El uso de software libre en los sitios *web* universitarios españoles

Carlos G. Figuerola, José L. Alonso Berrocal, Ángel F. Zazoy Emilio Rodríguez

Grupo de Investigación REINA, Universidad de Salamanca
reina@usal.es http://reina.usal.es/

Resumen El desarrollo de la tecnología *web* ha convertido a éste en uno de los principales medios de comunicación de muchas entidades. Éste es el caso de universidades y otras instituciones académicas, que han hecho del *web* el instrumento probablemente más importante de difusión de información.
En el presente trabajo se analizan datos exhaustivos obtenidos de los sitios *web* de 72 universidades españolas, y se ofrece una visión detallada sobre los sistemas operativos, versiones más usadas, distribuciones preferidas; y se ponen en relación con el tamaño real de los servidores *web* soportados. Otro tanto se hace con el software servidor *web*. Los lenguajes de programación *web* utilizados son analizados desde el punto de vista de la mayor o menor penetración del software libre, al igual que los distintos formatos utilizados para los diferentes tipos de información. Igualmente, y en la medida en que hay datos disponibles, al menos para algunas universidades, referidos a años anteriores, se intenta un estudio diacrónico que muestra la evolución de varios de los elementos estudiados. Todo ello permite ofrecer una visión bastante pormenorizada sobre la difusión y uso del software libre en los sitios o sedes *web* de las universidades epañolas, así como su evolución reciente.

Palabras clave: software libre, sitios *web*, universidades

1. Introducción

El desarrollo de la tecnología *web* ha convertido a éste en uno de los principales medios de comunicación de muchas entidades. Éste es el caso de universidades y otras instituciones académicas, que han hecho del *web* el instrumento probablemente más importante de difusión no sólo de informaciones administrativas, sino también de materiales docentes y resultados de la investigación. La facilidad de realización y publicación de páginas *web* ha contribuido sin duda a ello, propiciando la proliferación de servidores y sitios *web*.

De otro lado, parece de interés conocer datos acerca de la penetración y uso real del *sofware libre* en distintos ámbitos. Diversos trabajos han avanzado en esta línea, aplicando diferentes enfoques y metodologías; entre ellos, y por citar sólo algunos, cabe mencionar el conocido como *Libro Blanco* (2), referido sobre todo al ámbito empresarial; el informe sobre el *software libre* en Cataluña y en

España (15) o el informe REINA (7) (nada que ver con nuestro Grupo de Investigación, la homonimia es mera coincidencia) sobre el uso de la tecnología en la Administración Pública española. Algunos otros trabajos sobre distintos aspectos del *web* ofrecen también información sobre el uso del *software libre*, como el Informe sobre el *web* español de Baeza-Yates (4), o, anterior y más limitado, el estudio cibermétrico sobre los *webs* de algunas instituciones universitarias o relacionadas con la investigación científica (3).

En el transcurso de un estudio realizado durante 2006, subvencionado por el MEC, tuvimos ocasión de explorar de forma automática los dominios *web* de 72 universidades españolas (existen 74), prácticamente en su totalidad. Esto nos permitió recopilar abundantes datos sobre el desarrollo del *web* en las universidades españolas. El estudio cibermétrico citado antes, realizado en 2003, aunque mucho menos exhaustivo, también nos permite disponer de datos para abordar el uso y penetración del *software libre* en las universidades españolas.

Este trabajo está organizado como sigue: en la siguiente sección se exponen algunos problemas de orden metodológico encontrados en la realización de este estudio, al tiempo que se ofrecen los datos de tipo general en que se basa nuestro estudio. A continuación se analizan los sistemas operativos utilizados mayoritariamente, haciendo especial distinción entre los que son *soft libre* y los que no lo son. En la sección siguiente se analiza en parecida forma el *software* servidor *web*; y a continuación, los lenguajes de programación *web* empleados, para hacer, en la sección siguiente, un análisis de los formatos de ficheros más utilizados, siempre desde la perspectiva de observar el uso y penetración del *soft libre* y de los estándares abiertos. Finalmente, se extraen una serie de conclusiones.

2. Metodología

Como se ha dicho, los datos base proceden de la exploración automática del *web*. Dicha exploración se llevó a cabo mediante la utilización de un par de *spiders* de elaboración propia; cae fuera del ámbito de este trabajo la discusión de los problemas de diversa índole relacionados con el diseño y trabajo de los *spiders*; una discusión de este tipo puede encontrarse en (10). Pero sí cabe afirmar que se trata no de un muestreo más o menos amplio, sino de una recogida de datos cercana a la totalidad del *web* universitario español.

Universidades exploradas	72
Número total de hosts	6 392
Número total de páginas	4 199 081

Cuadro 1. Datos globales obtenidos en 2006

Los *spiders* recogen páginas *web*, de las cuales pueden extraerse una serie de datos de interés. Pero buena parte de la información interesante para nuestro propósito procede no de las páginas en sí, sino de las cabeceras devueltas por

los diferentes servidores en respuesta a la solicitud de las distintas páginas por parte de los clientes (navegadores o *spiders*). Estas están reguladas por el protocolo correspondiente ((8) y (14)) y, de forma opcional, ofrecen informaciones de diverso tipo, en especial en la cabecera *Server:*. Hay que advertir, sin embargo, que muchas de estas informaciones no son siempre fiables; en muchos casos, simplemente no existen; y, en otros, no siempre dicen la verdad (9), por diversas razones.

De otro lado, la falta de uniformidad en algunas de las informaciones más interesantes para nosotros conducen a una dispersión difícil de tratar. Esto hace que no siempre podamos establecer determinados aspectos (por ejemplo, el Sistema Operativo del *host*). Pero la exhaustividad de los datos recolectados puede suplir estas carencias.

3. Sistemas Operativos

Una de las informaciones interesantes es la referente al Sistema Operativo de los *hosts* que conforman los *webs* de las diferentes universidades. La dispersión de datos derivada de la variedad de sistemas y versiones es grande, pero algunos grandes bloques pueden distinguirse. Igualmente, hay un número importante de *hosts* que no ofrecen datos sobre este particular, así como de máquinas que ofrecen datos demasiado genéricos. Éste es el caso de los que señalan como Sistema Operativo Unix, sin mayor precisión; este dato es importante porque podemos sospechar que muchos de ellos son Linux, aunque no tenemos constancia fehaciente; también, porque determinadas variedades de Unix no pueden considerarse *software libre*.

Los bloques de Sistemas Operativos que hemos podido distinguir pueden verse en el Cuadro adjunto. De ellos son *soft libre* todos los del bloque Linux, que es el más numeroso. También lo es FreeBSD, aunque éste tiene una presencia testimonial. También lo son, probablemente, parte de los etiquetados como Unix, sin más, pero éstos no pueden ser cuantificados.

De esta forma, si atendemos al número de *hosts* que utilizan cada variedad de sistema operativo, dejando de lado aquéllos de los que no tenemos datos precisos, podemos apreciar una igualdad entre *software libre* y propietario.

Linux	1 856
Windows	1 766
Unix	1 344
MacOS / Darwin	77
Sun Solaris	45
OpenVMS	43
OS/2	2
FreeBSD	9
Otros / sin datos	1 250

Cuadro 2. Sistemas Operativos por número de hosts

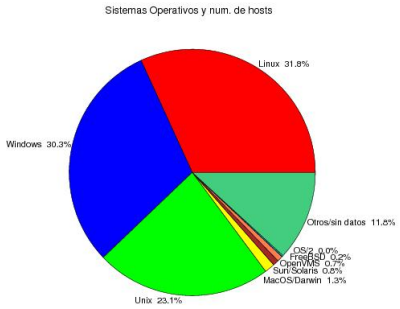

Figura 1. Sistemas Operativos por número de *hosts*

Sin embargo, si tomamos en cuenta no el número de *hosts*, sino el de páginas que albergan dichos *hosts* apreciaremos diferencias notables. La cantidad de páginas servidas en máquinas con Linux es abrumadoramente mayoritaria, y eso si no tomamos en cuenta que, como se ha dicho, parte de las máquinas son sistema operativo declarado como Unix probablemente son Linux también.

El seguidor más inmediato, el de los sistemas Windows, hospeda menos de la tercera parte que Linux. El resto de sistemas, de otro lado, tiene una presencia reducida, haciendo excepción de los *hosts* con sistema operativo de Sun; aunque siempre muy por debajo de los mayoritarios.

Linux	1 108 406
Unix	1 560 744
Windows	337 029
Sun-Solaris	64 912
MacOS-Darwin	4 634
OpenVMS	5470
FreeBSD	626
OS2	509
Otros/ s.d.	1 116 751

Cuadro 3. Sistemas Operativos por número de Páginas

Adicionalmente, siendo Linux el Sistema Operativo mayoritario, y la opción más importante dentro del *software libre* de este tipo, podemos preguntarnos acerca de las distintas variantes o distribuciones utilizadas por los servidores universitarios. Hay, como podía esperarse, también una dispersión notable, debido a la multiplicidad de versiones; pero podemos apreciar claramente que la distribución más utilizada es Debian, seguida a corta distancia de Red Hat/Fedora. Las demás se encuentran ya a considerable distancia.

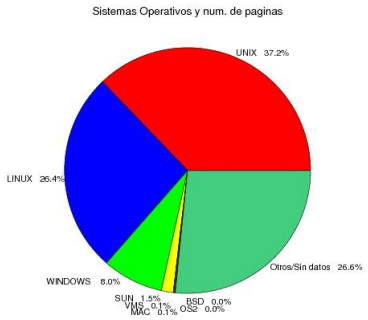

Figura 2. Sistemas Operativos por número de páginas

Debian	767
Red Hat/Fedora	721
Suse	199
Mandrake/Mandriva	79
Ubuntu	59
Gentoo	9

Cuadro 4. Distribuciones Linux por número de hosts

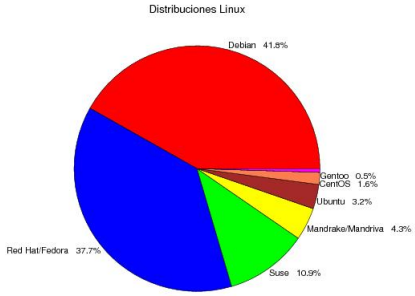

Figura 3. Distribuciones Linux y número de *hosts*

4. Servidores *Web*

Otra de las informaciones útiles que podemos obtener es el *software* servidor *web* que utilizan los diferentes *hosts* de las universidades. Como en otros elementos analizados, la dispersión de datos es importante, pero a pesar de ello podemos obervar tendencias generales, así como el mayor o menor grado de implantación de *software libre* frente a propietario.

Apache	4 170
Microsfot IIS	1 425
Netscape	144
Oracle	66
Zope	63
WASD OpenVSM	43
Sun-ONE-Web-S	41
Roxen	25
Lotus-Domino	24
SAMBAR	21
AOLserver	14
Otros	356

Cuadro 5. Software servidor *web* por número de hosts

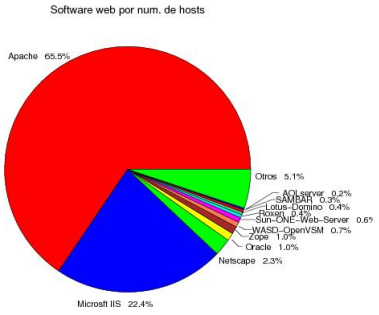

Figura 4. Software Servidor *Web* y número de *hosts*

Desde este punto de vista, cabe resaltar el dominio de Apache sobre todos los demás servidores. El más cercano, IIS de Microsoft, queda a gran distancia, y los demás a mucha mayor distancia todavía. Parece que, en lo que a servidores se refiere, las dos grandes opciones son Apache (*soft libre*) y MS IIS (propietario), con clara ventaja para el primero.

Si observamos el *soft* servidor desde el punto de vista de la cantidad de páginas albergadas, en lugar del número de *hosts*, la diferencia entre Apache y

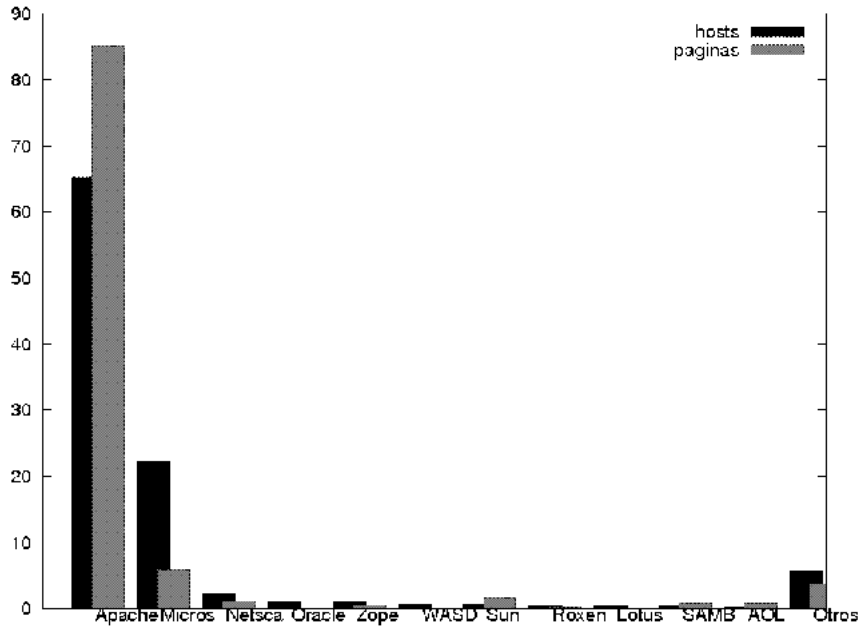

Figura 5. Servidores según número de *hosts* y páginas

Apache	3 574 881
Microsoft IIS	248 217
Netscape	47 836
Oracle	697
Zope	22 598
WASD OpenVSM	5 470
Sun-ONE-Web-S	64 912
Roxen	7 932
Lotus-Domino	766
SAMBAR	34 929
AOLserver	32 118
Otros	158 725

Cuadro 6. Software servidor *web* por número de páginas servidas

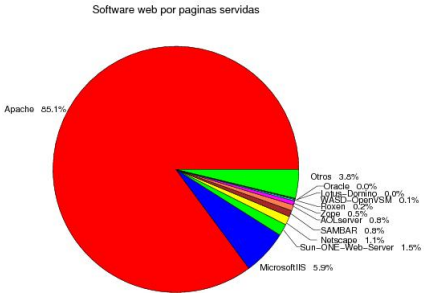

Figura 6. Software Servidor *Web* y número de páginas

MS IIS se acentúa notablemente a favor del primero. Parece claro que el grueso de las páginas que conforman el *web* universitario español está servido a través de Apache.

De otra parte, si relacionamos estos datos con los que vimos antes sobre sistemas operativos, parece claro que una parte importante de las instalaciones de Apache lo son sobre máquinas Windows.

Apache, de otro lado, tiene numerosas versiones y variantes, como es sabido. Apache 1.3 parece ser el más difundido; incluso más que Apache2; tal vez la mayor complejidad de éste último, junto con la estabilidad y buen rendimiento del primero, pueden explicar esto.

Apache 2.2	49
Apache 2.1	2
Apache 2.0	1 555
Apache 1.3	1 957
Apache 1.2	12
Apache Coyote	80
Otros	38
Sin especificar	477

Cuadro 7. Versiones de Apache

Otro asunto es el tamaño de los servidores, medido con las datos de que disponemos: el número de páginas. Aunque ya hemos comentado una visión global, que nos muestra a Apache como dominador absoluto, quizá una visión algo más pormenorizada pueda enseñarnos algo más. Para ello, hemos considerado varios grupos de servidores en función de su tamaño; a partir de ahí, vemos que, por lo que se refiere a Apache, más del 80 % de los servidores que lo usan tienen menos de 100 páginas. Sin embargo, para Microsoft IIS, esta cifra desciende a casi el 60 %.

Para el resto de los tamaños los datos siguen parecida tónica. Esto parece indicar que Microsoft se usa para servidores grandes con más asiduidad que Apache; tal cosa no debería extrañar, puesto que parece menos probable que alguien pague licencias de *soft* propietario para ofrecer pocas páginas.

	menos de 100	entre 100 y 1 000	entre 1 000 y entre 10 000	más de 10 000
Apache	81.01	13.6	4.36	1.03
Microsoft	59.92	28.69	9.21	2.18
Netscape	82.64	0.29	9.03	0
Oracle	96.97	3.03	0	0
Zope	69.84	20.63	9.52	0
WASD	0	81.82	18.18	0
Sun-ONE	85.37	7.32	0	7.32
Roxen	60	28	12	0
Lotus	92	8	0	0
Sambar	52.38	38.1	0	9.52
AOLServer	78.57	7.14	7.14	7.14

Cuadro 8. Software servidor *web* por tamaño de servidores (en %)

En este sentido, cabe destacar el caso del servidor WASD, cuyas pocas instalaciones paracen dedicarse a servidores de muchas páginas. Este caso, no obstante, es poco significativo, si tenemos en cuenta que este servidor se usa en sólo dos universidades; éstas parecen haber optado claramente por este servidor, puesto que contabilizan hasta 43 *hosts* entre ambas. Probablemente estamos ante máquinas relativamente potentes con varios *hosts* virtuales. WASD es un paquete para máquinas con VMS, que se distribuye mediante licencia GPL (6).

También podemos fijarnos en el servidor Roxen, usado en un par de universidades, pero con 25 *hosts*. Roxen, sin embargo, es *software libre* que corre en multitud de sistemas y arquitecturas, y que acompaña a un potente Sistema de Gestión de Contenidos (1).

Apache	67
Microsoft IIS	61
Netscape	18
Oracle	18
Zope	22
WASD OpenVSM	2
Sun-ONE-Web-S	15
Roxen	2
Lotus-Domino	8
SAMBAR	6
AOLserver	5

Cuadro 9. Software servidor *web* por universidades

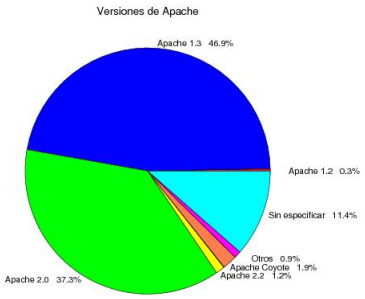

Figura 7. Versiones de Apache

Si observamos los datos recogidos en 2003 sobre esta particular, referidos sólo a número de hosts con cada *software* servidor, veremos que ya entonces el asunto se dirmía básicamente entre Microsoft y Apache, con clara predominancia de éste. Los servidores de Netscape, en sus distintas variantes, eran entonces bastante utilizados, puesto que más de un 13 % de los *hosts* lo utilizaban.

Apache	56.25
Microsoft IIS	16.86
Netscape	13.56
Oracle	0.65
Zope	0.13
Sun-ONE-Web-S	0.26
Roxen	0.13
Lotus-Domino	0.78
SAMBAR	1.84

Cuadro 10. Software servidor *web* en 2003 (porcentajes en num. de *hosts*)

5. Programación *web*

Una de las formas de aportar interactividad a las páginas *web* es el uso de *scripts* que corren en el servidor, y que reciben datos a través de formularios y similares. La tecnología clásica de este tipo se basa formularios CGI; es difícil, con los datos que tenemos, saber qué lenguajes hay del otro lado de esos CGI. En ocasiones los programas o *scripts* tienen la extensión *.pl*, pero todo el mundo sabe que Perl es uno de los lenguajes favoritos de los CGI. Los programas sin extensión, o con la extensión *.cgi* son, de lejos, los más abundantes, de manera que, por esa vía, no hemos podido avanzar demasiado.

Más moderno es el uso de intérpretes como PHP o ASP. Como es bien sabido, el primero es *soft libre* (12) mientras que el segundo es un ejemplo típico de *software propietario* de Microsoft (5).

Gracias a la formación de los URL, podemos comparar la cantidad de enlaces a *scripts* de uno u otro tipo. La diferencia es bastante clara a favor de PHP; naturalmente, puede pensarse en una asociación entre el *software* Servidor *Web* y el lenguaje utilizado, pero, en cualquier caso, la tendencia hacia el uso de soluciones de código libre es muy marcada en este campo.

	PHP	ASP
Núm. de Páginas	14 012 783	544 750

Cuadro 11. Uso de *scripting* en servidor

6. Formatos de ficheros

Las páginas *web* enlazan recursos de muy diverso tipo, y los formatos de dichos recursos son un exponente claro de la penetración o no de los formatos abiertos. Estos recursos enlazados van desde información institucional (planes de estudio, procedimientos administrativos, etc.) a material docente de las diversas asignaturas, pasando por resultados de la investigación, convocatorias, etc.

En su mayor parte, de trata de recursos que se desea difundir de forma general. Cuendo no es así, se utilizan protecciones con contraseñas u otros métodos y, en consecuencia, caen fuera del ámbito analizado por nostros. Naturalmente, tales recursos pueden haber sido elaborados con programas diversos, pero lo importante para este tabajo son los formatos con que se difunden.

En este sentido, la variedad de recursos y de formatos nos lleva nuevamente a una dispersión poco útil. Pero si nos centramos en algunos de los casos más comunes, como es el caso de los documentos de tipo básicamente textual (aunqaue puedan incorporar imágenes, etc.), veremos que PDF es el formato favorito, sin lugar a dudas; éste es un formato abierto (13), y lo mismo puede afirmarse del texto plano. Cosa distinta es el caso del RTF (16), que, sin ser abierto en sentido estricto, sí permite una relativa interoperabilidad. El formato del procesador de textos de Microsoft, plenamente propietario (*.doc*), queda en último lugar.

La comparación con los datos obtenidos en 2003 no es sencilla, toda vez que entonces no se recogieron datos sobre el uso de RTF, y sí sobre otros formatos, hoy de escasa significancia. Pero si tenemos en cuenta solamente los tres tipos (DOC, TXT y PDF), podremos extraer alguna conclusión. En primer lugar, el descenso de la proporción de documentos en texto plano, lo cual parece lógico, si tenemos en cuenta su pobreza visual. Más importante, la confirmación clara del decaimiento del formato propietario DOC en favor del PDF; hay razones técnicas para ello (por ejemplo, la no posibilidad de modificación de los PDF

colgados en la red), pero queremos ver también un escoramiento claro hacia el uso preferente de formatos abiertos.

Es preciso mencionar el detalle de que el número alto de ficheros RTF se debe en buena parte a una única universidad.

Por contra, recursos en formatos *Open Document* (11) son muy escasos, practicamente testimoniales. En sentido contrario, es relativamente abundante el uso de PPT *MS Power Point*.

Por lo que se refiere a formatos de imagen, la abundancia de JPEGs era, probablemente, esperada, como también la de GIFs. Ambos formatos, sin embargo, tienen usos diferentes, por lo que tal vez sea más oportuno comparar GIF (propietario) con PNG (formato abierto). En este sentido, parece clara la preponderancia de GIF.

pdf	993 536
doc	111 437
txt	278 255
rtf	189 628
xls	7 870
ppt	16 718
opendoc	129
jpg	494 186
gif	309 358
png	26 423

Cuadro 12. Formatos de ficheros enlazados

	2003	2006
.doc	23.83 %	8.06 %
.pdf	45.32 %	71.83 %
.txt	31.35 %	8.06 %

Cuadro 13. Evolución de formatos de texto más importantes

7. Conclusiones

Se han analizado datos obtenidos tras la exploración automática de la mayor parte del espacio *web* universitario español. A través de estos datos, hemos visto que el *software libre* tiene un grado notable de implantación a nivel de servidores o *hosts*. En lo que respecta a Sistemas Operativos, Linux es el más utilizado al igual que diferentes versiones de Unix sin identificar. En *software servidor*, el dominio del *soft libre* es patente, en especial a través de Apache, así como

en programación *web* la supremacía es para PHP, la opción libre, de forma abrumadora.

Las opciones propietarias están básicamente representadas por productos Microsoft (otras opciones parecen minoritarias); mientras que del lado del *soft libre* las opciones parecen concentrarse en Linux (especialmente Debian), Apache y PHP.

Esto, por lo que se refiere a servidores. Sin embargo, en lo que se refiere a formatos de ficheros, el Open Document resulta marginal. Se trata de especificaciones muy recientes, lo cual influye en su poco uso, pero cabe preguntarse si, en general, el uso de la *suite* ofimática propietaria de Microsoft sigue siendo mayoritaria, y esto repercute en los formatos de los ficheros que se cuelgan en la red. En sentido contrario, sin embargo, el aumento de PDFs y la disminución de .doc puede apuntar a una mayor conciencia acerca del uso de estándares abiertos para difundir información, independientemente de que, para su elaboración, se haya utilizado *software* no abierto.

De otro lado, a pesar de la precariedad de datos para épocas anteriores, la comparación con lo recogido en 2003 muestra que la situación actual, en lo que a penetración y uso de *software libre* no es sino una continuación de las tendencias apuntadas entonces, en el sentido de una mayor implantación general de las alternativas libres frente a las propietarias.

Bibliografía

[1] Roxen Internet Software AB. Roxen web server, 2007. http://www.roxen.com/products/webserver/.

[2] Alberto Abella García and Miguel Angel Segovia. Libro blanco del software libre, 2006. http://libroblanco.com/joomla/document/III_libro_blanco_del_software_libre.pdf.

[3] José Luis Alonso Berrocal, Carlos G. Figuerola, and Ángel F. Zazo. *Cibermetría: Nuevas Técnicas de Estudio Aplicables al Web*. Trea, Gijón, 2004. ISBN: 84-9704-114-3.

[4] Ricardo Baeza-Yates, Carlos Castillo, and Vicente López. Características de la Web de España. *El profesional de la información*, 15(1):6–17, enero-febrero 2006. http://www.catedratelefonica.upf.es/webes/2005/Estudio_Web_Espana.pdf.

[5] Microsoft Corporation. The official microsoft asp.net 2.0 site, 2006. http://www.asp.net.

[6] Mark G. Daniel. WASD. The only web environment implemented expresely for VMS, 2007. http://wasd.vsm.com.au/.

[7] Consejo Superior de Administración Electrónica. Informe REINA (Recursos Informáticos de la Administración del Estado), 2006. http://www.csi.map.es/csi/iria2006/iria_2006.pdf.

[8] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. RFC 2616: Hypertext Transfer Protocol–HTTP/1.1, 1999. ftp://ftp.rfc-editor.org/in-notes/rfc2616.pdf.

[9] Carlos G. Figuerola, José Luis Alonso Berrocal, Angel F. Zazo, and Emilio Rodríguez Vázquez de Aldana. Web page retrieval by combining evidence. In C. Peters, F. Gey, J. Gonzalo, H. Mueller, G.J.F. Jones, M. Kluck, B. Magnini, and M. de Rijke, editors, *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers*, volume 4022 of *Lecture Notes in Computer Science*, pages 880–887. Springer, 2006.

[10] Carlos G. Figuerola, José Luis Alonso Berrocal, Ángel F. Zazo Rodríguez, and Emilio Rodríguez Vázquez de Aldana. Diseño de spiders. Technical Report DPTOIA-IT-2006-002, Departamento de Informática y Automática - Universidad de Salamanca, March 2006. http://reina.usal.es/pub/figuerola2006diseno.pdf.

[11] Organization for the Advancement of Structured Information Standards. Open document format for office applications (opendocument) v1.1, 2007. http://docs.oasis-open.org/office/v1.1/OS/OpenDocument-v1.1-html/OpenDocument-v1.1.html.

[12] The PHP Group. Php: Hypertext processor, 2007. http://www.php.net.

[13] Adobe Systems Incorporated. Porqué pdf?, 2007. http://www.adobe.com/es/products/acrobat/adobepdf.html.

[14] R. Khare and S. Lawrence. RFC 2817: Upgrading to TLS within HTTP/1.1, 2000. ftp://ftp.rfc-editor.org/in-notes/pdfrfc/rfc2817.txt.pdf.

[15] Meritxell Roca Sales. El software libre en Catalunya y España, 2006. http://portal.uoc.edu/west/media/F-1214-1272.pdf.

[16] Microsoft Technical Support. Microsoft MS-DOS, Windows, Windows NT, and Apple Macintosh Applications. Rich Text Format (RTF Specifications, 2001. http://www.snake.net/software/RTF/RTF-Spec-1.7.pdf.