Maq. ENFOQUES-2 20/5/03 15:17 Página 1

Caracterización de una muestra de sedes Web españolas bajo dominio .es

Characterization of a Sample of Spanish Web Sites under the Top Level Domain .es

C.B. Amat

Resumen

Se ha llevado a cabo un análisis cuantitativo de las características de una muestra aleatoria de sedes Web españolas encuadradas en el dominio .es. Además del estudio de su accesibilidad directa, las variables analizadas incluían tamaño, estructura jerárquica, número y tipo de los elementos incluidos en las páginas, grado de conexión a través de enlaces y grado de interactividad mediante funciones. El modelo general resultante corresponde a una distribución exponencial de la práctica totalidad de las variables, lo que permite definir 3 grupos de sedes. Se discuten las implicaciones de estos hallazgos para los sistemas de recuperación de información distribuída en Internet y se sugiere que el estudio dinámico de esta u otras muestras permita seguir la evolución de las sedes Web españolas y mejorar su accesibilidad a través de los buscadores.

Palabras Clave: Sedes Web, caracterización del Web

Summary

A random sample of Spanish Web sites has been studied. Characteristics such as size, jerarchical structure, number of pages, number and type of text and graphical elements and links in and out of the sites show a similar statistical pattern, that of a power distribution. Three groups or levels of sites can be devised according to these results and it is suggested a dynamic study of the sites as the method best suited to ascertain their characteristics and to adjust search engines for retrieval of their content.

Keywords: Web Sites, Web Characterization

Se ha llevado a cabo un análisis cuantitativo de las características de una muestra aleatoria de sedes Web españolas encuadradas en el dominio .es

1.- Introducción

Los términos Web Demographics, Web Statistics, Web Metrics, Web Characterization, Cybermetrics e incluso Web Archaeology se han empleado desde hace tiempo para referirse a la variada gama de proyectos y trabajos de análisis cuantitativo de ese componente de Internet. Junto a los recuentos estadísticos realizados periódicamente por organismos reguladores, cabe encuadrar en este área las investigaciones cuantitativas dirigidas a la obtención de hallazgos sobre el volumen de este espacio informativo, su dinámica, su estructura y sus características. A grandes rasgos, se han identificado inicialmente tres líneas de investigación[1]:

1) el análisis del rendimiento y de los flujos de información y datos a nivel global
2) el análisis de contenido de los documentos distribuidos en el espacio Web
3) el estudio de la interrelación entre esos mismos documentos y sus usuarios.

Las investigaciones encuadradas en el primer grupo, de los que son buen ejemplo los trabajos de KC Claffy para el National Laboratory for Applied Network Research[2] y la Cooperative Association for Internet Data Analysis[3], se dirigen al estudio de los modelos de flujo informativo y sus resultados son aprovechables en la mejora de infraestructuras y de los sistemas de medición de audiencias. Los análisis estructurales del web considerado globalmente, que también podrían encuadrarse en esta línea, ofrecen además resultados directamente relacionados con la accesibilidad de la información en el Web. Es el caso del trabajo de Andrei Broder y colaboradores que, entre su lista de motivaciones, incluye "el diseño de estrategias de recopilación en el Web" y "la comprensión sociológica de la creación de contenidos en la Web"[4]. Otro tanto ocurre con trabajos, también eminentemente numéricos, que comienzan a ofrecer interesantes modelos predictivos[5].

Maq. ENFOQUES-2 20/5/03 15:17 Página 2

El diseño, la implantación y el mantenimiento de cualquier sistema de información ha de tener en cuenta las características de los documentos que incorpora

Buenos exponentes de los análisis sobre la interacción entre documentos y usuarios en el espacio Web, son, por una parte, los trabajos de James Pitkow y su grupo, iniciados en el Instituto Tecnológico de Georgia y continuados en Xerox, cuyo objetivo último pasa por dilucidar las propiedades del Web como "sistema ecológico del conocimiento" habi. Por otra parte, los ingentes análisis del grupo de investigación de Digital sobre las transacciones de Altavista[7], de IBM sobre las de WebCrawler[8] y de Jansen, Spink y Saracevic sobre Excite[9] concretan su tratamiento a las operaciones de recuperación de información de los usuarios. En fechas muy recientes, se han empezado a realizar estudios sobre la accesibilidad de las sedes para colectivos desfavorecidos (usuarios con algún tipo de minusvalía). Estos estudios emplean también análisis cuantitativos de determinados indicadores y elementos de la sintaxis de las páginas. Uno de los más recientes examina el ajuste a las recomendaciones técnicas del Consorcio Web de las sedes de las universidades españolas[10].

Uno de los más antiguos ejemplos de estudio del contenido de los documentos Web está contenido en la comunicación de Tim Bray a la quinta Conferencia Internacional sobre el WEB[11]. Su trabajo no sólo presenta una descripción estadística de variables como el tamaño de las páginas o su formato. Además, el recuento de los enlaces permite identificar las "mayores y más visibles sedes", dotando de un componente estructural a su análisis. En la misma conferencia Woodruff y colaboradores empleaban una muestra mucho mayor que la de Bray para analizar un completo conjunto de variables, que abarcaban desde el tamaño de los documentos al número de errores de código de las páginas[1]. Desde estos trabajos, dos elementos han animado la línea de caracterización cuantitativa del contenido de los documentos Web: el refinamiento metodológico y, especialmente, la puesta en relación de los hallazgos con la accesibilidad de los documentos en ese espacio.

Los estudios iniciales tomaban como muestra los documentos recopilados por sistemas de recuperación (OpenText, Inktomi) con apreciable intención promocional. Pero, a finales de los años 90, el grupo de investigación de los laboratorios NEC publicó unos resultados demoledores sobre la cobertura y el nivel de actualización de los mayores sistemas de búsqueda en Internet[12,13]. Por su parte, el grupo de investigación de Digital publicaba sus análisis sobre la cobertura de los sistemas[14], y el de Compaq aportaba importantes refinamientos metodológicos a la técnica de muestreo empleada por el equipo de NEC[15] que, previamente, el Web Characterization Project de OCLC había divulgado[16].

Los trabajos más recientes comienzan a atribuir una semántica a los enlaces entre sedes y páginas, de forma que la interconexión entre documentos puede contribuir a definir "temas comunes"[17]. Estos estudios "topológicos", claramente basados en conceptos y técnicas bibliométricas[18] apoyan la conveniencia del análisis estructural, que paulatinamente se encamina a posibilitar el análisis automático de contenido (clasificación e indización) basado en la interconexión entre documentos Web[19].

El diseño, la implantación y el mantenimiento de cualquier sistema de información ha de tener en cuenta las características de los documentos que incorpora. Mucho más si, como en el caso de los sistemas de recuperación del espacio Web, se ofrece la recuperación de documentos íntegros a través de la indización exhaustiva de su contenido. El volumen, el número de elementos o la jerarquía de páginas de una sede son factores con alguna incidencia en su accesibilidad a través de los sistemas de recuperación de información distribuída en Internet (SIDI). Otros, como el número de enlaces a otras sedes, la abundancia de textos en soportes indizables o la homogeneidad estructural de las páginas tienen una importancia crucial. La mera obtención de muestras del espacio Web es un factor ineludible en el cálculo de muchos indicadores empleados en la evaluación de los sistemas de recuperación de información distribuída (véase si no la tabla que aporta Ford[20]). El presente trabajo limita sus objetivos, precisamente, a la obtención y caracterización de una muestra de sedes Web españolas como base para la evaluación de los sistemas de recuperación de información distribuída en Internet (SIDI) en el entorno nacional.

Maq. ENFOQUES-2 20/5/03 15:17 Página 3

2.- Selección de las sedes y criterios de análisis

A lo largo del presente trabajo, se consideran equivalentes los términos *dominio de segundo nivel* y *sede*, que constituyen la unidad de análisis. A este respecto, se sigue la definición propuesta por la Web Characterization Activity del W3C: "Una sede Web es una colección de páginas Web interrelacionadas, que incluye una como principal, y residen en la misma localización de red" [21]. Aunque esta definición parece demasiado indinada hacia la concepción física de las sedes web, será matizada más adelante, siguiendo la propuesta terminológica inicial[22] que, curiosamente, se modificó en fecha posterior a la liquidación del grupo de W3C.

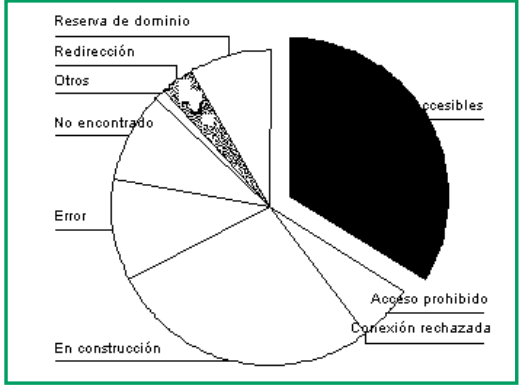
Las páginas web integradas en cada sede constituyen sus unidades de contenido. Sin embargo, cabe considerarlas como documentos compuestos de una serie de elementos:

1) indicadores de relación con otros documentos
2) llamadas de función
3) textos
4) elementos gráficos en diversos formatos que pueden presentarse aislados o en combinación con los anteriores.

A partir de la base de datos de ES-NIC, el Registro Delegado de Internet en España, se obtuvieron aleatoriamente 500 dominios de segundo nivel el día 4 de enero de 2001. En esa fecha, la base de datos contaba con 29.858 registros. Las 500 sedes estaban albergadas en 171 proveedores de dominio distintos (mediana = 25, máximo = 49).

La accesibilidad de cada sede se verificó mediante conexión directa. De las 500, sólo 168 presentaban información accesible. Las sedes en construcción (27,8%), las direcciones erróneas (19,6%) y las reservas de dominio, asociadas en la mayoría de los casos con marcas comerciales (8,2%) constituían los 3 grandes grupos cuya información resultó inaccesible, dos tercios del total (Figura 1).

FIGURA 1:
ACCESIBILIDAD DE LAS SEDES



En algunos casos se detectaron elementos inaccesibles. Correspondían en su mayor parte a ficheros gráficos cuyo análisis rebasaba el tiempo fijado por el programa de verificación (ver párrafos siguientes) a pesar de que cada sede problemática se conectó en 3 ocasiones diferentes. El resto de estos elementos no eran accesibles por la existencia de errores de ubicación, directorios inexistentes o encaminamientos fallidos.

También se verificó mediante conexiones directas la existencia de múltiples denominaciones o asignación a diferentes dominios de primer nivel de idénticos contenidos en los dominios de segundo nivel. 40 sedes ofrecían una doble denominación (.es y .com), 2 más ofrecían idénticos contenidos bajo .es y .net y una sede ostentaba triple asignación: .es, .com y .net. También se detectó denominación múltiple bajo dominio .es. Así, tpi.es ofrecía la misma información que paginas-amarillas.es en el momento de las conexiones.

También se verificó mediante conexiones directas la existencia de múltiples denominaciones o asignación a diferentes dominios de primer nivel de idénticos contenidos en los dominios de segundo nive

En el análisis se examinó la estructura en niveles de cada sede y se individualizaron los elementos integrantes, su tipo y volumen. También se cuantificaron los enlaces de cada sede a otras, así como los enlaces de partida de cada página

Para el análisis de las 168 sedes válidas, se recurrió a Xenu's Link Sleuth 1.1c, un programa de verificación de enlaces que, además, permite identificar algunas características de los dominios que analiza[23]. De esta forma se examinó la estructura en niveles de cada sede y se individualizaron los elementos integrantes, su tipo y volumen. También se cuantificaron los enlaces de cada sede a otras, así como los enlaces de partida de cada página.

3.- Caracterización

Las sedes analizadas contienen un total de 15.812 páginas que incluyen 47.735 elementos y ocupan 5,68 terabytes. Por término medio, cada sede distribuye la información en 90,4 páginas, estructuradas en algo menos de 6 niveles jerárquicos. Sin embargo, ni éstas ni las restantes variables estudiadas se ajustan a una distribución normal. Así, el número de niveles fluctúa entre 1 y 218, con desviación típica de 16,7, muy superior a la media de 5,66. La mediana de la distribución de esta variable se sitúa en 4. De igual modo, las medianas de las restantes distribuciones se alejan considerablemente de las medias: la del número de páginas se sitúa en 22,5 por sede, el volumen en 644,02 Kilobytes y el número de elementos en 81,5 por sede. La figura 2 muestra en a) la falta de ajuste a la normalidad de la variable número de elementos por sede y en b) el carácter exponencial de la distribución. Por otra parte, la existencia de contenidos duplicados (traducciones a varias lenguas de las mismas páginas) y de versiones de acceso diferentes (con "flash", sin marcos...) limita el valor de la variable referida al volumen.

FIGURA 2a

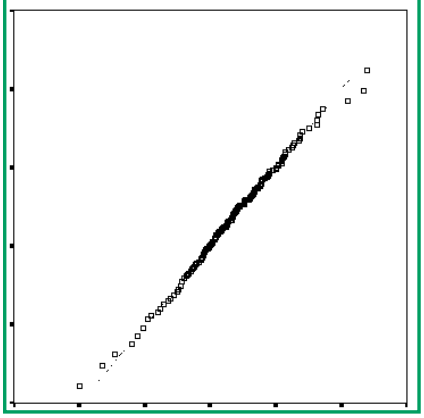
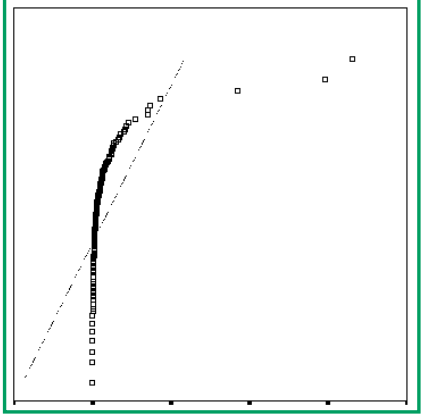


FIGURA 2b



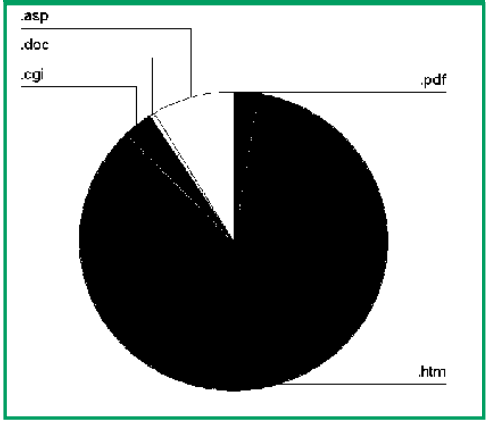
Se han individualizado los elementos textuales bajo 5 formatos diferentes: los integrados en páginas HTML, los documentos resultantes de aplicaciones diversas (Acrobat y MS Word) y los resultantes de consultas a bases de datos: ASP y CGI. La figura 3 muestra la proporción de cada uno de estos segmentos y la preponderancia de las páginas HTML estáticas sobre las páginas dinámicas y los archivos preparados con otras aplicaciones. Sólo 3 sedes empleaban documentos xml, si bien este formato representaba el 19% de las páginas de una de ellas.

Las páginas disponen de títulos significativos en muy alta proporción: en más de la mitad de las sedes, el número de páginas tituladas supera el 90%. Los elementos gráficos suponen el 56,72% del contenido de las sedes en términos de volumen. La mitad de las páginas contienen 1,7 archivos gráficos

Maq. ENFOQUES-2 20/5/03 15:17 Página 5

o menos. Existen, de nuevo, grandes desequilibrios en la distribución: sedes con gran riqueza gráfica y una con una media de 36 gráficos por página. Los formatos GIF y especialmente JPG son los dominantes, con proporciones que superan el 41 y el 58% respectivamente. Presencia testimonial cabe atribuir a los formatos PNG y BMP. De igual modo, el contenido de elementos audiovisuales es muy escaso: menos del 5% de las sedes incluyen archivos videográficos, una única sede contiene archivos sonoros y también es única la sede que ofrece representaciones de realidad virtual (archivos .vrml).

FIGURA 3: ARCHIVOS TEXTUALES



Las llamadas de función invocan operaciones a expensas de las acciones de los usuarios. Se han identificado operaciones de envío de correo y funciones javascript. Resulta chocante que 35 sedes no ofrezcan posibilidad alguna de contacto por correo. La mediana de esta distribución –de nuevo muy alejada de la normalidad– es de 1, y el valor máximo de 286. Los elementos JavaScript se utilizan sólo en el 30,5% de las sedes, con un máximo de 100 y, nuevamente, una distribución muy desequilibrada.

El conjunto de sedes analizadas muestran en total 207.208 enlaces de partida (links out), de los cuales 10.330 se dirigen a otras sedes. Las medianas de estas distribuciones son 218 y 2, respectivamente. 63 sedes, más de la tercera parte, no han establecido enlace alguno con otras.

La tabla que aparece a continuación presenta un resumen descriptivo de algunas de las variables estudiadas.

	Bytes ficheros textuales	Bytes ficheros gráficos	Niveles jerárquicos	Nº pp.	Nº elementos	Enlaces internos/pp.	Enlaces externos/pp.
Media	1.570.918,21	2.074.536,61	5,66	90,37	286,33	12,7716	,5073
Mediana	108.388,50	382.235,00	4,00	22,50	81,50	7,3962	,0737
Desv. típ.	5.958.887,956	8.301.555,391	16,698	209,199	787,656	18,90647	1,05015
Mínimo	427	0	1	1	1	,00	,00
Máximo	66.065.061	79.108.200	218	1.831	6.994	188,60	10,09

4.- Repercusión en los SIDI y panorámica

Se ha mencionado ya que el volumen de las sedes depende a partes iguales de los elementos textuales y gráficos. Cabe añadir que la existencia de documentos .pdf tiene una gran incidencia en el volumen de archivos textuales, si bien es cierto que el formato de intercambio de documentos puede combinar tanto elementos textuales como gráficos en su contenido. Desde el punto de vista de la recuperación de la información contenida en sedes Web, las variables que se refieren al volumen de

Desde el punto de vista de la recuperación de la información contenida en sedes Web, las variables que se refieren al volumen de las sedes no son las más importantes. Mayor relevancia tiene la estructuración y la distribución de los diversos tipos de documentos

Maq. ENFOQUES-2 20/5/03 15:17 Página 6

El procedimiento de análisis utilizado ofrece una visión estática, casi una instantánea de las sedes. Y, más que esta imagen sincrónica, que se limita al periodo de toma de datos, se requiere un seguimiento de la evolución de las sedes

las sedes no son las más importantes. Mayor relevancia tiene la estructuración de las sedes y la distribución de los diversos tipos de documentos. Aunque las sedes que contienen documentos en formato de intercambio (.pdf) apenas rebasan el 10%, la información que contienen no podrá recuperarse a menos que un SIDI disponga de mecanismos de indización de este tipo de archivos. Por otra parte, existe una fuerte correlación positiva entre el número de páginas .html de una sede y su estructuración jerárquica. Sin embargo, esta correlación disminuye mucho cuando las páginas son .asp o resultado de una consulta .cgi. Y, si bien es cierto que las páginas dinámicas así generadas representan una pequeña parte en la muestra analizada, también lo es que su naturaleza dinámica y "volátil" impide que los SIDI de recopilación automática tradicionales capturen, indicen y posibiliten la recuperación adecuada de su contenido. En todo caso, el hecho de que determinadas sedes aporten la mayor parte del contenido (medido en número de páginas o en volumen de textos) podría suponer un desequilibrio en la cobertura de los SIDI de recopilación automática.

Así mismo, la distribución de enlaces tiene una incidencia muy directa sobre la accesibilidad de la información contenida en las sedes a partir de los SIDI. De hecho, la mayoría de los modernos sistemas recurren a programas (robots) de recopilación automática que emplean los enlaces para desplazarse de sede en sede. Se ha evidenciado que la existencia de enlaces se justifica sobre todo por las necesidades de estructuración de las sedes. De hecho, es significativa ($p < 0,01$) la correlación entre el número de enlaces internos de las sedes y el correspondiente número de niveles jerárquicos. En general, la media de enlaces por página baja desde 12,77 enlaces internos (a otras páginas de la misma sede) hasta 0,5 enlaces externos (a otras sedes).

Sería aventurado esbozar un retrato robot de las sedes Web bajo dominio .es. Se ha demostrado que la naturaleza de las distribuciones lo desaconseja o lo imposibilita. En todo caso, sería posible definir tramos y adscribir una sede Web española al grupo dominante, al intermedio o a un tercer grupo residual. A la vista de los resultados globales, se puede decir que un alto número de páginas y elementos, con gran nivel de estructuración jerárquica o, alternativamente, el empleo significativo de procedimientos de generación dinámica de páginas (.asp, .php, .dtl u otros), formatos avanzados de representación (xml, css) y un gran número de enlaces a otras sedes caracterizan al grupo dominante de sedes. En la medida en que las cifras descienden, se halla un grupo intermedio cuya característica distintiva es la baja interactividad: pocos indicadores de relación (enlaces externos) y pocas llamadas de función (correo). Desde el punto de vista del contenido informativo, se dirían sedes "autosuficientes", poco colaboradoras o aisladas. El grupo restante de sedes se muestran en estado incipiente de desarrollo, apenas sobrepasado el nivel cero de las meras reservas de dominio.

Precisamente, el procedimiento de análisis utilizado ofrece una visión estática, casi una instantánea de las sedes. Y, más que esta imagen sincrónica, que se limita al periodo de toma de datos, se requiere un seguimiento de la evolución de las sedes. El ritmo de crecimiento del dominio .es no es muy acelerado (7.000 nuevos dominios registrados en 2001)[24] y el establecimiento de un observatorio permitiría un estudio evolutivo de las sedes en sus aspectos cuantitativos y también tecnológicos, relativos al tipo de aplicaciones empleadas para su mantenimiento y para el soporte de los contenidos que distribuyen. Por otra parte, se ha demostrado que el dominio .es sólo abarca una tercera parte del espacio Web correspondiente a España[25]. Nuevos procedimientos de selección de sedes podrían arrojar conclusiones mejor fundadas sobre el mismo.

No obstante todo lo anterior, el diseño, la implantación y la evaluación de SIDI ha de tener en cuenta los resultados de éste y otros trabajos de caracterización del espacio Web español, un espacio al que se deben ajustar al máximo sus características operativas, además de su cobertura[26].

Maq. ENFOQUES-2 20/5/03 15:17 Página 7

5.- Referencias

[1] Woodruff, A.; Aoki, P.M.; Brewer, E.; Gauthier, P.; Rowe, L.A.: An Investigation of Documents from the World Wide Web. *Computer Networks and ISDN Systems, 28*(7-11): 963-980, 1996.

[2] Monk, T.; Claffy, K.C.: A Survey of Internet Statistics/Metrics Activities. Technical Report, National Laboratory for Applied Network Research, June 1996. Accesible en http://www.caida.org/outreach/papers/1996/metricsurvey/metricsurvey.html [11 de abril de 2002]

[3] Claffy, K.C.: Measuring the Internet. *IEEE Internet Computing, 4*(1): 73-75, 2000.

[4] Broder, A. et al: Graph Structure in the Web. *Computer Networks, 33*(1-6) June 2000.

[5] Adamic, L.A.; Huberman, B.A.: The Web's Hidden Order. *Communications of the ACM, 44*(9): 55-59, September 2001.

[6] Huberman, B.A.; Pirolli, P.L.T; Pitkow, J.E.; Lukose, R.M.: Strong Regularities in World Wide Web Surfing. *Science, 280*(5360): 95-97, 3 de abril de 1998.

[7] Silverstein, C.; Henzinger, M.; Marais, H.; Moricz, M.: Analysis of a Very Large Altavista Query Log. SRC Technical Note 1998-14. Palo Alto, Systems Research Center, Digital Equipment Corporation, 1998.

[8] Zien, J.; Meyer, J.; Tomlin, J.; Liu, J.: Web Query Characteristics and their Implications on Search Engines. Tenth International World Wide Web Conference. May, 1-5, 2001, Hong Kong. Accesible en http://www.www10.org/cdrom/posters/1077.pdf [16 de Abril de 2002]

[9] Jansen, B.J.; Spink, A.; Saracevic, T.: Real Life, Real Users and Real Needs: a Study and Analysis of User Queries on the Web. *Information Processing and Management, 36*(2): 207-227, 2000.

[10] Térmens Graells, M.; Ribera Turró, Mireia; Sulé Duesa, A.: La universidad española en la web. Revista Española de Documentación Científica (en prensa).

[11] Bray, T.: Measuring the Web. Fifth International World Wide Web Conference. May, 6-10, 1996, Paris. Accesible en http://www5conf.inria.fr/fich_html/papers/P9/Overview.html [3 de abril de 2002].

[12] Lawrence, S.; Giles C.L.: Searching the World Wide Web. *Science, 280* (5630): 98-100, April 1998.

[13] Lawrence, S.; Giles C.L.: Accesibility of Information on the Web. *Nature, 400* (6740): 107-109, July 1999.

[14] Bharat, K.; Broder, A.: A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines. *Computer Networks and ISDN Systems, 30*: 379-388, 1998.

[15] Henzinger, M.R.; Heydon, A.; Mizenmacher, M.; Najork, M.: Measuring Index Quality Using Random Walks on the Web. *Computer Networks, 31*: 1291-1303, 1999.

[16] O'Neill, E.T.; McClain, P.D.; Lavoie, B.F.: A Methodology for Samplig the World Wide Web. *Annual Review of OCLC Research* 1997. Accesible en http://www.odc.org/research/publications/arr/1997/oneill/o%27neillar980213.htm [2 de mayo 2002]

[17] Eckmann, J.P.; Moses, E.: Curvature of Co-Links Uncovers Hidden Thematic Layers in the World Wide Web. *Proceedings of the National Academy of Sciences USA*, 99(9): 5825-5829, 2002.

[18] García Santiago, M.D.: Topología de la información en la World Wide Web: Modelo metodológico de visualización en una red hipertextual nacional. Tesis doctoral, Univ. de Granada, 2000.

Maq. ENFOQUES-2 20/5/03 15:17 Página 8

[19] Glover, E.J.; Tsioutsiouliklis, K.; Ñawrence, S.; Pennock, D.M.; Flake, G.W.: Using Web Structure for Classifiying and Describing Web Pages. Eleven International World Wide Web Conference. May, 7-11, 2002, Honolulu.

[20] Ford, G.: Theory and Practice in the Networked Environment: A European Perspective, en McClure C.R. y Carlo Bertot, J. (Edts): *Evaluating Networked Information Services. Techniques, Policy and Issues*. Medford, Information Today, 2001, Pags 1-22

[21] Lavoie, B.; Frystyk Nielsen, H. (edts): *Web Characterization Terminology & Definitions Sheet*. World Wide Web Consortium, 24 May 1999. Accesible en http://www.w3.org/1999/05/WCA-terms/ [3 de mayo de 2002].

[22] Web Characterization Project: *Web Stes: Concepts, Issues and Definitions (First Draft)*. OCLC Research, February 18 2000. Accesible en http://wcp.oclc.org/pubs/rn1-websites.html [6 de mayo de 2002].

[23] Hausherr, Tilman: Xenu's Link Sleuth v1.1c, 2001. Accesible en http://home.snafu.de/tilman/xenulink.html.

[24] *Estadísticas del ES-NIC. Dominios registrados en los últimos años*. Accesible en http://www.nic.es/estadisticas/index.html [13 de junio de 2002].

[25] Baeza Yates, R.: The Web of Spain. *UPGrade*, 3(3): 82-84, 2002.

[26] R Fornás (buscopio.net) y A. de la Cueva (Univ. de Valencia) aportaron acertadas críticas a este manuscrito. I. Aguilló (CINDOC-CSIC) proporcionó interesantes materiales cuantitativos. V. Castelo y D.R. López (RedIRIS-CSIC) facilitaron la extracción de la muestra aleatoria del registro.

Carlos Benito Amat
(carbea@iata.csic.es)
Unidad de Biblioteca y Documentación Científica
Instituto de Agroquímica y Tecnología de Alimentos
CSIC