Departamento de Biblioteconomía y Documentación
Universidad de Salamanca

Métodos y herramientas para la extracción de datos en Cibermetría. El software académico y comercial

Trabajo de grado

Natalia Arroyo Vázquez
Director: José A. Frías Montoya

Salamanca, 2004

El presente trabajo ha sido posible gracias a la financiación del proyecto del V programa marco WISER (INRP-2001-10002), en el marco del cual ha surgido, al apoyo y sugerencias de sus participantes, y muy especialmente a los integrantes del grupo de trabajo InternetLab (CINDOC-CSIC), mis compañeros, y su director, Isidro Aguillo, cuyos esfuerzos se ven aquí proyectados.

Índice

Introducción

El sistema de publicación de la ciencia moderna, basado en la revisión por pares, es el medio comúnmente aceptado por los científicos para dar a conocer sus hallazgos, siendo a su vez, según Maltrás (2003), uno de sus rasgos más característicos, y como tal ha ocupado un lugar privilegiado entre los estudios dedicados a la comunicación científica, al tratarse además de un canal formal algunos de cuyos aspectos resultan fáciles de cuantificar gracias a las bases de datos disponibles desde mediados del siglo pasado. De esta manera se han desarrollado una serie de indicadores de la producción científica que juegan un papel importante en la toma de decisiones por parte de los organismos que financian el funcionamiento de dicho sistema. Sin embargo, y aunque se emplea comúnmente la expresión "comunicación científica" para referirse a este proceso, existen otros medios para que los científicos puedan transmitir su conocimiento, llamados informales y que resultan más difíciles de medir por su falta de publicidad y las trabas a su acceso que ello conlleva.

La aparición de Internet y su posterior popularización ha supuesto un gran cambio en las relaciones entre los investigadores: el correo electrónico supone una forma mucho más rápida y barata de comunicación que la tradicional correspondencia, las listas de correo especializadas constituyen actualmente una importante forma de difusión e intercambio de noticias, eventos, nuevas publicaciones y conocimientos, el FTP se emplea, aunque cada vez con menor frecuencia, para el intercambio de ficheros, mientras que el World Wide Web es también empleado como medio para la publicación formal en revistas científicas electrónicas, reguladas, al igual que las tradicionales revistas en papel, por el sistema de revisión por pares, e informal por la presencia de contenidos científicos en las páginas web de algunos investigadores, grupos de investigación, etc., o incluso por la publicación de pre-prints, comunicaciones, artículos o cualquier otro tipo de documento con este tipo de información. Todas estas novedades con respecto de los medios habituales conllevan la aparición de nuevos frentes de investigación en los estudios sobre la ciencia y abren nuevos caminos para un mayor conocimiento y difusión de ésta.

Desde la popularización del Web a mediados de los noventa la presencia de diferentes entidades ha ido creciendo en los países más desarrollados hasta el punto de que en la actualidad existen dos mundos casi paralelos, siendo uno de ellos el Web y la vida real el otro. De esta manera, es también posible estudiar la presencia en el WWW de las instituciones dedicadas a la labor científica, las relaciones entre ellas a través de los hipervínculos que enlazan a unas con otras, los contenidos que alojan mediante el análisis de palabras clave, etc.

Sin embargo, todos estos estudios no están exentos de problemas metodológicos. Por una parte, resulta difícil definir una unidad de análisis básica, y, por otra, la ausencia de unas bases de datos que mantengan este tipo de información hace necesario buscar herramientas que puedan extraer los datos requeridos, y éstas se encuentran en ocasiones con una serie de limitaciones por su propia naturaleza o por la del mismo Web. Los motores de búsqueda han sido el método al que más se ha recurrido para la obtención de datos cuantitativos de este entorno, aunque no siempre son los más adecuados, especialmente cuando se trabaja a nivel *micro* por su falta de finura, por lo que además se han venido empleando programas comerciales y desarrollando programas académicos con tales fines.

Sobre el comportamiento y adecuación a unos casos u otros de este tipo de programas poco se conoce, por lo que se ha considerado necesario realizar un estudio comparativo con el fin de estudiar y evaluar varios de ellos, su forma de trabajar y de actuar ante determinadas situaciones, así como los resultados que ofrecen, para su posible uso en este sentido.

Para ello se comenzará introduciendo en un primer capítulo el área en el que se enmarca este tipo de estudios, la Cibermetría, partiendo de algunas nociones básicas y una revisión bibliográfica en la que se identificarán los principales frentes de trabajo, para pasar a presentar, en el capítulo 2, el software seleccionado para su estudio, haciendo hincapié en su funcionamiento, sus principales características, las opciones que el usuario puede seleccionar, y los resultados que de ellos se pueden esperar, así como sus puntos fuertes y débiles. En el tercer capítulo se presentarán los resultados de un trabajo empírico realizado con el mismo objetivo de estudiar el comportamiento e información que de ellos se pueden esperar. Y ya para terminar, se realizará una evaluación del software y se extraerán las conclusiones pertinentes al respecto.

Capítulo 1. Marco teórico

En el presente capítulo se intentará introducir brevemente la disciplina en la que se enmarca el estudio realizado. Para ello se comenzará con un rápido recorrido por los principios más básicos de la Bibliometría, Cienciometría e Informetría, áreas de las que surge la Cibermetría y a las que se encuentra muy ligada en la actualidad por lo reciente de su aparición, para después explicar los orígenes de esta última, sus principios básicos y los frentes abiertos hasta el momento. A pesar de la juventud de la disciplina, ya se han publicado varios trabajos de revisión que han servido como apoyo (Bar-Ilan, 2001; Bar-Ilan y Peritz, 2002).

1.1. Antecedentes: Bibliometría, Cienciometría, Informetría

La publicación en 1963 del trabajo de Price *Little Science, Big Science* marca el inicio de los estudios sobre la "ciencia de la ciencia", o Cienciometría, término acuñado por Nalimov y Mulchsenko para referirse a la aplicación de métodos cuantitativos a la investigación sobre el desarrollo de la ciencia considerada como un proceso de información. Fue en la década de los 60 cuando se produjo también un auge de los estudios bibliométricos, propiciado por la informatización de las bases de datos y una mayor demanda por parte de los responsables en política científica para evaluar los resultados generados (Sancho, 1990), y cuyo punto de partida puede situarse en la acuñación del término Bibliometría por parte de Pritchard (1969) para referirse a una ciencia que llevaba existiendo ya durante medio siglo bajo el nombre de bibliografía estadística. La relación entre ambas —Cienciometría y Bibliometría— e Informetría es tan estrecha que en ocasiones se emplea el término Bibliometría para referirse genéricamente a todas ellas.

El ámbito de cada una de ellas lo define Tague-Sutcliffe (1992) de la siguiente manera:

Bibliometría

> *…is the study of the quantitative aspects of the production, dissemination and use of recorded information.*

Informetría

> *…is the study of the quantitative aspects of information in any form, not just records, or bibliographies, and in any social group, not just scientists.*

Cienciometría

> *…is the study of the quantitative aspects of science as a discipline of economic activity.*

El descubrimiento de una serie de leyes métricas por las que se rigen los procesos de información y comunicación científicas vino a formar el corpus teórico de estas disciplinas. Así, fueron aplicadas las leyes formuladas por Lotka, Bradford y Zipf entre otras, como leyes básicas de la cienciometría.

- Según la ley de Lotka (1926) o ley cuadrática inversa, difundida por Price, el número de autores que produce un número determinado *n* de artículos es inversamente proporcional al cuadrado de *n*.
- La ley de Bradford (1934) o ley de la dispersión de la literatura científica se refiere a la distribución de la literatura científica como algo muy desigual, ya que la mayor parte de los artículos publicados se concentra en un pequeño número de revistas (núcleo), mientras que un pequeño número de artículos son publicados en una gran cantidad de revistas.
- La ley de Zipf (1949) o de la distribución de las frecuencias de la utilización de palabras en un texto considera que al ordenar las palabras de un texto en orden decreciente a su frecuencia de aparición en él, el producto de la multiplicación de las frecuencias de observación de las palabras de los textos por el valor numérico del rango que ocupan dichas palabras en una distribución de frecuencias de observación es constante.

Por otra parte, fue la aparición de las bases de datos de publicaciones científicas del ISI (*Institute for Scientific Information*) de Filadelfia —fundado por Garfield—, el *Science Citation Index* (SCI) en 1963 lo que aportó las bases fundamentales para llevar a la práctica todas estas teorías y desarrollar las técnicas que hicieran posible la evaluación de la ciencia y los científicos. Surgen así una serie de indicadores de la producción científica que Okubo (1997) divide en dos grupos: indicadores de la actividad científica y tecnológica, basados principalmente en recuentos (número de publicaciones, de citas, coautores, patentes, o citas a patentes), e indicadores relacionales (co-publicaciones, índice de afinidad, enlaces científicos medidos a través de citas, cocitas, copalabras, etc.).

Pero sin duda alguna, el indicador que más interés ha generado es el factor de impacto de Garfield (1976), que mide el interés que suscitan los trabajos publicados en una determinada publicación científica a través de sus citas, y se calcula relacionando el número de citas recibidas en un determinado año por los trabajos publicados en una revista durante los dos años anteriores con el total de artículos publicados en ella durante esos dos años anteriores. El factor de impacto de cada revista se publica anualmente en el *Journal Citation Reports* (JCR) del ISI.

Las relaciones entre las citas bibliográficas son medidas mediante el estudio de cocitas y las relaciones bibliográficas (*bibliographic coupling*). Las cocitas (Small, 1973) miden el número de veces en que dos artículos son citados simultáneamente en un mismo trabajo. A mayor número de citas conjuntas, más cerca se encontrarán dos artículos en un mapa de cocitas, que es el resultado visual de este tipo de análisis, generado mediante la creación de *clusters* de documentos relacionados, y a través de los cuales es posible identificar áreas y sub-áreas de la ciencia y su evolución en el tiempo. Por otra parte, se puede decir que dos publicaciones están bibliográficamente relacionadas (*bibliographic coupling*) cuando poseen una a más referencias comunes, perteneciendo así a un mismo campo temático, lo que permite también la creación de mapas de la ciencia (Kessler, 1963).

La introducción del análisis de co-ocurrencias de palabras, procedente de las teorías de las redes de actores (*Actor Network Theory*), supuso una novedad en los años 80 (Callon et al., 1983). La idea se basa en la presunción de que si dos palabras son empleadas conjuntamente en un mismo trabajo guardarán una cierta relación temática. Por lo tanto, también es posible crear mapas conceptuales de la ciencia a partir de *clusters* definidos según el índice de co-ocurrencia de determinadas palabras, consideradas clave.

1.2. Cibermetría

Los orígenes de la Cibermetría pueden situarse, según Aguillo (2000a), a mediados de los noventa. Marcada por un fuerte carácter descriptivo, sus principales objetivos consistían en principio en el estudio de aspectos como la evolución del tamaño del Web y la descripción de los primeros motores de búsqueda. Es el reconocimiento de las dos principales características del Web, su cobertura global y su naturaleza hipertextual, lo que propicia la publicación de una serie de trabajos en los que se aplican los principios de la Bibliometría, Cienciometría e Informetría al estudio del Web, así como la posterior aparición de la revista electrónica *Cybermetrics* en 1997, presentada en el primero de una serie de seminarios sobre la diseminación de resultados del análisis cuantitativo de Internet (Aguillo, 1997), que se celebran con carácter bianual en el marco de las conferencias de la *International Society for Scientometrics and Informetrics* (ISSI).

En un principio fueron propuestos varios términos para designar a la nueva disciplina, como señala Björneborn (2004): *netometrics* (Bossy, 1995), *webometry* (Abraham, 1998), *internetometrics* (Almind e Ingwersen, 1996), *web bibliometry* (Chakrabarti et al., 2002), aunque finalmente se adoptaron dos: *webometrics*, propuesto por Almind e Ingwesen (1997) y *cybermetrics*, que toma su

nombre de la revista electrónica. Para su traducción al español fueron adaptados literalmente del inglés ambos términos, pasando así a designarse en un principio estas disciplinas Cibermetría y Webometría; sin embargo, dada la malsonancia de este último ha sido sustituido por Webmetría.

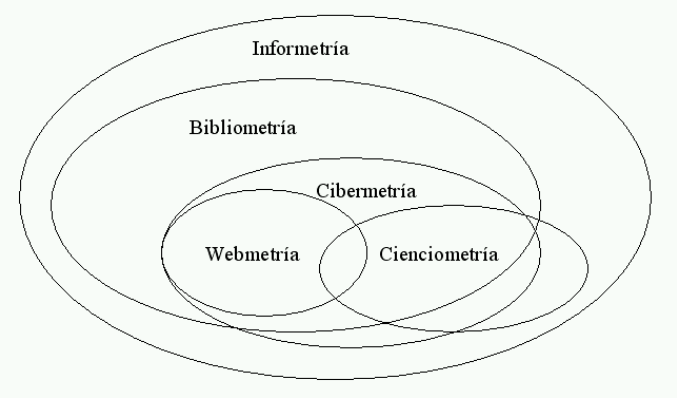


Figura 1.1. Relación entre Bibliometría, Cienciometría, Informetría, Webmetría y Cibermetría. Adaptado de Björneborn (2002).

Aunque en la práctica son casi empleados como sinónimos, existe un matiz que las distingue: su ámbito de actuación, tal y como queda reflejado en las definiciones que Björneborn (2004) propone. Según él, la Cibermetría puede ser entendida como

> *…the study of the quantitative aspects of the construction and use of information resources, structures and technologies on whole Internet, drawing on bibliometric and informetric approaches.*

Y la Webmetría como

> *…the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web, drawing on bibliometric and informetric approaches.*

De ambas —casi idénticas— cabe subrayar los siguientes elementos:

- Se refieren sólo al estudio de aspectos cuantitativos.
- Se centran en la construcción y uso de recursos de información, estructuras y tecnologías.
- Se basan en los enfoques de Bibliometría e Informetría.

Su relación con estas ciencias, a las que quizás se podría añadir la Cienciometría, es de solapamiento en algunas áreas, ya que comparten un mismo enfoque y un mismo ámbito, enmarcado

por la Informetría, y que es por lo tanto el de los estudios cuantitativos sobre la información (figura 1.1).

1.2.1. Leyes clásicas

El origen de la aplicación de las técnicas bibliométricas al Web está en la idea de que un hipervínculo apuntando a una página web determinada es equivalente a una cita bibliográfica. Así, desde esta perspectiva y por analogía con las citas bibliográficas comenzó a emplearse el término *sitations* (McKiernan, 1996; Aguillo, 1997; Rousseau, 1997), o "sitas" en español, para designar a los enlaces. Otros trabajos que iniciaron esta línea y sentaron algunas bases, de los que se irá hablando más adelante, son los de Bray (1996), Larson (1996), Abraham (1996), Abraham (1998), o Almind e Ingwersen (1997).

A partir de aquí, se trabajó sobre la siguiente pregunta, formulada por autores como Boudourides, Sigrist y Alevizos (1999) y Egghe (2000): ¿son las leyes clásicas de la Informetría válidas en Internet?. Si bien los primeros sólo plantearon esta pregunta, Egghe señala que sería posible que dichas leyes se dieran en el Web si se cumplen dos condiciones: que haya una fuente (revistas, autores, artículos…) y un item a medir (artículos, citas o referencias…), como sucede en la Cienciometría, lo cual no siempre es evidente —las páginas web no siempre tienen un autor explícito ni están publicadas en una revista—, pero sí en el caso de las revistas electrónicas, o cuando existe una página web que actúa como fuente y al menos un hipervínculo en ella, lo que sería un item, tal y como demostró Rousseau (1997).

En ese mismo trabajo se comprueba que la Ley de Lotka se da en la distribución de sitas a un sitio web y de dominios de primer nivel por sitio web. Incidiendo más tarde en ello, Rousseau y Rousseau (2000) diseñaron un programa para facilitar el cálculo del grado de ajuste de un conjunto de datos cualesquiera a este tipo de distribución. Esta misma ley fue observada por Boudourides y Antypas (2002) en la distribución del número de páginas por sitios web.

Sobre otras leyes clásicas de la Cienciometría incidieron Bar-Ilan (1997), quien aplica la Ley de Bradford a los grupos de noticias en un estudio sobre la enfermedad de las vacas locas empleando como método para la recogida de datos el motor de búsqueda AltaVista, llegando a la conclusión de que la distribución de Bradford se puede aplicar a dicho medio e identificar núcleos; o Faba, Gerrero y Moya (2003) en un estudio sobre el ajuste de los datos de sitas a esta misma

distribución, entre otros. También otras leyes exponenciales han sido aplicadas, entre otros, a enlaces entrantes y salientes (Albert, Jeong y Barabasi, 1999), número de páginas por sitio (Adamic y Huberman, 2000). Además, la Ley de Zipf ha sido extrapolada a la distribución de enlaces entrantes (Broder et al., 2000), la red que forman e-mails enviados y recibidos, la conectividad de los enrutadores en Internet (Adamic y Huberman, 2002), o ya más recientemente a los correos enviados a los grupos de noticias (Kot, Silverman, y Berg, 2003).

En lo que a las leyes de crecimiento exponencial propuestas por Price (1963) sobre la ciencia se refiere, se cumplen, tal y como Egghe identificó (2000), también en aspectos de la Red como el número de servidores (Zakon, 2003) o el número de hosts (Internet Software Consortium, 2000).

1.2.2. Otras técnicas

Otras técnicas procedentes de la Cienciometría que se han aplicado al estudio del Web son el análisis de citas, cocitas y de co-ocurrencia de palabras. El primero de ellos, el análisis de enlaces web —o sitas— como si de citas bibliográficas se tratara ha encontrado oposición entre algunos autores (Egghe, 2000; Van Raan, 2001), que consideran que la motivación es distinta y no tienen parangón, mientras que otros opinan que se ha encontrado un equivalente a las bases de datos tradicionales (Rodríguez i Gairín, 1997; Almind e Ingwersen, 1997).

El análisis de co-citas, basado en la premisa de que a mayor número de veces en que dos documentos son citados conjuntamente, mayor posibilidad existe de que su contenido esté relacionado, fue primeramente adaptado al Web por Larson (1996) con la creación de un mapa de co-citación en el área de ciencias de la Tierra mediante la técnica de escalamiento multidimensional (MDS). Posteriormente otros autores lo han empleado en este ámbito con diferentes fines: Dean y Hanzinger (1999) para la búsqueda de páginas relacionadas en el Web; Boudourides, Sigrist y Alevizos (1999) para el estudio de los las relaciones participantes en el proyecto SOEIS; Kumar et al. (1999), en la identificación de comunidades también en el Web; y ya más recientemente por Thelwall y Wilkinson (2004) para identificar sitios web similares.

Y por último, el análisis de co-ocurrencia de palabras, que parte de la premisa de que los términos que aparecen en un texto representan el contenido del mismo para así medir la ocurrencia conjunta de los términos que aparecen simultáneamente en varios documentos con el fin de generar

redes conceptuales de los diferentes campos, han sido aplicados al campo de la Cibermetría en estudios como los de Ross y Wolfram (2000), quienes lo emplearon para analizar pares de términos enviados a un motor de búsqueda, Excite; o Leydesdorff y Curran (2000), cuyo objetivo era conocer la relación entre los componentes de la "triple hélice" mediante la co-ocurrencia de los términos "*university*", "*industry*" y "*government*" en páginas de diferentes dominios.

Sobre las técnicas para la identificación de *clusters* ha aparecido recientemente una serie de trabajos en los que se proponen diversos métodos para la agrupación de páginas web, sitios web o palabras clave (Menczer, 2004; Cothey, Aguillo y Arroyo, 2004).

1.2.3. Indicadores web

Otro de los frentes abiertos en el campo de la Cibermetría es la creación de indicadores. De ellos el que más líneas ha ocupado, quizás por su indudable importancia para la Bibliometría, es el factor de impacto web (*Web Impact Factor* en la literatura en anglosajona). Aunque el concepto de impacto de la información en el Web fue en un principio introducido por Rodríguez i Gairín (1997), fue Ingwersen (1998) quien finalmente lo definió como

> *…the logical sum of the number of external- and self-link pages pointing to a given country or web site divided by the number of pages found in that country or web site —at a given point in time. The numerator thus consists of the number of link pages —not the number of links.*

El mismo Ingwersen habla además de un factor de impacto web *externo*, que calcula extrayendo las auto-sitas del recuento total, evitando así uno de los problemas del factor de impacto. Varios autores, como Egghe (2000), Li (2003), Smith (1999), Smith y Thelwall (2002), Thelwall (2002), Vaughan y Hysen (2002) entre otros, han continuado esta línea con diversas aportaciones de indudable interés pero cuya extensión sobrepasaría los límites de este apartado. Merece la pena sin embargo mencionar la correlación hallada entre los índices del *Research Assessment Exercise* (RAE) y los factores de impacto web de varios departamentos universitarios en el Reino Unido por Li et al. (2003).

La relevancia de los indicadores web fue puesta de manifiesto con el ya finalizado proyecto del V Programa Marco de Investigación y Desarrollo de la Comisión Europea EICSTES[1] (*European*

[1] http://www.eicstes.org/

Indicators, Cyberspace and the Science-Technology-Economy System), en el que fueron empleados con éxito en el análisis del sistema europeo de ciencia, tecnología y economía en Internet para establecer las relaciones entre el sector de I+D y los actores de la nueva economía, y parece consolidarse con WISER[2] (*Web Indicators forScience, Technology and Innovation Research*), también parte del V Programa Marco, que tiene por objetivo explorar las posibilidades y problemas en el desarrollo de una nueva generación de indicadores de ciencia y tecnología basados en el Web. En él se da la siguiente definición de indicador web:

> *…a policy relevant measure that quantifies aspects of the creation, dissemination and application of science and technology in so far as they are represented on the internet or the World Wide Web.*

Otros de los indicadores más característicos de la disciplina, definidos en el entorno de EICSTES, se resumen en la tabla 1.1.

Indicador	*Definición*
Profundidad	Numero de niveles de la estructura de un sitio web, situando la raíz en el nivel uno
Densidad	Número total de enlaces por página, incluyendo enlaces salientes internos y externos y enlaces dentro de una misma página
Conectividad	Número de enlaces diferentes en un sitio web, incluyendo enlaces salientes internos y externos, pero no los enlaces dentro de una misma página
Navegabilidad	Densidad de enlaces salientes internos en un sitio web, teniendo también en cuenta los enlaces repetidos
Endogamia	Porcentaje de enlaces salientes internos diferentes con respecto al número total de enlaces salientes diferentes
Luminosidad	Numero de enlaces salientes externos (enlaces desde un sitio web a otros diferentes)
Dispersión	Tipología y frecuencia de los enlaces salientes en un sitio web de acuerdo a diferentes criterios de distribución
Visibilidad	Numero de enlaces externos recibidos por un sitio web
Popularidad	Numero de visitas diferentes recibidas por un sitio web
Diversidad	Tipología y frecuencia de enlaces recibidos por un sitio web según distintos criterios de distribución

Tabla 1.1. Indicadores Web. Fuente: Proyecto EICSTES. Deliverable 8.1.

1.2.4. Herramientas para la extracción de datos

Hasta el momento las herramientas más empleadas por los cibermetras en la extracción de datos del Web han sido los motores de búsqueda. Numerosos estudios han venido investigando sus

[2] http://www.wiserweb.org/ y http://www.webindicators.org/

características —tamaño y cobertura (Bharat y Broder, 1998; Lawrence y Giles, 1998; Lawrence y Giles, 1999; Notess, 2003; Vaughan y Thelwall, 2004), estabilidad (Bar-Ilan, 1998), solapamiento (Larsen, 2002), entre otros— y siguiendo su evolución en el tiempo.

En un principio el ya desaparecido AltaVista constituyó el centro de atención por sus ventajas competitivas con respecto de los demás (recordemos el ya citado artículo de Rodríguez i Gairín que lo equipara al *Citation Index*), pero desde su aparición ha sido Google quien ha atraído la mayor parte de las miradas en un principio por la novedad de su PageRank (Brin y Page, 1998), que se fundamenta en la idea, muy afín a los postulados bibliométricos, de que el número de enlaces que apuntan a una página determinará su posición —de ahí el nombre de "posicionamiento web" que recibe el área dedicada al estudio de la visibilidad de los sitios web en los distintos motores de búsqueda, que en este momento son cinco: Google, Yahoo!, Microsoft Search, Teoma y Wisenut — en los resultados de una búsqueda, ocupando los primeros puestos si el número es mayor, pero también por la carrera que hasta el momento está ganando con respecto a otras herramientas en lo que a cobertura se refiere —actualmente tiene indexadas 4,285,199,774[3] de páginas web—, así como por toda la cultura que se está creando alrededor de él y los servicios que va añadiendo a su oferta[4].

Sin embargo, las limitaciones de los motores de búsqueda son muchas:

- No cubren todo el Web, ni siquiera la parte que es técnicamente indexable.
- Los algoritmos que emplean están protegidos por el secreto comercial, por lo que su funcionamiento es un misterio.
- La validez de los datos que arrojan es cuestionable, dadas las grandes diferencias encontradas incluso en datos recogidos en intervalos de tiempo mínimos (Bar-Ilan, 1998; Rousseau, 1999; Snyder y Rosenbaum, 1999), aunque en este sentido AltaVista ha mejorado notablemente (Thelwall, 2001a).
- Existen sesgos nacionales en la cobertura de los sitios, tal y como se ha demostrado recientemente (Vaughan y Thelwall, 2004).

Todas estas limitaciones y la búsqueda de métodos alternativos han llevado a algunos grupos de investigación a desarrollar sus propias herramientas, que permiten, como se demostrará más adelante, un mayor control sobre los resultados. Es el caso de los *crawlers* SocSciBot, diseñado por el

[3] Datos recogidos en junio de 2004.
[4] Actualmente se encuentra en etapa de prueba su servicio gratuito de e-mail, con una capacidad que supera a todos sus competidores.

Statistical Cybermetrics Research Group[5] de la Universidad de Wolverhampton (Thelwall, 2001b), y Webcount, uno de los resultados del proyecto EICSTES (Adams y Gilbert, 2003).

Otra alternativa, propuesta por Aguillo (1998b; 2000b), es el uso de herramientas de segunda generación, en concreto programas comerciales de bajo coste (*shareware*) como *mapeadores* o verificadores de enlaces, que han sido empleados para la extracción de datos de sedes web en varios proyectos de investigación. La diferencia de estas dos últimas soluciones reside en la perspectiva o nivel del análisis: si bien los motores de búsqueda son capaces de extraer información sobre dominios o sub-dominios, la presencia de ciertas palabras en el espacio web, o patrones lingüísticos (nivel *macroscópico*), los *crawlers* académicos y comerciales se basan en unidades de análisis menores, como sitios o *sedes web* (nivel *microscópico*), lo que hace más costoso su trabajo para grandes volúmenes de información.

Aunque el Web es el servicio más estudiado dentro de la Cibermetría, es de justicia mencionar también otros trabajos realizados sobre las redes de comunicación en otros ámbitos, como los grupos de noticias —recordemos de nuevo los trabajos de Bar-Ilan (1997) y Faba, Guerrero y Moya (2003) — o el e-mail (Drineas et al., 2004; Zelman y Leydesdorff, 2000).

1.2.5. El enfoque de otras disciplinas

Además de los enfoques próximos a la Informetría de los que se ha hablado hasta ahora, otras aportaciones desde diversos campos como la informática, las física, las matemáticas o las ciencias sociales han sido realizadas. Quizás la que más implicaciones ha supuesto, más bien por su valiosa aportación a los algoritmos de *crawleado*, búsqueda e identificación de comunidades, y como fenómeno sociológico, en palabras de Broder et al. (2000), es la aplicación de la teoría de grafos al análisis de la estructura del Web —aunque también a otros procesos de comunicación en Internet—, pudiendo ser este considerado un grafo dirigido en el que las páginas web son nodos o vértices y los hipervínculos que las relacionan los arcos o aristas. Uno de los primeros autores en explotar esta perspectiva fue Abraham (1996; Abraham y Foresta, 1996), cuyas estrategias de medida y visualización del Web está basada en la teoría de sistemas dinámicos complejos.

Sobre la estructura del WWW han sido extraídas interesantes conclusiones acerca de sus características: Broder et al. (2000) lo dibujan como una especie de lazo (*bow-tie*) formado por un

[5] http://cybermetrics.wlv.ac.uk/socscibot/

núcleo de páginas altamente interconectadas, unos lazos —con enlaces dirigidos al núcleo uno y que salen de él en el otro—, un delgado tubo que conecta ambos lazos, y algunos hilos sin procedencia definida. Recientemente una estructura similar, esta vez con forma de corona, fue hallada por Björneborn (2004) a partir de los datos del Web académico del Reino Unido (figura 1.2).

El diámetro web, entendido como el número mínimo de enlaces que hay que seguir para navegar desde un documento a otro, es, según Albert, Jeong y Barabasi (1999), de 19, un número sorprendentemente bajo que demuestra la alta conectividad que caracteriza a la Red. Kleinberg y Lawrence (2001) identificaron dos tipos de páginas web: *hubs* y *authorities*; los primeros son aquellos que apuntan a muchos *authorities*, y éstos, a su vez, los enlazados por muchos *hubs*. Los patrones de ambos, tal y como estos autores afirman, pueden ser empleados para determinar comunidades de páginas relacionadas por un mismo tema mediante técnicas de clusterización.

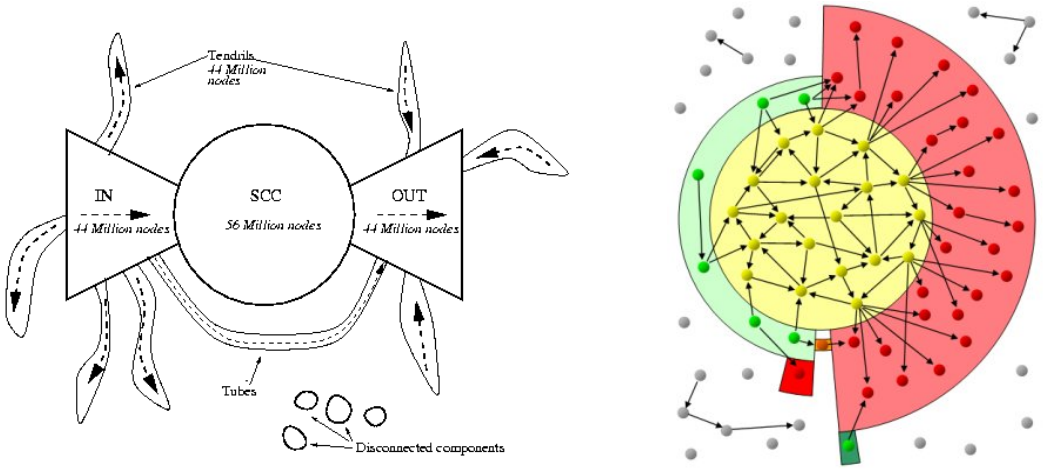


Figura 1.2. Representaciones del Web como grafo. Modelos *bow-tie* de Broder el al. y *corona* de Björneborn.

Fueron Watts y Strogatz (1998) quienes manifestaron que el Web cumple la característica de algunos grafos conocida como *small-world*, por analogía con el fenómeno así llamado (también conocido popularmente como “seis grados de separación”) que consiste en un estado intermedio entre la regularidad y la aleatoriedad en la distribución de los arcos de un grafo, lo que da lugar a un alto grado de agrupación o clusterización. Este trabajo abrió una línea de investigación continuada entre otros por Adamic (1999), Barabasi (2003) y Björneborn (2002; 2004).

Otras aproximaciones, algunas desde las teorías de redes, tales como el análisis de redes sociales (*Social Network Analysis*), —basada en que las relaciones o interacciones entre actores son los factores clave que sostienen y definen una red, como puede ser Internet (Wellman et al., 1996)—, Redes Neuronales Artificiales (RNA) —en la identificación de *clusters* y la creación de mapas (Kohonen, 1995; Polanco, François y Lamirel, 2001)—, teoría de redes complejas (Scharnhorst, 2004); y otras como las teorías del caos —el análisis caótico ha sido aplicado a las series temporales de datos obtenidos de Internet, tales como datos de tráfico de red o de datos de caché (Kugiumtzis y Boudourides, 1998)— y fractales —(Abraham, 1996; Abraham y Foresta, 1996; Egghe 1997), el último de los cuales hace un estudio de la dimensión fractal de los sistemas de hipertexto—, también han sido aplicadas al estudio del Web.

1.2.6. Problemática de la Cibermetría

Internet ha sido caracterizado en numerosas ocasiones como un medio dinámico y cambiante, rasgo que ha sido objeto de estudio en numerosas ocasiones, aunque ha sido Koehler quien más ha profundizado en este aspecto. Este investigador de la Universidad de Oklahoma establece dos tipos de cambios en los documentos web (Koehler, 1999b): persistencia —que define como la existencia o desaparición de páginas y sitios web— e intermitencia —variante de la persistencia definida como la desaparición y posterior reaparición de documentos Web, que considera se dan en un momento concreto sobre alrededor de un 5% de las páginas web. Posteriormente este autor ha venido desarrollando, en esta misma línea, otros indicadores del cambio en este ámbito, tales como consistencia y permanencia (Koehler, 1999a, 2002).

Las herramientas disponibles en la actualidad y la propia naturaleza de Internet hacen imposible determinar exactamente cuál es su tamaño. En lo que al tamaño físico se refiere existe información institucional muy precisa sobre la evolución de las estructuras en Internet, como el número de *hosts* (*Internet Systems Consortium, Inc.*[6], *RIPE Network Coordination Centre*[7]), o de dominios (*Country Registry Resources*[8]), pero sobre el número exacto de páginas sólo pueden realizarse estimaciones. Lawrence y Giles (1999) consideraron que en febrero de 1999 había unos 800 millones de páginas web distribuidos en un total de 3 millones de servidores, de los cuales tan sólo un 6% tenía contenidos científicos o educativos. En Julio de 2000 Moore y Murray (2000) hablan de un total de 2.200 millones de páginas web, con un incremento mensual de 7 millones de páginas. Hasta el

[6] http://www.isc.org/
[7] http://www.ripe.net/
[8] http://www.countrynics.com/

momento la única posibilidad es extraer datos sectoriales (por dominios o sub-dominios, lenguas, etc.) interrogando a los motores de búsqueda (Zakon, 2003).

Sin embargo, una de las limitaciones de estos, tal y como ya se señaló anteriormente, es su imposibilidad para cubrir todo el Web, dando así lugar a lo que se conoce como Web invisible —Internet invisible para referirse a todo el ciberespacio—, también llamada *deep web* (Bergman, 2001) o *dark matter* (Bailey, Craswell y Hawking, 1999), por no ser indexable para los motores de búsqueda, y cuyo tamaño ha sido estimado entre unas 400 y 550 veces mayor que el del Web visible, como si este último sólo fuera la punta de un iceberg (Bergman 2001). Sherman y Price (2001) la definen de la siguiente manera:

> *…text pages, files or other often high-quality authoritative information available via the World Wide Web that general-purpose search engines cannot, due to technical limitations, or will not, due to deliberate choice add to their indices of Web pages.*

Dichos autores distinguen cuatro tipos de invisibilidad:

- Web opaco, que estaría formado por aquello ficheros que podrían estar incluidos en los índices de los motores de búsqueda pero no lo están, en su mayoría por motivos económicos.
- Web privado. Lo forman los ficheros que han sido deliberadamente excluidos de los motores de búsqueda por sus creadores, bien mediante el protocolo de exclusión de robots (robots.txt) o la meta etiqueta correspondiente (noindex), bien mediante la asignación de un password.
- Web propietario es la parte del Web invisible a la que los motores de búsqueda no pueden acceder porque sólo se permite su acceso a personas que han aceptado previamente una serie de condiciones.
- El Web realmente invisible lo forman aquellas páginas que no pueden ser indexadas por los motores de búsqueda por motivos técnicos, debido a diferentes razones, como estar desconectadas, consistir en su mayor parte en imágenes, audio o vídeo, así como Flash, Shockwave, Postscript, ejetucables, comprimidos, etc., incluir contenidos en bases de datos, que el contenido sea cambiante, o generado dinámicamente.

Acerca de la problemática que los métodos de extracción de datos existentes plantean ya se habló en el correspondiente apartado, especialmente en los que a los motores de búsqueda se refiere. Si bien sobre los *crawlers* que soportan los motores de búsqueda se ha realizado algún que otro estudio (Cothey, 2003; Pant, Srinivasan y Menczer, 2004), los trabajos que se adentran en el

conocimiento de *crawlers* académicos y comerciales, objeto del presente trabajo, son casi intexistentes, aunque algunos los han tocado de forma coyuntural. Pero dejemos este aspecto para hablar de él más adelante.

Otro punto también mencionado es el referente a las unidades de medida, en el que nadie parece ponerse de acuerdo por la dificultad que entraña el ajustar la diversidad de criterios que los encargados de crear y mantener páginas web adoptan a unos patrones determinados. Algunos de los trabajos publicados se basan en el recuento de páginas web (Almind e Ingwersen, 1997), mientras que otros prefieren emplear el concepto, un tanto confuso por su falta de conceptualización, de sitio web.

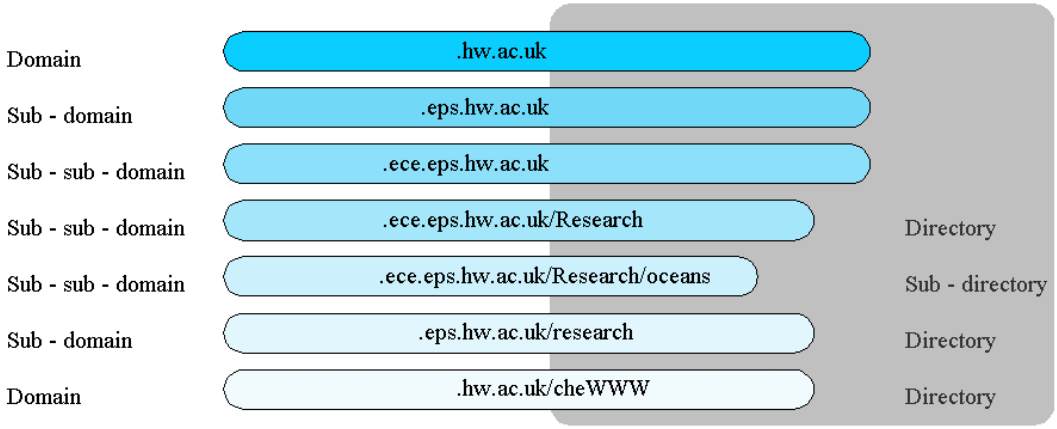


Figura 1.3. Notación de las URLs de sedes web. La superficie de las barras representa el tamaño del sitio (un dominio incluirá varios sub-dominios, e incluso directorios o sub-directorios dependientes de él), y el área gris divide visualmente la parte correspondiente a los dominios del área de los directorios y sub-directoriosque cuelgan de él.

Por ello, y desde un punto de vista más documental, Aguillo (1998a) propuso el concepto de sede web definido por Arroyo y Pareja (2003) como

> *...página web, o conjunto de páginas web ligadas jerárquicamente a una página principal, identificable por una URL y que forma una unidad documental reconocible e independiente de otras bien por su temática, bien por su autoría, o por su representatividad institucional. Teniendo en cuenta este último aspecto se reconocerían tres tipos de sedes web: institucionales, temáticas y personales.*

El inconveniente que plantea este concepto radica en la dificultad para su identificación y extracción (Arroyo, Pareja y Aguillo, 2003), cuya automatización no es lo suficientemente operativa

—puesto que no siempre es posible identificar sedes web léxicamente, esto es, por su URL, ya que no siempre se cumplen los mismos patrones (figura 1.3)— como para poder prescindir del componente humano, con la gran cantidad de esfuerzo que esto conlleva. A pesar de ello, en España existe una larga experiencia en su utilización por su aplicación en varios proyectos de investigación que lo avalan (Pareja, González y Aguillo, 1999; Aguillo, 2000a).

Desde una perspectiva cercana Facil Ayan, N., Li, W. y Kolak, O. (2002) proponen el concepto de dominio lógico, definido como

> *…a group of pages that have a specific semantic relation and a syntactic structure that relates them to each other, as opposed to a physical domain, which is identified simply by domain name.*

La gran ventaja que supone este tipo de unidades es la facilidad que presenta para automatizar su identificación y extracción del Web, puesto que serían fácilmente identificables con dominios y sub-dominios. Sin embargo, al navegar por el Web se comprueba que bajo este patrón es imposible ajustar todos los casos.

1.2.7. La Cibermetría en España

En España la Cibermetría goza en la actualidad de un sorprendente desarrollo gracias a la amplia labor de difusión realizada, tal y como puede constatarse tras la aparición de varios artículos publicados en revistas nacionales e internacionales, dos tesis doctorales (Alonso, 2002; Faba, 2002), que además han dado lugar a sendas monografías recientemente publicadas (Alonso, Figuerola y Zazo, 2004; Faba, Guerrero y Moya, 2004), y la contribución en varios proyectos de investigación nacionales y europeos (ICYTNET, EICSTES, WISER), así como su enseñanza en algunas universidades, como es el caso de la Universidad de Extremadura.

Todos estas cuestiones y algunas otras constituyen algunas de las líneas de investigación abiertas en el campo objeto de estudio. El desarrollo y empleo de otras técnicas, así como los problemas y limitaciones que plantean, constituyen el marco del trabajo que se presenta. Sobre ellos, *crawlers* académicos y comerciales, se pretende aclarar algunos puntos acerca de su funcionamiento, limitaciones y posibilidades que permitan asentar algunas bases para el desarrollo de futuras investigaciones.

Capítulo 2. El software

En este capítulo se presentará el software seleccionado como candidato para la cuantificación automatizada de sedes web, haciendo especial hincapié en los rasgos que pueden afectar a la obtención de unos resultados u otros. Dicha selección ha sido realizada fundamentalmente en función de la disponibilidad de los programas —la mayor parte de ellos, sobre todo en lo que al software comercial se refiere, son versiones demo o *freeware*— y compatibilidad con el equipamiento técnico del que se dispone —varios PCs funcionando bajo el sistema operativo Windows 2000—, pero también por su adecuación a las funciones que de ellos se requieren.

Esta es sin duda la sección más técnica, por lo que se ha incluido como apoyo un glosario explicando el significado de aquellos términos de entre los manejados que puedan plantear más dificultades. Además se han insertado imágenes de los interfaces del software y los *outputs* que generan (anexo 1) y se adjunta un CD con los programas evaluados para una mejor comprensión de los mismos.

De todas estas herramientas, dos fueron diseñadas dentro del ámbito académico (*Webcount* y *SocSciBot*) específicamente con fines cibermétricos, mientras que el resto son programas comerciales: un verificador de enlaces (*Xenu Link Sleuth*) y varios *mapeadores* (*Astra SiteManager*, *COAST WebMaster*, *Funnel Web Profiler*, *Microsoft Site Analyst* y *Content Analyzer*, *WebKing* y *Web Trends*). Éstos últimos ya fueron propuestos por Aguillo (1998b) para tales fines, tal y como se avanzó en el capítulo anterior, y existe una amplia experiencia en el uso de algunos de ellos, aunque muy poco se ha escrito sobre su aplicación y funcionamiento (Arroyo y Pareja, 2003).

2.1. Astra SiteManager

Astra SiteManager es una herramienta desarrollada por Mercury Interactive Corporation[9] para la gestión y control de sitios web. Al igual que otros programas comerciales de este tipo, permite obtener un esquema visual de las sedes analizadas, detectar errores, comparar resultados con otros anteriores, analizar páginas generadas dinámicamente, o generar informes con algunas de las estadísticas más relevantes. Para este estudio ha sido escogida una versión de prueba de la última en

[9] http://www.mercuryinteractive.com/

aparecer, la 2.0, que puede descargarse gratuitamente del Web —aunque previamente es necesario completar un formulario con algunos datos personales— para ser probada durante algún tiempo.

Su funcionamiento, como el de otras herramientas de este tipo, es sencillo: basta con introducir la página principal o *home page* de la sede a analizar y configurar una serie de opciones para que *Astra SiteManager* comience a *mapearla*, —hacer un mapa de ella—. Tras un intervalo de tiempo que puede resultar más o menos largo dependiendo del tamaño de la sede, la memoria del equipo, y las tareas adicionales que éste esté realizando simultáneamente, dicho proceso finaliza y, además de un mapa (en forma de grafo) del sitio, se ofrece al usuario la posibilidad de generar varios informes con los resultados obtenidos. Una de sus grandes ventajas, especialmente cuando el volumen de trabajo es más grande, consiste en que esta tarea de análisis puede programarse para ser realizada en una fecha y hora determinadas, lo cual permite al usuario delegar en la máquina todo el proceso.

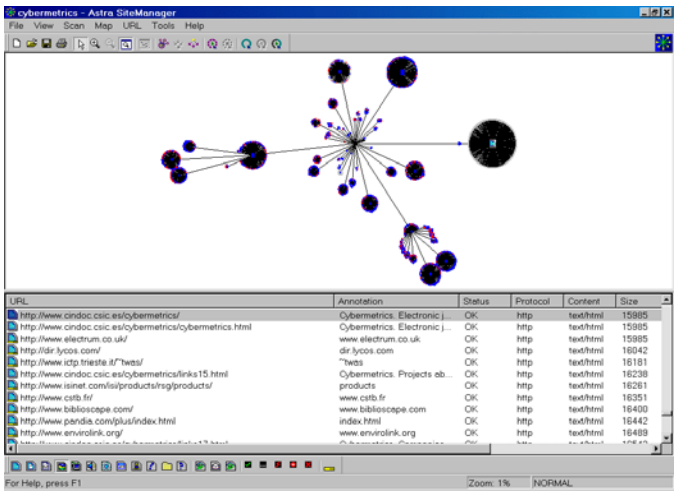


Figura 2.1. *Astra SiteManager*.

Sin duda alguna es la elección de opciones a configurar del programa la parte más complicada por su variedad, pero a la vez una de las más relevantes, puesto que de ella dependen tanto la obtención de resultados óptimos como una correcta interpretación de los mismos, por lo que es imprescindible detenerse previamente en su comprensión. Si bien es verdad que muchas de estas opciones no llegan a cambiar los resultados, sino que no son necesarias en determinadas situaciones o sólo afectan a cuestiones de visualización, otras pueden llegar a invalidar un análisis o determinar el sentido de un estudio concreto. Esta es la razón última por la que se hace necesario su enumeración.

- Definición del *browser* a través del cual se explorará el Web. La importancia de este punto radica en la necesidad por parte de algunas páginas de un explorador específico para su visualización, lo cual puede incidir en los resultados, aunque de forma mínima, pues la mayor parte de los diseñadores Web intentan que sus páginas puedan ser visualizadas por el mayor número de usuarios posible. Al ser uno de los más extendidos, en este caso se ha seleccionado el empleado por defecto, en este caso Microsoft Internet Explorer 6.0.
- Definición del editor de HTML. Ante la obligación de seleccionar uno, se opta por el empleado por defecto por *Astra SiteManager*: *Notepad*, aunque este dato no es relevante en este caso, al no afectar directamente en la consecución de unos datos u otros.
- La configuración del *proxy* se refiere a la posibilidad de especificar el servidor que actuará como mediador entre el programa y otros servidores externos. Al no ser necesario el uso de intermediarios para el acceso a Internet no serán marcadas estas opciones.
- El chequeo de enlaces externos sólo será empleado cuando se desee verificar enlaces a otros servidores; en caso contrario sólo servirá para ralentizar el trabajo del software, como sucede con *Astra*.
- Las opciones relativas a formularios se refieren al envío de datos a través de éstos cuando sea requerido. En el presente estudio han sido obviadas, pues queda fuera su alcance la interacción con otros servidores, tal y como se explicará en el capítulo siguiente.
- La limitación de las invocaciones a CGI scripts (*Astra SiteManager* ofrece hasta un máximo de 100) permite evitar demoras y fallos de programa innecesarios cuando no existe un interés por que sean analizados. Esta opción no será seleccionada para permitir así estudiar cómo son tratados por el software.
- La identificación del cliente HTTP se realizará por defecto.
- Las opciones de análisis de páginas dinámicas permiten avisar al usuario antes de entrar en el modo de análisis dinámico. Para evitar la ralentización del proceso han sido excluidas.
- Opciones de autentificación de páginas HTTP. Se ofrecen dos posibilidades: continuar con la exploración al no especificar información de este tipo cuando sea requerida (la seleccionada en este caso), o preguntar al usuario al respecto.
- Con el fin de explorar las sedes completamente no se limitará el alcance en la profundidad del análisis, entendida como el nivel de la ruta en el examen, situando en un primer nivel la *home page*, en el segundo los hipervínculos hallados a partir de estas, y así sucesivamente.
- Exclusión o inclusión de ciertas URLs. Esta opción puede ser gran utilidad para evitar aquellas páginas problemáticas que provocan errores de programa, aunque no ha sido necesario en ninguno de los casos.

- Definición del servidor. Se muestra automáticamente, sin posibilidad de modificación, el nombre del servidor que contiene al sitio a analizar.
- Ruta del documento. Define la ruta de red en la que *Astra SiteManager* busca documentos HTML cuando se desea analizar sitios alojados en un directorio local.
- Indicación del alias del directorio especificado, pues a determinados sitios es posible acceder desde varias direcciones.
- La distinción entre mayúsculas y minúsculas[10] en los nombres de fichero hace referencia a una de las características del servidor. Esta opción ha sido siempre marcada para evitar posibles variantes.
- Nombres de fichero por defecto. Este programa, como algunos otros (anexo 2), al empezar a trabajar, busca los nombres de la página de inicio de la sede por defecto, y de ahí que pida su identificación previa. Se ofrece una lista con los nombres más comunes (index.htm, index.html, homepage.html, welcome.html...), y en el caso de que no se encuentre entre ellos deberá ser introducido manualmente.
- Autenticación de áreas HTTP. Se definen las áreas protegidas por contraseña para poder acceder a ellas. Al trabajar con sitios ajenos a los que no se tiene acceso estas áreas quedarán siempre fuera del análisis, pasando a formar parte del Web invisible.

Una gran parte de estas opciones es común a muchos de los programas analizados, por lo que en adelante sólo serán enumeradas. En el anexo 2 se incluye un cuadro resumen de todas ellas para favorecer las comparaciones.

El resultado final que *Astra SiteManager* produce, a petición del usuario es el siguiente:

1. Informe de cambios, que compara el mapa resultante con otros generados en fechas anteriores para la misma sede.
2. Informe "*Link Doctor*", en el que se enumeran los vínculos rotos hallados.
3. Informe sobre el tiempo de descarga, que describe el tamaño y el tiempo invertido en cada página del sitio. Esta información resulta de gran utilidad a los webmasters interesados en facilitar al usuario la navegación, aunque no sea el caso.

[10] La traducción de la expresión inglesa *case sensitive*, relativa a la propiedad de diferenciar mayúsculas de minúsculas, parece ser un tema en el que nadie se pone de acuerdo. Aunque también se puede encontrar como "sensible a la caja", "dependiente de caja" o "distingue caja" (en referencia a la caja tipográfica) se ha optado por "distingue entre mayúsculas y minúsculas" por cuestiones de claridad (Pozo, 2001).

4. Y por último, el informe de páginas, sin duda el más interesante para el tema del que se trata, puesto que detalla una serie de datos que pueden resultar de gran utilidad para algunos estudios cibermétricos, por lo que se analizará a continuación en detalle.

El informe de páginas contiene datos identificadores sobre la sede en cuestión, como su nombre (extraído de la etiqueta META correspondiente) y URL, así como la fecha y hora en que fue generado, y lo que el denomina *Page Status Summary*, que informa sobre las siguientes estadísticas:

- Número de páginas, locales por un lado (refiriéndose a aquellas que forman parte de la sede) y externas (que no forman parte de ella) por otro, entendidas como el número de URLs en el mapa del sitio que pueden ser consideradas páginas actuales, y no recursos o páginas dinámicas; quedan incluidas páginas y formularios HTML.
- Número de recursos, incluido texto plano, imágenes, vídeo, audio, otras aplicaciones, java, FTP, e-mail, y otros.
- Y número de páginas dinámicas, que es el número de URLs CGI, tal y como las especificaciones del propio programa definen.

Finalmente se detalla un listado de todas las URLs incluidas en el análisis, tanto locales como externas, lo cual dificulta la comprensión del funcionamiento de este software, que puede ser observado a través de los resultados y dicho listado.

Los grandes inconvenientes de *Astra SiteManager* residen en la presentación, demasiado general, de unas estadísticas que no distinguen tipos de ficheros (especialmente entre recursos como imágenes, audio, aplicaciones, etc.), y en el error del dato de paginas locales que, aunque se define como el número de URLs que no son recursos (texto plano, imágenes, etc.) ni páginas dinámicas, en la práctica se trata de una suma total de los hipervínculos (no repetidos) que apuntan hacia otros recursos de la sede, tal y como —ya se adelanta— ha demostrado el experimento llevado a cabo.

2.2. COAST WebMaster

COAST WebMaster es un producto desarrollado por COAST Software[11] que permite administrar sitios web o intranets, organizando carpetas y ficheros, verificando enlaces, haciendo un seguimiento de sus cambios y actualizaciones, o examinando sus problemas. Puede conseguirse una

[11] http://www.coast.com/

versión de prueba 6.0 (la última, aparecida en 2002) que caduca a los 15 días. Uno de sus últimos éxitos, tal y como reza en su sitio web, es haber sido elegido recientemente por el Ministerio de Defensa de los Estados Unidos como gestor de contenidos de las páginas web de la marina, las fuerzas armadas y el ejército del aire para su seguridad.

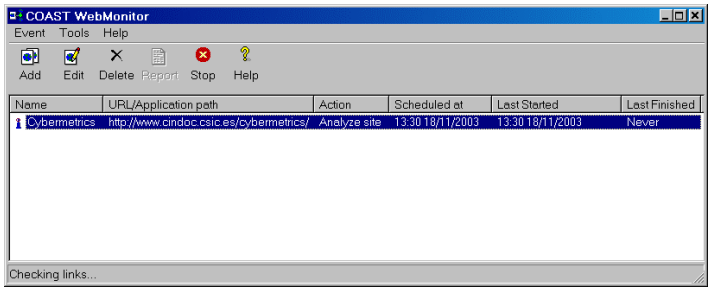


Figura 2.2. *COAST WebMaster*.

Aunque sus utilidades son varias —realizar copias del sitio a un disco local, comparar mapas, comprobar la respuesta de un determinado servidor, o lanzar alguna otra aplicación—, la que permitirá obtener datos cuantitativos sobre la sede es la encargada del análisis de sitios web. Su puesta en marcha puede incluso programarse mediante una herramienta auxiliar, el *COAST WebMonitor*, que funciona de la misma forma que el *Scheduler* de *Astra SiteManager*.

El funcionamiento de *COAST* es muy similar a cualquiera del resto de los programas de características similares, así como las opciones que permite configurar, aunque en este caso la variedad es mayor que en otras herramientas. Veamos las más interesantes de ellas, el resto quedan desglosadas en el anexo 2.

- Configuración del explorador, al igual que *Astra SiteManager*.
- Examinar el sitio completo. Por defecto *COAST* señala esta opción, pero también se puede limitar la profundidad del análisis hasta un nivel determinado.
- Verificar enlaces externos.
- Distinción entre mayúsculas y minúsculas.
- *Honor Robot Protocol*. Los ficheros robots.txt son usados por los *webmasters* para proteger sus sitios de *crawlers* y otros agentes puedan explorarlos. Esta opción, que se mantendrá marcada, hace recaer en el usuario la responsabilidad de observar o no dicho protocolo.
- Leer datos de imágenes. Esta opción debe ser seleccionada cuando sea necesario para analizar las imágenes del sitio, pero puede retrasar el proceso.

- Analizar páginas de Java Server (JSP), siempre que se incluyan en el sitio a analizar.
- Analizar los enlaces en documentos de Word.
- Gestión de *cookies*. Esta opción debe seleccionarse si el sitio analizado emplea *cookies* para generar o dirigir el contenido de las consultas recibidas.
- El número máximo de conexiones se refiere a la cantidad de enlaces que el programa puede seguir simultáneamente.
- Inclusión y exclusión de URLs.
- Reconocimiento de alias.
- Introducir contraseña para áreas protegidas.
- Configuración del *proxy*.
- Opciones de servidores. Permiten especificar servidores adicionales a examinar.
- Identificación de sesión.
- *COAST* también permite establecer acceso FTP a sitios web.
- La opción "definición de extensiones" permite al usuario establecer los tipos de ficheros que serán considerados como páginas HTML, lo que aumenta el control del proceso por parte del usuario. Siempre que sea permitida esta opción se señalarán en adelante las siguientes extensiones: htm, html, shtm, shtml, sml, stm, stml, htp, cfm, asp, jsp y php.
- Ficheros huérfanos son aquellos que no son enlazados por ninguna otra de las páginas del sitio por varios motivos:
    a. Haber quedado obsoletos y haber sido olvidados.
    b. Existir enlaces hacia ellos pero que han sido creados por medio de *scripts*, por lo que no pueden ser detectados mediante un análisis normal.
    c. Haber sido aislados u ocultados deliberadamente.

  COAST WebMaster permite analizarlos, aunque no haya enlaces a ellos, siempre que sea fuera de línea.
- Las opciones de visualización permiten escoger la forma en que los datos se dispondrán visualmente.

El resultado final son una serie de informes —cuyos datos puede seleccionar el usuario—, en mayor o menor número dependiendo del tamaño de la sede a analizar, en formato HTML y relacionados unos con otros mediante enlaces. La información que aportan va desde un resumen sobre el análisis y los errores hallados, tales como enlaces rotos, páginas que se descargan con demasiada lentitud, o páginas desaparecidas, detallados también en otro informe, hasta el desglose de las estadísticas de la sede tal y como se puede apreciar en el anexo 1.2.

En este último informe llaman la atención algunos datos que sólo *COAST* de entre los programas analizados calcula, como el número de ficheros Flash o de Microsoft Word, páginas que contienen *applets* de Java, Java *scripts*, controles ActiveX, VisualBasic *scripts*, formularios y marcos. Pero por otro lado se echan de menos otros más comunes: imágenes (sólo se desglosan formatos GIF y JPEG), ficheros de sonido u otros multimedia, así como ficheros ricos (pdf, Postscript, doc, rtf, xls, ppt), que pueden ser considerados un buen indicador de contenidos científicos.

Aunque no en un formato exportable —lo cual sería de gran utilidad en posteriores estudios de análisis de sitas—, esta información puede ser segregada y visualizada dentro de la ventana del programa, lo cual ayuda a profundizar en el conocimiento del mismo.

En definitiva, el enfoque de *COAST WebMaster* hacia programaciones dinámicas, cuyos resultados habrá que evaluar posteriormente, junto con el gran número de opciones que permite configurar al usuario, lo convierten en una de las herramientas más completas de las seleccionadas.

2.3. Funnel Web Profiler

Funnel Web Profiler es una herramienta de gestión de sitios web comercializada por Quest Software[12] que permite el análisis, identificación de problemas de diseño e implementación, control de cambios, e integración de datos de tráfico web mediante una aplicación complementaria, *Funnel Web Analyzer*. En el presente estudio ha sido empleada la versión 2.0, actualizada en 2002, en su edición *Personal*, más limitada que la *Profesional* en aspectos como el número de "items" (enlaces únicos que apuntan a la misma sede referentes a páginas HTML, imágenes y otros) que permite analizar en cada sitio (hasta 1.000 en la *Personal* y 10.000 en la *Profesional*), la capacidad para exportar datos a otros formatos (Excel, HTML, texto o CSV) o de programar tareas, o el soporte técnico. A pesar de todo ello, esta versión ha sido la escogida por su fácil disponibilidad —puede descargarse gratuitamente del sitio web de la compañía, previo envío de algunos datos personales—, y bajo coste.

Las opciones de programa permitidas son muy similares a las de otras herramientas de las mismas características (anexo 2). Las que marcan la gran diferencia en este sentido son las dirigidas a

[12] http://www.quest.com/

integrar datos de visitas, a través de los cuales puede medirse la popularidad[13]. Original, cuanto menos, con respecto al otros programas, y muy útil en algunas ocasiones, resulta además la posibilidad de tratar por igual las URLs finalizadas con barra oblicua y las que no. Otras opciones ofertadas son:

- Verificación de enlaces externos.
- Copia en caché de los ficheros durante el análisis.
- Distinción entre mayúsculas y minúsculas.
- Inclusión y exclusión de URLs en el análisis.
- Aceptación o rechazo de *cookies*.
- Introducción de contraseñas.
- Configuración del *proxy*.
- Disponibilidad de aplicaciones auxiliares desde *Funnel Web* para la visualización ficheros.
- Identificación del cliente.
- Opciones de visualización, como la selección de la etiqueta que aparece en las páginas web (URL o título).

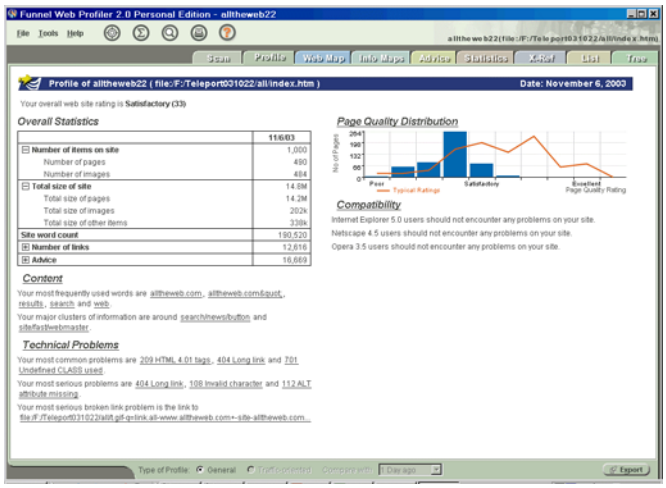


Figura 2.3. *Funnel Web Profiler*.

Una vez finalizado el proceso de análisis se puede acceder a los resultados en varias pestañas dentro de la misma ventana del programa, lo cual impide que puedan ser manipulados,

[13] La popularidad puede definirse como el número y distribución de las visitas recibidas en un período de tiempo dado. Siempre que se disponga de los ficheros log se pueden emplear programas de estadísticas de visitas para su análisis; en caso contrario se puede recurrir a Alexa.com.

guardados en otro lugar, copiados o exportados a otros formatos, función que sí incorpora la versión *Professional*. Tales resultados se presentan de la siguiente manera:

- En la pestaña *Scan* se muestra el estado de las fases del análisis.
- La pestaña *Profile* resume las estadísticas obtenidas durante el análisis, más bien de forma sucinta:
    - Número de items en el sitio (que no podrá nunca superar los 1.000), distinguiendo número de páginas e imágenes.
    - Tamaño total del sitio en MB o KB, según el tamaño.
    - Número de palabras contenidas en la sede.
    - Número de enlaces, especificando aparte tanto los enlaces rotos como los externos.
    - Asesoramiento. En este apartado se especifica una serie de sugerencias, ya tipificadas y clasificadas.
- En la siguiente pestaña, *Web Map*, se representa el mapa de la sede analizada mediante un grafo en el que cada página es un nodo y los enlaces entre ellos, arcos (anexo 1.3).
- Muy visual resulta la representación gráfica, en forma de mapa topográfico (anexo 1.3) del contenido de una determinada sede mediante el análisis de las palabras que incluye. En dicho mapa las palabras más recurrentes aparecen en negrita y en zonas de mayor altitud, mientras que las menos repetidas se encuentran al nivel del mar y en tipografía de menor tamaño.
- *Advice*. En esta hoja se detallan los consejos para la optimización de la sede resumidos anteriormente y debidamente codificados.
- En la pestaña *Statistics* se muestra una serie de gráficos elaborados a partir de datos como el tamaño individual de las páginas, el tamaño de las descargas, la última fecha de actualización de cada página, su fecha de caducidad, las calificaciones recibidas de acuerdo con su calidad, o la proporción del tipo de items (páginas, imágenes u otros).
- X-ref. Distribución de los recursos por página HTML, tales como fuentes, formularios, hojas de estilo, *scripts*, imágenes, *plug-ins*, etc.
- List. Listado de los enlaces encontrados en la sede, tanto internos como externos, con algunos de sus datos.
- Tree. Representación en forma de árbol de la estructura de la sede.

A pesar de las limitaciones de *Funnel Web* incluso en su versión *Professional*, que han impedido obtener datos completos de algunas de las sedes seleccionadas y provocado un considerable retardo en el proceso al generar tantos análisis y representaciones, sus característicos gráficos y la gran cantidad de información que procesa y ofrece la dotan de gran interés.

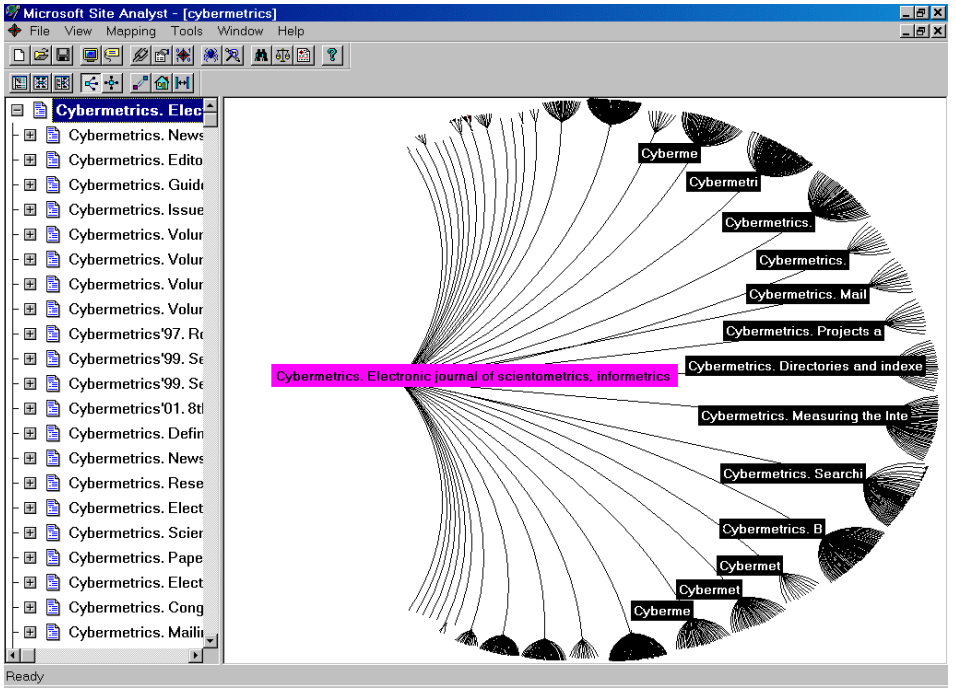
2.4. Microsoft Site Analyst y Content Analyzer

Tanto *Site Analyst* como *Content Analyzer* son programas de los llamados *mapeadores* por permitir crear mapas de sitios web. Desarrollados por Microsoft, sus últimas versiones, 2.0 y 3.0 respectivamente, resultan antiguas en el entorno de Internet, (datan de 1997 y 1998). La de *Content Analyzer* se comercializa como una de las herramientas del paquete Site Server 4.0, y necesita una licencia para entrar en funcionamiento. Sin embargo, la versión de *Site Analyst* empleada es una demo que caduca a los 30 días pero que puede ser instalada una y otra ver en un mismo equipo sin limitación alguna, pudiendo así utilizarse indefinidamente sin licencia con esta pequeña molestia. El interfaz de ambos es casi idéntico, así como las utilidades que permiten, las opciones a configurar, o los resultados que de ellos pueden obtenerse, salvo algunas excepciones insignificantes.

Además del análisis y gestión de sitios web, otras de sus funcionalidades son la verificación de listados de enlaces, la realización de copias de sitios web a un sitio local, o la comparación de informes con otros generados anteriormente, pero los objetivos de este estudio lo enfocarán hacia el primero de ellos, el análisis de sedes web, por ser el que arroja los resultados que lo hacen idóneo para su aplicación para la extracción de datos cuantitativos en el área de Cibermetría.

Dichos resultados consisten en un mapa *ciberbólico* en el que se representa la sede analizada como un grafo dirigido (digrafo), y una serie de informes, en mayor o menor medida dependiendo del tamaño de la sede, puesto que el contenido será mayor también. Estos informes (anexo 1.4) contienen una serie de estadísticas sobre la sede, relativas a enlaces, páginas y recursos incrustados en ellas, tales como imágenes, ficheros de audio, vídeo o texto, aplicaciones, o pasarelas. Cada uno de estos datos se detallan en otros ficheros, incluyendo listados de los hipervínculos que los enlazan, que pueden ser fácilmente exportados a otros formatos (como xls o txt). De especial interés es el que indica los enlaces que apuntan a otras sedes para la realización de estudios de *sitas*.

Las opciones que ambas herramientas de Microsoft permiten configurar al usuario son idénticas, y no destacan por su originalidad o novedad con respecto a cualquiera del resto de los programas, si bien aparecen en disposición y orden diferentes, manera que se enumerarán conjuntamente.



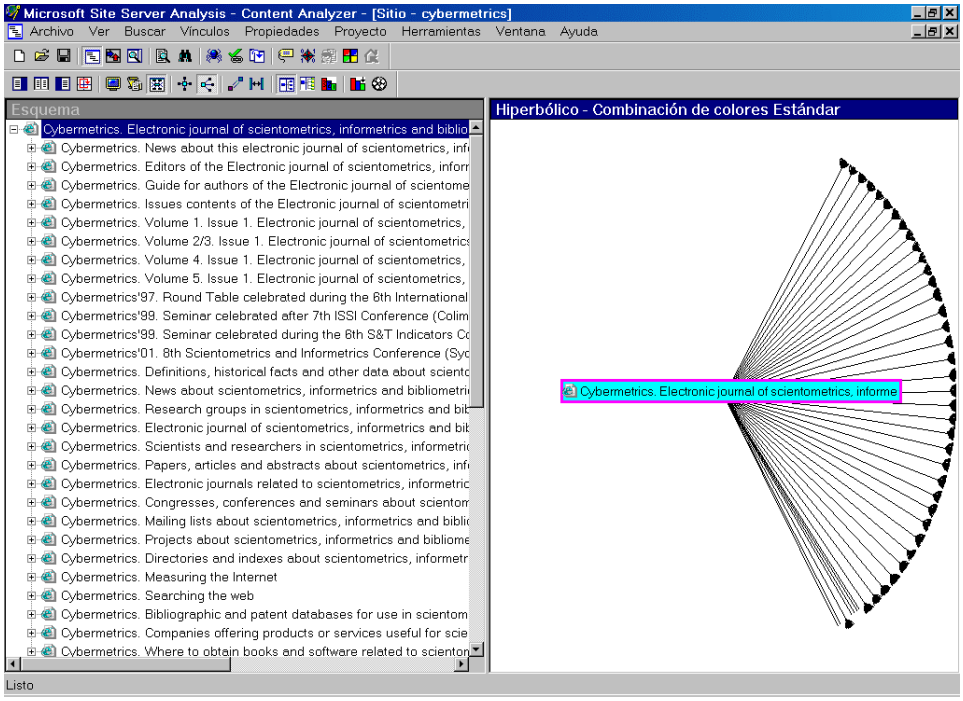


Figura 2.4. *Site Analyst* y *Content Analyzer*.

- Exploración de todo el sitio.
- Definición de rutas por jerarquía en dirección URL.
- Generación de informes de sitios.
- Distinción entre mayúsculas y minúsculas en direcciones URL.
- Comprobar enlaces fuera del sitio.
- Seguimiento del protocolo de robot.
- Especificación del agente de usuario o *browser*.
- Introducción de extensiones y restricciones.
- Configuración del *proxy*.
- Aplicaciones auxiliares.
- Opciones de visualización del mapa *ciberbólico*.
- Introducción de contraseñas.
- Copia del sitio.

En lo que respecta a su funcionamiento, varias ventanas pueden abrirse simultáneamente en un mismo equipo, el único elemento que las limita es la gran cantidad de recursos que requieren, en especial de memoria RAM. Esto puede llevar a que en sitios de gran tamaño el proceso se retrase o paralice, llegando incluso a provocar el bloqueo del programa.

Site Analyst y *Content Analyzer* han probado su utilidad en la obtención de datos cuantitativos por su participación en varios proyectos de I+D citados anteriormente. Sin embargo, la relativa antigüedad de este software junto con la rapidez con que Internet evoluciona llevan a plantearse la posibilidad de que Microsoft se esté quedando atrás en algunos aspectos con respecto a otros competidores, en este terreno. A esta pregunta se intentará responder en el capítulo siguiente.

2.5. SocSciBot

SocSciBot es un *crawler* creado por el *Statistical Cybermetrics Research Group*[14], el grupo de investigación en Cibermetría de la universidad inglesa de Wolverhampton, con fines académicos, del que se puede disponer de forma gratuita acreditando la condición de investigador o doctorando. Trabaja sobre Windows 95 o versiones posteriores, y las tareas que desarrolla consisten, básicamente, en analizar sitios web y copiar su estructura hipertextual en un fichero, así como las palabras que

[14] http://cybermetrics.wlv.ac.uk/socscibot/

aparecen en cada página del sitio, por lo que resulta de gran utilidad en el análisis de citas y de contenidos. Durante la redacción de este trabajo ha aparecido una nueva versión, la 1.8.108[15], que incorpora nuevas opciones avanzadas y aumenta su capacidad.

Su funcionamiento es bien sencillo: al ejecutar el programa es necesario indicar, en este mismo orden, la carpeta en que se guardarán los informes generados, la página principal de la sede que se desea analizar, el nombre que será asignado a los informes que contienen los resultados, y el dominio que las URLs que irá encontrando en su camino deben incluir para ser consideradas internas.

En este último punto conviene detenerse, ya que no se trata de una simple característica que lo distingue de los demás, sino que de su correcta elección depende la validez de los resultados. *SocSciBot*, durante el análisis, compara cada una de las URLs que va encontrando con la introducida en esta casilla y las va agrupando en dos categorías: internas —que forman parte de la sede— y externas —que forman parte de otras sedes—, de manera que si la URL no es la adecuada se sumarán datos al lado contrario, invalidando los resultados. La selección de una URL u otra dependerá siempre de los objetivos del estudio que se desee realizar. En este caso se tratará de homogeneizar, siempre en la medida de lo posible dadas las grandes diferencias, los criterios seguidos en la recogida de datos, y por lo tanto se introducirá la misma URL que la página principal hasta el último directorio o sub-directorio posible (*véase* figura 2.3).

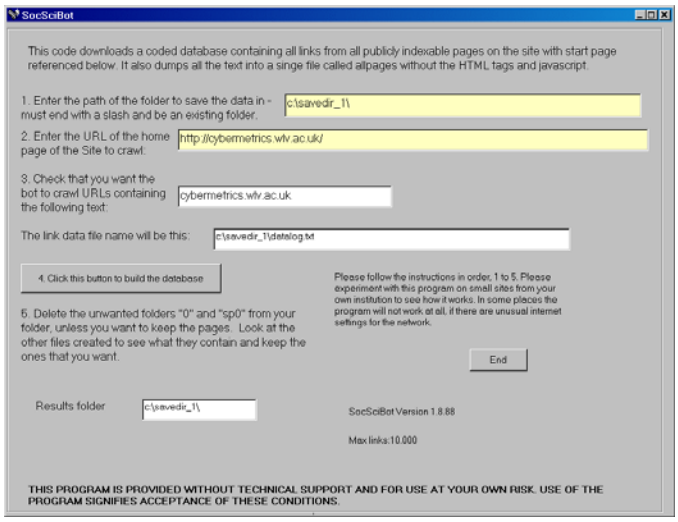


Figura 2.5. *SocSciBot.*

[15] El 6 de mayo de 2004.

Sin embargo *SocSciBot* realiza siempre, por defecto, un truncamiento en dicha URL, de manera que toma como páginas internas aquellas alojadas en sub-dominios de la propia sede. Esto no constituiría ningún inconveniente, puesto que podrían ser consideradas sedes anidadas, como sucede con las alojadas en sub-directorios adyacentes, si el resto del software escogido para este estudio funcionara de la misma forma. Por lo tanto, y ante la imposibilidad de controlar este punto, será necesario tenerlo en cuenta en la interpretación de los resultados, puesto que será la causa de cifras demasiado altas.

Tal y como se puede observar, el número de opciones permitidas al usuario es escaso, se limita a aspectos básicos, dejando recaer, de esta forma, toda la responsabilidad en el propio programa y sus características internas. Estas, por su parte, no parecen estar restringidas, sino que simplemente se limita a recoger todos los enlaces que encuentran a su paso.

El proceso de análisis puede variar entre apenas unos segundos para las sedes de menor tamaño y hasta algo más de una hora para las mayores, lo que lo convierte en uno de los más rápidos. Dada su sencillez y escaso consumo de recursos, permite además que varias ventanas están abiertas en un mismo equipo al mismo tiempo sin retardar el funcionamiento ni generar errores o bloqueos del programa, que no han sido experimentados, que no se han dado en ninguno de los casos.

El resultado final son varios ficheros en formato txt (anexo 1.5), por lo que el espacio que ocupan en disco es relativamente reducido. Los más útiles por su información son:

- El fichero con la estructura hipertextual de la sede, cuyo nombre asigna el usuario en una de las pocas opciones permitidas, la cuarta.
- El fichero denominado “allpages.txt” contiene un volcado del texto de todas las páginas que forman parte de la sede.
- El fichero “summary.txt” informa de algunas de las estadísticas de la sede. Sin embargo sus datos no son fiables para esta versión debido a las sucesivas actualizaciones que el programa ha venido sufriendo sin que haya sido modificada esta parte (Thelwall, 2003), por lo que se hace necesaria la utilización de un programa auxiliar que extrae información de "datalog.txt" y "allpages.txt".
- Fichero “index.txt”, en el que se enumeran las URLs de las páginas web de la sede ordenadas por nivel de profundidad.

La gran limitación que esta herramienta presenta está el número de páginas web que permite analizar, nunca más de 10.000 por sede, que lo inhabilita para grandes sedes. Este problema ha sido subsanado en gran medida en la nueva versión, que amplía a 50.000 el margen.

2.6. Webcount

Webcount es un programa desarrollado para EICSTES por Adams y Gilbert (2003) mediante el lenguaje Ruby (Matsumoto, 2002), un lenguaje interpretado, que permite la programación orientada a objetos, con la finalidad de recoger estadísticas de sitios web.

Para iniciar la cuantificación de una sede web cualesquiera basta con ejecutar *Webcount*, señalar la *home page* de dicha sede y, en los casos requeridos, añadir los correspondientes comandos marcando las opciones deseadas, referidas más bien a utilidades complementarias, por lo que no guardan relación alguna con las ofrecidas por los programas presentados, tal y como se puede comprobar en la figura 3.6. Estas utilidades complementarias permiten, entre otras cosas:

- Cuantificar varias sedes anotadas en un fichero de texto.
- Modificar el tiempo de espera en la conexión a páginas web.
- Añadir sub-dominios como parte de la misma sede, tal y como hace *SocSciBot* por defecto.
- Prescindir de las estadísticas de WayBack Machine[16]. Este motor de búsqueda permite la interrogación a una base de datos que contiene versiones antiguas de sitios web, desde 1996 hasta la actualidad, por lo que puede considerarse un archivo de Internet. Los datos que *Webcount* obtiene al interrogarlo son el número de páginas distintas, que forman parte del sitio, guardadas en cualquier fecha durante un año. Dadas las limitaciones técnicas del *Internet Archive* que, por motivos obvios, hacen imposible hacer una copias exactas y actualizadas de todos los sitios en Internet, dichos datos deben ser cuidadosamente interpretados.
- Prescindir de las estadísticas de Alltheweb, que por defecto se recogen, relativas al número de enlaces recibidos por la sede.

[16] http://www.archive.org/

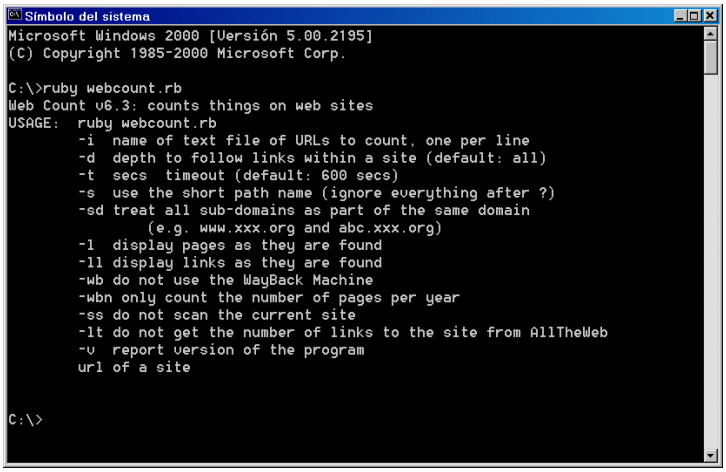


Figura 2.6. *Webcount.*

De todas ellas, sólo la tercera será seleccionada, evitando así alargar el proceso y aumentar el volumen de los ficheros con tareas innecesarias. De este proceso, que apenas dura unos segundos, se obtienen como resultado varios ficheros de texto con dos tipos de información: errores encontrados y estadísticas. En los segundos (figura 3.7) se informa de datos tales como:

- Número de páginas analizadas. La diferencia entre enlaces y páginas está en que el número de enlaces incluye duplicados.
- Número de palabras.
- Número de enlaces externos, que son aquellos que apuntan a páginas alojadas en otro servidor distinto de la página que lo enlaza.
- Enlaces internos, aquellos que apuntan a páginas dentro de la sede.
- Número de páginas externas.
- Número de páginas internas.
- Hash links, son las llamadas anclas o enlaces con la forma “#name” y que envían a otra parte de la misma página.
- Imágenes. Es el número de referencias a ficheros dentro de la etiqueta <img>, que presumiblemente serán imágenes.
- CGI. Son enlaces que hacen referencia a una URL que incluyen la secuencia “cgi” en la dirección. Estos enlaces también se incluyen en los recuentos de enlaces y páginas, tanto internos como externos.
- Mail. Es el número de enlaces que incluyen la secuencia “mailto” en su dirección. Estos enlaces no quedan incluidos en el recuento de páginas y enlaces, bien sean internos o externos.

- M-Media es el número de etiquetas del tipo “bgsound” (background sound), “embed” (Flash, Shockwave, Quicktime y ficheros de *plug-ins* similares) y “applet” (Java programs).
- Profundidad de la sede. Es el número máximo de páginas a las que hay que acceder para llegar desde la página principal a cualquier otra página del sitio.
- Fecha y hora de los datos y el tiempo que ha durado el proceso de análisis.

Web Count v6.3 at Tue Sep 30 11:23:40 Hora de verano romance 2003. Options:

URL	Year	Pages scanned	Words	External links		External pages		Internal links		Internal pages	
http://www.cindoc.csic.es/cybermetrics	Links to the site			1,198							
http://www.cindoc.csic.es/cybermetrics	TOTAL	1996	0	0	0	0	0	0	0	0	0
http://www.cindoc.csic.es/cybermetrics	TOTAL	1997	0	0	0	0	0	0	0	0	0
http://www.cindoc.csic.es/cybermetrics	TOTAL	1998	23	5466	135	133	370	66	1	82	
http://www.cindoc.csic.es/cybermetrics	TOTAL	1999	30	13115	676	645	636	82	0	115	
http://www.cindoc.csic.es/cybermetrics	TOTAL	2000	49	43909	1569	1488	1441	78	21	385	
http://www.cindoc.csic.es/cybermetrics	TOTAL	2001	82	104844	2227	2043	2530	137	104	969	
http://www.cindoc.csic.es/cybermetrics	TOTAL	2002	45	57492	562	491	1245	102	50	343	
http://www.cindoc.csic.es/cybermetrics	TOTAL	Now	95	118451	2384	2019	3480	233	219	1306	

Finished at Tue Sep 30 11:47:32 Hora de verano romance 2003. Run time: 1432 secs.

Figura 2.7. Informe generado por Webcount.

Algunos de los problemas que plantea, señalados por sus creadores, son:

- Códigos fuente pobres o demasiado complejos.
- Enlaces ocultos en ficheros JavaScript, que ningún *crawler* puede penetrar.
- El trabajo con sub-dominios, por lo que se añadió la opción para incluir sus páginas en el recuento de las sedes.
- El trabajo con bases de datos.
- Errores por agotamiento del tiempo, que parecen suceder con mayor frecuencia en Windows XP que en Mac. Esta es la razón por la que fue introducida una opción para aumentar el tiempo de espera.
- Algunos sitios web no responden a no ser que se envíen las *cookies* adecuadas.
- Redireccionamientos, que, como sucede con otras herramientas, no son reconocidos.

Pero la gran limitación de este programa reside en que no puede copiar la estructura hipertextual de las sedes analizadas, lo que lo dificulta enormemente la comprensión de su funcionamiento por parte de los usuarios.

2.7. WebKing Lite

WebKing es el software desarrollado por ParaSoft[17] para prevenir y detectar errores en el desarrollo de aplicaciones web de cualquier nivel, enfocado tanto a páginas estáticas (HTML) como dinámicas. Para ello consta de varios módulos que se reparten distintas funciones:

- White-box Testing permite comprobar la construcción de sitios web.
- Black box Testing, encargado de chequear su funcionalidad.
- Regression Testing se dedica al mantenimiento de la integridad del sitio.
- Web-Box Testing™, que posibilita la comprobación de páginas dinámicas, una a una.

Existen dos versiones de WebKing 2.0, *Lite Mode* y *Standard Mode*, diferentes por su disponibilidad —el primero puede instalarse libremente, mientras que para acceder al segundo es necesario disponer de una licencia válida para Windows o Linux— y sus prestaciones —el *Standard Mode* permite el uso de cualquiera de ellas, pero con *Lite Mode* quedan restringidas algunas funciones, como la creación y comprobación de páginas dinámicas, la prevención y detección de errores en *applets*, CSS o JavaScript, la publicación, etc.—. Teniendo en cuenta las necesidades de este estudio, *Lite Mode* resulta adecuado, ya que la tarea que será requerida de él, el análisis de determinadas sedes web para la obtención de estadísticas, puede ser desempeñada sin restricción alguna.

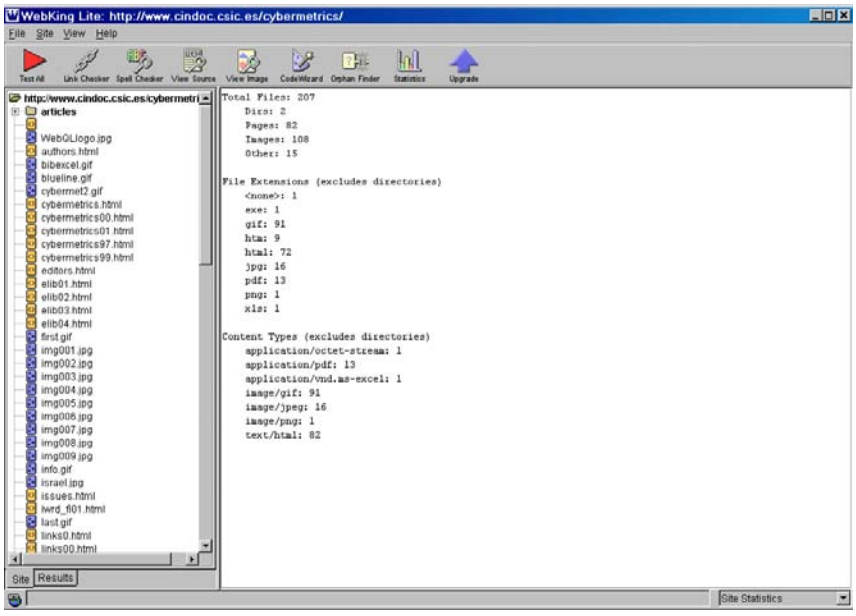


Figura 2.8. *WebKing*.

[17] http://www.parasoft.com/

Dicho análisis se realiza siguiendo pasos muy similares a los de otros programas —*Astra SiteManager*, *COAST Webmaster*, *Funnel Web*, o *Site Analyst* y *Content Analyzer*—, habiendo seleccionado previamente un conjunto de opciones casi idénticas, entre las que se encuentran la autenticación de contraseñas, distinción entre mayúsculas y minúsculas, obediencia a restricciones de robot, gestión de *cookies*, especificación de páginas por defecto, selección del *browser*, o verificación de enlaces externos. Una de las más útiles para evitar ruido en los resultados es aquella que permite seleccionar las estadísticas que se mostrarán y las que no.

El resultado final podría calificarse de austero, puesto que, huyendo de gráficos y mapas, tan característicos de *Funnel Web*, presenta, junto a un esquema en forma de árbol de los recursos visitados (páginas, imágenes, etc.), una serie de estadísticas sobre la sede analizada, organizadas por extensión de ficheros por un lado, y por tipo MIME después, además de un somero resumen previo (figura 3.8). Esta organización y la posibilidad de escoger, sin restricciones del software, los tipos de datos adecuados a cada ocasión convierten a *WebKing* en una herramienta de gran utilidad. Si a esto se añadiera un listado de enlaces exportables a otros formatos, quizás su mayor limitación, estaríamos ante una herramienta excepcional para el campo que se trata.

2.8. Web Trends

Web Trends 7.0 es la última versión de la herramienta comercializada por NetIQ[18], de cuyo sitio web puede ser descargada una demo que caduca a los 30 días, habiéndose registrado previamente. Las funciones que incorpora son:

- Análisis de tráfico web, mediante la cual puede medirse la actividad de un determinado sitio, incluyendo las visitas recibidas. Ello se lleva a cabo mediante el análisis de ficheros log.
- Análisis e información de la actividad del servidor *proxy* mediante el uso de los ficheros log, para conocer, por ejemplo, el uso que los empleados de una organización hacen de Internet.
- Análisis de enlaces, que consiste en el examen del sitio web especificado y la verificación de los hipervínculos del mismo.

[18] http://www.netiq.com/

- Alerta y control, de las cuales es la encargada del análisis de enlaces la que permite obtener los datos necesarios para este estudio.

A pesar del interés que pueden tener todas ellas, incluso en el área de Cibermetría, ya que el análisis de ficheros log o visitas es la fuente primaria para calcular la popularidad, es el análisis de enlaces el que se ajusta a los objetivos de este estudio, por lo que en adelante habrá que detenerse en él.

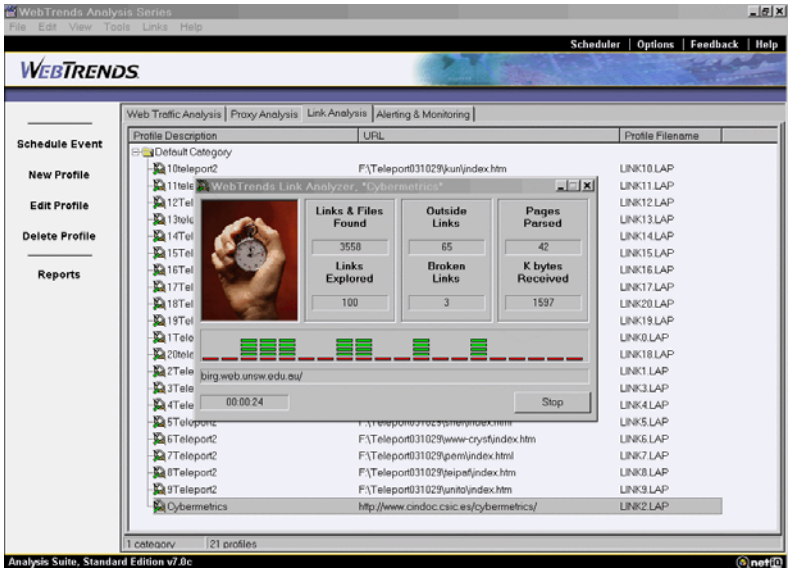


Figura 2.9. Funcionamiento de *Web Trends Link Analyzer*.

Las opciones que *Web Trends 7.0* ofrece al usuario son las siguientes:

- Especificación del número de conexiones simultáneas permitidas, siendo 60 el máximo.
- Delimitación del tiempo máximo por conexión, en segundos, que el programa esperará antes de entender que se ha producido un error.
- Configuración del número de segundos a esperar antes de abortar una transmisión o recepción de datos. Cuanto más alto es el valor el análisis será más lento, ya que toma su tiempo para comprobar si los enlaces están rotos.
- Identificación del agente usuario.
- Retraso por respuesta, es el número de milisegundos de retraso permitidos entre sucesivas peticiones al servidor remoto.
- Lenguaje en que se expresarán los informes.

- Tipos de fichero. Esta opción permite elegir las extensiones que se desean incluir en cada tipo de fichero, al igual que *COAST*, pero con la diferencia de que este último sólo permite la definición de páginas HTML, lo que convierte a *Web Trends* en una herramienta muy flexible y controlable.
- La opción de verificar enlaces internos supone una novedad con respecto al resto del software, que lo hace por defecto.
- Verificar enlaces externos, pero sólo a un nivel de profundidad por debajo.
- Configuración del *proxy*.
- Sitio web que distingue entre mayúsculas y minúsculas.
- Ignorar las URLs que se activan solo con pasar el ratón.
- Filtros. Se refiere a la exclusión de URLs.

La puesta en funcionamiento de esta tarea de *Web Trends* se caracteriza por la gran cantidad de errores producidos, incluso en sedes que normalmente no son problemáticas y de tamaño pequeño o medio (anexo 3), especialmente trabajando en línea.

El resultado final son una serie de informes en formato HTML, cuyo lenguaje puede ser seleccionado por el usuario, que contienen estadísticas básicas sobre el sitio —tales como número de páginas HTML, tamaño de la sede en KB, número de enlaces, tanto internos como externos, por tipo (HTTP, HTTPS, archivo, FTP, e-Mail, Telnet, Gopher, etc.), número de archivos por tipo (páginas HTML, imágenes, ejecutables, texto, Java, documentos, bases de datos, ficheros comprimidos, almacenados, de audio, vídeo, y ShockWave) y su tamaño—, errores hallados, y algunas sugerencias para su optimización.

2.9. Xenu Link Sleuth

Xenu's Link Sleuth es una herramienta sencilla y fácil de manejar, presentada en un interfaz intuitivo para el usuario, y efectiva a la vez. Creada por Tilman Hausherr en 1997, su éxito le ha llevado a introducir continuas modificaciones y actualizaciones, la última de las cuales, la 1.2e, apareció en septiembre del 2003, puede ser descargada de forma totalmente gratuita del Web[19].

[19] http://home.snafu.de/tilman/xenulink.html

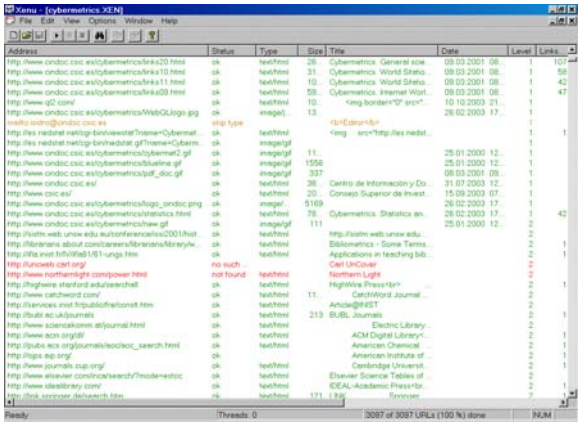


Figura 2.10. Interfaz de *Xenu*.

Aunque su función principal es verificar enlaces (no sólo HTML, sino también imágenes, marcos, *plug-ins*, *backgrounds*, mapas de imágenes locales, hojas de estilo, *scripts* y *applets* de java) y detectar aquellos que no funcionan indicando el motivo, *Xenu* permite:

- Verificar de nuevo los enlaces rotos para detectar errores temporales de red. Siempre se recomienda volver a llevar a cabo esta opción tras un primer análisis, puesto que se ha comprobado que el ratio de enlaces rotos disminuye tras posteriores comprobaciones.
- Examinar parcialmente sitios FTP y Gopher.
- Detectar las URLs que re-direccionan a otra nueva.
- Generar informes del sitio en formato HTML.

Su puesta en funcionamiento es bastante intuitiva, basta con introducir la página principal de la sede a analizar (siempre con la barra invertida después del último directorio) y configurar algunas opciones básicas, como las siguientes (figura 2.11):

- Verificación de enlaces externos.
- Inclusión o exclusión de URLs que serán tratadas como parte de la sede en el análisis.
- Número de conexiones paralelas permitidas, hasta un máximo de 100.
- Máximo nivel de profundidad en el análisis, hasta un total de 999, que es más que suficiente en la mayor parte de las sedes.
- Preguntar la contraseña cuando sea requerido. Es preferible desactivar esta opción para evitar que aparezcan continuamente molestas ventanas de diálogo.
- Consideración de los redireccionamientos como errores.

- Datos a incluir en el informe. Es recomendable seleccionar sólo aquella información que vaya a ser de utilidad, ya que la aplicación que genera estos informes, *Volcanoe*, suele bloquearse al trabajar con demasiados datos.

El tiempo que *Xenu* se toma para cada análisis depende, al igual que sucede con el resto del software expuesto, del tamaño de la sede o lista de enlaces por analizar, de la memoria del equipo, y de la actividad adicional que este realice simultáneamente. Una vez finalizado el proceso de comprobación es posible generar un informe con los resultados obtenidos, que el mismo usuario puede seleccionar (anexo 1.7), tal y como se ha explicado anteriormente, según sus necesidades. Para el caso que nos ocupa resultan imprescindibles las estadísticas del sitio y el mapa del mismo.

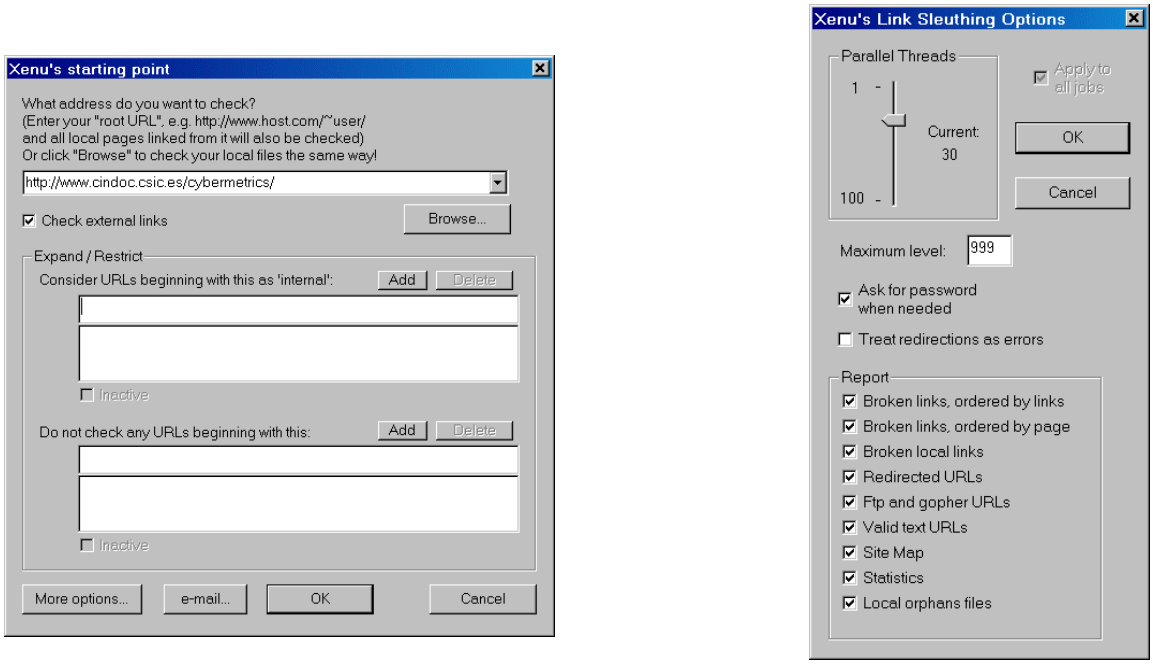


Figura 2.11. Opciones de *Xenu*.

Xenu es, como se ha podido comprobar, una de las herraqmientas más sencillas, intuitivas y fáciles de manejar, lo cual, unido a la simplicidad de los resultados la convierten en una potente herramienta especialmente útil para la verificación de hipervínculos, cuya única limitación está en los sitios web de mayor tamaño.

Capítulo 3. Trabajo empírico

En el capítulo que comienza se describen los diferentes pasos en la realización de un trabajo empírico cuyo fin es la obtención de datos cuantitativos a partir de los cuales se efectuará la evaluación del software presentado en el anterior capítulo. Para comenzar, se explicará la metodología diseñada para la recogida de datos, así como los incidentes observados en esta primera fase. A continuación se expondrán los resultados obtenidos, que plantean una serie de cuestiones a las que se intentará dar respuesta. Finalmente, se evaluará el software seleccionado teniendo en cuenta varios criterios designados en función de su adecuación a los fines perseguidos.

3.1. Diseño de la investigación y recogida de datos

Una vez estudiados el funcionamiento y las posibilidades del software seleccionado, se ha llevado a cabo una investigación que tiene por objetivo el análisis de los resultados que a través de dicho software pueden obtenerse para una interpretación correcta de los mismos. Conviene puntualizar que en ningún momento se pretende comparar los resultados de diferentes programas, puesto que se incurriría en un error metodológico dadas las grandes diferencias entre ellos, que podrán ser constatadas más adelante, sino más bien obtener puntos de referencia para la evaluación de los mismos, y no de las sedes analizadas.

El método seguido en dicho experimento consiste en la extracción de estadísticas, mediante cada uno de los programas presentados, de 19 sedes web seleccionadas de entre el espacio web académico europeo, atendiendo a los siguientes criterios:

- Se han seleccionado sedes de la mayor parte de los países de la ya antigua Europa de los 15, quedando así representados Alemania, Austria, Bélgica, Dinamarca, Finlandia, Grecia, Holanda, Italia, Suecia, Reino Unido, y, más ampliamente, España.

- Sedes de diferentes tipos de instituciones, tales como universidades (*Salamanca* y *San Pablo CEU*), facultades (*Faculteit der Filosofie* y *School of Clinical Dentistry*), departamentos (*Dipartimento di Scienze Giuridiche, Departamento de Economía*) grupos de investigación, centros e institutos de investigación (*Instituto de Tecnología Eléctrica*, *Institut für*

Humangenetik, *Centre for Language Technology*), incluso una revista electrónica (*Cybermetrics*) han sido escogidas.

- Varias áreas académicas quedan representadas, tanto científicas o tecnológicas (odontología clínica, cristalografía y biocomputación, tecnología de los materiales, tecnología eléctrica, genética...) como de humanidades y ciencias sociales (filosofía, ciencias jurídicas, economía...).

- El tamaño de las sedes es variable, desde las más pequeñas, de menos de 100 páginas —*Cybermetrics*, *School of Clinical Dentistry*, *Institut für Humangenetik*, *Instrumentelle Analytik Umweltanalytik*—, hasta otras de tamaño considerable —*Abteilung für Informationswirtschaft*, de más de 10.000 páginas—, pero siempre se ha intentado evitar las de gran tamaño para evitar los limitaciones que en este sentido presentan algunos programas, bloqueos del software y tiempos demasiado largos en el proceso de análisis.

- Especial importancia se le ha conferido a los distintos lenguajes de programación web empleados en su diseño (véase más adelante tabla 3.3), de manera que se ha intentado reunir a varios de ellos, tanto estáticos —HTML, como las sedes de *Cybermetrics* o el *Institut für Humangenetik*— como dinámicos —del lado del cliente, JavaScript, como el *Institutionen för Materialtekniks* o la *Escuela Técnica Superior de Ingenieros Agrónomos*, y Flash, como el *Instituto de Tecnología Eléctrica*; y del servidor ASP, como la *Universidad San Pablo CEU*, y CGI, *Crystallography and Biocomputing*—, para estudiar el comportamiento del software al enfrentarse a cada uno de ellos.

Cada una de estas sedes fue analizada con cada uno de los programas seleccionados en dos fechas diferentes, en primer lugar el 22 de octubre de 2003, y en segundo lugar el día 29 del mismo mes, justo una semana después, para tener en cuenta posibles oscilaciones temporales que se presupone deberían estar motivadas por los cambios sufridos en el web y no por el propio software. Cada uno de estos dos muestreos ha sido realizado en el menor intervalo de tiempo posible para evitar dichos cambios, un día para el grueso de las sedes, aunque algunas, por su tamaño y algunos problemas técnicos, no han podido concluir hasta el día siguiente.

Antes de comenzar cada uno de los muestreos se procedió a realizar un volcado de cada una de las sedes a un disco duro local con espacio suficiente como para albergar tal cantidad de información (varios GB), de manera que finalmente se ha podido disponer de dos volcados, uno

realizado el 22 de octubre de 2003 y otro una semana después. La intención de este procedimiento es doble: por un lado, tener un punto de referencia que no admita variaciones temporales sobre el contenido de la sede para poder comparar los resultados de cada programa, y por otro estudiar el comportamiento de dichos programas al trabajar fuera de línea.

	Institución	*URL*
1	Cybermetrics	www.cindoc.csic.es/cybermetrics
2	Instituto de Tecnología Eléctrica	www.ite.upv.es
3	Departamento de Economía	www.upct.es/~de
4	Universidad San Pablo CEU	www.ceu.es
5	School of Clinical Dentistry	www.shef.ac.uk/dentalschool
6	Crystallography and Biocomputing	www-cryst.bioc.cam.ac.uk
7	Pembroke College	www.pem.cam.ac.uk
8	Technological Educational Institute of Patras	www.teipat.gr
9	Dipartimento di Scienze Giuridiche	www.dsg.unito.it
10	Faculteit der Filosofie	www.kun.nl/phil
11	Institutionen för Materialtekniks	www.mat.chalmers.se
12	Centre for Language Technology	cst.dk
13	Département d'Electricité, Electronique et Informatique	www.montefiore.ulg.ac.be
14	Abteilung für Informationswirtschaft	wwwai.wu-wien.ac.at
15	Instrumentelle Analytik Umweltanalytik	www.uni-saarland.de/fak8/iaua
16	Institut für Humangenetik	www.medizin.uni-greifswald.de/humangen
17	Matematiika	www.math.jyu.fi
18	Escuela Técnica Superior de Ingenieros Agrónomos	www.etsia.upv.es
19	Universidad de Salamanca	www.usal.es

Tabla 3.1. Sedes web analizadas.

Por ello, una vez finalizados ambos muestreos en línea, se han efectuado otros dos fuera de línea sobre las copias locales realizadas previamente, esta vez sin prisas pues el peligro de que se produzca algún tipo de cambios en las sedes desaparece. Por lo tanto, finalmente para cada una de las sedes habrá 4 muestras, dos *online* (una del 22 y otra del 29) y otras dos *offline* (también de ambas fechas).

Para la descarga de las sedes fue empleado un programa volcador, *Teleport Pro* (figura 3.1), en su versión 1.29.1981, que funciona de la siguiente manera: comienza por la página principal de una determinado sitio web, dada por el usuario, la examina cuidadosamente y extrae todos los enlaces y referencias a datos incrustados en sus páginas. A continuación los compara con los tipos de fichero

previamente especificados; en caso contrario *Teleport Pro* recupera todos los ficheros. Finalmente, los almacena en la carpeta del proyecto. La limitación de esta versión, que puede bajarse del Web[20], reside en el número de veces que puede emplearse, sólo 40, y el número de ficheros que permite volcar, hasta un máximo de 65.000. Este tamaño ha sido sólo superado por la sede *Abteilung für Informationswirtschaft*, que no ha podido ser copiada totalmente, lo cual no afecta a los objetivos de este estudio, que como se recordará, consisten en estudiar el comportamiento del software seleccionado, siendo las estadísticas tan sólo un apoyo para tal fin.

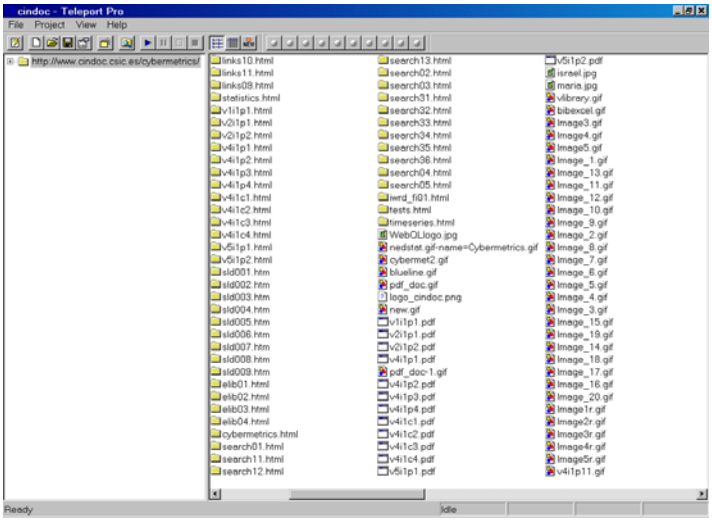


Figura 3.1. *Teleport Pro*.

Durante la recogida de datos se registraron varias incidencias. En un primer momento se seleccionó la sede web de la universidad de La Rioja para formar parte del experimento, pero durante el primer volcado el departamento informático de dicha universidad interpretó que estaban siendo víctimas de un ataque DoS, por lo que el acceso a su sede web se vio restringido y hubo de ser excluida del estudio. Esta es una muestra de cómo, una vez más, la implantación de una política de seguridad demasiado restrictiva puede llegar a afectar a la libertad de los usuarios.

En la actualidad, y por incidencias propias del WWW, no es posible conectar con la sede del *Instituto de Tecnología Eléctrica* de la Universidad Politécnica de Valencia, que, por otro lado, no parece haber desaparecido ni cambiado su URL. Para su visualización puede ser consultado el Internet Archive[10].

[20] http://www.tenmax.com/teleport/pro/

Todos los programas comerciales de entre los seleccionados permiten el análisis de sitios almacenados localmente, pero no sucede así con los académicos, por lo que fue necesario para llevar a cabo el muestreo fuera de línea copiar las sedes volcadas en un servidor web, y así emular un trabajo en línea. Por otro lado se dieron ciertos problemas técnicos en el funcionamiento de *WebKing* y *Funnel Web*, que se bloqueaban al hacer cualquier intento de trabajo fuera de línea, razón por la que fue imposible obtener ningún resultado de cualquiera de ellos.

Uno de los problemas encontrados son los errores de programa, que provocan el aborto del proceso de análisis. Siempre que se ha dado esta situación se ha realizado un segundo intento, en alguno de los casos con éxito, pero no siempre, de manera que este tipo de errores ha quedado señalado en la tabla de datos con un guión. Los motivos de este incidente pueden estar en el gran tamaño de algunas sedes —es el caso de la *Universidad San Pablo CEU* y *Abteilung für Informationswirtschaft*, tal y como se desprende de la figura 3.2) o en una determinada configuración —*Crystallography and Biocomputing*, *Centre for Language Technology*— de las mismas, que algunos programas tienen la capacidad para resolver favorablemente pero otros no.

Los programas más sensibles a cualquier problema parecen ser *WebTrends*, con una tasa total de error que supera el 76%, y en segundo lugar, ya a una gran distancia, *Xenu*, con un 28% (figura 3.3). En el lado contrario se sitúan *Content Analyzer* y *COAST WebMaster*, que no han presentado error alguno. En cualquier caso, se puede observar cómo las posibilidades de error parecen aumentar al trabajar en línea, mientras que disminuyen con el trabajo fuera de línea.

Conviene recordar que, tal y como se comentó en el capítulo anterior, *SocSciBot* y *FunnelWeb* tienen un límite en lo que al tamaño de las sedes a analizar se refiere: el primero sólo puede procesar hasta 10.000 hipervínculos, el segundo hasta 1.000 “items”. Esta es la razón por la que alguno de los datos no es totalmente correcto, incidencia que se marca en la tabla del anexo 3, que recoge todos los resultados obtenidos, con la cifra en cuestión consignada entre paréntesis.

Y por último, cabe constatar otro problema, el de las páginas que envían a otra página. Es el caso que se da en la sede web de la *Universidad de Salamanca*, que al teclear la URL de su página principal, www.usal.es, envía directamente a la página www.usal.es/webusal/Principal.htm, por lo que, para evitar una cuantificación errónea cuyo resultado sería una sola página, se ha escogido la segunda para ser lanzada al software como punto de partida, aun a riesgo de encontrar recursos que no

se encuentren en el directorio /webusal/, sino en otro al mismo nivel, en cuyo caso no serían reconocidos como propios de la sede, ha sido preferida esta solución por acercarse más a la realidad.

En el apartado que sigue se pasará a comentar y analizar los resultados obtenidos del experimento llevado a cabo.

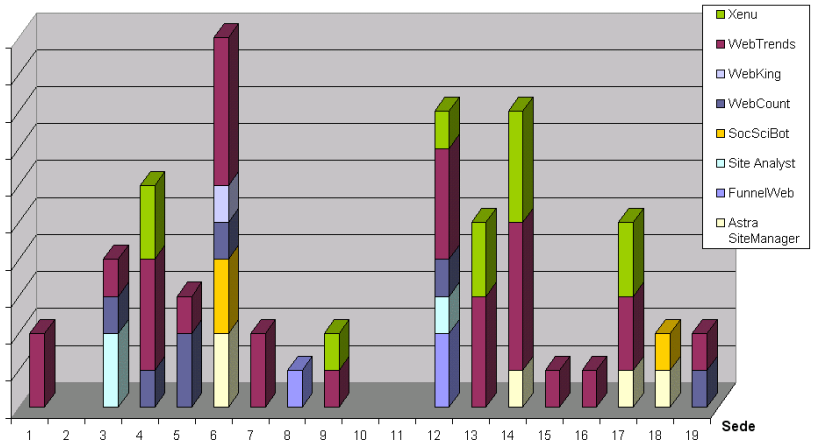


Figura 3.2. Número de errores de programa por sede. En el eje de las X se representa el número asignado a cada sede en la tabla 3.1.

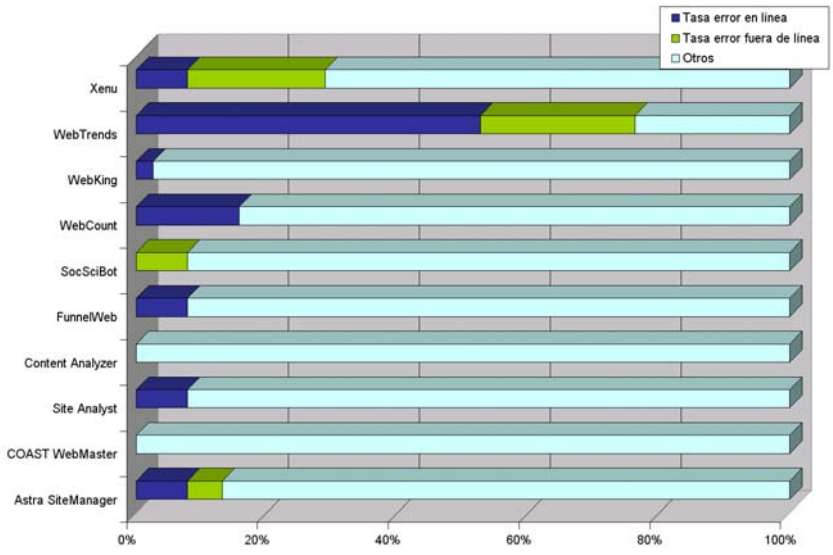


Figura 3.3. Tasas de error (%) del software empleado.

3.2. Análisis de datos y resultados

3.2.1. Descripción general

Los resultados que este tipo de programas arroja son amplios y de lo más variado, tal y como se ha podido constatar durante el capítulo anterior, pero la naturaleza y alcance del presente trabajo no permite abarcarlos a todos. Por ello, teniendo en cuenta el objetivo general marcado —el estudio del comportamiento del software seleccionado—, y para una mayor simplificación en el análisis, se emplearán los datos cuantitativos para comprender el alcance de cada una de las herramientas, y más en concreto los datos sobre el número de páginas de cada una de las sedes, que actuarán como indicador de la cobertura en el mapeo de cada uno de los programas, puesto que sólo a partir de ellas es posible acceder a otros recursos.

También es sabido que la obtención de unos resultados u otros depende, en gran medida, de las opciones de programa seleccionados por el usuario. En el caso que nos ocupa se ha intentado obtener la máxima homogeneidad posible en este sentido, con el fin de sintetizar y así evitar confusiones, aunque se ha asumido que es imposible al cien por cien, y que los datos no son en modo alguno comparables, tal y como ya se adelantó.

En este sentido, se pueden agrupar los programas a evaluar**,** según el control por parte del usuario de los resultados cuantitativos relativos al número de páginas de la sede, de la siguiente manera:

GRUPO 1. Aquellos que permiten al usuario su definición como una opción de programa más, como *COAST WebMaster* o *Web Trends*, en cuyo caso se han seleccionado los siguientes tipos de ficheros: html, htm, shtm, shtml, sml, stm, stml, htp, cfm, asp, jsp y php.

GRUPO 2. Los que presentan los resultados por tipos de fichero y por lo tanto sólo es posible para el usuario definirlos a posteriori, durante la etapa de recogida de datos (*SocSciBot, WebKing, Xenu* y *Teleport Pro*). En este caso se han sumado todos los tipos de ficheros definidos en el grupo anterior bajo el epígrafe “páginas”, excepto para los resultados de *SocSciBot*, que por problemas técnicos sólo incluye ficheros htm y html.

GRUPO 3. Aquellos que no permiten al usuario control alguno sobre los resultados, como *Astra SiteManager, Funnel Web, Site Analyst* y *Content Analyzer*, así como *Webcount*.

Un problema añadido a la diversidad de criterios empleados son las diferencias en la terminología manejada por cada programa y que puede dificultar la recogida e interpretación de los resultados. Por ello se incluye a continuación una tabla que recoge las expresiones empleadas por cada software.

Software	*Terminología*
Astra SiteManager	Local pages
COAST WebMaster	Total number of HTML pages
Site Analyst	Onsite pages
Content Analyzer	Onsite pages
Funnel Web	Number of pages
Webcount	Internal pages
WebKing	Pages
Web Trends	HTML páginas
Xenu	Text/html

Tabla 3.2. Terminología empleada por cada programa para denotar el número de páginas por sede.

La primera impresión que se obtiene al observar la tabla de datos (anexo 3) es que existen grandes diferencias entre la mayor parte de los resultados de cada programa para una misma sede, incluso entre los obtenidos en una misma fecha. Esto se explica en parte por la diferencia de criterios a la hora de entender lo que son páginas y lo que no lo son, pero también por una serie de características de cada software protegidas por el secreto comercial. Además, en contra de lo esperado, se observa que los resultados de los muestreos en línea difieren también de los resultados obtenidos fuera de línea.

Las mayores coincidencias se dan, como se puede comprobar, en las sedes 2 y 11, que incluyen un fichero Flash y un menú generado con JavaScript, ya que ninguno de estos programas puede penetrarlos en busca de enlaces. Pero incluso entre los resultados de estas dos sedes destacan los de *Astra SiteManager* y *Webcount* por no coincidir con la gran mayoría. Sobre el primero de ellos se deduce al comparar los datos obtenidos por las copias de las sedes web efectuadas mediante *Teleport Pro* que bajo el epígrafe “local pages” se suman no sólo enlaces a páginas HTML sino todos los

hallados dentro de la sede, incluidos aquellos que envían a cualquier otro recurso incrustado, ya sean imágenes, ficheros ricos o media, etc.

Pero sobre *Webcount* es difícil explicar cual es la razón de sus desacuerdos, que, por otra parte no se reducen sólo a estas dos sedes, sino a todas las que forman parte de la muestra. Esta dificultad estriba en que en los informes que genera no se muestra la relación de los hipervínculos visitados, por lo que es casi imposible para el usuario extraer conclusiones sobre su funcionamiento. Este mismo problema se da con *Astra SiteManager*, pero no por no listar los enlaces visitados, sino por las razones expuestas en el párrafo anterior.

Por otra parte, las mayores coincidencias se observan entre las sedes web con páginas no dinámicas, es decir, las construidas con el lenguaje HTML, como 1, 5, 15 y 16. Además se da la casualidad de que todas ellas tienen un tamaño pequeño que no supera el centenar de páginas. También esta vez *Astra SiteManager* y *Webcount* son las únicas excepciones que confirman la regla, por los motivos anteriormente explicados.

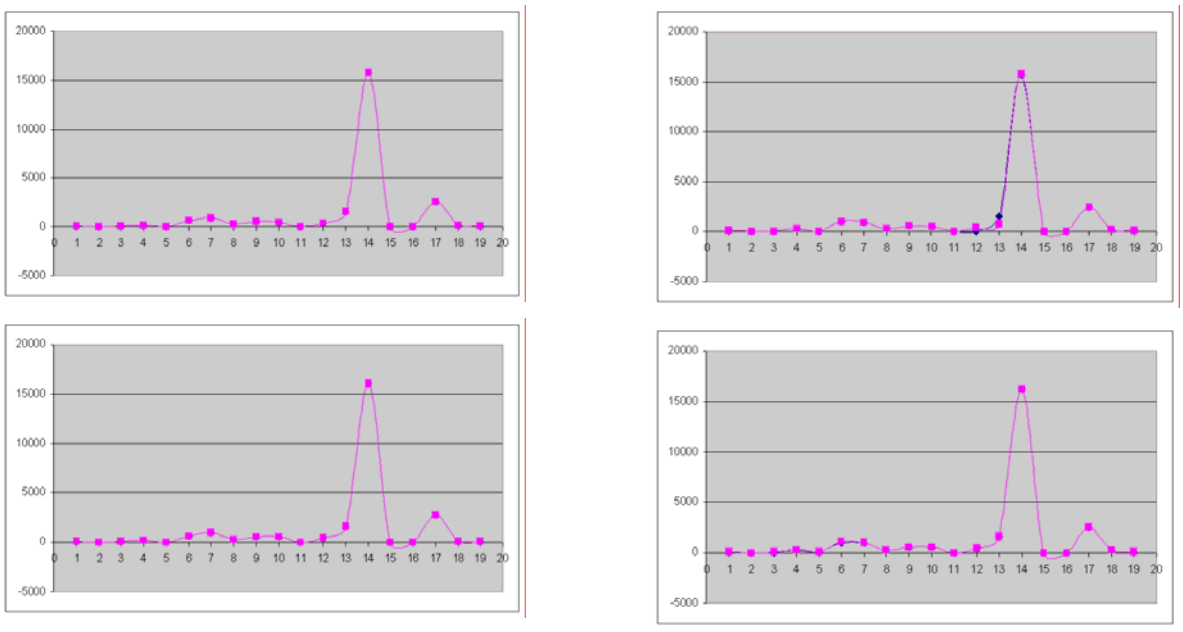
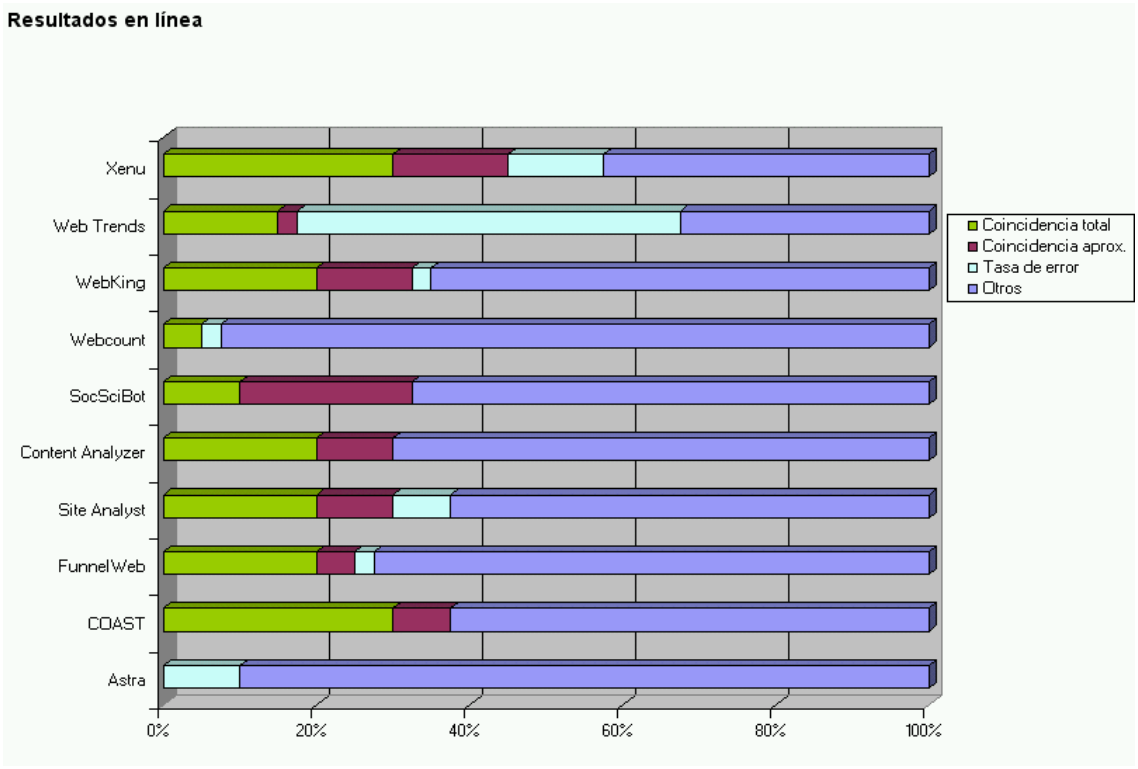
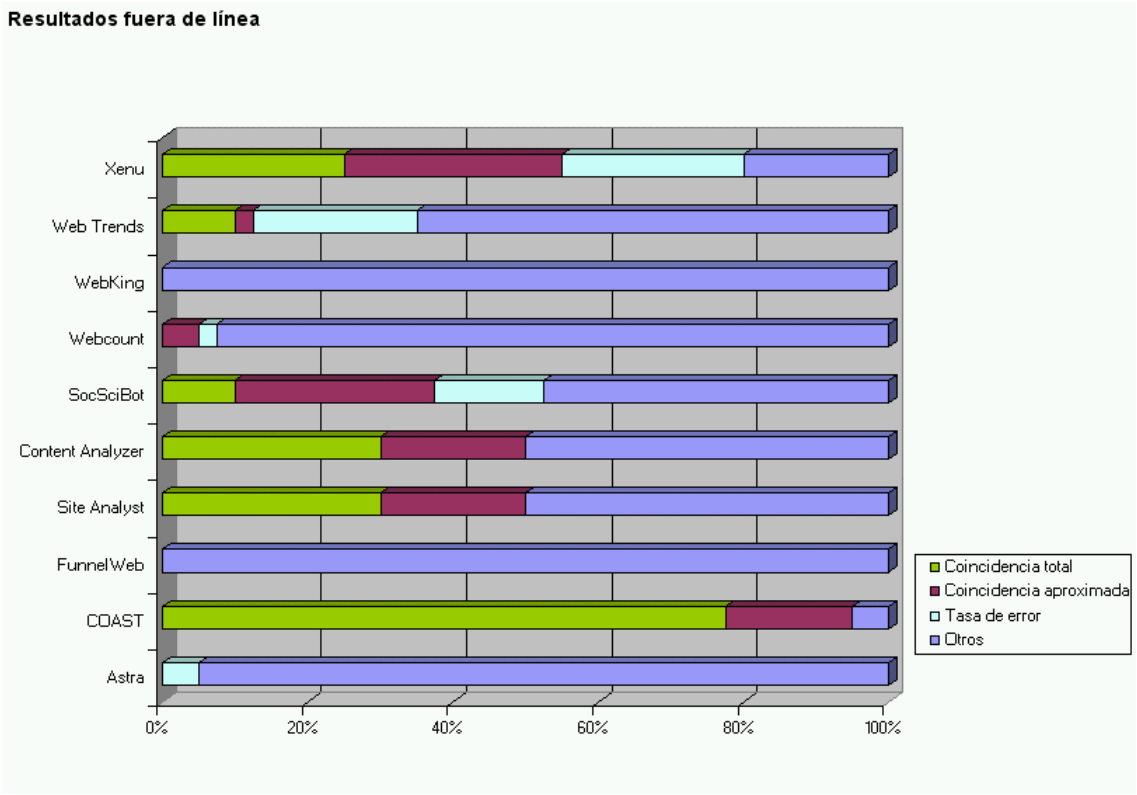


Figura 3.4. Comparación entre los resultados de *Site Analyst* y *Content Analyzer*. Los gráficos superiores corresponden a cada uno de los muestreos en línea, el de la izquierda es el del día 22 y al de la derecha del 29, mientras que los inferiores representan los datos obtenidos a partir de los muestreos fuera de línea. En cada uno de ellos los puntos de forma cuadrada corresponden a *Content Analyzer*, los rombos más oscuros a *Site Analyst*; el eje de las X recoge el número de cada sede web y el de las Y el valor que toman.





Figuras 3.5 y 3.6. Grado de coincidencia entre los resultados obtenidos por cada programa y los volcados de *Teleport Pro*, tanto en línea como fuera de línea.

Al comparar los datos del primer y segundo muestreo, en línea por un lado y fuera por otro, no se perciben apenas oscilaciones temporales, y aquellas que se dan se deben más bien a los cambios naturales que el Web sufre, que parecen ser mayores en las sedes más grandes y generadas con un cierto grado de dinamismo.

Se perciben además grandes similitudes entre los datos de las herramientas de Microsoft, *Site Analyst* y *Content Analyzer*, tal y como se aprecia en los gráficos superiores (figura 3.4). Sólo en el primer muestreo, el del día 22 realizado en línea, se manifiestan resultados dispares para las sedes web números 3, 12 y 13, a pesar de que también hay diferencias, que además se dan en los resultados en línea del 29, menores en las sedes 4, 6, 7, 9, 14 y 18. Resulta significativo que sean precisamente éstas las que incluyen algún otro lenguaje aparte de HTML en su código fuente, lo que indica que ambos programas se comportan de forma distinta ante este tipo de lenguajes. Más adelante se estudiarán los detalles.

Sin embargo en los resultados fuera de línea las coincidencias son casi totales en todas las sedes (gráficos inferiores de la figura 3.4). Esta misma pauta se da en el resto de los programas, que tienen un comportamiento más predecible cuando trabajan fuera de línea.

Esta tendencia es palpable además en las figuras 3.5 y 3.6, que muestran el grado de coincidencia[21] entre los resultados obtenidos por cada software y los ficheros descargados por *Teleport Pro*. Tal y como se puede apreciar, el grado de coincidencia total en los resultados de los muestreos en línea sólo alcanza un porcentaje del 30% en *Xenu* y *COAST WebMaster*, manteniéndose en el resto por debajo del 20%. El grado de coincidencia aproximado, sin embargo, ya hace a algunos programas remontarse hasta cotas superiores al 35%, sin alcanzar nunca el 50%. El caso de *Astra SiteManager* es especial: no se da coincidencia alguna entre sus resultados y los de *Teleport Pro*, debido a que, como se señaló anteriormente, los resultados que se han podido obtener sobre *Astra* se refieren al número total de recursos, no sólo páginas HTML, dentro de la sede.

Pasando al segundo de los gráficos, el relativo a los muestreos fuera de línea, se deduce que la tasa de coincidencia total aumenta para algunos programas, llegando hasta el 75% de coincidencia total con COAST o casi el 30% con *Site Analyst* y *Content Analyzer*, que ganan especialmente en aproximación (hasta casi el 50% en total). Sin embargo otros pierden, como *Webcount*, *Web Trends* y

[21] El grado de coincidencia ha sido calculado mediante la suma de los datos en línea del 22 de octubre y los del 29 por un lado, y de los datos fuera de línea del 22 y 29, por otro.

Xenu, que gana en aproximación pero pierde en coincidencia total y aumenta también el número de errores registrados. Sólo *SocSciBot* se ha mantenido en la misma línea, aunque con más errores también.

En ambos gráficos se demuestra, una vez más, la similitud entre los datos de *Site Analyst* y *Content Analyzer*, casi iguales, aunque en el primero, el relativo a los muestreos en línea, se registra el error de programa en la sede 3.

3.2.2. Algunos interrogantes

Tras esta primera aproximación a los resultados surgen más preguntas que conclusiones. Todas ellas se pueden resumir en una fundamental: ¿cuál es la causa de las grandes diferencias que se han podido constatar en los resultados?. Para abordar esta cuestión se dividirá el problema en dos, teniendo en cuenta el tipo de diferencias: en primer lugar se intentará dar una explicación a las diferencias en los resultados para una misma sede web de cada unos de los programas, y en segundo lugar a las diferencias entre los resultados en línea y fuera de línea.

3.4.2.1. Diferencias entre programas

Las razones de las grandes diferencias en los resultados obtenidos de cada uno de los programas deben buscarse en:

- Las distintas opciones de programa que cada software permite configurar al usuario, y que son por lo tanto controlables pero que es necesario tener en cuenta, ya que pueden hacer variar los resultados. Cuando se realizan estudios con varios programas, como es el caso, se pueden intentar homogeneizar los resultados seleccionando opciones similares, pero dicha homogeneidad nunca será total, debido a las grandes diferencias entre programas, que ya quedaron exhaustivamente explicadas en el capítulo anterior.

- Las características del software ocultas por el secreto comercial, que son aquellos rasgos que influyen en el funcionamiento de los programas y que el programador diseña pero que quedan completamente fuera del alcance del usuario.

Sin embargo algunas de estas últimas características se pueden intuir, en mayor o menor medida dependiendo del grado de opacidad o transparencia de cada programa, a través de un análisis de los resultados. Este es el punto en que se incidirá a continuación, ya que resulta clave para entender el comportamiento de cada programa y los resultados que cabe esperar de cada uno de ellos.

Tal y como se ha constatado, las mayores diferencias entre el software se manifiestan especialmente en las sedes web dinámicas, mientras que en las estáticas parece haber mayor uniformidad de criterios a la hora de ser analizadas. Por ello son este primer tipo de sedes web en las que nos vamos a detener a continuación con el fin de comprender mejor cómo se comportan ante ellas los diferentes programas.

3.2.2.2. Las grandes diferencias: páginas dinámicas

Los objetos dinámicos, por oposición a los estáticos, son generados por un servidor como respuesta a una petición de un cliente, lo que implica una necesidad de información por parte de este. Sin embargo, al no poder ningún *crawler* completar esta tarea (excepto con la información con la que el propio explorador web identifica por defecto al usuario), se provoca una situación de ausencia de interacción con el usuario y por lo tanto un error por parte del servidor.

A esta problemática, identificada por Cothey, se añade la formación de lo que él denomina agujeros negros. Al modificar algunos objetos dinámicos su URL con un apéndice indicando el momento en el que se ha accedido a ellos, y tener por lo tanto una misma página URLs diferentes según el momento en el que se accede a ella, el *crawler* no es capaz de diferenciarlas y las analiza una y otra vez recuperándolas así un número arbitrario de veces. Esta situación sólo puede evitarse limitando el número de URLs a recoger o la profundidad del análisis.

El empleo de lenguajes dinámicos en el Web ha ido en aumento en los últimos años, tal y como señalan Arroyo, Pareja y Aguillo (2003) a partir de una muestra recogida entre los años 2001 y 2002[22]. Entre ambos años el uso de lenguajes de programación dinámicos del lado del servidor creció en un 2’37 % en el ámbito académico universitario de la Europa de los 15, mientras que para lenguajes que se ejecutan del lado del cliente supuso más de un 26% Esto constituye, por lo tanto, un reto para este tipo de software.

[22] Resultados extraídos de una muestra de 25.416 sedes web del ámbito académico universitario de la Europa de los quince.

Dejando a un lado las sedes web con ficheros Flash y JavaScript, puesto que ya se demostró en el capítulo anterior la imposibilidad de que ningún *crawler* se adentre en ellos en busca de hipervínculos, se estudiarán en adelante aquellos ejemplos de sedes web que contienen lenguajes de programación que se ejecutan del lado del servidor (tabla 3.3). Tres de ellos han sido empleados en las páginas de algunas de las sedes web de la muestra: ASP (*Active Server Pages*), PHP (*Hipertext Preprocesor*) y CGI (*Common Gateway Interface*), de los cuales el segundo de ellos no ha sido empleado realmente en la construcción de la sede web, sino que aparece enlazado y puede por lo tanto servir para extraer algunas conclusiones.

PROGRAMACIÓN WEB ESTÁTICA	HTML, HTM...		(1) www.[illegible]c.csic.es/cybermetrics
			(3) www.upct.es/~de
			(5) www.shef.ac.uk/dentalschool
			(9) www.dsg.unito.it
			(10) www.kun.nl/phil
			(13) www.montefiore.ulg.ac.be
			(15) www.uni-saarland.de/fak8/iaua
			(16) www.medizin.uni-greifswald.de/humangen
PROGRAMACIÓN WEB DINÁMICA	CLIENT-SIDE	Flash	(2) www.ite.upv.es
		JavaScript	(7) www.pem.cam.ac.uk
			(8) www.teipat.gr
			(11) www.mat.chalmers.se
			(14) wwwai.wu-wien.ac.at
			(19) www.usal.es
	SEVER-SIDE	ASP	(4) www.ceu.es
		PHP	(18) www.etsia.upv.es
		CGI	(6) www-cryst.bioc.cam.ac.uk
			(12) cst.dk
			(14) wwwai.wu-wien.ac.at
			(17) www.math.jyu.fi

Tabla 3.3. Clasificación de las sedes web de la muestra según los lenguajes de pr[illegible]nación empleados.

Para ello se ha recurrido a los informes generados por cada uno de los programas, la mayoría de los cuales incluyen la estructura hipertextual de los sitios analizados, lo que permite comprobar qué tipo de ficheros han sido examinados. Esta información no es facilitada por *Webcount*, que se limita, como ya se apuntó, a ofrecer las estadísticas de la sede, ni por *COAST WebMaster*, que tan sólo genera listados de aquellas páginas que plantean algún tipo de problema, lo cual supone una información, aunque pobre, para extraer algunas conclusiones. *Astra SiteManager*, por su parte, sí ofrece listados de los ficheros reconocidos, pero los muestra mezclando enlaces internos con externos sin orden alguno, lo que dificulta en gran medida su exploración.

También los ficheros descargados por *Teleport Pro* serán analizados para comprobar el alcance de este software, ya que de él dependen los resultados procedentes de los muestreos *offline* de otros programas. Es decir, si *Teleport Pro* no es capaz de volcar determinados tipos de ficheros, entonces los *crawlers* lanzados sobre dichas descargas no podrán analizarlos en los muestreos *offline* por razones obvias. Este punto servirá como base para el desarrollo del siguiente apartado.

Durante el examen de los informes generados se repite la problemática identificada por Cothey como *ausencia de interactividad con el usuario* en aquellas sedes web en las que es necesario introducir unos parámetros para generar contenidos mediante la interrogación a una base de datos. Buenos ejemplos se encuentran en las sedes cst.dk y www-cryst.bioc.cam.ac.uk, que utilizan el sistema CGI, en las que se encuentran sendas casillas de búsqueda. Cuando se interroga a cualquiera de ellas sin introducir parámetro alguno el resultado es una página de error. Esto se plasma en los informes obtenidos por *COAST WebMaster*, que incluye las páginas en las que se insertan todas estas casillas de búsqueda en el listado de vínculos rotos.

En el caso de la primera sede —cst.dk—, los resultados cuantitativos que los distintos programas arrojan no son muy diferentes entre sí, excepto para *SocSciBot*, que suma además las páginas alojadas en otros sub-dominios (como www.cst.dk) por el truncamiento que, como ya se explicó en el capítulo 2, realiza por defecto en la parte izquierda de la URL.

Con la segunda sede web, www-cryst.bioc.cam.ac.uk, sucede que, aparte de ser imposible para ningún *crawler* generar contenidos a partir de la interrogación a una base de datos, se encuentran hipervínculos directos a páginas que son generadas dinámicamente por lo que los *crawlers* son capaces de acceder a ellas y sumarlas en sus recuentos. Programas como *FunnelWeb*, *SocSciBot*, *WebTrends* y *Xenu* (cuyo resultado es tan elevado, si se compara con el resto, porque ha incluido en el recuento páginas externas a la sede) son capaces de reconocer ese tipo de páginas CGI.

Por otro lado, las sedes que emplean lenguajes de programación como ASP o PHP pueden ser *mapeadas* en mayor o menor grado dependiendo de la capacidad del software para reconocerlas o no, siempre que sus páginas sean enlazadas desde otras y no se requiera una mayor interactividad por parte del usuario. Así, se podrá clasificar al software analizado según su cobertura, entendida como la capacidad para reconocer y analizar un mayor o menor número de páginas web, en tres grupos:

1. Amplia cobertura, que incluiría a aquellos programas que son capaces de adentrarse en la estructura del sitio, independientemente de cuál sea el lenguaje de programación empleado en su diseño.
2. Media cobertura, en referencia al software que sólo permite analizar una mínima parte de las páginas en ciertos lenguajes de programación web, normalmente las de inicio.
3. Falta de cobertura, cuando no se han encontrado indicios de ninguna página en determinados lenguajes. El único caso hallado que se ajusta a estos parámetros es *Teleport Pro* al analizar pasarelas.
4. Sin clasificar, por falta de evidencias que indiquen su posición, como ocurre con *Webcount* —que no ofrece listado alguno de las páginas analizadas— u otros programas de los que no se han podido obtener datos para sedes concretas por errores de programa. A veces incluso, a pesar de disponer de un listado de las páginas visitadas y no haberse producido error de programa alguno, no se ha encontrado ninguna evidencia que lleve a situarlo en un grupo u otro. En la tabla 3.4, que recoge esta clasificación, han sido marcados con un interrogante.

Sin embargo el software parece comportarse de forma diferente según el lenguaje al que se enfrenta, por lo que hacer este tipo de agrupación no resulta fácil excepto en unos pocos casos —*FunnelWeb*, *SocSciBot* y *Xenu*—. Por ello, al realizar la mencionada clasificación resulta necesario tener además en cuenta esta tercera variable.

	ASP	*PHP*	*CGI*
Astra SiteManager	Alta	Media	?
Site Analyst	Media	Media	Media
Content Analyzer	Media	Media	Media
COAST WebMaster	Alta	?	Media
Funnel Web	Alta	Alta	Alta
SocSciBot	Alta	Alta	Alta
WebKing	Alta	?	Media
Web Trends	Media	?	Alta
Webcount	?	?	?
Xenu	Alta	Alta	Alta
Teleport Pro	Alta	Alta	Ninguna

Tabla 3.4. Cobertura del software según el lenguaje de programación.

Tal y como se puede observar, algunos programas, como *FunnelWeb*, *SocSciBot* y *Xenu*, cubren ampliamente las páginas de las sedes analizadas que emplean cualquiera de los tipos de

lenguaje dinámico representados, por lo que resultan muy apropiados para el análisis de ese tipo de objetos. Otros programas sin embargo sólo alcanzan una cobertura media para cualquiera de los lenguajes analizados. Es el caso de las herramientas desarrolladas por Microsoft, *Site Analyst* y *Content Analyzer*, que, como se recordará, no han sido actualizadas desde 1997 y 1998 y parecen estar quedándose atrás en este sentido. Y por último existe un tercer grupo de programas cuya cobertura es amplia para algún lenguaje en concreto pero media para otros, como sucede con *Astra SiteManager*, *COAST WebMaster*, *WebKing*, *WebTrends* o *Teleport Pro*.

Un importante aspecto a tener en cuenta es el apartado en el que se incluye cada recurso, es decir, cómo es considerado por parte de cada programa. Es el caso de las páginas PHP, que no han sido definidas internamente por *Site Analyst* y *Content Analyzer* como extensiones reconocidas y, por lo tanto, al enfrentarse a ellas son incluidas en el apartado *Otros recursos*, un cajón de sastre que recoge los documentos con extensiones que el programa no identifica. Esto sucede también con *Astra SiteManager*, que tiene un apartado llamado *Dynamic* bajo el cuál no queda muy claro qué tipos de ficheros se agrupan exactamente por la opacidad, que ya ha sido señalada en otras ocasiones, de este programa. La solución a esta disgregación de los resultados está, pues, en sumar estos datos al número total de páginas en este caso.

Desde el punto de vista de los resultados cuantitativos, y teniendo en cuenta que estos no coinciden en ninguno de los casos para este tipo de sedes, e incluso varían enormemente (en especial en los muestreos en línea) y que la estructura hipertextual de las sedes web analizadas es desconocida para nosotros, resulta prácticamente imposible determinar exactamente cuál es el software que más se acerca a la realidad de cada una de las sedes. Sin embargo, sí es posible hacerse una idea, en términos generales, de la cobertura que pueden abarcar, excepto en los casos en que no hay evidencias por falta de datos debido a errores de programa o ausencia de pruebas.

Según todos los indicios y tal y como se comprueba en la tabla, podría decirse que son *Funnel Web*, *SocSciBot* y *Xenu* las soluciones más apropiadas para el análisis de sitios web dinámicos por la cobertura que de ellos hacen, independientemente de otras limitaciones que puedan presentar. Sin embargo existen otros programas que pueden también ser empleados, siempre que sean tenidas en cuenta sus restricciones.

3.2.2.3. Diferencias entre los resultados online y offline

Las diferencias entre los resultados de los muestreos *online* y *offline*, de los que ya se habló anteriormente, se ponen de manifiesto en la figura 3.7, donde se puede apreciar cómo las coincidencias sólo alcanzan un 42% en el mejor de los casos (*Xenu*) y un 21% en el peor (*Astra SiteManager*), dándose casi siempre dichas coincidencias en las sedes web con lenguajes de programación estáticos únicamente (*véase* tabla 3.4), mientras que las divergencias suponen, por lo tanto más de las mitad de los casos. Pero si se profundiza en el sentido en el que dichas divergencias se dan, es decir, si en los muestreos *online* los resultados son mayores cuantitativamente que en los muestreos *offline*, se comprueba que en algunos programas priman los primeros, mientras que en otros predominan los segundos, pero sorprendentemente no siempre en el mismo sentido para una misma sede.

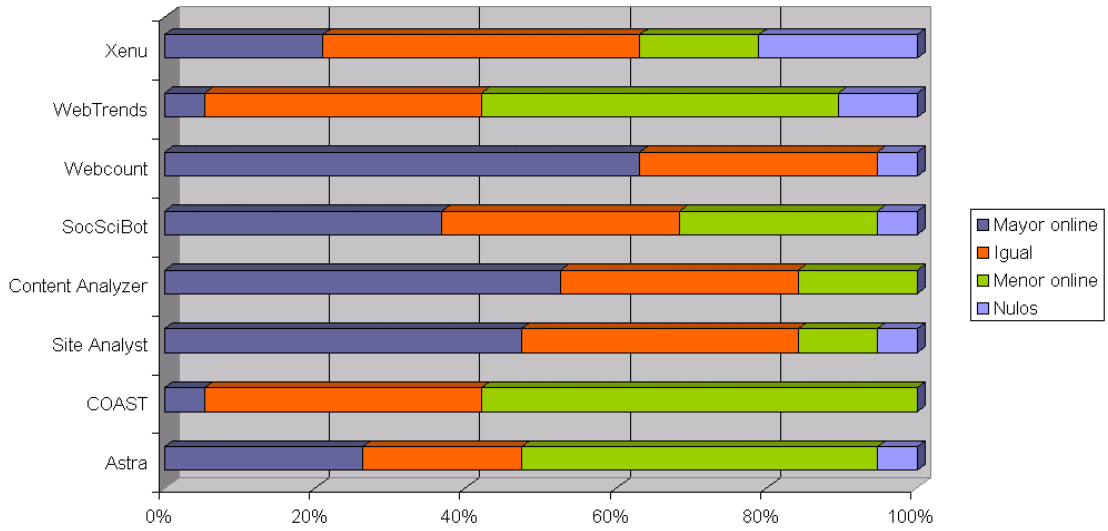


Figura 3.7. Tasa de coincidencia (%) entre los resultados de los muestreos *online* y *offline*. Esta tasa se ha calculado a partir de la diferencia entre la media de los valores de los dos muestreos en línea, por un lado, y la media de ambos muestreos fuera de línea, por otro. Los valores nulos corresponden a los errores de programa.

En el apartado anterior ya se adelantó cómo en los datos *offline* los resultados obtenidos dependían en gran medida de la forma en que *Teleport Pro* se comporta, que pueden originar discordancias entre las características reales de sede web y el volcado de *Teleport Pro*, provocando así diferencias entre los resultados en línea y fuera de línea. Esta conclusión se desprende de las figuras 3.5 y 3.6, en las que se exponía el grado de coincidencia entre los datos de los muestreos *online* por

una parte y *offline* por otra, con los ficheros descargados mediante *Teleport*, y que se comprueba es bastante mayor en los segundos. También se observa una mayor coincidencia entre los resultados de las sedes web sin programación dinámica (anexo 3).

Se puede concluir, por lo tanto, que las diferencias entre los resultados de los muestreos en línea y fuera de línea están motivados por dos factores:

- La forma en que *Teleport Pro* se comporta, que es difícil de detallar sin conocer la estructura real de la sedes de la muestra, pero que, con la información con que se cuenta se puede pensar que no plantea problema alguno con los lenguajes analizados en el apartado anterior, excepto con CGI.
- El lenguaje de programación empleado en el diseño de la sede web en cuestión, puesto que el grado de coincidencia aumenta siempre que se enfrenta a sedes web con páginas dinámicas.

Capítulo 4. Conclusiones

4.1. Evaluación del software

En este apartado se evaluará el software examinado en función de una serie de criterios enfocados a los objetivos del trabajo, esto es, su uso con fines cibermétricos. Estos criterios son los siguientes:

- Consumo de recursos, es decir, requerimientos del sistema sobre el que funcionan.
- Cobertura de páginas dinámicas, tema que acaba de tratarse en el apartado anterior. Para su evaluación se emplearán los estándares definidos en la tabla 3.4.
- Dificultades encontradas en su puesta en marcha y funcionamiento.
- Entorno gráfico, que sirve de apoyo a la información textual o numérica y que facilita su visualización.
- Limitaciones a su uso.
- *Output* o resultados producidos a partir del análisis.
- Opciones de programa, expuestas en el capítulo 2.
- Transparencia u opacidad de los resultados, o, lo que es lo mismo, facilidad para estudiar el funcionamiento del programa en cuestión. Para ello es de gran ayuda disponer de un esquema de la estructura hipertextual de la sede web analizada.
- Utilidades adicionales, que pueden servir en el desempeño de otras tareas.

Como apoyo para la evaluación, a modo de resumen y con el fin de facilitar las comparaciones, se ha incluido en el anexo 4 una tabla con anotaciones básicas sobre cada programa en este sentido.

Astra SiteManager, muy similar a *Site Analyst* y *Content Analyzer* en lo que al entorno gráfico se refiere, resulta un programa un tanto confuso en la exposición de resultados, ya que no queda claro qué se recoge bajo cada epígrafe ni la forma en que esta herramienta se comporta. Por otra parte, la información ofrecida en sus informes resulta un tanto pobre, tan sólo muestra datos generales de páginas, recursos y objetos dinámicos, sin desglosar los tipos que los integran. Por lo demás, *Astra* es un programa en la línea de otros gestores de sitios web.

Por su parte, *Site Analyst* y *Content Analyzer* son casi idénticos en la forma en que se comportan y los resultados que arrojan. A pesar de su relativa antigüedad, que les ha hecho quedarse

atrás en lo que al análisis de sedes web dinámicas se refiere —que puede paliarse en la etapa de recogida de datos extrayendo determinados ficheros de ciertas categorías para después sumarlos a otras más adecuadas a las necesidades del usuario— los informes que generan son bastante satisfactorios en cuanto a su facilidad para ser exportados a cualquier otro formato, los datos que contienen, y el hecho de que se enumeren los enlaces analizados, lo que les convierte en herramientas con la transparencia suficiente como para adentrarse en la forma en la que se comportan.

La cantidad de opciones de programa que *COAST WebMaster* permite seleccionar al usuario, que le dotan de una gran flexibilidad, los datos que arroja como resultado de su análisis —HTML dinámico, número de ficheros Flash, etc.—, que son únicos en comparación con otros programas comerciales de sus características, y la forma, previsible para el usuario, en la que analiza los sitios especificados —es el que más se aproxima a los datos del volcado realizado con *Teleport Pro*, con un 80% de coincidencia en los muestreos fuera de línea— los convierten en un programa con un gran potencial para la obtención de datos para su uso con fines cibermétricos. Sin embargo, dos grandes inconvenientes empañan su utilidad: por un lado se echan en falta datos sobre ciertos tipos de ficheros como imágenes y ficheros ricos, y por otro, listados de las URLs analizadas, que ayudan a comprender el funcionamiento del programa y de gran utilidad en la realización de estudios sobre los sitios enlazados desde la sede objeto de análisis.

El gran punto a favor de *Funnel Web Profiler* es la cantidad de gráficos que genera y que son de una gran utilidad para visualizar los resultados; especial atención merece el mapa de palabras contenidas en cada sede web, en forma de mapa topográfico, que supone un útil análisis de contenidos para algunos casos. Pero en el platillo contrario de la balanza hay que situar sus grandes limitaciones: el número de items que es capaz de analizar (tan sólo 1.000 para esta versión) y los escasos datos cuantitativos que ofrece.

Pero sin duda alguna uno de los rasgos más deseables para este tipo de software es la capacidad para analizar páginas web dinámicas de una forma lo suficientemente apropiada. En este sentido son *Funnel Web*, *SocSciBot* y *Xenu* las herramientas más apropiadas. De ellas esta última, *Xenu*, es un software de gran sencillez, fácil de entender y emplear por el usuario, y cuyos resultados son muy completos y permiten una gran flexibilidad en su interpretación, ya que son agrupados por tipos MIME.

SocSciBot, por su parte, es una buena herramienta para el análisis de sedes web dinámicas y además aporta los datos necesarios para el análisis de las relaciones de la sede web en cuestión con

otros sitios fuera de ella y el estudio de contenidos, lo cual, unido a la sencillez de su interfaz y los objetivos con los que fue diseñado, la convierten en uno de los programas más adecuados para los propósitos establecidos. En su contra cabe señalar, sin embargo, la limitación en lo que al número de enlaces por analizar se refiere y la forma diferente que tiene de separar las páginas fuera y dentro de la sede (incluyendo entre estas últimas aquellas que forman parte de otras sedes alojadas en un sub-dominio de la primera). Esta última cuestión debe ser tenida siempre bien presente a la hora de interpretar los resultados obtenidos, puesto que la concepción es bien distinta de otras herramientas.

La mayor objeción a *Web Trends* es el gran número de errores de programa que genera, que llega a superar el 76% de los lanzamientos en el experimento llevado a cabo, y que constituyen un grave inconveniente. Algo similar cabe decir sobre *Webcount*, una herramienta en la que es difícil confiar por los datos que arroja, demasiado diferentes al resto, incluso al analizar sedes web estáticas, y completa falta de transparencia.

Y ya por último, *Web King* es, a pesar de las limitaciones propias de la versión empleada, una herramienta bastante completa y con unos resultados óptimos, que se disponen, al igual que lo hace *Xenu*, por tipo MIME. Quizás su mayor carencia sea la imposibilidad para desglosar la estructura hipertextual de las sedes web analizadas.

4.2. Discusión y futuras líneas de investigación

En el presente trabajo se ha profundizado en el funcionamiento de varios *crawlers* académicos y comerciales, lo que sienta las bases para su empleo en futuras investigaciones en las que se pretenda extraer datos con fines cibermétricos. Para ello simplemente es necesario tener en cuenta las diferencias que se dan entre los distintos programas —debidas no sólo a sus características, sino también a su comportamiento ante diferentes situaciones—, así como las limitaciones y puntos fuertes expuestos sobre cada uno de ellos, y muy especialmente los resultados que se pueden esperar.

Tal y como se ha observado, el principal problema detectado es la forma que tienen este tipo de *crawlers* de enfrentarse a las sedes web con páginas dinámicas, que parecen proliferar con el tiempo, siendo cada vez más frecuentes en el entorno académico, donde cada vez se presta mayor atención al diseño, especialmente en aquellas instituciones más grandes como universidades o grandes centros de investigación. De esta manera, en el momento en que estas u otras páginas dejan de ser visibles para los *crawlers* académicos y comerciales, pasan a formar parte de lo que se ha dado en

llamar el *Web invisible*, si se aplica la misma terminología empleada para los motores de búsqueda, haciéndose así necesario definir de la forma más aproximada posible en qué consiste esta parte. Esta cuestión, junto con el uso que se pueda hacer de estos programas y el estudio del funcionamiento de este software ante otros tipos de ficheros, tales como imágenes, fichero multimedia, ficheros ricos, etc, diferentes de las páginas web, constituirán el punto de partida para futuras investigaciones.

Referencias

Abraham, R. H. (1996) Webometry: measuring the complexity of the World Wide Web. World Futures 50: 785-791. Disponible en: http://www.ralph-abraham.org/articles/MS%2385.Web1/[23].

Abraham, R. H. (1998) Webometry: measuring the synergy of the World Wide Web. Biosystems. 46(1-2):209-212.

Abraham, R. H.; Foresta, D. (1996) Webometry: chronotopography of the World Wide Web. Disponible en: http://www.ralph-abraham.org/articles/MS%2389.Web3/.

Adamic, L. A. (1999) The small world Web. Proceedings of ECDL'99. Lecture Notes in Computer Science. 1696:443-452.

Adamic, L. A.; Huberman, B. A. (2000) Power-law distribution of the World Wide Web. Science. 287:2115-2116.

Adamic, L. A.; Huberman, B. A. (2002) Zipf's law and the Internet. Glottometrics. 3:143-150.

Adams, K.; Gilbert, N. (2003) Indicators of intermediaries' role and development. Deliverable 6.2. Proyecto EICSTES.

Aguillo, I. (1997) A new electronic journal devoted to Scientometrics. 6th Conference International Society for Scientometrics and Informetrics; Jerusalem.

Aguillo, I. (1998a) Hacia un concepto documental de sede web. El Profesional de la Información. 7(1-2):45-46.

Aguillo, I. (1998b) Herramientas de segunda generación. Anuario SOCADI. 85-112.

Aguillo, I. (2000a) Indicadores: hacia una evaluación no objetiva (cuantitativa) de sedes web. Jornadas Españolas de Documentación - FESABID 2000. 7:233-248.

Aguillo, I. (2000b) A new generation of tools for search, recovery and quality evaluation of World Wide Web medical resources. Online Information Review. 24(2):138-143.

Albert, R.; Jeong, H.; Barabasi, A. L. (1999) Diameter of the World Wide Web. Nature. 401(6749):130-131.

Almind, T.; Ingwersen, P. (1996) Informetric analysis on the World Wide Web: a methodological approach to "internetometrics". Centre for Informetric Studies, Royal School of Library and Information Science (CIS Report 2).

Almind, T. C.;Ingwersen, P. (1997) Informetric analyses on the World Wide Web: metodological aproaches to "Webometrics". Journal of Documentation. 53(4):404-426.

[23] Todas las URLs a las que se remite en esta lista de referencias han sido verificadas y actualizadas en julio de 2004.

Alonso, J. L. (2002) Cibermetría: análisis de los dominios web españoles. Salamanca: Universidad de Salamanca. [Tesis doctoral].

Alonso, J. L.; Figuerola, C. G.; Zazo, A. F. (2004) Cibermetría: nuevas técnicas de estudio aplicables al Web. Gijón: TREA.

Arroyo, N.; Pareja, V. M. (2003) Metodología para la obtención de datos con fines cibermétricos. III Taller de Indicadores Bibliométricos; Madrid. Disponible en: http://internetlab.cindoc.csic.es/varios/Metodolog%EDa%20datos%20ciberm%E9tricos.pdf.

Arroyo, N.; Pareja, V. M.; Aguillo, I. (2003) Description of web data in D3.1. Proyecto EICSTES. Deliverable 3.2.

Bailey, P.; Craswell, N.; Hawking, D. (2000) Chart of darkness: mapping a large intranet. CSIRO CMIS Technical Report.

Bar-Ilan, J. (1997) The "mad cow disease", usenet groups and bibliometric laws. Scientometrics. 39(1):1997.

Bar-Ilan, J. (1998) Search engine results over time: a case study on search engine stability. Cybermetrics. 2/3(1). Disponible en: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html.

Bar-Ilan, J. (2001) Data collection methods on the Web for informetric purposes: a review and analysis. Scientometrics. 50(1):7-32.

Bar-Ilan, J.; Peritz, B. C. (2002) Informetric theories and methods for exploring the Internet: an analytical survey of recent research literature. Library Trends. 50(3):371-392.

Barabasi, A. L. (2003) Linked: how everything is connected to everything else and what it means for business, science, and everyday life. New York: Plume.

Bergman, M. K. (2001) White paper. The deep Web: surfacing hidden value. [Página web] Disponible en: http://www.brightplanet.com/pdf/deepwebwhitepaper.pdf.

Bharat, K.; Broder, A. (1998) Measuring the Web. [Página web] Disponible en: http://www.research.compaq.com/SRC/whatsnew/sem.html.

Björneborn, L. (2002) Small-world link structures on the Web. [Presentación] Disponible en: http://www.db.dk/lb/2002smallworld.pps.

Björneborn, L. (2004) Small-world link structures across an academic web space: a library and information science approach. Copenhagen: Department of Information Studies, Royal School of Library and Information Science. [Tesis doctoral] Disponible en: http://www.db.dk/lb/phd/phd-thesis.pdf.

Bossy, M. (1995) The last of the litter: "Netometrics". Solaris Information Communication. (2):245-250.

Boudourides, M.; Antypas, G. (2002) A simulation of the structure of the World Wide Web.

Sociological Research Online. 7(1).

Boudourides, M. A.; Sigrist, B.; Alevizos, P. D. (1999) Webometrics and the self-organization of the European information society. Proyecto SOEIS, Rome meeting. Disponible en: http://hyperion.math.upatras.gr/webometrics/.

Bradford, S. C. (1934) Sources of information on specific subjects. British Journal of Engineering. 137:85-85.

Bray, T. (1996) Measuring the Web. Fifth International World Wide Web Conference; Paris.

Brin, S.; Page, L. (1998) The anatomy of a large-scale hypertextual Web search engine. Computers Networks and ISDN Systems. 30(1-7):107-117.

Broder, A.; Kumar, R.; Maghoul, F.; Raghavan' P.; Rajagopalan, S.; Stata, R.; Tomkins A.; Wiener, J. (2000) Graph structure in the Web. Computer Networks. 33(1-6):309-320.

Callon, M.; Courtial, J. P.; Turner, W. A.; Bauin, S. (1983) From translation to problematic networks: an introduction to co-word analysis. Social Science Information. 22:191-235.

Chakrabarti, S.; Joshi, M. L.; Punera, K.; Peenock, D. M. (2002) The structure of broad topics on the Web. WWW2002 Conference.

Cothey, V, (2003) Web-crawling reliability. 9th ISSI International Conference on Scientometrics and Informetrics; Beijing.

Dean, J.; Hanzinger, M. (1999) Finding related pages in the World Wide Web. Computers Networks and ISDN Systems. 31:389-401.

Development of Web-Indicators (2003). Proyecto EICSTES, Deliverable 8.1.

Diccionario de Internet y redes de Microsoft. (2003) Aravaca: McGrw-Hill. Interamericana de España.

Drineas, P.; Krishnamoorthy, S.; Sofka, M. D.; Yener, B. (2004) Studying e-mail graphs for intelligence monitoring and analysis in the absence of semantic information. Symposium on Intelligence and Security Informatics (ISI'04), junio.

Egghe, L. (1997) Fractal and informetric aspects of hypertext systems. Scientometrics. 40(3):455-464.

Egghe, L. (2000) New informetric aspects of the Internet: some reflections, many problems. Journal of Information Science. 26(5):329-335.

Faba, C. (2002) Análisis cibermétrico de la información web: el caso de Extremadura en Internet.Granada: Universidad de Granada. [Tesis doctoral].

Faba, C.; Guerrero, V. P.; Moya, F. (2003) "Sitation" distributions and Bradford's law in a closed web space. Journal of Documentation. 59(5):558-580.

Faba, C.; Guerrero, V. P.; Moya, F. (2004) Fundamentos y técnicas cibermétricas. Mérida: Junta de Extremadura.

Garfield, E. (1976) A bibliometric analysis of references. Journal Citation Reports. Annual V.9. ed..

Filadelfia: Institute for Scientific Information.

Hispabyte.com. [Página web] Disponible en: http://hispabyte.com/.

Ingwersen, P. (1998) The calculation of Web Impact Factor. Journal of Documentation. 54(2).

Internet Software Consortium (2000) Internet Domain Survey [Página web].Disponible en: http://www.isc.org/index.pl?/ops/ds/hosts.php.

Kessler, M. M. (1963) Bibliographic coupling between scientific papers. American Documentation. 1963; 14:10-25.

Kleinberg, J. M.; Lawrence, S. (2001) The structure of the Web. Science. 294:1849-1850.

Koehler, W. (1999a) An analysis of web page and web site constancy and permanence . Journal of the American Society for Information Science and Technology. 50(2):162-180.

Koehler, W. (1999b). Digital libraries and World Wide Web sites and page persitence. Information Research. 4(4).

Koehler, W. (2000) Web page change and persistence. Journal of the American Society for Information Science and Technology. 53(2):162-171.

Kohonen, T. (1995) Self-organizing maps. Berlin: Springer.

Kot, M.; Silverman, E.; Berg, C. A. (2003) Zipf's law and the diversity of biology newsgroups. Scientometrics. 56(2):247-257.

Kugiumtzis, D.; Boudourides, M. A. (1998) A chaotic analysis of Internet ping data: just a random number generator?. SOEIS project, Bielefeld conference.

Kumar, R.; Raghauam, P.; Rajagopalan, S.;Tomkins, A. (1999) Crawling the Web for emerging cyber-communities. Computer Networks. 31(11):1481-1493.

Larsen, B. (2002) Exploiting citation overlaps for information retrieval: generating a boomerang effect from the network of scientific papers. Scientometrics. 54(2):155-178.

Larson, R. R. (1996) Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of the cyberspace. Proceedings of the 59th Annual Meeting; Baltimore, Maryland.

Lavoie, B.; Nielsen, Henrik F. (eds.). (2002). Web characterization terminology and definitions sheet. World Wide Web Consortium. [Página web] Disponible en: http://www.w3.org/1999/05/WCA-terms/.

Lawrence, S.; Giles, C. L. (1998) Searching the World Wide Web. Science. 280:98-100.

Lawrence, S.; Giles, C. L. (1999) Accesibility of information on the Web. Nature. 400:107-109.

Leydesdorff, L.; Curran, M. (2000) Mapping university-industry-government relations on the Internet: the construction of indicators for a knowledge-based economy. Cybermetrics. 4(1).

Li, X. (2003) A review of the development and application of the Web Impact Factor. Online Information Review. 2003; 27(6):407-417.

Li, X.; Thelwall, M.; Musgrove, P.; Wilkinson, D. (2003) The relationship between the WIFs or

inlinks of computer science departments in UK and their RAE ratings or research productivities in 2001. Scientometrics. 57(2):239-255.

Lotka, A. J. (1926) The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences. 16:317.

Maltrás, B. (2003) Los indicadores bibliométricos: fundamentos y aplicación al análisis de la ciencia. Gijón: TREA.

Matsumoto, Y. (2002) Ruby. [Página web]. Disponible en: http://www.ruby-lang.org/en/.

McKiernan, G. (1996) CitedSites (sm): citation indexing of web resources. [Página web] Disponible en http://www.public.iastate.edu/~CYBERSTACKS/Cited.htm.

Menczer, F. (2004) Lexical and semantic clustering by web links. JASIST. [Por publicar] Disponible en: http://informatics.indiana.edu/fil/Papers/JASIST-04.pdf.

Moore, A. and Murray, B. H. (2000) Sizing the Web. Cyveillance, Inc. White Paper. Disponible en: http://www.cyveillance.com/web/downloads/Sizing_the_Internet.pdf.

Facil Ayan, N.; Li,W.; Kolak, O. (2002) Automating extraction of logical domains in a web site. Data & Knowledge Engineering. 43(2):179-205.

Notess, G. R. (2003) Search engine statistics: database total size estimates. [Página web] Disponible en: http://www.notess.com/search/stats/sizeest.shtml.

Okubo, Y. (1997) Bibliometric indicators and analysis of research systems. STI working papers. OECD.

Pant, G.; Srinivasan, P.; Menczer, F. (2004) Crawling the Web. En: Levene, M.; Poulovassilis, A., eds.: Web Dynamics. Springer.

Pareja, V. M.; González, A.; Aguillo, I. (1999) Ciencia y tecnología españolas en Internet: valoración a través de la presencia de organismos públicos españoles y de sus revistas electrónicas. Arbor; CLXII(639):367-390.

Polanco, X.; François, C.; Lamirel, J. C. (2001) Using artificial neural networks for mapping of science and technology: a multi-self-organizing-maps approach. Scientometrics. 51(1):267-292.

Pozo, J. R. (2001) Traducción de la especificación HTML 4.01 al castellano. [Página web] Disponible en: http://html.conclase.net/w3c/html401-es/progreso.html.

Price, D. J. de S. (1963) Little science, big science. New York: Columbia University Press.

Pritchard, A. (1969) Statistical bibliography or bibliometrics?. Journal of Documentation. 24:348-349.

Rodríguez i Gairín, J. M. (1997) Valoración del impacto de la información en Internet: Altavista, el "Citation Index" de la Red. Revista Española de Documentación Científica. 20(2):175-181.

Ross, N. C. M.; Wolfram, D. (2000) End user searching on the Internet: an analysis of term pair topics submitted to the Excite search engine. Journal of the American Society for Information

Science and Technology. 51(10):949-958.

Rousseau, B.; Rousseau, R. (2000) Lotka: a program to fit a power law distribution to observed frequency data. Cybermetrics. 4(1).

Rousseau, R. (1997) Sitations: an exploratory study. Cybermetrics. 1(1). Disponible en: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

Rousseau, R. (1999) Daily time series of common single word searches in AltaVista and NorthernLight . Cybermetrics. 2/3. Disponible en: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

Sancho, R. (1990) Indicadores bibliométricos utilizados en la evaluación de la ciencia y la tecnología. Revisión bibliográfica. Revista Española de Documentación Científica. 13:842-865.

Scharnhorst, A. (2004) Conceptualization of the Web in complex network theory. 4S-EASST Meeting. Paris, 25-28 agosto.

Sherman, C.; Price, G. (2001) The invisible Web. Uncovering information resources search engines can't see. Medford, New Jersey: Information Today Inc.

Small, H. (1973) Co-citation in the scientific literature: a new measure of the relationship between two documents. Journal of the American Society for Information Science and Technology. 24(4):265-269.

Smith, A. G. (1999) A tale of two web spaces: comparing sites using web impact factors. Journal of Documentation. 55(5):577-592.

Smith, A.; Thelwall, M. (2002) Web Impact Factors for Australasian universities. Scientometrics. 54(3):363-380.

Snyder, H.; Rosenbaum, H. (1999) Can search engines be used as tools for web-link analysis?: a critical view. Journal of Documentation. 55(4):375-384.

Tague-Sutcliffe, J. (1992) An introduction to Informetrics. Information Processing and Management. 28(1):1-3.

Thelwall, M. (2001a) The responsiveness of search engine indexes. Cybermetrics. 5(1). Disponible en: http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html.

Thelwall, M. (2001b) A web crawler design for data mining. Journal of Information Science. 27(5):319-325.

Thelwall, M. (2002). A comparison of sources of links for academic Web Impact Factor calculations. Journal of Documentation. 58(1):60-72.

Thelwall, M. (2003) SocSciBot. 30 octubre. E-mail a Natalia Arroyo

Thelwall, M.; Wilkinson, D. (2004) Finding similar academic Web sites with links, bibliometric couplings and colinks. Information Processing & Management, 40(3), 515-526. [Preprint].

Van Raan, A. F. J. (2001) Bibliometrics and Internet: some observations and expectations.

Scientometrics. 50(1):59-63.

Vaughan, L.; Hysen, K. (2002) Relationship between links to journal web sites and Impact Factors. Aslib Proceedings. 2002; 54(6):356-361.

Vaughan, L.; Thelwall, M. (2004) A fair history of the Web?: examining country balance in the Internet Archive. Library & Information Science Research.

Vaughan, L.; Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. Information Processing & Management, 40(4), 693-707.

Watts, D. J.; Strogatz, S. H. (1998) Collective dinamics of "small-world" networks. Nature. 393:440-442.

Webopedia (s.d.). [Página web] Disponible en: http://www.webopedia.com/.

Wellman, B.; Salaff, J.; Dimitrova, D.; Garton, L.; Gulia, M.; Haythornwaite, C. (1996) Computer networks as social networks: collaborative work , telework. and virtual community. Annual Review of Sociology. 22:213-238.

Zakon, R. H. (2003) Hobbes' Internet Timeline v6.0 [Página web]. Disponible en: http://www.zakon.org/robert/internet/timeline/.

Zelman, A.; Leydesdorff, L. (2000) Threaded e-mail messages in self-organization and science and technology studies oriented mailing lists. Scientometrics. 48(3):361-380.

Zipf, G. K. (1949) Human behavior and the principle of last effort: an introduction to human ecology. Cambridge, MA, Addison-Wesley.

Glosario

No se ha pretendido, al componer este glosario, asentar una terminología, sino más bien aclarar una serie de términos que pueden resultar extraños a lectores ajenos al área. Por ello han sido extraídas sus definiciones de varias fuentes. Se trata fundamentalmente de términos informáticos, de procedencia anglosajona todos ellos, cuya absorción ha sido tan rápida que su traducción al castellano puede resultar en ocasiones extraña, como es el caso de la palabra *home page*. Sobre otros ni siquiera los expertos se ponen de acuerdo en cuál sería el término más adecuado en nuestro idioma; es justamente lo que sucede con la cualidad *case-sensitive* aplicada a un servidor, por lo que se ha optado por el que se considera más apropiado. Por todo ello se ha decidido añadir, junto al término en español, que será siempre la entrada aceptada, el término en inglés, independientemente de la predominancia de uno u otro. La excepción se halla en los términos que han sido totalmente asumidos al español en la misma forma que en inglés (como *cookies*, por ejemplo), que se indicarán sólo una vez para evitar repeticiones innecesarias.

Applet. Programa que puede descargarse a través de Internet y ejecutarse en la máquina receptora. Están a menudo escritas en lenguaje de programación Java y se ejecutan dentro de un software explorador, siendo usadas principalmente para personalizar o añadir elementos interactivos a una página web.

ASP. Acrónimo de Active Server Pages. Es una tecnología orientada al Web desarrollada por Microsoft y que está diseñada para permitir el desarrollo de *scripts* del extremo servidor (en lugar de scripts del extremo cliente). Las páginas ASP son archivos de texto que no sólo pueden contener texto y etiquetas HTML, como sucede en los documentos web estándar, sino también comandos escritos en un lenguaje de desarrollo de *scripts* (como VBScript o JavaScript) y que pueden ser ejecutados en el servidor, lo que permite desarrollar interactividad a un documento o personalizar la visualización o el suministro de información al cliente sin preocuparse de la plataforma que este esté utilizando.

CGI. Acrónimo de Common Gateway Interface. Se trata de una especificación para la transferencia de información entre un servidor World Wide Web y un programa CGI, que es cualquier programa, escrito en cualquier lenguaje de programación (como C, Perl, Java o Visual Basic) diseñado para aceptar y devolver datos de acuerdo con la especificación CGI. Estos programas suelen ser la forma más común de interactuar dinámicamente con los usuarios; muchas páginas HTML que contienen formularios, por ejemplo, usan un programa CGI para procesar los datos recibidos.

Ciberespacio. Es el conjunto de contenidos accesibles en formato electrónico. La condición de accesibilidad universal de Internet aconseja utilizar el término como sinónimo de la Internet de los contenidos, fundamentalmente pero no exclusivamente, el webespacio. (Aguillo, 2003)

Cookie. Bloque de datos enviados por un servidor web a un sistema cliente. El explorador guarda estos datos localmente en un fichero de texto para después devolverlo al servidor cada vez que el explorador le pide una página.

Crawler. También conocidos como *spiders*, son programas que recogen automáticamente páginas web, y suelen estar detrás de los motores de búsqueda.

DoS. (Denial of Service) Tipo de ataque informatizado, usualmente planeado, que trata de interrumpir el acceso a través del Web. La forma más común consiste en abrumar a un servidor Internet con solicitudes de conexión que no pueden ser satisfechas.

Explorador (*browser*). Software que permite a un usuario visualizar documentos HTML y acceder a archivos y a software relacionados con estos documentos. Se basa en el concepto de hipervínculos, que permiten al usuario hacer clic con el ratón para saltar entre un documento y otro en el orden que deseen.

Formulario (*form*). Documento en línea que contiene espacios en blanco para que el usuario los rellene con la información solicitada y que puede ser enviado a través de una red a una organización que solicita la información. En el Web suelen estar codificados para su procesamiento mediante *scripts* CGI.

Protocolo de exclusión de robots (*Robots exclusion protocol*). Es el método que permite a los administradores de sitios web indicar a los robots visitantes qué partes de su sitio no pueden visitar[24].

Java. Lenguaje de programación orientado a objetos desarrollado por Sun Microsystems, Inc. Similar a C++, Java es más pequeño, portable y fácil de usar por ser más robusto y ejecutar su propia gestión de memoria. Java se emplea para programar aplicaciones de pequeño tamaño o *applets* y para el World Wide Web, así como para crear aplicaciones de red distribuidas. Las *applets* Java se utilizan en el Web para añadir efectos multimedia e interactividad a las páginas, como música de fondo, presentaciones de vídeo en tiempo real, animaciones, calculadoras y juegos interactivos. Su mayor contribución al Web ha sido en forma de Java Server Pages.

JavaScript. Lenguaje de *script* desarrollado por Netscape Communications y Sun Microsystems y que está vagamente relacionado con Java. Sin embargo no es un lenguaje verdaderamente orientado a objetos y su rendimiento es más limitado porque no es un lenguaje compilado. Gracias a JavaScript se puede añadir a las páginas web funciones y aplicaciones básicas basadas en red, aunque en menor número y de menor complejidad que las disponibles con Java. El código JavaScript se incluye en la página junto con el código HTML y se considera más fácil de escribir, especialmente para programadores principiantes. La versión de Microsoft es JScript.

JSP. Acrónimo de Java Server Pages. Tecnología creada por Sun Microsystems para permitir el desarrollo de aplicaciones basadas en el Web independientes de la plataforma. Utilizando etiquetas HTML y XML y scriptles Java, JSP ayuda a los desarrolladores de sitios web a crear programas interplataforma. Los *scriplets* JSP se ejecutan en el servidor, no en el explorador web, y generan

[24] http://www.robotstxt.org/

contenido dinámico para las páginas web, con la posibilidad de integrar contenido procedente de una diversidad de orígenes de datos, como bases de datos, archivos y componentes JavaBean.

Lenguaje interpretado (*interpreted languaje*). Característica de un lenguaje de programación que consiste en que el código fuente del programa se lee y procesa mientras se está ejecutando (en tiempo de ejecución), es decir, que van leyendo las instrucciones una a una y las van ejecutando según las decodifican. Son independiente de la máquina en que esté funcionando. El programa es traducido a código máquina cada vez que se ejecuta. Los programas escritos en estos lenguajes suelen llamarse *scripts*.

MIME. Acrónimo de Multipurpose Internet Mail Extensions. Es un protocolo ampliamente empleado en Internet para permitir la transmisión de datos mediante correo electrónico sin tener que traducirlos primero los datos a formato ASCII. Esto se lleva a cabo utilizando los tipos MIME que describen el contenido de un documento.

Página principal (*home page*). Página principal de un sitio web, que suele servir de índice o tabla de contenidos para acceder a otros documentos que forman parte del sitio.

Página web (*web page*). Documento contenido en la WWW que se compone de un archivo HTML con una serie de archivos asociados que almacenan gráficos o *scripts* y que está situado en un directorio concreto de una máquina determinada y que, por lo tanto, es identificable mediante una dirección URL.

Página web dinámica (*dynamic web page*). Página web que tiene una forma fija, pero contenido variable, que permite adaptarla a los criterios de búsqueda de un cliente.

Pasarela (*gateway*). Dispositivo que conecta redes que utilizan diferentes protocolos de comunicaciones de modo que la información pueda fluir entre una y otra. Una pasarela se encarga tanto de transferir la información como de convertirla a una forma compatible con los protocolos utilizados en la red de destino.

Servidor proxy (*proxy server*). Componente cortafuegos que gestiona el tráfico internet entrante y saliente de una red de área local y que puede proporcionar otras funciones, tales como las de almacenamiento en caché de documentos y el control de accesos.

Script. Programa compuesto por una serie de instrucciones dirigidas a una aplicación o a otro programa de utilidad. En el WWW se utilizan comúnmente para personalizar o añadir interactividad a las páginas web.

Sede web (*institutional web site*). Página web, o conjunto de páginas web ligadas jerárquicamente a una página principal, identificable por una URL y que forma una unidad documental reconocible e independiente de otras bien por su temática, bien por su autoría, o por su representatividad institucional. Teniendo en cuenta este último aspecto se reconocerían tres tipos de sedes web: institucionales, temáticas y personales.

Sitio web (*web site*). Colección de páginas web interrelacionadas, que incluyen una principal, y residen en el mismo sitio de red.

Anexos

Anexo 1. Informes generados por el software

1.1. Astra SiteManager

Mercury Interactive Astra SiteManager Pages Report

Site Name: Cybermetrics. Electronic journal of scientometrics, informetrics and bibliometrics

Site URL: http://www.cindoc.csic.es/cybermetrics/

Generated: 22/10/2002 11:36:01

The map is not saved to disk.

Page Status Summary

	Pages	Resources	Dynamic		Pages	Resources	Dynamic
Local	**201**	**0**	**0**	**External**	**2760**	**0**	**64**
OK	200	0	0	OK	1689	0	62
Unread	0	0	0	Unread	466	0	1
Inaccessible	0	0	0	Inaccessible	200	0	0
Access Denied	0	0	0	Access Denied	70	0	0
Not Found	1	0	0	Not Found	335	0	1

Annotation	Last Modified	File Size	Load Size	Incoming Links	Outgoing Links	Broken Links
faq.html	16/10/2003 1:08:02	18204	18204	1	0	0
www.metaplus.com	16/10/2003 1:08:02	0	0	2	0	0
www.ademe.fr	02/06/2003 9:39:44	1843	1843	1	0	0
www.yahoo.dk	02/06/2003 9:39:44	0	0	1	1	0
cronin-achilles.html	02/06/2003 9:39:44	0	0	1	1	1
siteserver3.asp	02/06/2003 9:39:44	0	0	1	1	0
www.dimdi.de	02/06/2003 9:39:44	0	0	1	1	0
www.lib.ox.ac.uk	02/06/2003 9:39:44	0	0	1	0	0
rsghp.html	02/06/2003 9:39:44	0	0	1	1	1
intres.cgi	02/06/2003 9:39:44	0	0	1	1	0
www.profusion.com	02/06/2003 9:39:44	33920	33920	1	0	0
www.raging.com	02/06/2003 9:39:44	0	0	1	1	0
www-centrim.bus.bton.ac.uk	23/04/2003 9:03:29	11713	11713	1	0	0
index.html	23/04/2003 9:03:29	0	0	1	1	0

1.2. COAST WebMaster

Executive Summary

for http://www.cindoc.csic.es/cybermetrics

Executive Summary Problem Reports Site Properties

This report highlights key problem areas on http://www.cindoc.csic.es/cybermetrics as detected by COAST WebMaster.

Scan Details	Analysis completed on 22/10/2003 17:22:53 using COAST WebMaster.
Broken Links	778 (29.76%) of the 2614 links on this site are broken.
Pages with Broken Links	49 (59.76%) of the 82 pages on this site have links that are broken.
Broken Anchors	9 (0.34%) of the 2614 links on this site are broken.
Pages with Broken Anchors	1 (1.22%) of the 82 pages on this site have links that are broken.
Slow Pages	10 (12.20%) of the 82 pages on the site have download times exceeding 50s at 28.8 modem speeds.
Permanently Moved Pages	1 (0.04%) of the 2614 links are permanently moved. Links to each permanently moved page should be updated to reflect its new location.
Failed PageRules	No relative pagerules were defined for this scan. No global pagerules were defined for this scan.

Created by COAST WebMaster on Wed Oct 22 17:23:05 2003
Visit COAST Software Inc. at http://www.coast.com/ for information on COAST WebMaster.

Site Properties

for http://www.cindoc.csic.es/cybermetrics

Executive Summary Problem Reports **Site Properties**

http://www.cindoc.csic.es/cybermetrics - Site Properties
Scan Start Time: 22/10/2003 17:10:00
Server Type: Apache/1.3.27 (Unix) Midgard/1.4.4/SG mod_python/2.7.6 Python/1.5.2 mod_ssl/2.8.13 OpenSSL/0.9.7a DAV/1.0.3 PHP/4.3.1 mod_perl/1.26

General Site Information

Scan status:	Scan complete
Total number of files:	205
Number of levels read on site:	4
Total site size:	7.92 MB
Scan end time:	22/10/2003 17:22:53

Site Totals

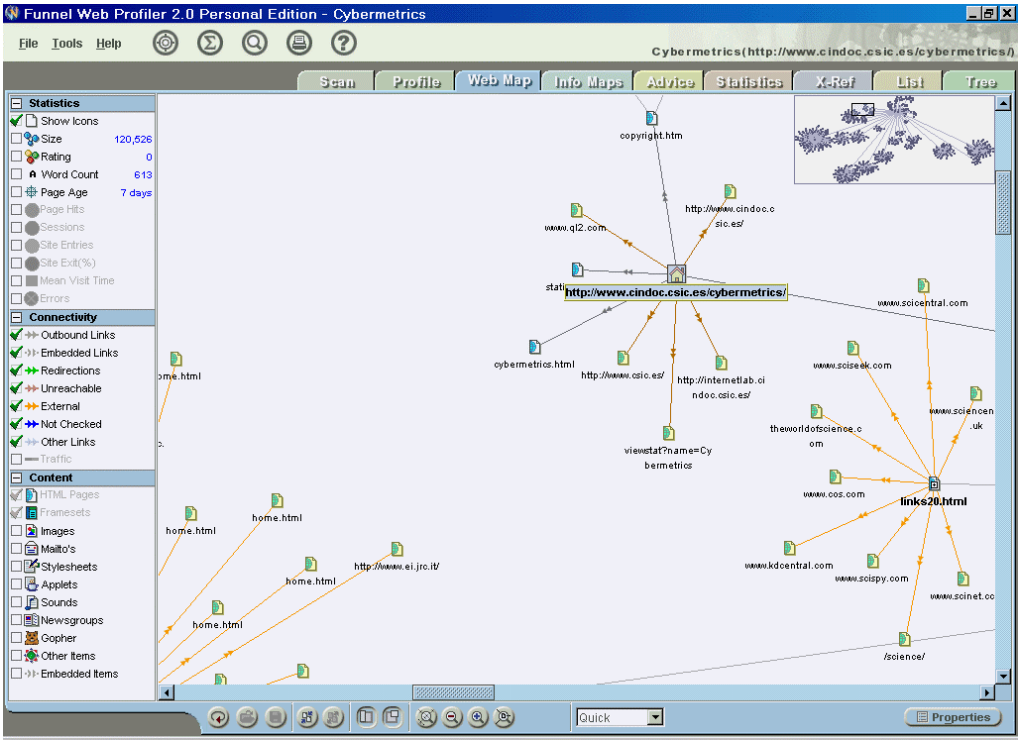
Total number of broken links:	778
Total number of broken anchors:	9
Total number of pages with failed PageRules:	0
Total number of internal links:	353
Total number of external links:	2261
Total number of unverified links:	0
Total number of HTML pages:	82
Total number of dynamic HTML pages:	0
Total number of Flash pages:	0
Total number of Microsoft Word pages:	0
Total number of GIF files:	91
Total number of JPEG files:	16
Total number of mailto links:	148

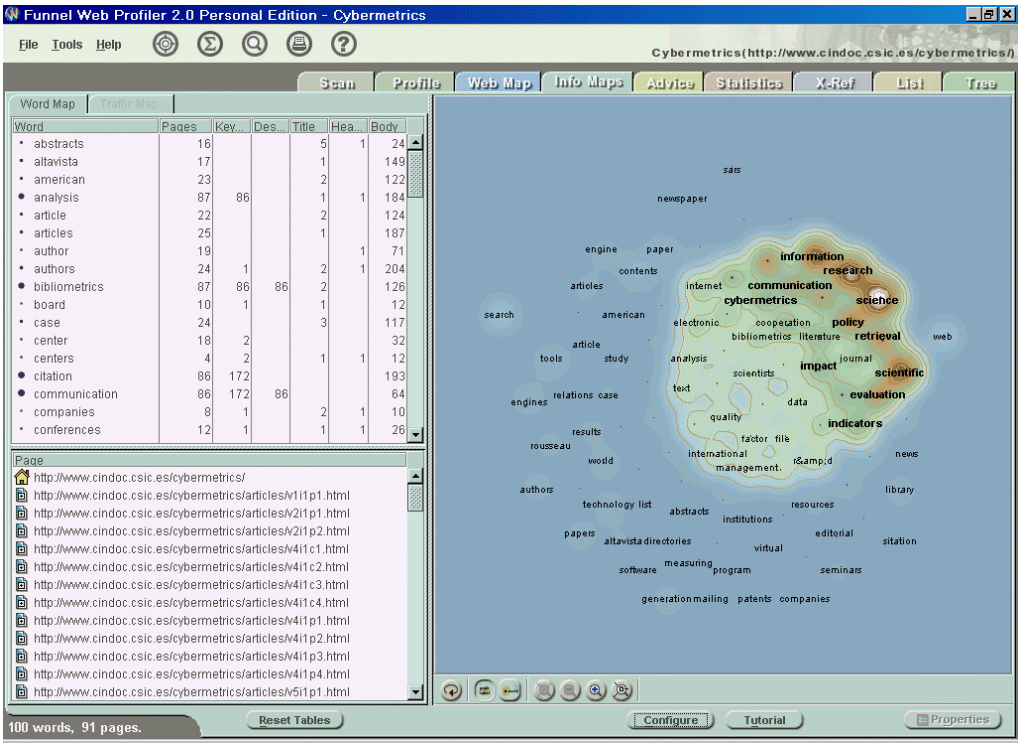
HTML Page Information

Total number of pages with Java applets:	0
Total number of pages with JavaScript:	2
Total number of pages with ActiveX controls:	0
Total number of pages with VB script:	0
Total number of pages with frames:	0
Total number of pages with forms:	0

Created by COAST WebMaster on Wed Oct 22 17:23:05 2003
Visit COAST Software Inc. at http://www.coast.com/ for information on COAST WebMaster.

1.3. Funnel Web Profiler





1.4. Microsoft Site Analyst y Content Analyzer

Microsoft Site Analyst

Site Summary Report for www.cindoc.csic.es/cybermetrics/

Site Summary Pages Hierarchy Images Media Gateways Help
Error Report Internet Duplicates Offsite InLinks Unexplored Index

WebMap for www.cindoc.csic.es

Object Statistics		
Type	Count	Size
Pages	1874	3553314
Images	109	1226935
Gateways	17	N/A
Internet	465	N/A
Java	0	0
Applications	153	3511344
Audio	0	0
Video	0	0
Text	3	0
WebMaps	0	0
Other Media	7	0
Totals	2628	8291593

Status Summary		
	Objects	Links
Onsite	205	3967
OK	204	3964
Not Found (404)	1	3
Other Errors	0	0
Unverified	0	0
Offsite	2423	2697
OK	0	0
Not Found (404)	0	0
Other Errors	0	0
Unverified	2423	2697
Totals	2628	6664

Map Statistics	
Map Date	Oct 22 10:11 2003
Levels	7
Avg Links/Page	81

Microsoft Site Analyst

Server Summary	
Domain:	www.cindoc.csic.es
Server Version:	Apache/1.3.27 (Unix) Midgard/1.4.4/SG mod_python/2.7.6 Python/1.5.2 mod_ssl/2.8.13 OpenSSL/0.9.7a DAV/1.0.3 PHP/4.3.1 mod_perl/1
HTTP Version:	1.1

Microsoft Site Analyst

Explored Onsite Page Report for www.cindoc.csic.es/cybermetrics/

Site Summary Pages Hierarchy Images Media Gateways Help
Error Report Internet Duplicates Offsite InLinks Unexplored Index

WebMap for www.cindoc.csic.es

Page Status Summary					
	Pages	Links		Pages	Links
Onsite	82	2881	Offsite	1792	1951
OK	81	2878	OK	0	0
Not Found (404)	1	3	Not Found (404)	0	0
Other Errors	0	0	Other Errors	0	0
Unverified	0	0	Unverified	1792	1951

Name	Level	Last Modified	Size	Load Size	Links on Page	Offsite Links	InLinks	Broken Links
Cybermetrics. Electronic journal of scientometrics, informetrics and bibliometrics	1	N/A	15985	N/A	51	7	2	0
Cybermetrics. News about this electronic journal of scientometrics, informetrics and bibliometrics	2	N/A	13902	27378	47	6	69	0
Cybermetrics. Editors of the Electronic journal of scientometrics	2	N/A	21506	36092	78	27	69	0

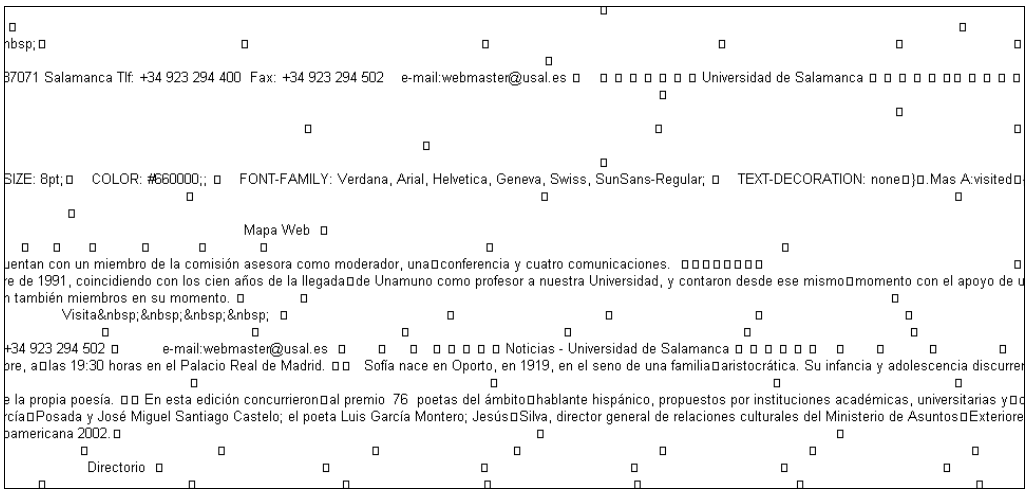
1.5. SocSciBot

URL	Links	outLinks	HTML downloads	complete HTML pages	pages without duplicates
http://www.usal.es/	2	0	2	1	1
URL	Links	outLinks	HTML downloads	complete HTML pages	pages without duplicates
http://www.usal.es/webusal/Principal.htm	103	559	73	68	68

Informe *summary.txt*

0	http://www.usal.es/webusal/Principal.htm
1	http://www.usal.es/webusal/Gobierno/Organos.htm
2	http://www.usal.es/webusal/AccesoUniversidad/AccesoUniversidad.htm
3	http://www.usal.es/webusal/Representacion/Organos.htm
4	http://www.usal.es/webusal/Titulaciones/Titulaciones.htm
5	http://www.usal.es/webusal/Legislacion/Legislacion.htm
6	http://www.usal.es/webusal/Doctorado/Doctorado.htm
7	http://www.usal.es/webusal/ServiciosUniversitarios/ServiciosUniversitarios.htm
8	http://www.usal.es/webusal/Postgrado/Postgrado.htm
9	http://www.usal.es/webusal/Gestion/Gestion.htm
10	http://www.usal.es/webusal/CursosExtraordinarios/CursosExtraordinarios.htm
11	http://www.usal.es/webusal/RelacionesInternacionales/RelacionesInternacionales.htm
12	http://www.usal.es/webusal/CursosInternacionales/CursosInternacionales.htm
13	http://www.usal.es/webusal/Investigacion/Investigacion.htm
14	http://www.usal.es/webusal/Centros/Centros.htm
15	http://www.usal.es/webusal/Fundaciones/Fundaciones.htm
16	http://www.usal.es/webusal/Bibliotecas/Bibliotecas.htm
17	http://www.usal.es/webusal/Novedades/Novedades.htm
18	http://www.usal.es/webusal/virtual/Virtual.htm
19	http://www.usal.es/webusal/Novedades/noticias/noticia1.htm
20	http://www.usal.es/webusal/Novedades/noticias/noticia2.htm
21	http://www.usal.es/webusal/Novedades/noticias/noticia3.htm
22	http://www.usal.es/webusal/Novedades/noticias/propuesta_expositiva.htm
23	http://www.usal.es/webusal/Novedades/noticias/programa_opera.htm
24	http://www.usal.es/webusal/Novedades/noticias/espacio_europeo.htm
25	http://www.usal.es/webusal/PrincipalIng.htm
26	http://www.usal.es/webusal/visita/index.htm
27	http://www.usal.es/webusal/Recursos/Recursos.htm
28	http://www.usal.es/webusal/mapa/index.htm
29	http://www.usal.es/webusal/Enlaces/universia.htm
30	http://www.usal.es/webusal/Titulaciones/simultaneos.htm
31	http://www.usal.es/webusal/Gobierno/gerencia/gerente.htm
32	http://www.usal.es/webusal/Gobierno/OrganosIng.htm
33	http://www.usal.es/webusal/AccesoUniversidad/AccesoUniversidadIng.htm
34	http://www.usal.es/webusal/Representacion/OrganosIng.htm
35	http://www.usal.es/webusal/Titulaciones/TitulacionesIng.htm
36	http://www.usal.es/webusal/Legislacion/LegislacionIng.htm
37	http://www.usal.es/webusal/Doctorado/DoctoradoIng.htm
38	http://www.usal.es/webusal/ServiciosUniversitarios/ServiciosUniversitariosIng.htm

Informe *index.txt*



Informe *allpages.txt*

1.6. Web Trends

Contenido

WEBTRENDS

usal

Estadísticas generales

El http://www.usal.es/webusal/Principal.htm sitio de la Web fue analizado el **Thu, Oct 23, 2003** a las **13:20**. WebTrends Link Analyzer encontró **12 HTML páginas** en este sitio. **8 de estas páginas** contienen un total de **8 enlaces interrumpidos**.

Problemas encontrados en este sitio

Enlaces interrumpidos (404 errores)
De los **291 enlaces**, **8** se refieren a páginas que no existen. Consulte la sección de Enlaces interrumpidos para obtener más detalles.

Otros errores
¡Enhorabuena! El Analizador de enlaces de WebTrends no encontró más errores en este sitio.

Errores de sintaxis de URL

Páginas interrumpidas
Felicidades! WebTrends Link Analyzer no encontró páginas interrumpidas en este sitio.

Sugerencias para mejoras

De las **12 páginas** en este sitio, **12** podrían ser mejoradas añadiéndoles títulos o atributos ALT, o atributos de altura o anchura a las imágenes. Los atributos ALT proveen una descripción de la imagen para los visitantes que visualizan la página en el modo de texto. (Esta descripción también se muestra mientras la página se está transfiriendo, antes de que la imagen aparesca). Los atributos de altura o anchura permiten a los exploradores transferir imágenes más rápidamente. Consulte la sección Sugerencias para mejoras para obtener más detalles.

Estadísticas del sitio

Tipos de enlaces
De los **291 enlaces totales** en este sitio, **97%** son HTTP, **0%** son FTP, los restantes son de correo electrónico, noticias, gopher o de otros tipos. Consulte la sección Estadísticas de enlace para obtener más detalles.

Tipos de archivos
52 del total de archivos en este sitio, **24%** son páginas HTML, **59%** son imágenes, y el resto son de varios otros tipos de archivos. Consulte la sección Estadísticas de los archivos para obtener más detalles.

Páginas más grandes (más lentas)
Este sitio contiene **12 HTML páginas**,de un tamaño combinado total de **1567 K Bytes** (incluyendo todos los gráficos y otros archivos enlazados). Entre más grandes sean las páginas, más lenta será la transferencia, lo cual puede presentar un problema si la mayoría de sus visitantes utilizan conexiones de acceso telefónico. Consulte la sección Estadísticas de los archivos para obtener más detalles.

Páginas más antiguas y más nuevas
Las páginas de HTML más antiguas fueron modificadas por última vez el **Thu May 15 18:47:03 2003** . El cambio más reciente fue hecho el **Thu Oct 23 09:46:37 2003** en la página **http://www.usal.es/webusal/Principal.htm**. Consulte la sección Páginas más antiguas y más nuevas para obtener más detalles.

Imágenes gráficas

WebTrends Link Analyzer encontró **32 imágenes** en este sitio, usando un total de **686 K Bytes** . Consulte la sección Imágenes gráficas para obtener más detalles.

Catalogo de imágenes
El Catalogo de imágenes muestra en miniatura todas las imágenes encontradas en este sitio. Este catalogo le puede ayudar a identificar rápidamente duplicados, o simplemente revisar el contenido gráfico de este sitio.

WEBTRENDS

Este informe fue generado por WebTrends Analysis Suite v7.0c (Build 1467)

1.7. Xenu

Xenu's broken link report

Created on May 13, 2004 at 12:43:41

Root URL: http://www.cindoc.csic.es/cybermetrics/

Table of contents

Statistics for managers

Correct internal URLs, by MIME type:

text/html	82 URLs	3569299 Bytes (3485 KB)	40.00%
image/jpeg	16 URLs	781535 Bytes (763 KB)	7.80%
image/gif	91 URLs	440231 Bytes (429 KB)	44.39%
image/png	1 URLs	5169 Bytes (5 KB)	0.49%
application/pdf	13 URLs	1866613 Bytes (1822 KB)	6.34%
application/octet-stream	1 URLs	1603771 Bytes (1566 KB)	0.49%
application/vnd.ms-excel	1 URLs	40960 Bytes (40 KB)	0.49%
Total	205 URLs	8307578 Bytes (8112 KB)	100.00%

All pages, by result type:

ok	2058 URLs	66.45%
skip type	149 URLs	4.81%
no such host	109 URLs	3.52%
not found	361 URLs	11.66%
timeout	52 URLs	1.68%
forbidden request	22 URLs	0.71%
no connection	9 URLs	0.29%
auth required	16 URLs	0.52%
temporarily overloaded	1 URLs	0.03%
server error	3 URLs	0.10%
cancelled / timeout	1 URLs	0.03%
extended error	315 URLs	10.17%
no info to return	1 URLs	0.03%
Total	3097 URLs	100.00%

Return to Top

This report has been produced by Xenu's Link Sleuth

Anexo 2. Opciones de programa

	Astra	COAST	FunnelWeb	Microsoft	SocSciBot	WebKing	WebTrends	WebCount	Xenu
Autenticación (contraseñas)	X	X	X	X		X			X
Configuración del browser	X	X	X	X		X	X		
Distinción mayúsculas/minúsculas	X	X	X	X		X	X		
Verificar enlaces externos	X	X	X	X		X	X		X
Nombres de fichero por defecto	X					X			
Definición de extensiones		X					X[25]		
Incluye el dominio...					X				
Exclusión / inclusión de URLs	X	X	X	X			X		X
Formularios	X	X	X						
Programas auxiliares	X	X	X	X		X			
Protocolo de robot		X		X		X			
Limitación de CGI	X					X			
Gestión de cookies		X	X			X			
Analizar el sitio completo		X		X					
Definición del proxy	X	X	X	X			X		
Datos de los informes		X				X			X
Profundidad del análisis	X	X							X
Conexiones simultáneas		X					X		X
Copia del sitio			X	X					
Control del tiempo			X				X	X	
Opciones de visualización		X		X			X		

[25] Además permite seleccionar todos los tipos de ficheros, al contrario que el resto, que sólo permite HTML.

Anexo 3. Resultados

		Teleport Pro				Astra Site Manager				COAST WebMaster				FunnelWeb		SocSciBot				WebKing(3)			
		031022		031029		Online		Offline		Online		Offline		Online		Online		Offline		031022		031029	
		HTML(1)	Pags.(2)	HTML	Pags.	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	HTML	Pags.	HTML	Pags.
1	www.cindoc.csic.es/cybermetrics	82	82	82	82	201	201	200	200	82	82	82	82	76	79	80	80	81	81	81	81	81	81
2	www.ite.upv.es	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	www.upct.es/~de	510	510	510	510	212	212	344	344	70	70	510	510	312	321	68	68	107	107	69	69	68	68
4	www.ceu.es	8965	8966	8959	8960	4367	4366	11836	11821	9037	9035	8966	8960	(390)	(392)	187	187	(5093)	(3460)	260	249	260	249
5	www.shef.ac.uk/dentalschool	58	58	58	58	143	143	131	131	59	59	58	58	58	58	56	56	57	57	57	57	57	57
6	www-cryst.bioc.cam.ac.uk	884	884	880	880	—	—	3104	3098	629	631	884	880	119	118	538	469	—	—	—	—	568	568
7	www.pem.cam.ac.uk	971	971	968	968	1885	1885	1940	1940	946	946	971	968	12	12	1017	1018	905	900	902	902	898	898
8	www.teipat.gr	345	645	649	649	776	776	681	683	617	618	643	647	—	314	330	326	376	311	314	314	314	314
9	www.dsg.unito.it	543	544	537	538	982	981	1143	1134	493	493	543	537	102	103	452	453	415	413	496	492	433	429
10	www.kun.nl/phil	511	511	511	511	783	783	767	765	511	511	511	511	351	352	495	495	486	485	507	507	507	507
11	www.mat.chalmers.se	3	3	3	3	7	7	7	7	3	3	3	3	3	3	3	3	3	3	3	3	3	3
12	cst.dk	461	461	460	459	888	888	859	859	411	411	461	460	—	—	714	1017	419	417	120	120	6	6
13	www.montefiore.ulg.ac.be	1612	1612	1704	1704	5306	5347	5122	5626	1399	1401	1611	1703	(380)	(380)	(1680)	(3371)	(4063)	—	1448	1442	1444	1438
14	wwwai.wu-wien.ac.at	19258	19258	19284	19284	50073	50790	53212	—	14353	14667	19257	19283	(501)	(504)	(5106)	(5051)	(3668)	(3673)	14628	14628	13150	13150
15	www.uni-saarland.de/fak8/iaua	30	30	30	30	121	121	135	135	30	30	30	30	29	30	28	28	29	29	29	29	29	29
16	www.medizin.uni-greifswald.de/humangen	14	14	14	14	27	27	28	28	14	14	14	14	14	14	12	12	13	13	13	13	13	13
17	www.math.jyu.fi	3267	3267	3715	3716	10360	—	9168	12504	2633	2767	3266	3714	310	313	1437	1462	4946	—	1634	1623	1659	1648
18	www.etsia.upv.es	181	181	28470	28487	815	815	753	—	180	180	181	28470	161	160	164	164	135	—	179	175	179	175
19	www.usal.es	553	553	104	104	240	241	822	226	86	86	553	104	83	83	83	3514	478	1	83	83	83	83

(1) Número de ficheros htm y html.
(2) Suma total del número de ficheros htm, html, shtm, shtml, sml, stm, stml, htp, cfm, asp, jsp y php.
(3) Los datos que se muestran corresponden exclusivamente al muestreo en línea, el único que ha sido técnicamente posible.

		Microsoft Site Analyst				Microsoft Content Analyzer				WebCount				WebTrends				Xenu			
		Online		Offline		Online		Offline		Online		Offline		Online		Offline		Online		Offline	
		031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029	031022	031029
1	www.cindoc.csic.es/cybermetrics	82	82	81	81	82	82	81	81	232	232	37	37	—	—	2491	2491	82	82	81	81
2	www.ite.upv.es	2	2	2	2	2	2	2	2	1	1	1	1	2	2	2	2	2	2	2	2
3	www.upct.es/~de	—	—	109	109	69	69	109	109	—	23	4	4	70	0	118	118	70	68	109	109
4	www.ceu.es	309	309	204	204	293	293	204	204	—	2182	5	5	516	—	—	—	8833	8844	—	—
5	www.shef.ac.uk/dentalschool	71	71	57	57	71	71	57	57	—	—	14	14	70	—	157	157	58	58	57	57
6	www-cryst.bioc.cam.ac.uk	1019	1002	613	613	1049	1049	613	613	25	25	5	5	—	—	—	—	6783	6435	639	639
7	www.pem.cam.ac.uk	980	980	970	965	978	974	970	965	19	19	12	12	—	—	1333	1328	950	950	970	965
8	www.teipat.gr	321	321	313	315	321	321	313	315	6	6	6	6	5	5	14	14	314	314	312	314
9	www.dsg.unito.it	541	540	543	537	544	544	543	537	38	38	25	25	187	—	729	718	537	537	—	536
10	www.kun.nl/phil	528	526	511	511	528	527	511	511	3	3	3	3	539	551	1385	1388	515	515	511	511
11	www.mat.chalmers.se	3	3	3	3	3	3	3	3	2	2	2	2	3	3	3	3	3	3	3	3
12	cst.dk	—	431	425	425	428	428	426	425	—	557	15	15	621	—	—	0	512	512	—	457
13	www.montefiore.ulg.ac.be	1601	1592	1608	1671	779	1612	1601	1664	32	32	27	27	—	—	—	3935	—	—	1582	1647
14	wwwai.wu-wien.ac.at	15713	16208	15787	16037	15792	16224	15786	16036	89	89	12	12	—	—	—	—	10534	—	—	—
15	www.uni-saarland.de/fak8/iaua	30	30	30	30	30	30	30	30	379	380	9	9	32	—	110	110	31	31	30	30
16	www.medizin.uni-greifswald.de/humangen	13	13	14	14	13	13	14	14	8	8	8	8	14	—	19	19	14	14	14	14
17	www.math.jyu.fi	2469	2518	2551	2828	2466	2520	2551	2828	2739	2879	13	13	—	—	5397	5901	2858	614	—	—
18	www.etsia.upv.es	259	259	178	134	248	248	178	130	11	11	11	11	330	327	709	141	181	181	181	78
19	www.usal.es	89	89	96	83	89	89	96	83	—	78	32	30	12	—	34	62	155	130	516	83

Anexo 4. Evaluación del software

	Astra SiteManager	**Site Analyst y Content Analyzer**	**COAST WebMaster**	**FunnelWeb**
Consumo de recursos	Gran consumo, especialmente RAM	Gran consumo, especialmente RAM	Gran consumo	Gran consumo, especialmente al analizar datos y generar gráficos
Cobertura de páginas dinámicas	ASP: alta PHP: media CGI: ?	Media	ASP: alta PHP: ? CGI: media	Alta
Dificultades	Errores de programa en sitios conflictivos	Errores de programa en sitios conflictivos		
Entorno gráfico	Vista *ciberbólica*	Vista *ciberbólica*	No	Gran importancia del entorno gráfico
Limitaciones	Falta de claridad de los datos	Los grandes sitios web provocan el bloqueo del programa	Caduca a los 15 días	Sólo 1.000 items por sede No guarda los informes Sólo una ventana por PC
Output	Informes que incluyen algunas estadísticas (páginas, recursos y objetos dinámicos, bien externos o locales) y estructura hipertextual	Informes que incluyen completas estadísticas (páginas, links, objetos y tipo de objetos, y estructura del sitio	Estadísticas (HTML dinámico, imágenes GIF y JPG, Flash, ficheros internos y externos…)	Vista de grafo, análisis de palabras, estructura hipertextual, algunas estadísticas: items, tamaño, enlaces (totales, rotos y externos) y palabras
Opciones de programa	Similares a otros programas de análisis de sitios web	Similares a otros programas de análisis de sitios web	Gran abanico de opciones Definición de extensiones de fichero	Similares a otros programas de análisis de sitios web
Transparencia	No	Si	No genera estructura hipertextual del sitio, sólo para problemas hallados	Si
Utilidades adicionales	Programador de tareas	Copia del sitio Verificador de enlaces	Programador de tareas	Estadísticas de tráfico Análisis de palabras Evaluación de la calidad del sitio

	SocSciBot	WebKing			WebTrends			WebCount	Xenu
Consumo de recursos	Apenas	Si			Si			No	Sólo grandes sitios y especialmente al generar grandes informes
Cobertura de páginas dinámicas	Alta	ASP alta	PHP ?	CGI media	ASP media	PHP ?	CGI alta	?	Alta
Dificultades	Truncamiento de la URL, por lo que no cuenta como los demás Extracción de estadísticas				Gran número de errores de programa			Los resultados son difíciles de interpretar	Se bloquea al generar grandes informes
Entorno gráfico	No	No			No			No	No
Limitaciones	Sólo hasta 10.000 enlaces Sub-dominios incluidos como recursos internos	Restricciones de la versión LITE mode							
Opciones de programa	Simplicidad de las opciones	Similares a otros programas de análisis de sitios web			Similares a otros programas de análisis de sitios web Incluye definición de extensiones			Simplicidad de las opciones	Simplicidad de las opciones
Output	Informes TXT que incluyen la estructura del sitio, contenido y estadísticas	Completas estadísticas por extensiones de ficheros			Informes HTML que incluyen algunas estadísticas (ficheros HTML, enlaces, tipo de ficheros, tamaño y errores)			Informes TXT que incluyen algunas estadísticas (páginas y enlaces internos y externos; palabras; imágenes; CGI…) y errores	Informes HTML que incluyen algunas estadísticas (URLs internas por tipo MIME) y estructura del sitio
Transparencia	Si	Si			No queda claro			No	Si
Utilidades adicionales		Verificador de enlaces			Análisis de ficheros log Sugerencias para la mejora del sitio			Extrae datos de WayBack Machine	Verificador de enlaces

Anexo 5. Software

En el CD adjunto se incluyen los *crawlers* evaluados en el presente trabajo, aptos todos ellos para Windows 98 y posteriores, algunos incluso para versiones anteriores.

Content Analyzer no ha podido ser incluido por tratarse de una versión con licencia. Sin embargo, cabe recordar que su diseño y opciones son, por otra parte, casi idénticos a los de *Site Analyst*.

Antes de la ejecución de cualquiera de los programas es necesaria su instalación.