X CONGRESO ISKO CAPÍTULO ESPAÑOL (Ferrol, 30 junio - 1 julio 2011)

ANÁLISIS DE LA WEB Y USABILIDAD: PRUEBA DE FUNCIONAMIENTO DE *MBOT WEBCRAWLER*

Manuel Blázquez Ochando.
Departamento de Biblioteconomía y Documentación, Facultad de Ciencias de la Documentación, Universidad Complutense de Madrid.
manuel.blazquez@pdi.ucm.es

Esmeralda Serrano Mascaraque.
Departamento de Ciencias Sanitarias y Medico-sociales, Facultad de Documentación de la Universidad de Alcalá.
esmeralda.serrano@uah.es

RESUMEN:
La presente investigación tiene como objetivo el desarrollo de un programa de análisis de la web, denominado *mbot,* que facilite la elaboración de estudios cibermétricos y en especial la obtención de datos sobre la tipología de los contenidos de un determinado área de conocimiento en la web, distribución o estratificación según los niveles de análisis empleados, número total de sitios, dominios y páginas analizadas, tamaño o volumen de los contenidos. Para demostrar su funcionamiento y comprobar sus capacidades se han llevado a cabo diversas pruebas entre las que destaca el análisis exhaustivo de dos centros de investigación internacionales, NASA y ESA. En ambos casos se extrae y cuantifica toda la información disponible, a fin de elaborar un estudio comparativo sobre su topografía, clasificación, tipos documentales, recursos, vínculos y contenidos. Por otro lado, se analizará cómo el factor de usabilidad, accesibilidad y arquitectura de la web afectan en el reconocimiento de patrones, en el código fuente de las páginas analizadas.

PALABRAS CLAVE:
Recuperación de información, webcrawler, mbot, cibermetría, indexación, parser, DOM, accesibilidad web, usabilidad web

TITLE: Web analysis and usability: running trial of *mbot webcrawler*

ABSTRACT:
This research has as primary goal the development of a new web analysis application, called *mbot*, that may help in the cybermetric studies and specially in the obtaining of data about the documental typology contained in a given knowledge area in the Web, its distribution or stratification according to the different analysis levels used, total amount of sites, domains and pages analyzed and size or volume of contents. To prove its operational capabilities we have run several trials, among which we want to underline the exhaustive analysis of two international research centres, ESA and NASA. In both cases we have extracted and quantified all available information, in order to perform a comparative study of their topography, classification, documentary types, resources, links and contents. On the other side the factor usability is taken into consideration, including accessibility and architecture of the web, as long as they have any incidence in the recognition of patterns in the source code of the researched pages.

KEYWORDS:
Information retrieval, webcrawler, mbot, cybermetrics, indexing, parser, DOM, web accesibility, web usability

Introducción

El análisis de la web es posible gracias al empleo de programas webcrawler capaces de analizar una lista de enlaces, que de partida son utilizados para extraer la información de cada dominio, sitio y página web (BAEZA YATES, R. y GRAELLS, E., 2008, pp.29-44). Este proceso automatizado permite obtener información relativa a la interrelación de hipervínculos en la red, la referenciación de los contenidos y, en definitiva, la tela de araña que articula la información que cualquier usuario utiliza. Si bien los beneficios derivados del uso de este tipo de programas son ampliamente reconocidos, al ayudar en la recuperación, indexación y control de los contenidos publicados, también es cierto que resultan complicados de gestionar y de aplicar extensivamente. Todo ello depende de la disposición o no de los soportes tecnológicos y del conocimiento técnico adecuados.

El objetivo que se plantea es acercar la tecnología webcrawler al documentalista, de tal manera que pueda aplicarlo extensivamente en su actividad profesional e investigadora. En este sentido se viene desarrollando el programa *mbot*, cuyos parámetros de funcionamiento resultan similares a los que cualquier otro webcrawler utilizaría; evitando las dificultades que entraña la programación de tales variables. Un ejemplo de ello es el programa *Nutch* sobradamente reconocido por su capacidad, pero cuyos requisitos de instalación y configuración (BLÁZQUEZ OCHANDO, M., 2009) implican el empleo de un sistema operativo Linux, la instalación de Java SDK, la configuración manual de *classpath* de java, la edición manual de los archivos de filtrado y ejecución del webcrawler, y el empleo de comandos para la obtención de resultados. Concretamente, el proceso de configuración del filtrado y las reglas de ejecución, véase *(figura1)*, constituyen el apartado más complicado para el usuario, siendo relativamente fácil dar al traste con el funcionamiento del sistema. Esta situación se genera al comprobar que los actuales manuales divulgados acerca de *Nutch* (Nutch Wiki, 2011), establecen una guía de instalación básica y no disponen de información suficiente para solucionar los habituales problemas del filtrado por extensiones, la configuración del análisis multinivel de la web de un determinado dominio, o la selección del tipo de contenidos que debe analizar y extraer, dificultando la labor del no especialista.

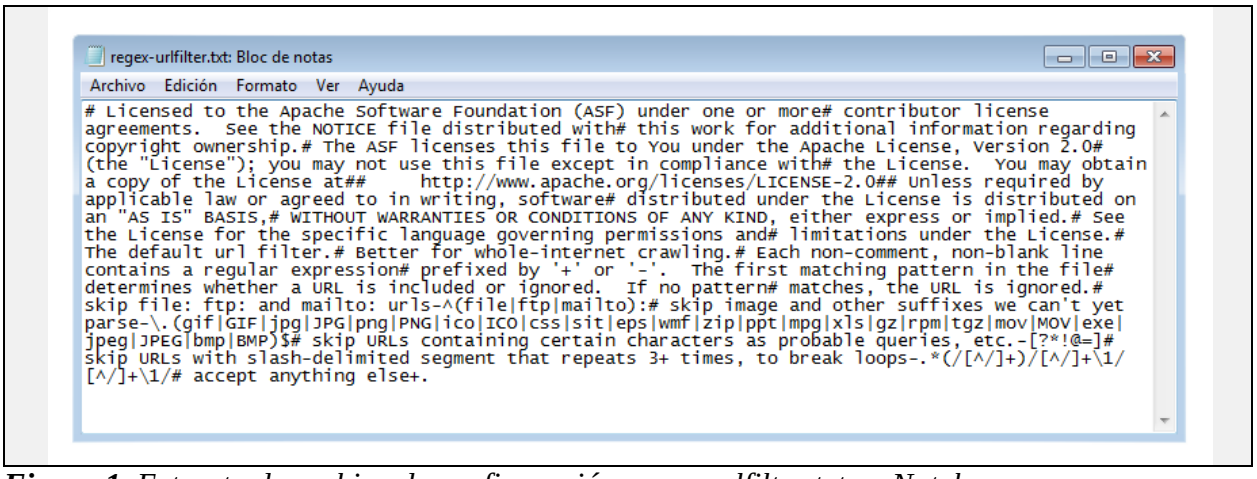

Figura 1. *Extracto de archivo de configuración regex-urlfilter.txt en Nutch*

Metodología y funcionamiento

Las complicaciones anteriormente mencionadas son solucionadas en parte por el webcrawler *mbot* desde el momento en que posibilita la configuración de dichas opciones con un interfaz gráfico sin necesidad de editar manualmente ningún archivo, véase *(figura2)*. Aspectos tales como el nivel de profundidad del análisis, la extracción de metadatos, meta-etiquetas, canales de sindicación, imágenes, documentos, archivos multimedia, código fuente, texto completo, el filtrado por sitio web por extensiones, así como los parámetros de ejecución de la herramienta para la transferencia de datos vía http, pueden ser modificados a voluntad, sin temor de fallos en la ejecución del programa.

Figura 2. *Pantalla de configuración de mbot*

En cuanto al método de funcionamiento de mbot, resulta muy similar al de cualquier otro webcrawler (CASTILLO, C., 2004, pp.4-8), siendo la principal diferencia la plataforma y tecnología utilizada en su diseño. Es decir, su instalación en soportes Apache, PHP y MySQL y el empleo de funciones cURL, DOM y XPath. Estas características permiten su instalación en la mayoría de servidores web.

El punto de partida, en la ejecución de las rutinas del webcrawler, comienza una vez se ha rellenado la muestra o semilla de enlaces preliminar. Dichos vínculos constituyen el objeto de análisis del programa, para los que desencadena su depuración, consistente en la eliminación de espacios, sustitución de caracteres especiales y finalmente la comprobación de la existencia o funcionamiento de la página o sitio web enlazado. Este proceso se lleva a cabo mediante conexión por socket de datos utilizando el puerto 80; de tal manera que si el recurso estuviera fuera de línea, éste sería retirado y reemplazado por el siguiente en la cola de análisis. Una vez hechas estas comprobaciones, se envía una petición de retorno de contenidos mediante el empleo de cURL, obteniendo el código fuente completo de la página web enlazada. A continuación, ésta es convertida en un objeto DOM, que permite su manejo y consulta mediante sentencias XPath que marcan la ruta de selección de los elementos que se desean analizar y extraer. La información es almacenada inicialmente por medio de *arrays* que posteriormente es volcada a la tabla de la base de datos en

MySQL indexando sus contenidos textuales y permitiendo a la postre su tratamiento y recuperación, véase *(figura3)*.

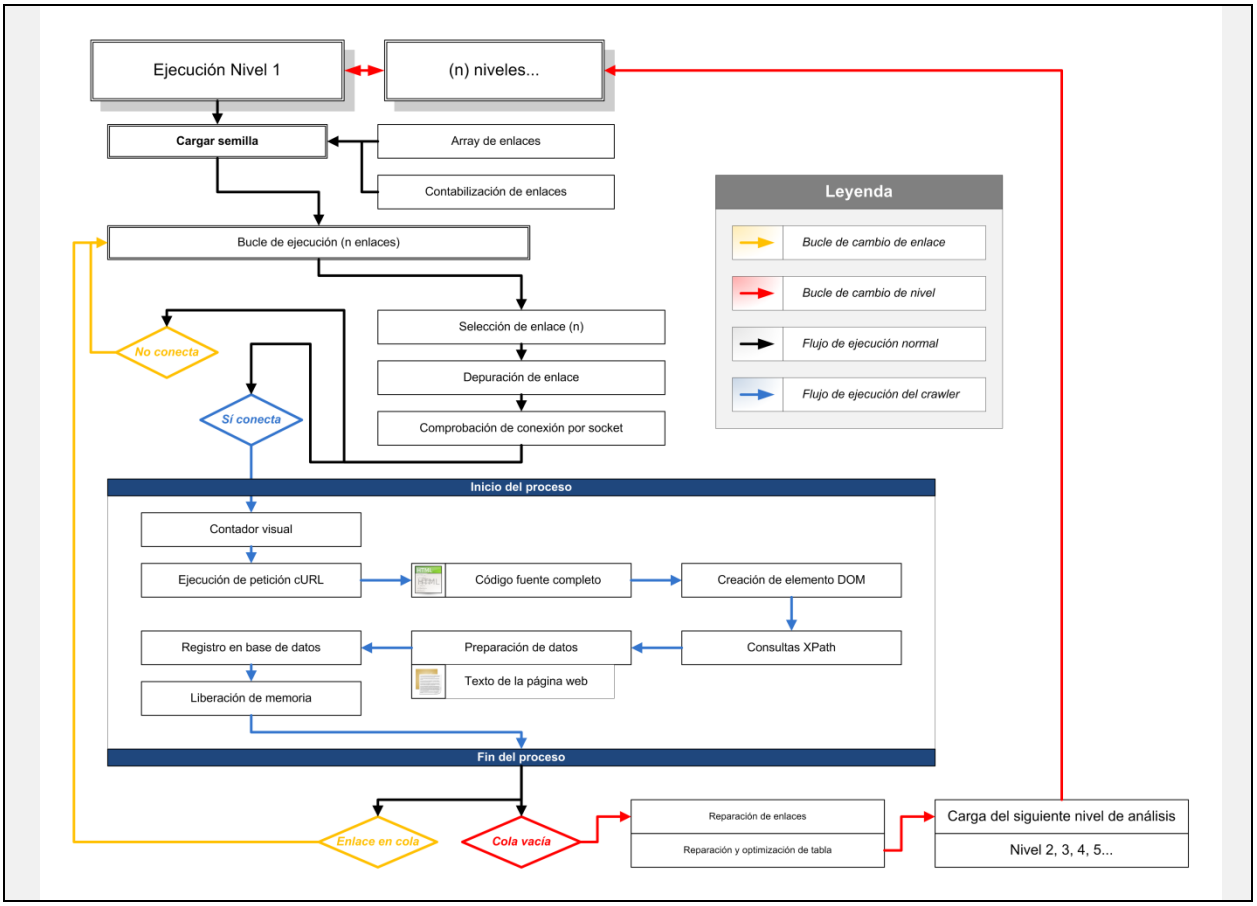

Figura 3. *Esquema de funcionamiento de mbot.*

Para ensayar el concepto del programa *mbot*, se ha desarrollado una prueba en red (BLÁZQUEZ OCHANDO, M., 2011) en la que es posible analizar cualquier sitio o página web que el usuario determine, obteniendo en el momento todos los contenidos, enlaces y código fuente perfectamente identificados. Este proceso no se desarrolla más allá de un único nivel de análisis y ello conlleva que no se inspeccionen terceros enlaces o páginas. El resultado obtenido en tal caso demuestra la validez del esquema de funcionamiento presentado, pero no permite conocer cómo se desenvuelve el programa cuando se enfrenta a un análisis multinivel, característica clave para cualquier webcrawler.

En este sentido se aborda una segunda prueba de funcionamiento, que es la que se presenta en este artículo. La finalidad de la investigación consiste en obtener resultados cuantitativos en relación a dos centros de investigación internacionales, en este caso la NASA (National Aeronautics and Space Administration) y la ESA (European Space Agency). La elección de tales instituciones se debe a la concordancia de la temática investigadora, al gran volumen de datos o contenidos que pueden albergar, a la variedad documental y a la posibilidad de establecer comparaciones cibermétricas entre dos centros con finalidades y objetivos similares. Además, se pone a prueba el análisis multinivel del webcrawler, configurando un proceso de tres niveles de profundidad, entendiendo que cada nivel representa el análisis de los enlaces extraídos en el nivel anterior, de forma concatenada. De los sitios web analizados se obtendrán cifras generales de sus contenidos, tiempo de ejecución del programa, desglose de contenidos por cada nivel de análisis y finalmente problemas de accesibilidad, arquitectura y usabilidad de las páginas analizadas por el sistema.

Análisis multinivel de la NASA y ESA

Los resultados obtenidos demuestran un mayor tamaño del sitio web de la NASA con 32.865 enlaces, frente a la ESA con 28.232, véase *(tabla1)*. Es destacable que las páginas web de la ESA utilizan en mayor medida los metadatos y meta-etiquetas con un ratio de 3,04 etiquetas por página analizada, frente al 0,24 de la NASA. En relación a la cantidad de documentos, la NASA consta de 882 frente a 304 de la ESA, lo que supone un 65,5% más de contenidos. Los canales de sindicación tienen una relevancia destacada al superar los 300 en el caso de la NASA y los 100 en el caso de la ESA; ello significa la importancia que tales instituciones conceden a la redifusión de sus contenidos, noticias y resultados para su divulgación.

	NASA	ESA
Enlaces analizados	2.680	2.521
Metadatos	49	3.296
Meta-etiquetas	613	4.365
Enlaces extraídos	32.865	28.232
Canales de sindicación	341	161
Documentos	882	304
Imágenes	9.330	6.248
Multimedia	73	202
Tamaño de datos	130MB	91.7MB

Tabla 1. *Cifras totales del estudio*

En cuanto a la ejecución del webcrawler, pueden comprobarse unos tiempos proporcionales al número de páginas analizadas, aunque notablemente superiores en el caso de la NASA, véase *(tabla2)*. El tiempo medio de análisis en el caso de la ESA es de 1,6 páginas por segundo y en la NASA de 1,08. Esta diferencia se debe a múltiples factores como el mayor número de documentos, enlaces e imágenes extraídas. En tales casos, el programa ha tenido que recopilar los enlaces, comprobar, mediante patrones en expresiones regulares, la extensión de los documentos enlazados, así como su clasificación y tipología.

	Indicador	NASA	ESA
Nivel1	Inicio	2011-03-27T18:28:09+02:00	2011-03-27T15:37:27+02:00
	Fin	2011-03-27T18:28:09+02:00	2011-03-27T15:37:27+02:00
	T. Parcial	*0.833479881287 segundos*	*0.425271987915 segundos*
Nivel2	Inicio	2011-03-27T18:33:22+02:00	2011-03-27T15:45:54+02:00
	Fin	2011-03-27T18:35:05+02:00	2011-03-27T15:46:39+02:00
	T. Parcial	*1 minuto 43 segundos*	*45 segundos*
Nivel3	Inicio	2011-03-27T19:36:34+02:00	2011-03-27T17:11:32+02:00
	Fin	2011-03-27T20:15:49+02:00	2011-03-27T17:37:02+02:00
	T. Parcial	*39 minutos 15 segundos*	*25 minutos 30 segundos*
Tiempo total		**40 minutos 59 segundos**	**26 minutos 16 segundos**

Tabla 2. *Tiempos de ejecución*

A este proceso se le une otra dificultad. A la hora de limpiar el código fuente de las páginas para la indexación de sus contenidos, una arquitectura web más compleja y menos accesible afecta inexorablemente al tiempo de ejecución del programa. Este último aspecto es clave a la hora de conseguir unos tiempos de funcionamiento más reducidos, como se demostrará más adelante.

En cuanto al análisis multinivel, disponible en la *(tabla3),* se puede comprobar una evolución de los contenidos en ambas instituciones. En el primer nivel se analiza la página de portada que, en ambos casos, proporciona un número de imágenes y enlaces parejos, ligeramente superior en el caso de la NASA. En cuanto a meta-descripción, la ESA ya incorpora metadatos y meta-etiquetas, así como canales de sindicación accesibles desde el navegador web. Esta tendencia se ve refrendada en el segundo nivel de análisis, en el que se comprueba un gran crecimiento en el número de documentos de la NASA con 141 frente a 20. En cuanto al número de imágenes también se detecta un fuerte incremento en la NASA asciende a 1117, que supone más del doble de las que se obtuvieron en la Agencia Europea. Sin embargo, su meta-descripción con un total de 358 etiquetas para 71 recursos analizados es muy superior que la NASA, en la que apenas se emplean 8 etiquetas para 108 páginas analizadas.

	NASA	ESA
Nivel 1	**1**	**1**
Metadatos	0	3
Meta-etiquetas	0	4
Enlaces extraídos	115	82
Canales de sindicación	0	2
Documentos	2	0
Imágenes	36	34
Multimedia	0	0
Nivel 2	**108**	**71**
Metadatos	0	160
Meta-etiquetas	8	198
Enlaces extraídos	2919	2817
Canales de sindicación	9	31
Documentos	141	20
Imágenes	1117	554
Multimedia	4	4
Nivel 3	**2571**	**2449**
Metadatos	49	3133
Meta-etiquetas	605	4163
Enlaces extraídos	29831	25333
Canales de sindicación	332	128
Documentos	739	284
Imágenes	8177	5660
Multimedia	69	198

Tabla 3. *Análisis multinivel de contenidos*

Esta tendencia se repite en el tercer nivel de análisis. Se constatan las cifras generales en cuanto a número de metadatos y meta-etiquetas, muy superior en la ESA con respecto a la NASA. En cuanto a la cantidad de documentos, imágenes y canales de sindicación la NASA aporta cifras superiores a la Agencia Europea del Espacio, véase *(figura4)*. Finalmente, es destacable el crecimiento en el número de archivos multimedia que incluyen audio y video, con 198 recursos en el caso de la ESA, más del doble que la NASA con 69.

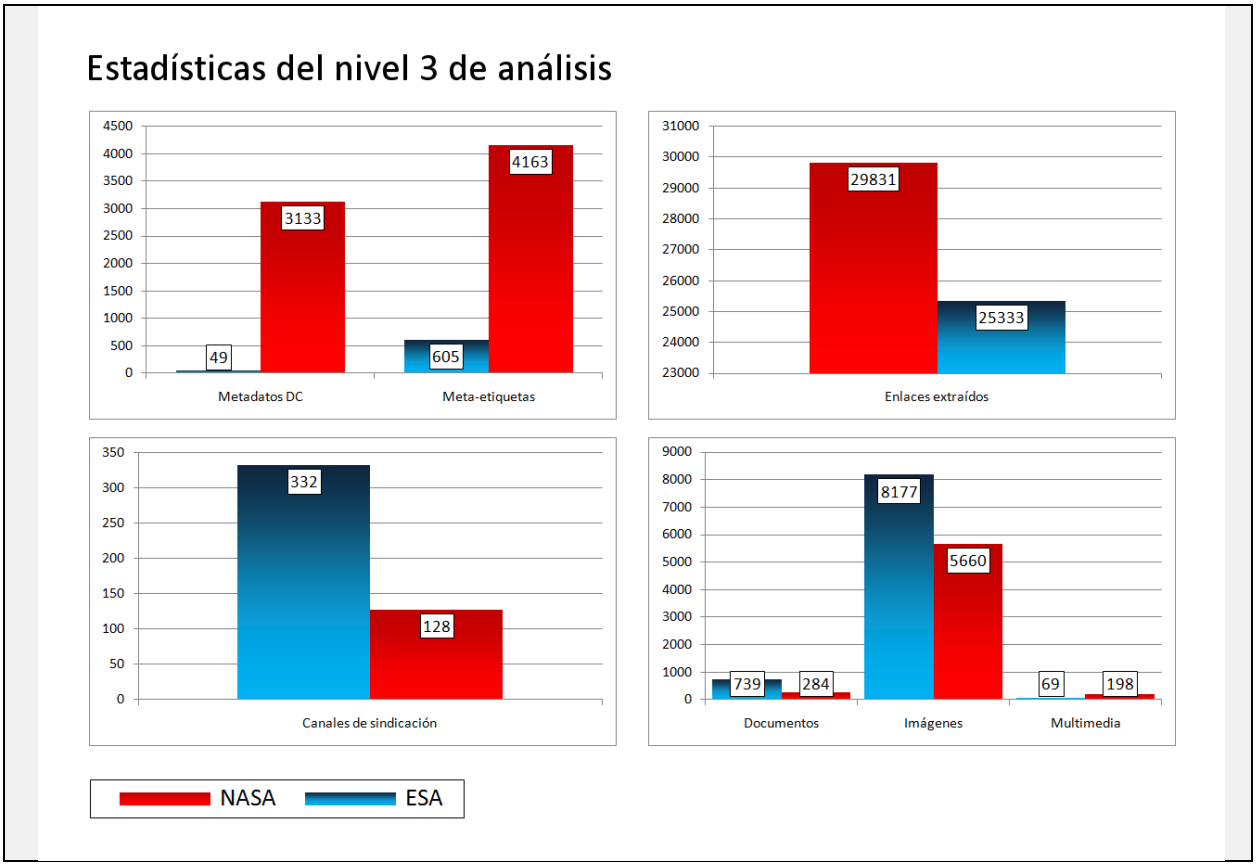

Figura 4. *Estadísticas del nivel3 de análisis. Véase diagrama completo disponible en:* http://www.mblazquez.es/documents/articulo-pruebas2-mbot-estadisticas.png

En cuanto a la tipología de los contenidos, véase *(tabla4)*, destacan los documentos PDF tanto en la NASA como en la ESA con 862 y 295 archivos respectivamente. Los tipos multimedia más frecuentes son los vídeos en formato WMV y MP4, que en la ESA suman hasta 167 recursos. Las imágenes con extensiones JPG, GIF y PNG son predominantes a lo largo de todo el análisis. Constituyen un activo importante en ambas instituciones científicas, que en suma aúnan 13.837 imágenes. También se demuestra un uso reducido o nulo de los tipos de archivo de presentación PPT, hojas de cálculo XLS, archivos de audio WAV y MP3, archivos de video FLV, AVI, MPG y SWF, así como de imágenes en formato BMP.

	NASA			ESA		
Niveles	**N1**	**N2**	**N3**	**N1**	**N2**	**N3**
Documentos	**2**	**141**	**739**	**0**	**20**	**284**
.doc	0	1	13	0	1	7
.pdf	2	140	720	0	19	276
.ppt	0	0	3	0	0	0

.xls	0	0	3	0	0	1
Audios	**0**	**0**	**15**	**0**	**4**	**15**
.wav	0	0	0	0	0	4
.mp3	0	0	15	0	4	11
Videos	**0**	**4**	**54**	**0**	**0**	**183**
.wmv	0	1	14	0	0	140
.mp4	0	3	35	0	0	27
.flv	0	0	0	0	0	7
.avi	0	0	3	0	0	0
.mpg / .mpeg	0	0	1	0	0	4
.swf	0	0	1	0	0	5
Imágenes	**36**	**1117**	**8177**	**34**	**554**	**5660**
.jpg	26	854	5437	30	457	4152
.png	0	31	721	0	0	404
.gif	10	218	1606	4	97	940
.bmp	0	0	7	0	0	9

Tabla 4. *Tipología de los contenidos, documentos y archivos multimedia.*

Problemas de accesibilidad, arquitectura y usabilidad

Considerado como parte de la prueba, los problemas de accesibilidad, arquitectura y usabilidad de los sitios web de la NASA y ESA, se ha equipado al programa *mbot* con un módulo especializado en la visualización de todos los textos indexados durante el proceso de recuperación. De esta forma es posible analizar la calidad del proceso de análisis y extracción de textos del código fuente de cada página web y compararlo con su código fuente original. Los resultados obtenidos muestran diversos fallos que en algunos casos son achacables al webcrawler y en otros a una deficiente programación de las páginas web.

Según se muestra en la *(tabla5)*, se han encontrado etiquetas que dificultan el proceso de extracción e indexación de textos. En el caso de la NASA, se detectan 22 etiquetas de las que al menos 11 de ellas no son HTML normalizado, otras no están correctamente construidas como es el caso de *</+script>* y 10 son etiquetas HTML válidas pero incorrectamente cerradas o utilizadas al constituir estructuras con más de 6 anidamientos entre capas de tipo * *. En los resultados obtenidos con la ESA, se presentan 10 etiquetas HTML con problemas de anidamiento y cierres incorrectos, lo que supone una importante diferencia a la hora de ser procesadas si se compara con los errores obtenidos en la arquitectura web de la NASA. Estos datos suponen que las páginas web de la NASA tienen el doble de fallos que las de la ESA, lo que concuerda con el tiempo de ejecución de casi el doble, 40 minutos frente a 26, tal como se especificó anteriormente en la *(tabla2)*.

	NASA		ESA
Etiquetas que dificultan la indexación	– – <di> – <table> – <tbody> – <noscript> – <font> – <base> – </+script> – <wbr> – <area> – <layer>	– <map> – <na> – <small> – <time> – <noindex> – <nobr> – <o:p> – <fb:like> – <asx> – <fb:fan> – <spacer>	– – <di> – <table> – <basefont> – <frameset> – <tbody> – <fieldset> – <legend> – <base> – <sup>
Caracteres especiales y construcciones defectuosas	– → – › – <! – ›> – <!>		No detectados
Arquitectura web	– Javascript embebido dentro de etiquetas table en HTML		– Capas dentro de tablas anidadas en terceras capas. – Varios niveles de anidamiento con cierres de etiquetas incorrectos. – Javascript embebido dentro de etiquetas table y div HTML
Estilo de programación	– No se delimita correctamente los atributos de las etiquetas HTML, dificultando la declaración de valores. – Eventos de javascript en enlaces por ejemplo: onmouseover, onclick, etc. – Menús y rutas de acceso poco accesibles, por ejemplo: *nasa directorates and offices › aeronautics research → › exploration systems › science → › space operations → › chief financial officer*		– Uso de etiquetas en mayúsculas – Etiquetas sin cerrar – Estilos CSS sin delimitar con etiquetas <style> – Eventos de javascript en enlaces por ejemplo: onmouseover, onclick, etc.

Tabla 5. *Problemas de arquitectura y accesibilidad web detectados con mbot*

Por otro lado, la NASA emplea caracteres especiales de difícil traducción o conversión para el webcrawler. Tal es el caso de las flechas utilizadas en los menús de tipo breadcrumb, en vez de ser imágenes o código ASCII equivalente *→*. En conclusión, la arquitectura web es en ambos casos mejorable. En las páginas de la NASA es frecuente encontrar códigos javascript embebidos dentro de etiquetas <*table*> en HTML. Este hecho es similar al dispuesto en las páginas de la ESA

que además presentan problemas en el anidamiento de tablas y capas <*div*> en varios niveles. Estos aspectos ofuscan al webcrawler en la detección de etiquetas, haciendo que la posibilidad de cometer errores en la indexación aumente. En cuanto al estilo de programación, se detecta el empleo de eventos javascript dentro de los enlaces, la incorrecta delimitación de los atributos dentro de etiquetas HTML, etiquetas sin cerrar, estilos CSS sin delimitar con etiquetas <*style*> y el empleo de mayúsculas en la edición del código fuente.

Para confirmar y ampliar la información obtenida durante el análisis, se someten los sitios web de ambas instituciones a un análisis heurístico de usabilidad utilizando una escala likert del 1 al 5 para valorar aspectos tales como la legibilidad, la presencia de secciones de la página, el árbol de navegación, los enlaces titulados, ayudas de navegación, elementos que dificultan la navegación o el acceso directo a los contenidos presentes en el trabajo de (NIELSEN, J., 2005) y (SERRANO MASCARAQUE, E. et al., 2010, pp.341-396). A continuación se aplican dos análisis automáticos de accesibilidad (TAW3 Test accesibilidad web, 2011) y HERA (BENAVÍDEZ, C. et al., 2005) en los que se obtendrá el número de errores detectados ya sean automáticos o manuales, dando como resultado una puntuación final que permita su comparación, tal como se aclara en el trabajo de (SERRANO MASCARAQUE, E., 2009, pp.61-103).

Análisis de usabilidad			
		NASA	**ESA**
Legibilidad		1	1
Presencia de secciones de la página		4	4
Árbol de navegación		1	3
Enlaces titulados		0	0
Ayudas de navegación		0	0
Elementos que dificultan la navegación		1	1
Acceso directo a la información		1	3
Elementos que distraen la navegación		0	0
Puntuación total		**9**	**12**
Análisis de accesibilidad			
		Errores NASA	**Errores ESA**
WCAG 1.0 TAW	A	1 automático, 259 manuales	0 automáticos, 5 manuales
	AA	91 automáticos, 238 manuales	6 automáticos, 8 manuales
	AAA	3 automáticos, 52 manuales	1 automático, 9 manuales
HERA	P1	3	3
	P2	8	10
	P3	5	4
Puntuación total		**660 errores**	**46 errores**

Tabla 6. *Análisis de usabilidad y accesibilidad complementarios.*

Como se expone en la *(tabla6)*, la usabilidad de ambos sitios web es similar, aunque ligeramente superior en el caso de la Agencia Espacial Europea con 12 puntos; ello se debe a un mejor árbol de navegación y acceso directo a la información. Son mejorables los aspectos de legibilidad ya sea por el tamaño de la fuente, como por su falta de redimensionamiento con unidades relativas de tipo *em*. Apenas se presentan ayudas a la navegación, careciendo en ambos casos de titulación de enlaces. El

empleo de javascript en gran parte de los menús de contenidos, dificulta el acceso a la información y genera artificios que dificultan la navegación del usuario.

Los resultados del análisis automático de la accesibilidad, con TAW y HERA, demuestran que el sitio web de la NASA con 660 errores, sobrepasa con diferencia a los 46 obtenidos por la ESA. Tales errores lo constituyen en su mayoría los códigos javascript y applets de java, considerados como objetos no accesibles, con más de 36 en todo el código, véase una muestra en la *(tabla7)*. Todo ello viene a confirmar los datos obtenidos por el programa *mbot* en los que se detecta una mayor tasa de errores y tiempo de análisis para el procesamiento de la información, que debe ser filtrada con mayor detenimiento hasta obtener los contenidos propiamente textuales.

```
…
Línea 97: <script src="/js/191731main_prototype.js" type="text/javascript" language="javascript">
Línea 97: <script src="/js/191696main_builder.js" type="text/javascript" language="javascript">
Línea 97: <script src="/js/191704main_effects.js" type="text/javascript" language="javascript">
Línea 97: <script src="/js/191701main_dragdrop.js" type="text/javascript" language="javascript">
Línea 97: <script src="/js/191738main_slider.js" type="text/javascript" language="javascript">
Línea 97: <script src="/js/191739main_sound.js" type="text/javascript" language="javascript">
Línea 97: <script src="/js/191698main_controls.js" type="text/javascript" language="javascript">
Línea 97: <script src="/js/194940main_RoundedMorph.js" type="text/javascript" language="javascript">
Línea 97: <script src="/js/191745main_transitions.js" type="text/javascript" language="javascript">
Línea 97: <script src="/js/191713main_global.js" type="text/javascript" language="javascript">
Línea 97: <script src="/js/191734main_ScrollRegion.js" type="text/javascript" language="javascript">
Línea 97: <script src="/js/191730main_PrettyCounter.js" type="text/javascript" language="javascript">
Línea 97: <script src="/js/191737main_SkinnedSelect.js" type="text/javascript" language="javascript">
Línea 97: <script src="/js/191695main_Blinds.js" type="text/javascript" language="javascript">
Línea 97: <script src="/js/191742main_Tabs.js" type="text/javascript" language="javascript">
Línea 97: <script src="/js/4564main_index.js" type="text/javascript" language="javascript">
Línea 97: <script src="/js/196917main_flash_rd.js" type="text/javascript" language="javascript">
Línea 97: <script src="/js/207088main_club2.js" type="text/javascript" language="javascript">
…
```

Tabla 7. *Muestra de elementos no accesibles en el sitio web de la NASA*

Conclusiones

1. El programa *mbot* puede efectuar análisis de la web hasta 3 niveles de profundidad, obteniendo metadatos, meta-etiquetas, canales de sindicación, enlaces, documentos, imágenes, archivos multimedia, código fuente y texto indexado.

2. El sitio web de la NASA es el más extenso con 32.865 páginas y 10.285 contenidos frente a 28.232 y 6.754 recursos del sitio web de la ESA. El número de canales de sindicación también resulta muy superior, con más de 300, lo que significa una mayor redifusión y alcance de la información publicada.

3. Las páginas web de la ESA están mejor descritas, ya que emplean una media de 3 meta-descripciones por página frente a 0,3 de la NASA. Esto es un total de 7.661 etiquetas con respecto a 662, lo que permite una mejor recuperación de los contenidos de la Agencia Europea en relación con los de su homóloga Norteamericana.

4. El tiempo de ejecución para analizar el sitio web de la NASA (40 minutos y 59 segundos) es un 36% superior al que se empleo con el sitio web de la ESA (26 minutos y 16 segundos).

Ello puede deberse a los errores de arquitectura y accesibilidad detectados en mayor medida en las páginas web de la NASA.

5. Los test de accesibilidad y usabilidad, así como los resultados obtenidos por mbot, permiten establecer una correlación entre una buena programación y arquitectura de la web que incide directamente en el tiempo de procesamiento y correcto tratamiento de la información con programas webcrawler.

Bibliografía

- BAEZA YATES, R. y E. GRAELLS. 2008. *Características de la Web Chilena 2007*. [online]. [Consultado el 19 Mar 2011]. Disponible en: http://www.ciw.cl/caracterizacion-web/estudio2007/estudio.pdf

- BENAVÍDEZ, C., E. GUTIÉRREZ Y RESTREPO, y C. MCCATHIE NEVILE. 2005. *Hera 2.1*. [online]. [Consultado el 19 Mar 2011]. Disponible en: http://www.sidar.org/hera/

- BLÁZQUEZ OCHANDO, M. 2009. *Instalación y primera preparación de Nutch1.0 en Ubuntu 9.10*. [online]. [Consultado el 19 Mar 2011]. Disponible en: http://mblazquez.es/documents/blazquez-M_2009_manual-de-instalacion-de-nutch.pdf.pdf

- BLÁZQUEZ OCHANDO, M. 2011. *Primeras pruebas del mbot webcrawler*. [online]. [Consultado el 19 Mar 2011]. Disponible en: http://www.mblazquez.es/documents/articulo-pruebas1-mbot.html

- CASTILLO, C. 2004. *EffectiveWeb Crawling*. Santiago de Chile: Dpto. de Ciencias de la Computación, Universidad de Chile.

- NIELSEN, J. 2005. *Heuristic Evaluation*. [online]. [Consultado el 19 Mar 2011]. Disponible en: http://www.useit.com/papers/heuristic/

- *Nutch Wiki*. 2011. [online]. [Consultado el 24 Mar 2011]. Disponible en: http://wiki.apache.org/nutch/

- SERRANO MASCARAQUE, E. 2009. Accesibilidad vs. usabilidad Web: evaluación y correlación. *Investigación Bibliotecológica*. **23**(48), pp.61-103.

- SERRANO MASCARAQUE, E., A. MORATILLA OCAÑA, y I. MORATILLA OCAÑA. 2010. Métrica para la evaluación de la accesibilidad en Internet: propuesta y testeo. *Revista española de Documentación Científica*. **33**(3), pp.341-396.

- *TAW3 Test accesibilidad web*. 2011. [online]. [Consultado el 19 Mar 2011]. Disponible en: http://www.tawdis.net/