17,40%) incluyen un fichero robots.txt válido, y apenas 130 (3,16%) ofrecieron una respuesta válida a la solicitud de *sitemaps*.

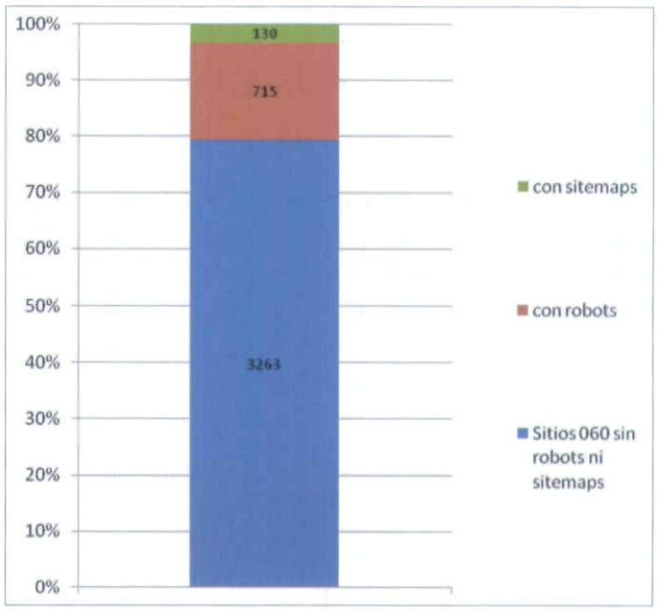


Figura 4. Resumen básico de resultados

En el caso de los *sitemaps*, el número de falsos positivos se debió fundamentalmente a que los servidores devolvían un documento html de "página no encontrada". Según el nombre, de los 130 ficheros con marcado xml válidos el 77,69% (101) aparecen como *sitemap.xml* y el 11,53% (15) como un fichero comprimido con el nombre de *sitemap.xml.gz*. El resto aparece como *sitemap.gz*, *sitemap_index.xml*, *sitemap_index.gz* o *sitemap_index.xml.gz*. Los errores detectados con mayor frecuencia son: declaraciones no válidas de tipo de documentos (34,04%), declaración del elemento "media" cuyo atributo "type" era 'text/xml' en vez de 'application/xml' (23,04%), o falta de declaración de algún espacio de nombres (14,89%).

En los ficheros robots.txt, el campo "disallow" aparece 6.166 veces, lo que da una media de 8,62 directorios o subdirectorios restringidos para un crawler, en cada uno de los sitios. Más escaso aún es el campo

"sitemap" en los ficheros robots.txt analizados, apenas 11 (0,01%). La existencia de este campo facilita a los *crawlers* encontrar el nombre y la ubicación del fichero *sitemap* en el sitio web.

En cuanto al campo "user-agent", el valor más utilizado es el comodín o *, que indica que las directivas se aplican para todos los *crawlers*: aparece en el 98,60% de los ficheros analizados, con menciones a 344 robots distintos. Los dos más mencionados son: *Googlebot*, en 112 ocasiones y *Googlebot-Image*, en 54. Aparecen también 50 referencias a *psbot* (*crawler* del buscador *iSearch.com*, especializado en la búsqueda de personas), mientras otros más clásicos, como *MSNbot* (de *Microsoft*) y *Slurp* (de *Yahoo*), aparecen sólo unas 7 veces cada uno. También aparecen mencionados en más de cinco ocasiones extractores de sitios web como *ExtractorPro*, *WebZIP*, *WebCopier*, etc.

El campo "disallow" muestra una gran variedad de valores. En las tablas siguientes se muestran los nombres de los directorios cuya frecuencia de exclusión supera la cincuentena.

En cuanto a los 130 *sitemaps* válidos encontrados (dejando aparte 16 sitios con *sitemaps* como ficheros comprimidos, que no pudieron ser validados), más del 90% son ficheros directos en los que se referencia el contenido que tiene que ser indexado por los motores de búsqueda, siendo escaso el uso de archivos de índice de *sitemaps* (14,61%). Los espacios de nombre indican la versión de *sitemap* empleada:

Directorio	Frecuencia
/	421
(Vacío)	239
/cgi-bin/	220
/includes/	215
/directorio.htm/	191
/templates/	157
/cache/	157
/images/	152
/modules/	145
/language/	125
/components/	113
/media/	110
/admin/	108
/administrator/	107
/installation/	106
/js/	70
/editor/	65
/mambots/	64
/help/	64
/tmp/	61
/css/	60
/plugins/	53

Tabla 1. Valores del campo "disallow"

Xmlns	Frecuencia
http://www.sitemaps.org/schemas/sitemap/0.9	52
http://www.google.com/schemas/sitemap/0.84	39
http://www.w3.org/2001/XMLSchema-instance	3 (error)
http://www.google.com/schemas/sitemap/0.9	1 (error)

Como se puede comprobar, es mayor la referencia al nuevo protocolo de *Sitemaps.org* frente al antiguo (aunque originario) de *Google*. En el caso de los índices de *sitemaps* empleados, por el contrario, la mayoría corresponde a este protocolo antiguo.

Xmlns	Frecuencia
http://www.google.com/schemas/sitemap/0.84	17
http://www.sitemaps.org/schemas/sitemap/0.9	1
http://www.google.com/schemas/sitemap/0.9	1 (error)

"El uso de ficheros robots.txt en la administración española está lejos de la media de organismos gubernamentales de otros países, como Estados Unidos"

Conclusiones

El uso de ficheros robots.txt (17,40%) en los sitios web de la administración española está bastante lejos de la media del uso de estos ficheros en organismos gubernamentales de otros países, como Estados Unidos, donde ronda el 44% (**Sun**, 2007). Ello demuestra una escasa preocupación por parte de los administradores de esos sitios web españoles respecto a la actividad de los *crawlers*.

De nuestro análisis se concluye, como en **Sun** (2007b), que en los ficheros robots.txt existe cierto sesgo hacia determinados motores, como *Google*. El 98,60% de los sitios declara directivas para excluir de la indexación ciertos directorios a todos los motores (*) pero, aplicando directivas concretas a algunos *crawlers*; en particular, *Googlebot* (112 veces) y *Googlebot-Ima-*

ge (54 veces) son los más citados. Esto significa que los administradores de los sitios piensan en favorecer la actividad de estos dos rastreadores, porque han detectado que son los que más veces visitan su sitio y quieren evitar una sobrecarga de actividad, o porque quieren asegurarse de que ciertos contenidos no aparezcan en el índice del motor de búsqueda más popular (ficheros de instalación y administración del sistema, espacios privados de la intranet, directorio de almacenamiento de aplicaciones cgi-bin, plantillas y ficheros temporales, etc.).

El elevado número de referencias a *psbot* (50 veces) puede estar provocado por la alta tasa de sobrecarga que provoca en algunos sitios web. Están más acordes con otros estudios las referencias a *MSNbot* (*Microsoft*) y *Slurp* (*Yahoo*) y a la aparición de numerosos robots extractores, ya que los administradores excluyen este tipo de robots cuando descubren que han provocado una sobrecarga al intentar bajarse todo un sitio para navegarlo de forma off-line.

"Es sorprendentemente alto el número de webs que quieren que el *crawler* no indexe nada y, por tanto, que sus contenidos no sean accesibles por motores de búsqueda"

Resulta sorprendente también el alto número de sitios (421, un 58,88%) que utilizan como criterio de exclusión ("disallow") el directorio raíz de un sitio (/), es decir, que el *crawler* no indexe nada y, por tanto, que sus contenidos no sean accesibles a través del motor de búsqueda. Asimismo, parece un tanto extraño que se aplique también a cualquier agente de usuario (*) puesto que significa que se desea que ni *Google*, ni *Yahoo* ni *MSN* indexen absolutamente nada del sitio. Otra combinación bastante frecuente es el uso de cualquier agente de usuario (*) con el campo de exclusión (disallow) vacío (239 casos, un 33,42%), lo que significa que la directiva no tiene ningún efecto sobre los *crawlers*.

El uso de *sitemaps* es indicador del grado de actualización de un sitio y también de cierto grado de preocupación de los administradores por la actualización de sus conocimientos. También aquí las webs de la administración española que contienen *sitemap*, un 3,16%, se encuentran por debajo de la media internacional, un 6,3% según datos de **Wilde** (2009). En el 77,69% de los casos son ficheros *sitemap* en formato xml; el resto son también ficheros *sitemap* en formatos comprimidos, fundamentalmente gz. Resulta significativo en cuanto al uso del protocolo *sitemap* que tan sólo encontremos dos únicos casos en toda la muestra estudiada donde se está haciendo un uso realmente correcto del archivo de índice de *sitemaps*: el portal *AyuntaWeb* (ayuntamientos de España e Hispanoamérica) y el del *Ayuntamiento de San Sebastián*.

http://www.ayuntaweb.info/sitemap.xml
http://www.donostia.org/sitemap.xml

Bibliografía

Ajay, S.; **Ekanayake, J.** "Analysis of the usage statistics of robots exclusion standard". En: *Iadis Intl Conf WWW/Internet*, 2006. *http://grids.ucs.indiana.edu/ptliupages/publications/IADISConferenceRobtoExclusion.pdf*

Kolay, S.; **D'Alberto, P.**; **Dasdan, A.**; **Bhattacharjee, A.** "A larger scale study of robots.txt". En: *Intl Conf on World Wide Web*, 2008, pp. 21-25.

Schonfeld, U.; **Shivakumar, N.** "Sitemaps: above and beyond the crawl of duty". En: *Intl Conf on World Wide Web*, 2009, pp. 991-1000.

Sun, Y.; **Councill, I. G.**; **Giles, C. L.** "A large scale study of robots.txt". En: *Intl Conf on World Wide Web*, 2007a, pp. 1123-1124.

Sun, Y.; **Zhuang, Z.**; **Councill, I. G.**; **Giles, C. L.** "Determining bias to search engines from robots.txt". En: *Proc of Intl Conf on Web*, 2007b, pp. 149–155.

Thelwall, M.; **Stuart, D.** "Web crawling ethics revisited: cost, privacy, and denial of service". *Journal of the American Society for Information Science and Technology*, 2005, v. 57, n. 13, pp. 1771–1779.

Wilde, E.; **Roy, A.** Web site metadata: UCB ISchool report 2009-028. *http://dret.net/netdret/publications#wil09b*

***Bonifacio Martín-Galán, Tony Hernández-Pérez, David Rodríguez-Mateos y Daniel Peña-Gil**, Universidad Carlos III, Departamento de Biblioteconomía y Documentación.*
bmartin@bib.uc3m.es
tony@bib.uc3m.es
pirio@bib.uc3m.es
dpgil@db.uc3m.es