Herramientas de segunda generación

Isidro F. Aguillo
CINDOC, CSIC (Madrid)
isidro@cindoc.csic.es

RESUM

Característiques i aplicacions de les principals eines de segona generació per a la localització i recuperació d'informació a Internet. S'inclou una anàlisi comparativa de prestacions.

RESUMEN

Características y aplicaciones de las principales herramientas de segunda generación para la localización y recuperación de información en Internet. Se incluye un análisis comparativo de prestaciones.

Introducción

El crecimiento de la Internet física parece haber desacelerado en los últimos meses. A finales de 1997 el número de ordenadores conectados a la red no alcanza los 30 millones, lo que significa que es el primer año de esta década en que no se ha producido un incremento superior al 100%. Tal previsión ya ha sido adelantada por Lottor en su «Internet Domain Survey» semestral correspondiente a enero de 1998 <www.nw.com/zone/WWW/report.html>.

Sin embargo, para el profesional de la información resulta mucho más interesante el crecimiento de la Internet de los contenidos, y en especial el incremento de las páginas Web, una unidad documental más relevante. El futuro de nuestra profesión puede encontrarse en la disponibilidad universal de la información en Internet. Ello ocurrirá una vez que el volumen de páginas alcance un nivel práctico de utilización, si no lo ha alcanzado ya.

Es difícil estimar tanto el número de páginas como el incremento anual. No obstante, parece probable que estemos por encima de los 300 millones de URLs (un concepto más amplio que el de página) y que el crecimiento de la Web implique la duplicación cada 173 días tal como señala Nielsen en Julio de 1997 (*Growth of the Web*: <www.useit.com/alertbox/9509.html>).

En los últimos años las bases de datos de los principales buscadores apenas recogían un pequeño porcentaje de todos esos recursos. Es posible que esta situación haya variado sustancialmente en el cuarto trimestre de 1997, cuando al menos Altavista <www.altavista.digital.com> y Hotbot <www.hotbot.com > superan holgadamente los 100 millones de URLs indexadas.

1. Herramientas de primera generación

Hasta hace relativamente poco tiempo las herramientas de localización y recuperación de información en Internet eran únicamente *server-side*. El mecanismo de recuperación de datos estaba instalado en un ordenador remoto, al que se debía acceder en las condiciones y limitaciones que marcaba tal sistema remoto, el servidor. Tal organización implica una cierta pérdida de flexibilidad, la imposibilidad de personalizar opciones y la dificultad de automatizar tareas. A ese conjunto de herramientas, basadas en el servidor, se las denomina de *primera generación.* Hasta hace relativamente poco tiempo eran las únicas disponibles, constituyendo de hecho el principal instrumento documental de la red. Aparte de limitaciones evidentes, su mera existencia ha permitido cierta crítica a los que consideraban caótica la organización de la información en Internet.

Se asume ahora un concepto dinámico de la Internet de los contenidos, en la que la información existe y es útil en la medida en que es posible localizarla y recuperarla (fundamental e indisociable binomio que ha resultado muy importante desde el punto de vista documental). Sin embargo, como queda dicho, las limitaciones de esta primera generación de herramientas son muy conspicuas y ha tenido que recurrirse a diferentes estrategias para tratar de mejorar su rendimiento documental. La estrategia más acertada hasta la fecha ha sido recurrir a las distintas herramientas de forma conjunta. Según las necesidades las herramientas se utilizan de forma diferenciada, pero se pueden complementar los resultados atendiendo a las características de cada uno de los servicios. Esto se ha utilizado como criterio taxonómico y así podemos distinguir dos grandes categorías de herramientas: Índices y Buscadores.

1.1. Índices

Los índices son listados de recursos organizados según un criterio determinado. De acuerdo a dichos criterios se establecen categorías que permiten la recuperación por navegación a través de las mismas. Los índices son realizados por humanos, como meros directorios, que pueden incluir anotaciones y metacategorías para ayudar en la localización de información.

Estas herramientas son más precisas pero menos exhaustivas que los buscadores. Ello es debido tanto al esfuerzo necesario para su mantenimiento, como a la metodología utilizada, que implica la reducción del contenido mediante adjudicación de descriptores (indización).

Los índices se dividen atendiendo al criterio clasificador en *índices geográficos*, como los populares Donde <donde.uji.es> o el Virtual Tourist <www.vtourist.com/webmap> e *índices temáticos* construidos sobre árboles de conocimiento más o menos ramificados. También es posible realizar una clasificación atendiendo a los niveles de profundidad con que se organiza el criterio temático, y así tenemos:

- *Directorio de directorios* (nivel supercero). Como su propio nombre indica recopilaciones de directorios, de los que los dos ejemplos más significativos de este grupo son la Virtual Library <vl.stanford.edu/Overview.html> y el Argus Clearinghouse <www.clearinghouse.net>

- *Directorios* (nivel cero). Interesantes listados de enlaces con o sin el título que acompañan a muchas páginas institucionales o personales apoyando los contenidos de estas como una bibliografía complementaria.

- *Indices anotados* (nivel uno). El especialista añade un comentario más o menos amplio, lo cual proporciona valor añadido al listado, pero no profundiza estableciendo subdirectorios a los que se acceda mediante enlaces (páginas diferentes).

- *Superíndices* (dos o superiores). Algunos de los índices más populares como LookSmart <looksmart.com>, con cerca de 250.000 sedes analizadas o los españoles Olé <www.ole.es> y Ozú <www.ozu.es>. Estos son los auténticos índices, pues su estructura invita a la navegación, como método de recuperación de la información.

Por ultimo, podríamos hablar de *Metaíndices* como en el caso de Yahoo <www.yahoo.com>. Dos son las características de este recurso «estrella», con tres cuartos de millón de sedes, que nos sugieren esta clasificación singular. Yahoo incluye para cada categoría temática tanto recursos como metarecursos. Eso significa que funciona como un índice y, además, como directorio de directorios. Aunque ambos grupos están segregados, separados físicamente con una línea horizontal, toda la información aparece cómodamente en la misma página. Además todas las disciplinas están interconectadas, de forma que un mismo recurso, o incluso, una categoría puede accederse desde diferentes vías. El motor de búsqueda interno proporciona flexibilidad al conjunto.

1.2. Buscadores

Los buscadores, o motores de búsqueda, extraen los recursos de una base de datos generada como resultado de la indexación a texto completo de los contenidos de las páginas Web (a veces, de otros recursos adicionales). Los correspondientes índices inversos se construyen a partir de todos los contenidos textuales o multimedia o de una parte significativa (título, dirección, descriptores de las etiquetas META, primeras líneas de las páginas) a las que se tiene acceso de forma automática.

Dicha labor es realizada por un programa llamado robot que explora, como quedó indicado, automáticamente los servidores WWW a nivel mundial, salvo exclusión definida –temporal, geográfica, número IP, etc.–, acordada o impuesta (si existe un fichero robots.txt que lo prohiba).

Al construirse la base de datos de manera automática, lo que permite una mayor cobertura de la misma, los buscadores son más exhaustivos tanto en número de sedes investigadas como en volumen de información dentro de cada sitio Web. En cambio resultan notablemente menos precisos que los índices debido a las limitaciones que imponen sus lenguajes de interrogación. La incorporación de la lógica difusa y de ciertos mecanismos de inteligencia artificial no han tenido aún impacto significativo en este tipo de herramientas. No obstante, esto podría no ser cierto en un futuro no muy lejano.

La recuperación corre a cargo de un sistema de gestión de bases de datos. Este es capaz, no solo de interrogar la base, sino también, mediante un complejo algoritmo, de determinar la pertinencia de los resultados en función de la estrategia de búsqueda. Esta puede realizarse con ayuda de la lógica «booleana» soportada de forma

explícita o implícita por el sistema. Aunque en este último caso tal estrategia oculta un mecanismo imperfecto por definición o incapaz por ahorro (escasez) de recursos de computación.

Los motores más sofisticados permiten además el uso de delimitadores por campos virtuales (título, dirección, fecha, contenido de las etiquetas META, etc.), objetos tales como ficheros multimedia (imágenes, sonido, vídeo) o enlaces hipertextuales.

Aunque bien conocidos, citaremos los más relevantes en un orden que pretende reflejar, de forma subjetiva, la calidad y utilidad relativa:

Motor (URL)	Tamaño	Velocidad	Potencia	Pertinencia	Actualidad
Altavista <www.altavista.digital.com>	5	5	5	4	4
Hotbot <www.hotbot.com>	5	4	4	4	4
Northern Light <www.nlsearch.com>	4	4	3	4	4
Infoseek <www.infoseek.com>	4	4	4	3	4
Excite <www.excite.com>	4	2	3	4	2
Lycos <www.lycos.com>	2	1	3	4	2
OpenText <index.opentext.net>	2	3	3	3	2
WebCrawler <webcrawler.com>	2	3	3	4	2
Magellan <www.mckinley.com>	1	3	3	5	2

Los valores asignados, de 1 (peor) a 5 (mejor) corresponden a una apreciación personal del tamaño (número de páginas indizadas), velocidad (media de acceso al buscador a lo largo del día), potencia (valorando los operadores «booleanos», los delimitadores y mecanismos anejos), pertinencia (adecuación de los resultados a la estrategia) y actualidad (frecuencia de actualización).

Por último, la aparición de los *metabuscadores* ha supuesto una notable aportación, al incrementar considerablemente el valor añadido de la presentación de resultados. Distinguimos dos grupos bajo este epígrafe:

- Los *metabuscadores monomotores*, ofrecen, bien la posibilidad de perfilar la búsqueda ofertando una serie de palabras candidatas asociadas al perfil de búsqueda por un algoritmo que valora su presencia en términos fundamentalmente cuantitativos, o bien agrupan los resultados según su origen, utilizando a tal fin como criterio jerárquico la dirección (URL) de los recursos. Estas son opciones que ofrecen, entre otros, Excite o Altavista en el primer caso o Northern Light e Infoseek en el segundo caso.

- Los *metabuscadores multimotores* han evolucionado partiendo del *multibuscador*, herramienta simple que ejecuta la estrategia contra varios motores simultáneamente y presenta los resultados sin más organización que la derivada de la velocidad de respuesta de cada buscador. Ahora existen los verdaderos *metabuscadores* que pueden eliminar duplicados, agrupar resultados y generar nuevos valores de pertinencia para ordenar los dichos resultados.

Entre los multibuscadores señalaremos *All-in-One* <www.albany.net/allinone>; *Avenue* <www.avenue.com>; *Digiway* <www.digiway.com/digisearch>; *FrameSearch* <www.w3com/fsearch>; *Superseek* <www.superlibrary.com/superseek>; *The Internet Sleuth* <www.isleuth.com> o *Metasearch* <www.metasearch.com>.

Una lista no exhaustiva de auténticos metabuscadores incluye en la actualidad *Cyber411* <www.cyber411.com>; *MetaCrawler* <www.metacrawlwe.com>; *Inference Find* <www.interference.com/ifind>; *Dogpile* <www.dogpile.com>; *MetaFind* <www.metafind.com>; *Insane Search* <www.cosmix.com/motherload/insane>; *Fusión* <lorca.compapp.dcu.ie/fusion>; *Savvy Search* <www.cs.colostate.edu/~dreiling/smartform.html>; *IntelliScope* <wizard.inso.com> y *Mamma* <www.mamma.com>.

1.3. Limitaciones de las herramientas de primera generación

A pesar de las numerosas y diversas posibilidades que nos ofrecen las herramientas citadas y de que su uso combinado puede ayudar en casos extremos, aun existen problemas documentales para cuya resolución se quedan cortas de prestaciones, incluyendo:

- El citado volumen de la Internet de los contenidos, con la existencia de 2 millones de servidores Web y más de 20 millones de sedes, lo que supone 300 millones de URLs y su crecimiento exponencial.

- El dispar el tamaño de las unidades documentales en el Web, ya sean sedes (algunas de ellas enormes, que superan los cuatro millones de páginas como GEOCITIES) o páginas que en algunos casos alcanzan los centenares de «pantallazos» o varios Megabytes de tamaño.

- Diferentes periodos de actualización de los contenidos de la red, que pueden generar una alta obsolescencia para sedes que se revisan poco. En el extremo contrario las sedes con periodos de actualización inferiores a las 24 horas que generan una notable inexactitud de los contenidos de índices y bases de datos.

- Algunos estudios hablan de una vida media de las páginas Web inferior a los 44 días, mientras que otros señalan que un robot puede tardar varios meses en revisar los contenidos de un buscador. Algunos de estos solo añaden nuevos recursos en periodos fijos de uno o dos días o incluso hasta una semana (Excite).

Asimismo hay que citar los problemas prácticos para la utilización de los resultados derivados de:

- El excesivo «ruido documental» de las respuestas obtenidas consecuencia de una inadecuada valoración de la pertinencia. Ello obliga a un ajuste por parte de un operador humano, que implica que los resultados han de ser clasificados e indizados una vez recuperados para adecuarlos a las necesidades del usuario. La necesidad de convertir los registros a formatos compatibles con las herramientas ofimáticas habituales. Este paso empieza a resultar imprescindible dado el gran volumen de información que se puede obtener, muy difícil de editar manualmente. Añádase además la gran variedad de formatos y la imposibilidad por parte de algunos servidores de exportar de forma continua los registros.

2. Herramientas de segunda generación

Nos encontramos con un conjunto totalmente nuevo de herramientas, diferenciadas de las anteriores porque son *client-side*. Se trata, por tanto, de programas totalmente independientes que se instalan en el ordenador cliente, lo que redunda en un mayor control y personalización de sus funciones. El hecho de que, a veces, algunas de estas herramientas pueden funcionar de forma autónoma respecto al cliente en el que estén instaladas ha llevado a que incorrectamente se generalice el nombre de *agente* o *bot*, que podría identificar sólo a alguna de ellas, no a todas.

En general, el conjunto resulta relativamente heterogéneo lo cual permite construir una clasificación muy descriptiva. Además, puesto que algunos de los mecanismos son paralelos a los que existen como servidores, dicha segregación resulta especialmente útil y admite análisis comparativos de prestaciones. Sin embargo, el valor añadido de alguno de ellos no se restringe únicamente a un incremento de la capacidad de automatización de tareas y personalización, sino que ofrecen posibilidades totalmente novedosas. Algunas de opciones inéditas resultan imposibles de implementar desde un servidor.

Entre las novedades más singulares destacaremos:

- La posibilidad de extraer información de la Internet invisible (*infranet*), el conjunto de registros de bases de datos o catálogos de biblioteca accesibles mediante formularios Web, pero que no son indizadas por los motores.

- El uso de los verdaderos agentes, que de forma autónoma, mediante mecanismos inteligentes pueden recorrer la red, extraer información e incluso «aprender» con ayuda del operador humano. La mayoría de los programas revisados son productos comerciales disponibles bajo el sistema *shareware* (evaluar antes de adquirir), lo que significa que puede obtenerse una copia de los mismos, más o menos operativa, de la red Internet. El precio no es excesivamente caro y son, precisamente, los programas más sofisticados los de mayor coste. Lamentablemente, para este tipo de programas apenas se ofrece soporte técnico y no es infrecuente que algunos títulos desaparezcan con gran rapidez.

A continuación presentamos una clasificación comentada de las citadas herramientas utilizando como criterio sistematizador las potencialidades y aplicaciones documentales de las mismas. Este criterio excluye otros programas, relativamente numerosos en la actualidad, a veces reunidos bajo la categoría de «utilidades de Internet» que son potencialmente interesantes. El interés de los mismos, fundamentalmente informático, puede ser más evidente en un futuro no muy lejano. Atendiendo a los usos documentales distinguimos cinco grandes grupos, por orden de complejidad: Clientes Z39.50, Volcadores, Metabuscadores, Indizadores y Mapeadores.

Caso aparte lo constituyen las herramientas *canalizadoras*, que tienen un carácter mixto. Basadas en la tecnología *push* podríamos calificarlas de híbridas, al requerir tanto de una instalación cliente como de una servidor. La incorporación de este tipo de servicios a los clientes universales (Netscape y Explorer) nos ha llevado finalmente a excluir estas interesantes herramientas de nuestra clasificación, donde previamente las considerábamos volcadores sofisticados. Se puede estar al corriente de las principales novedades de este tipo de programas visitando periódicamente alguno de los principales depósitos de *software* en la Internet. Por diferentes razones prácticas recomendamos utilizar los *mirrors* españoles de la red TUCOWS <tucows.arrakis.es o tucows.mundivia.es>.

2.1. Agentes de la Infranet (clientes Z39.50)

Se trata de un grupo de programas multibuscadores que extraen información de las bases de datos y catálogos de biblioteca que forman la Internet invisible. La utilización de pasarelas Web a diversas bases de datos ha permitido el acceso a un gran volumen de información, que aún estando disponible a través de la red, resulta invisible (no indizada) para los motores de búsqueda.

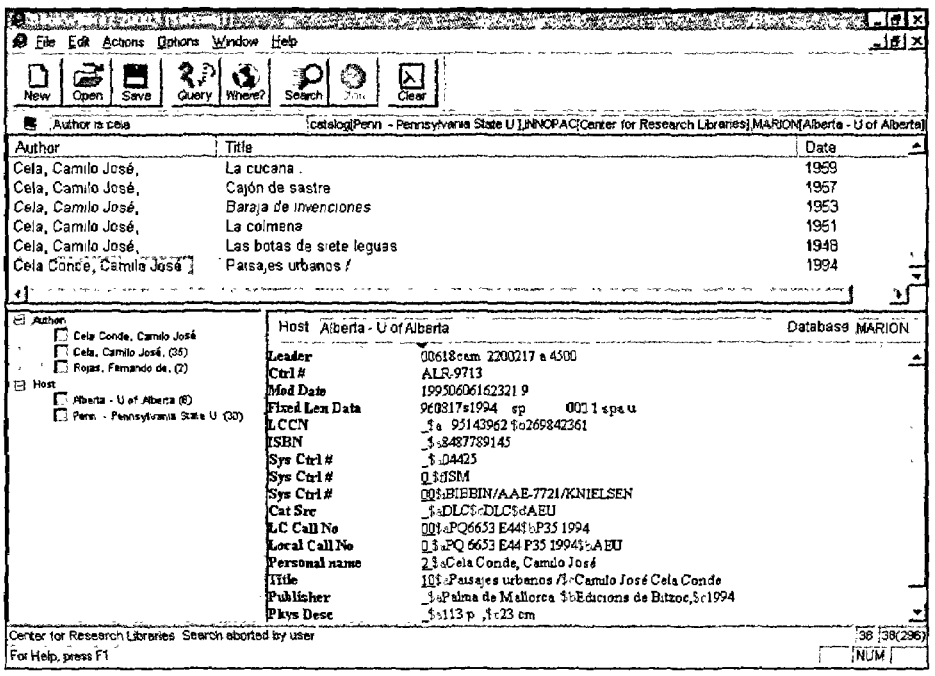
Una parte relevante de esa información corresponde a registros bibliográficos de las principales bibliotecas y grandes bases de datos de todo el mundo. Desde hace algunos años se ha venido trabajando en la homologación de un sistema de acceso universal a dicha información independientemente de los programas de automatización y gestión que se utilice en cada institución.

A finales de los años ochenta nació la norma Z39.50 (ahora ISO 23950) que permite la utilización de un interfaz, lenguaje de interrogación y formato de presentación de datos únicos. La implantación de esta norma en Internet aprovechando la arquitectura cliente-servidor supone un salto cualitativo, al permitir la interrogación simultánea de varias sedes independientes. Puesto que una parte cada vez más relevante de bibliotecas en todo el mundo soportan esta norma es posible extraer registros con un alto grado de precisión y exhaustividad. Recientemente el gobierno federal ha generalizado el uso de esta norma para el acceso a todos los catálogos de sus documentos oficiales (GILS), lo que aumenta el número de bases de datos utilizables. Se puede obtener información adicional de la sede mantenida por la Biblioteca del Congreso de Washington <lcweb.loc.gov/z3950>, donde figura una lista actualizada de bibliotecas con catálogos que tienen instalado el módulo Z39.50.

Las pasarelas Z39.50 originales diseñadas para acceder desde el Web a un único catálogo han dado paso a los clientes Z39.50 que permiten el acceso a gigantescos catálogos virtuales fruto de la suma de registros de todos aquellas bases de datos que soporten esta norma.

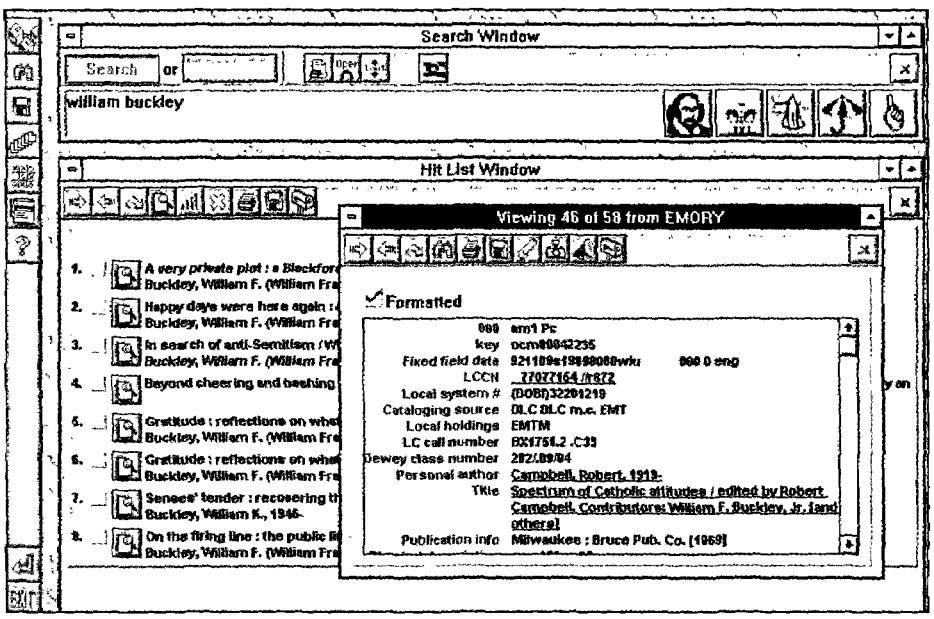
Este tipo de programas suelen ser comerciales, aunque es posible obtener de la red versiones *shareware* o de demostración, con algún tipo de limitación de las verdaderas posibilidades del programa número de registros). Aunque la mayoría son programas multibuscadores, en la medida que abren sesiones distintas que admiten estrategias de búsqueda independientes, también existen algunos clientes más sofisticados que podríamos llamar metabuscadores. Estos son capaces tanto de lanzar una única estrategia común, como integrar los resultados en una única pantalla, proporcionado datos adicionales resultado de la extracción y clasificación de algunos campos segregados (autores, bibliotecas).

Hay todavía muy pocos clientes, de forma que los programas que revisamos a continuación son realmente los únicos de esta categoría. La empresa Seachange <www.bookwhere.com> produce los dos clientes más interesantes:

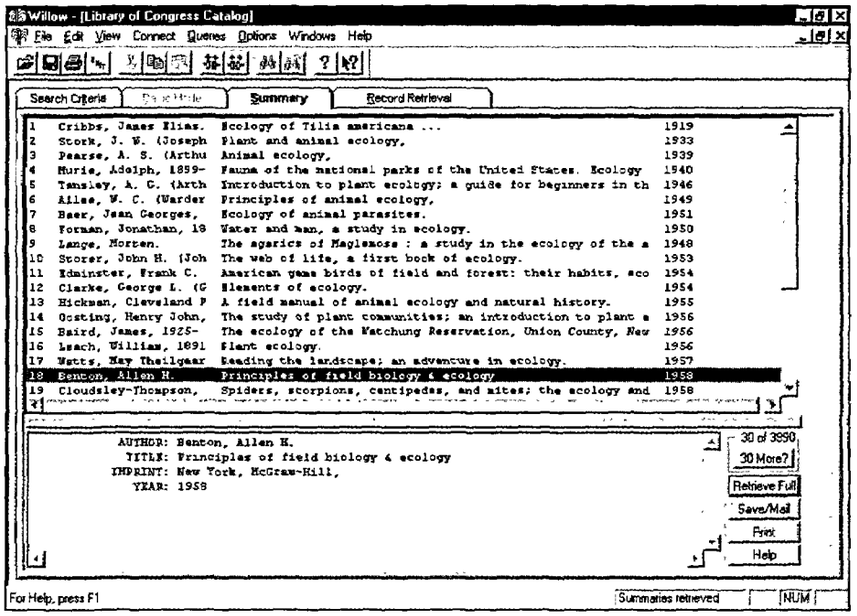


Bookwhere Pro (versión 2.1) que funciona tanto sobre Windows 3.x como sobre Windows 95 y *Bookwhere 2000* que solo opera bajo W95. Además de una presentación más organizada, la principal diferencia entre ambos programas es que el segundo soporta la versión 3 de Z39.50 y, por tanto, GILS. No obstante, parece menos robusto en estas primeras versiones y tiende a ser inestable.

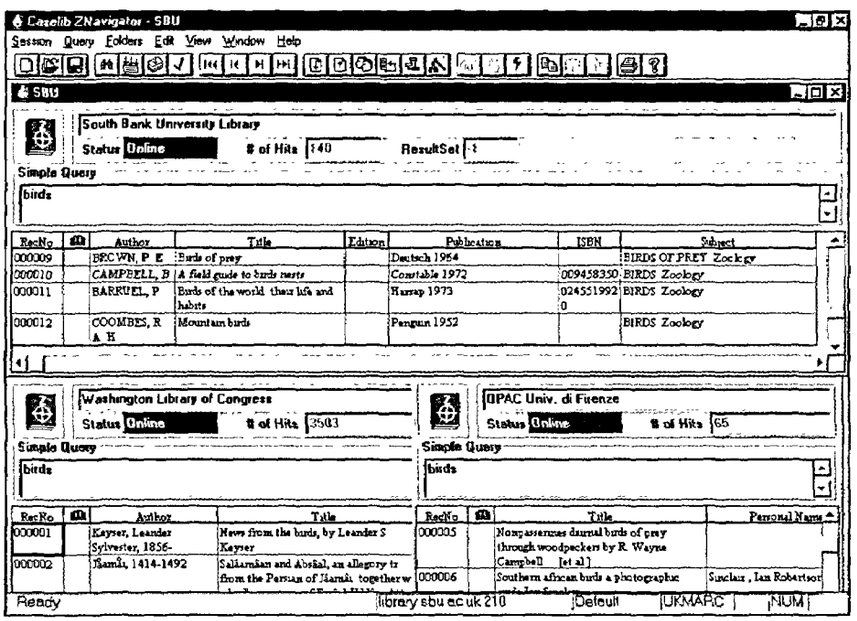
La principal ventaja de Bookwhere es su capacidad integradora de resultados, de forma que todos los registros se ofrecen en una única base de datos. Los registros (MARC) pueden ser exportados en una gran variedad de formatos, incluyendo una pasarela directa a ProCite y Reference Manager (excelentes sistemas de gestión bibliográfica, capaces de eliminar duplicados de forma automática). Es un programa muy flexible y fácil de configurar para nuevos catálogos.



La casa Sirsi <www.sirsi.com> ofrece una versión demo de su producto comercial *Vizion Pro* (v. 2.0), que resulta imposible de evaluar en profundidad. No obstante, puede ser programa interesante como se deduce de la inspección de las pantallas y la consulta de la ayuda.



La Universidad de Washington <www.washington.edu/willow> ha desarrollado *WinWillow* (versión 1.42) que corre bajo plataformas Windows (3x y 95) y que resulta especialmente interesante cuanto la consulta se realiza a través del servidor de dicha Universidad. Ello permite evitar algunos cuellos de botella que afectan a ciertos servidores a determinadas horas. Sin embargo, la flexibilidad del producto final se resiente y sólo resulta aconsejable su uso cuando los catálogos que nos interesan no son alcanzables de otro modo.



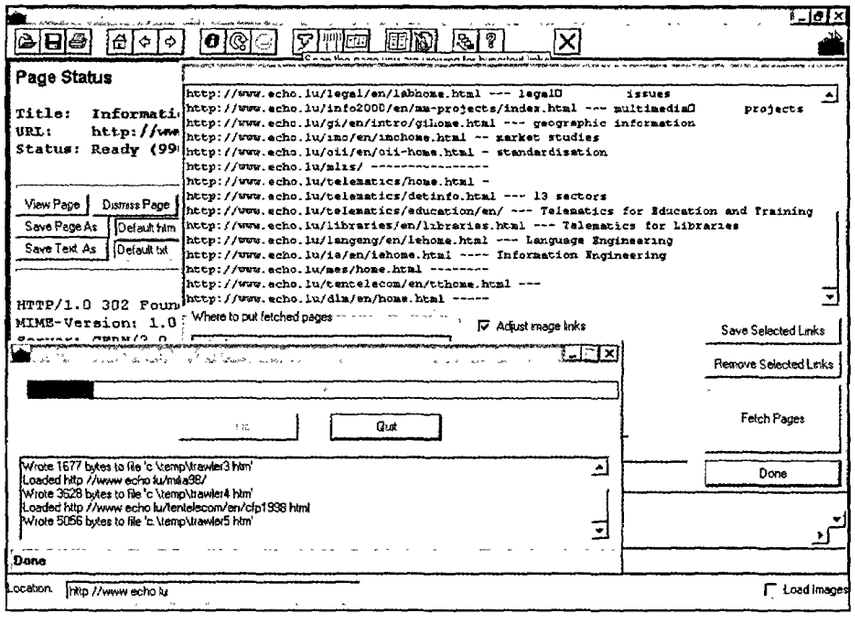
El proyecto CASELIB es una de las muchas iniciativas sobre Z39.50 desarrolladas al amparo del Programa de Telemática para Bibliotecas de la Comisión Europea, pero es la única que ha desarrollado un cliente autónomo operativo, Znavigator (versión 1.0h). Este proyecto, en el que participa la Biblioteca de la Universidad de Alcalá de Henares <www.alcala.es/biblio/caselib.htm>, tiene como socio comercial a la empresa española EnWare. Sin embargo, todavía no se ha decidido la explotación del programa, que resulta especialmente potente.

Al contrario que Bookwhere, Znavigator funciona con sesiones independientes para cada catálogo en un entorno multiventana. En cada ventana se puede ejecutar una estrategia distinta, lo que eventualmente puede resultar confuso. No obstante, las opciones de manipulación de registros son superiores a las ofrecidas en Bookwhere.

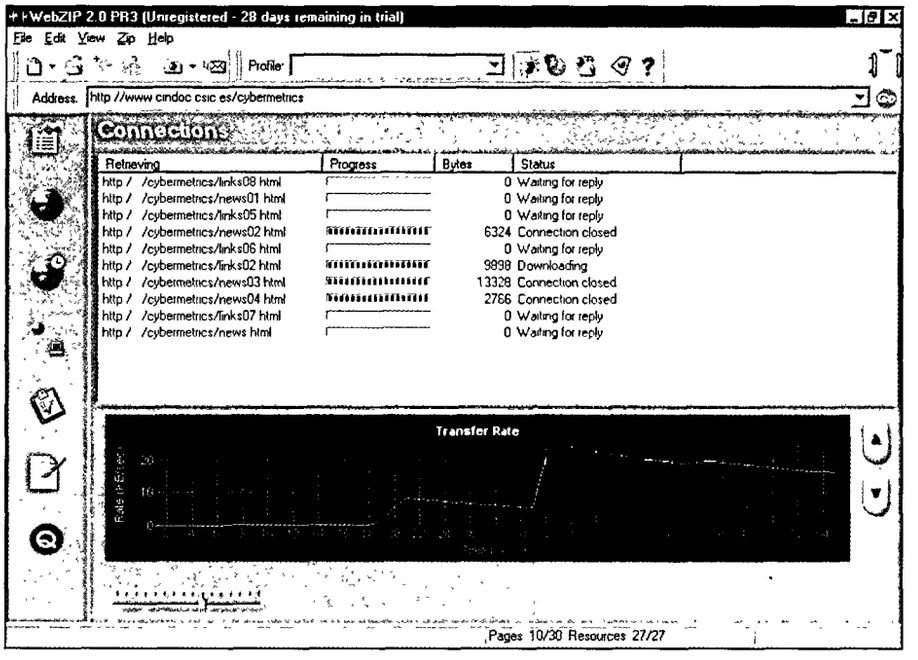
2.2. Programas volcadores y navegadores *offline*

Programas que permiten el volcado y archivo eficaz de sedes Web independientemente del tipo y volumen de los ficheros involucrados. Los distintos productos ofrecen diferentes prestaciones que permiten agilizar y automatizar esta tarea, muy penosa por otros medios.

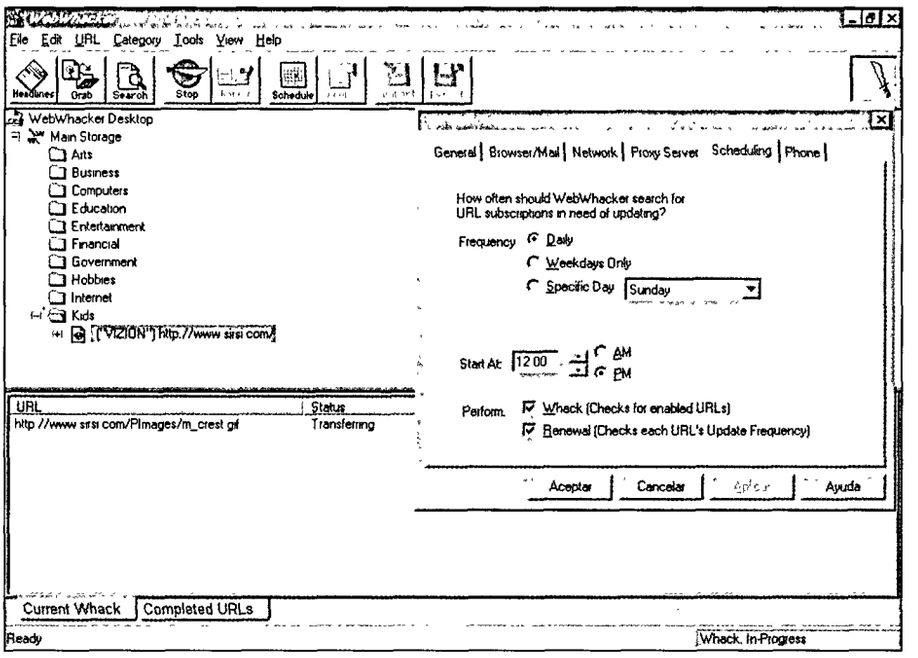
El número de programas *shareware* en esta categoría muy notable, por lo que sólo citaremos aquellos que, por una u otra razón, ofrezcan alguna prestación destacable. En general los precios de estos programas son muy bajos y la rentabilidad bastante elevada. Asimismo tampoco son demasiado exigentes en cuanto a prestaciones y la mayoría funcionan tanto bajo Windows 95 como en las diferentes versiones 3x.



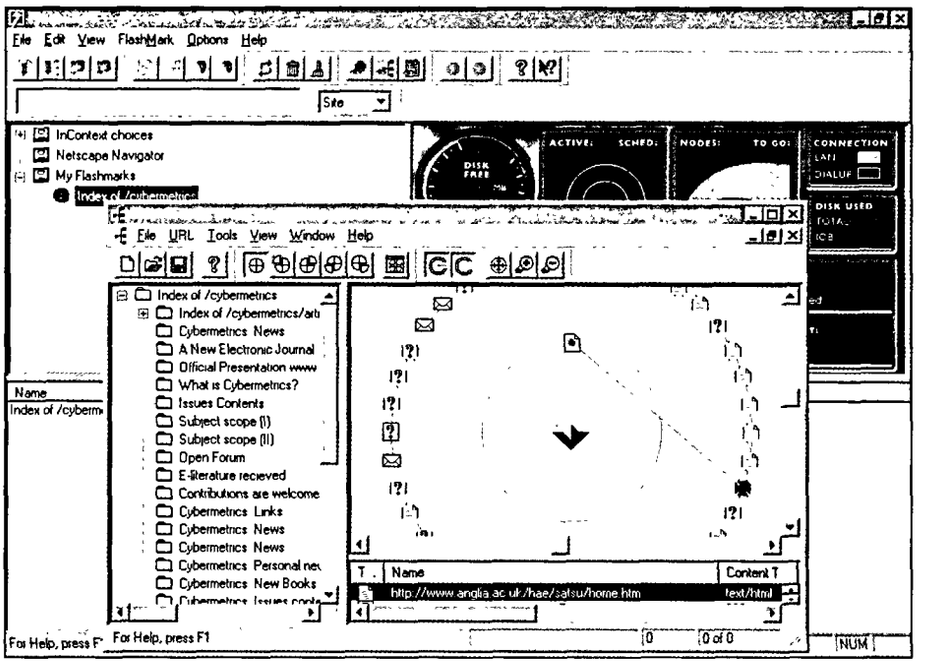
Con un sencillo interfaz, pero capaz de realizar volcados exhaustivos de sedes identificadas según los resultados de una búsqueda en alguno de los grandes motores *Arf* (versión 3.22) es distribuido por Bitsafe <dwave.net/~bitsafe>. Esta empresa también produce *Trawler* (versión 1.8), más configurable y capaz de trabajar sobre varias bases simultáneamente. Aunque podría considerarse una versión sofisticada (tiene algunas mecanismos de extracción de direcciones y filtros configurables) de Arf, ambos programas pueden utilizarse para tareas complementarias o, incluso, diferentes.



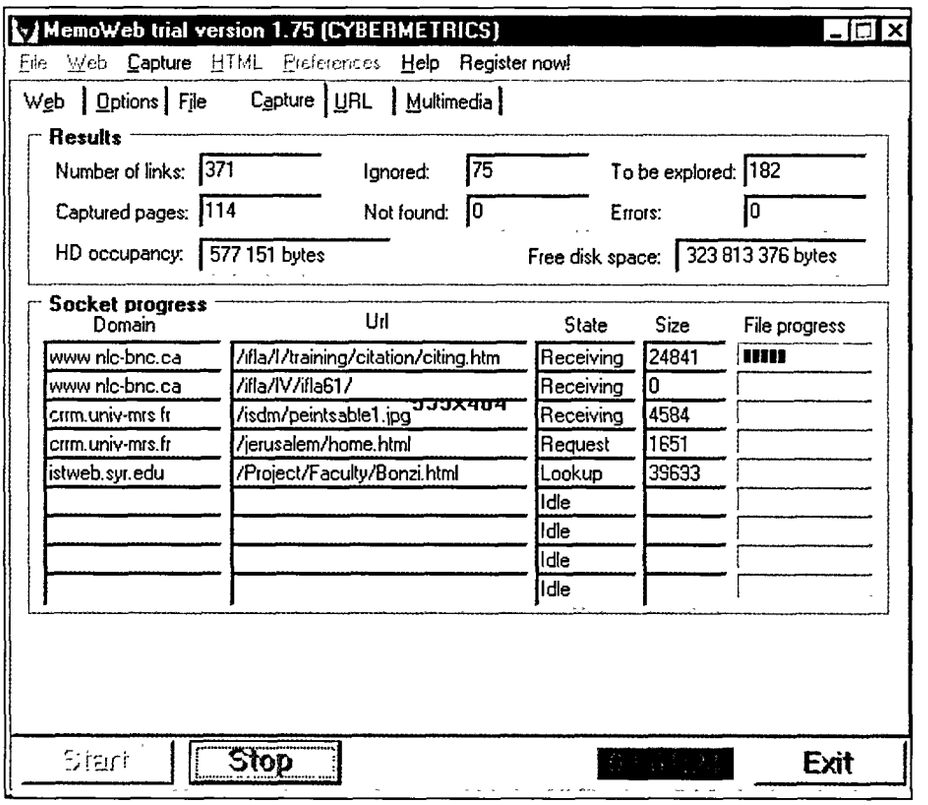
El interesante *WebZip* <www. spidersoft.com>, ya en su versión 2.0, es recomendable para construir archivos de sedes. Puesto que los ficheros se guardan comprimidos, resulta muy útil si se pretende que ocupen poco espacio de disco duro, realizar grandes volcados, o hacerlo con una elevada frecuencia. Posee algunas de las pantallas mejor estructuradas, más informativas y gráficas de este grupo y todos los procesos son transparentes. Además se pueden programar las sesiones.



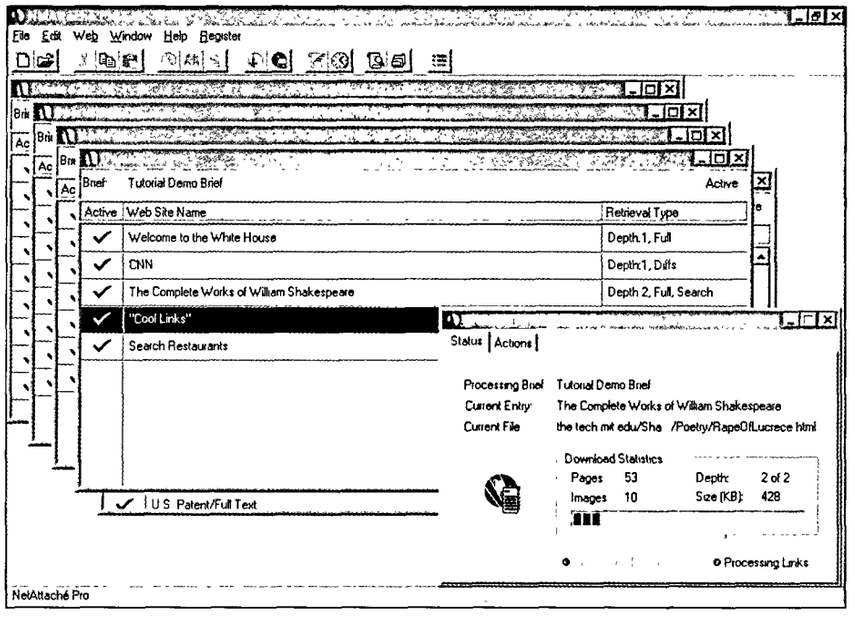
WebWhacker (v. 1.0 para W3x ó 3.2 para W95) es uno de los dos productos comerciales más interesantes de este grupo, aunque se puede evaluar durante quince días recuperándolo de la sede de Blue Squirrel <www.bluesquirrel.com>. Es, posiblemente, el más completo y configurable de los volcadores y está pensado para aquellas personas que dispongan de conexiones lentas (vía módem). En efecto, los volcados se pueden programar a horas de baja congestión y el programa es capaz incluso de descolgar y colgar la línea telefónica de forma automática.



FlashSite 1.02 (W95) es el otro programa comercial (www.everyware.com> que recomendamos. Funciona tanto volcando como mostrando los resultados, es muy gráfico y puede programarse como en el caso anterior. Además dispone de un índice (un gestor de *bookmarks* en realidad) muy cómodo para guardar el registro de los trabajos efectuados.

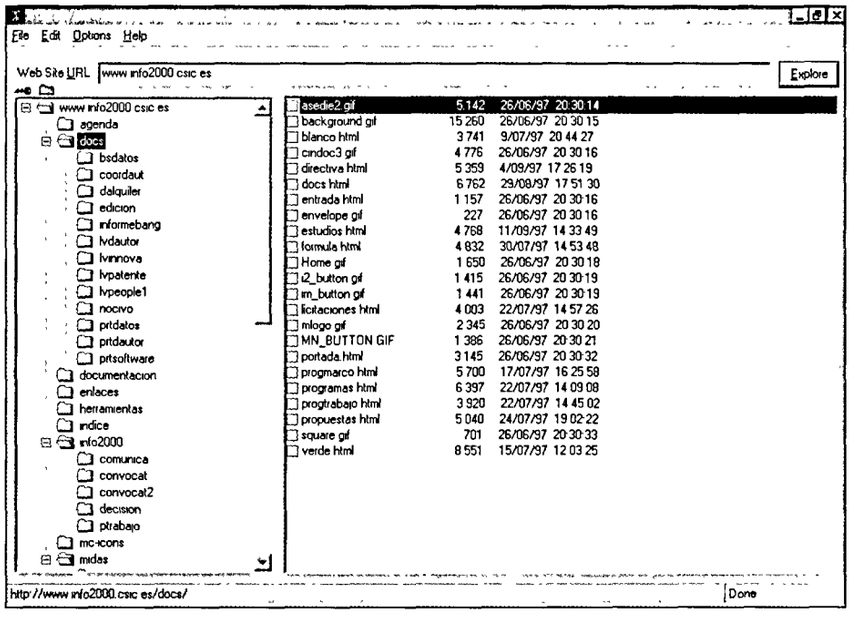


MemoWeb dispone de dos versiones (1.75) distintas para W3x y W95 <www.memoweb.com>. Esta última soporta plenamente la multitarea lo que hace de este programa uno de los más rápidos en su trabajo, además de contar con un interfaz moderno muy configurable. Entre otros aspectos destacables figura una sofisticada opción de filtros.

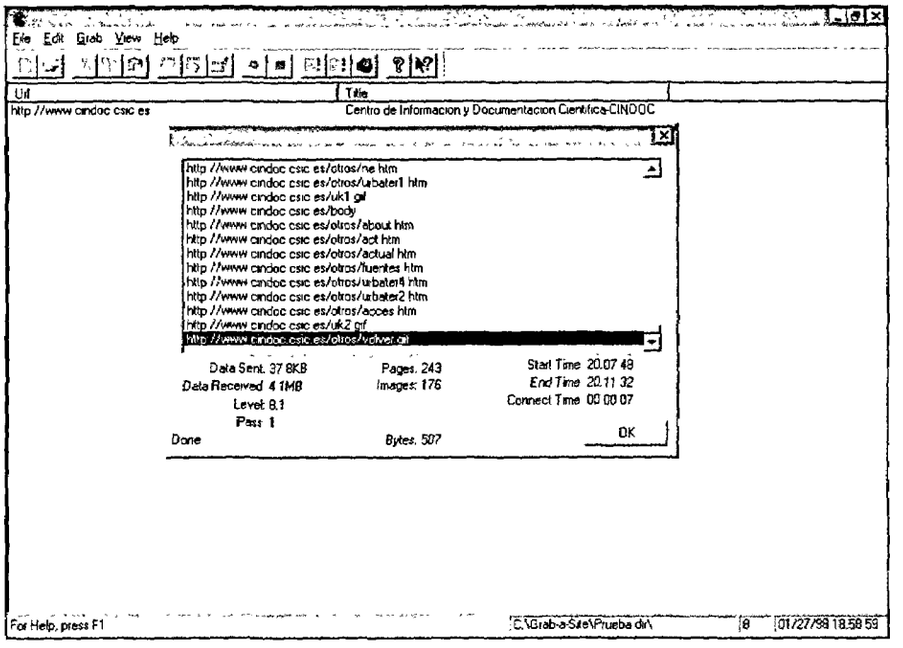


NetAttaché Pro 2.50e (W3x/ W95) de Tympani Developments <www.tympani.com> ofrece todas las funciones citadas para otros productos, pero presenta un sistema de organización de las sedes volcadas un poco confuso. La idea es excelente puesto que cubre todo el proceso, desde la búsqueda en un motor hasta la navegación *offline* de los volcados. Sin embargo, dada su estructura visual sólo resulta aconsejable si se pretende mantener un elevado número de sedes visitadas. En ciertas circunstancias, los programas basados en tecnología *push* resultan más apropiados que este que revisamos.

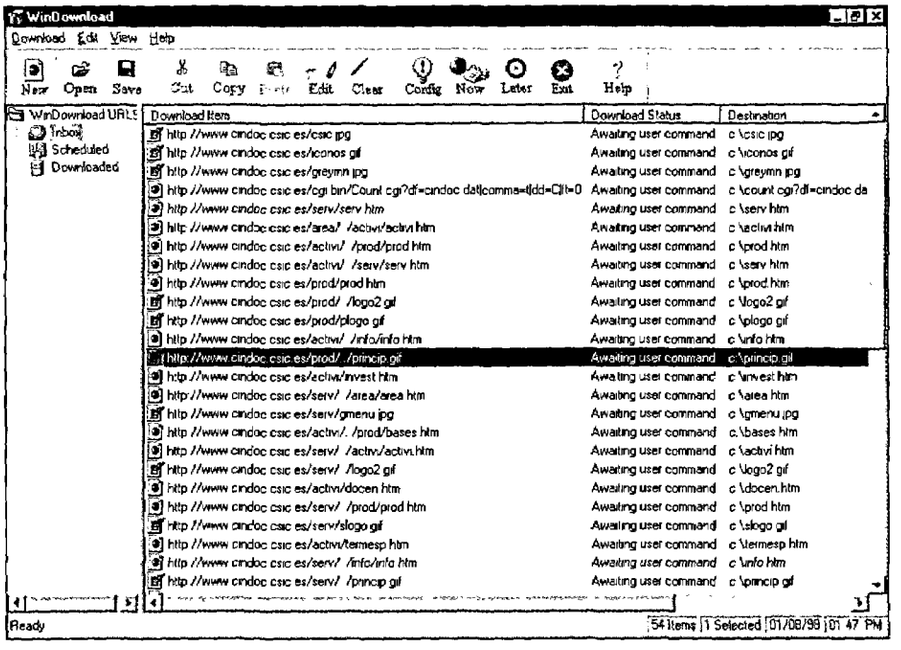
Entre los programas más sofisticados se encuentran los dos siguientes que sólo funcionan bajo Windows 95:



BlackWidow está en continua renovación, estando disponible la versión 3.4 <www.softbytelabs.com>. Es capaz de mostrar la estructura completa de una sede, pudiendo posteriormente individualizarse los ficheros que se desean volcar. Lo más destacable es que define como unidad de trabajo el fichero y no el directorio, lo que ofrece una gran flexibilidad. Una ventaja adicional es que estos perfiles pueden ser grabados para volcados periódicos.



Grab-a-Site (v. 3.0) es otro producto de Blue Squirrel, que además de volcar se puede utilizar como visualizador *offline*. El modo de funcionamiento es similar al anterior, puesto que también reconstruye primero la estructura de la sede. Sin embargo, el volcado se realiza a través de filtros, de forma que podemos personalizar que tipo de ficheros queremos recuperar.

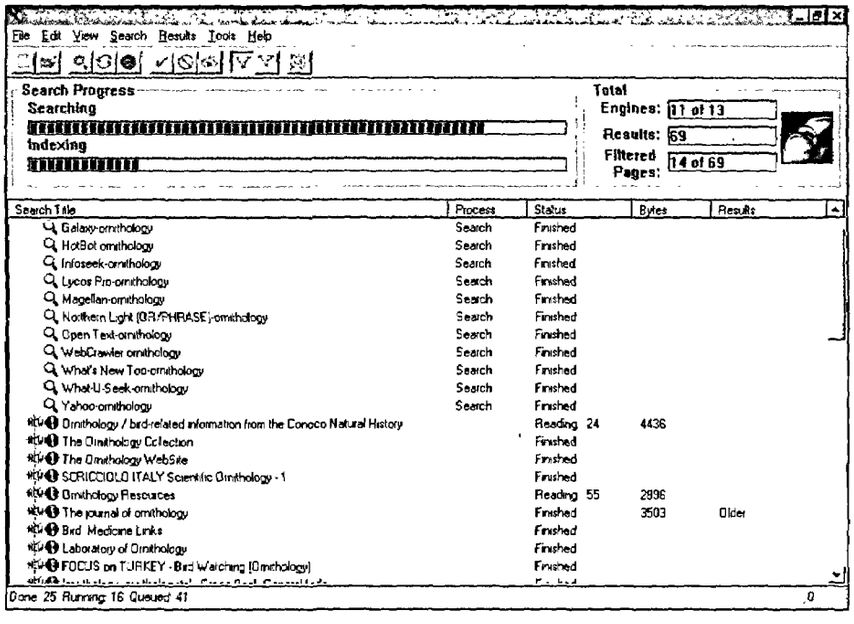


Una de las últimas incorporaciones es la versión 4.01 de *WinDownload* <mason.gmu.edu/~rmcclana/windownload.html>, que es uno de los representantes con más opciones de una categoría de programas ligeramente diferentes a los volcadores. En realidad, se trata de clientes ftp sofisticados, con opciones que permiten programar las transferencias y recuperar los volcados incluso tras varios episodios de corte de conexión. La mayoría de estos programas (*ftpeadores*), que no han sido incluidos, sólo pueden trabajar monosesiones. Citamos este por sus opciones avanzadas y sus posibilidades multisesión.

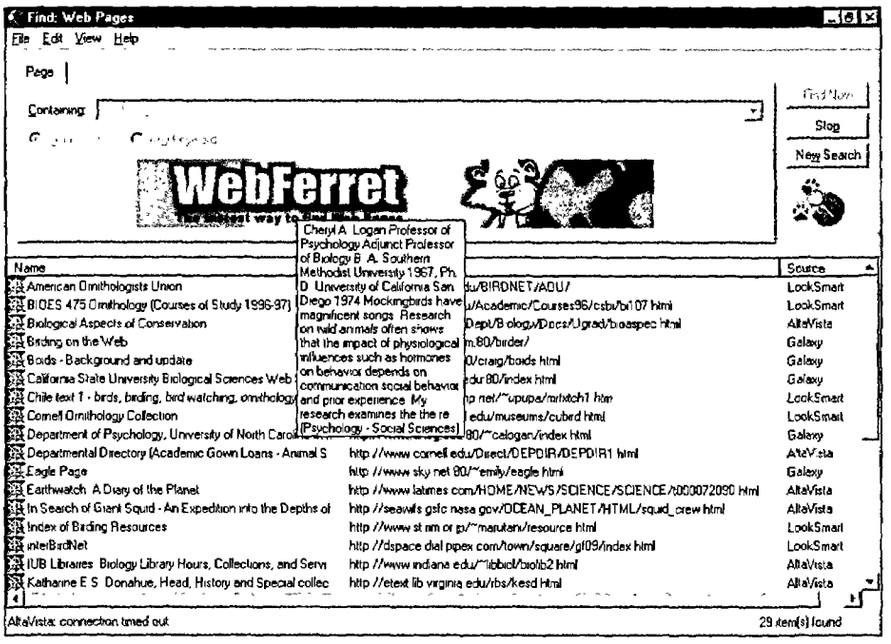
2.3. Multibuscadores y agentes metabuscadores

La versión cliente de los que hemos visto en el grupo de primera generación, con importantes novedades, entre las que destacamos sus capacidades de personalización, programación y exportación. Esta última opción supone una ventaja considerable a favor de los multibuscadores de segunda generación.

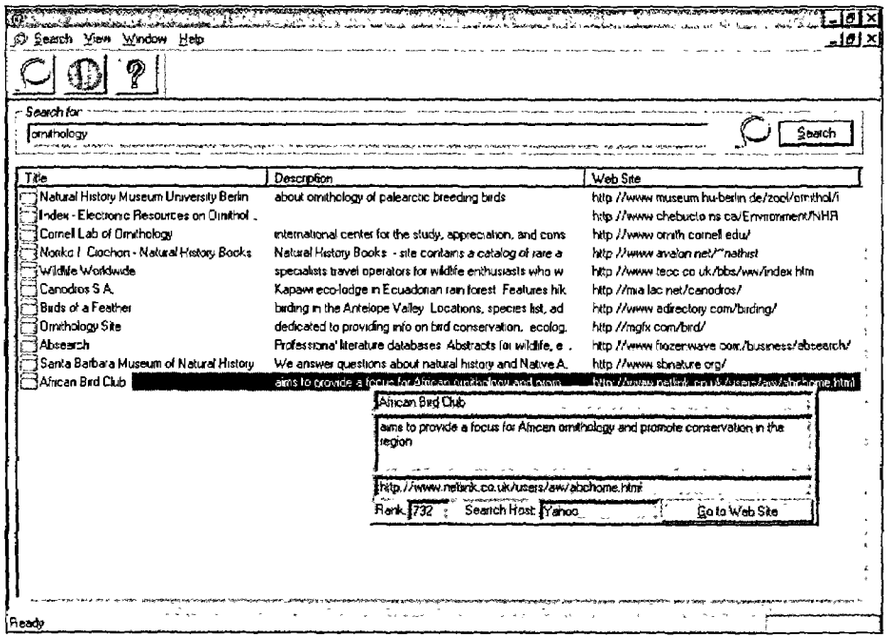
Sin embargo, lo verdaderamente llamativo es la posibilidad de lanzar trabajos *offline* realizados por agentes «inteligentes». En realidad, el usuario filtra manualmente los primeros resultados presentados por los programas lo que ayuda a perfilar la estrategia de búsqueda. Este proceso remeda bastante convincentemente al del aprendizaje, de forma que los resultados obtenidos tras unos pocos filtrados resultan sorprendentes. Este último grupo de programas recibe el nombre de *metabuscadores*.



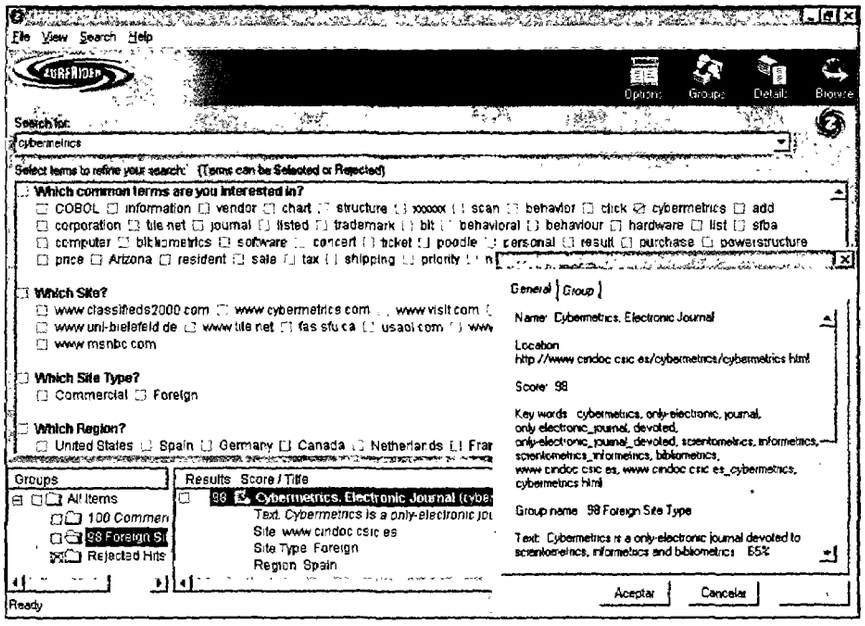
Iniciaremos la revisión con un programa netamente comercial, de la ya citada empresa inglesa Blue Squirrel <www. bluesquirrel.com>. Se trata de *WebSeeker* (v. 3.3), capaz de realizar la búsqueda en 100 motores (incluyendo Northern Light), con una presentación y unas prestaciones *muy cuidadas. A destacar las posibilidades de eliminar duplicados* y revisar periódicamente los resultados, proceso que se puede automatizar. En el extremo opuesto, sus limitadas capacidades de exportación de registros.



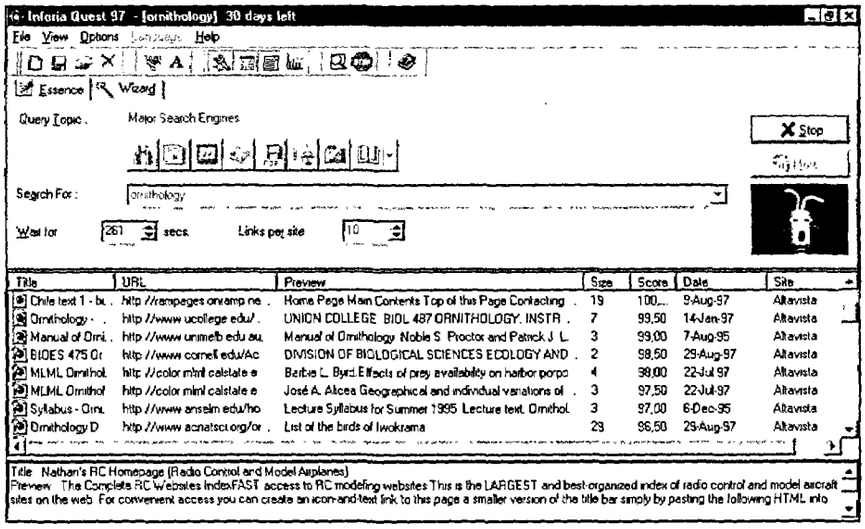
Ferretsoft <www.ferretsoft. com> lleva ya bastante tiempo perfilando las características de su *WebFerret Pro* (v. 2.01), aunque lo mantiene como un programa pequeño, diseñado más como auxiliar en la navegación. Asumiendo esos fines modestos cumple holgadamente su misión



De características muy similares al anterior *Lazo 2.0* de VaultBase <www.vaultbase. com>, ofrece las respuestas distribuidas por campos, lo que puede ser más útil a la hora de exportarlas.

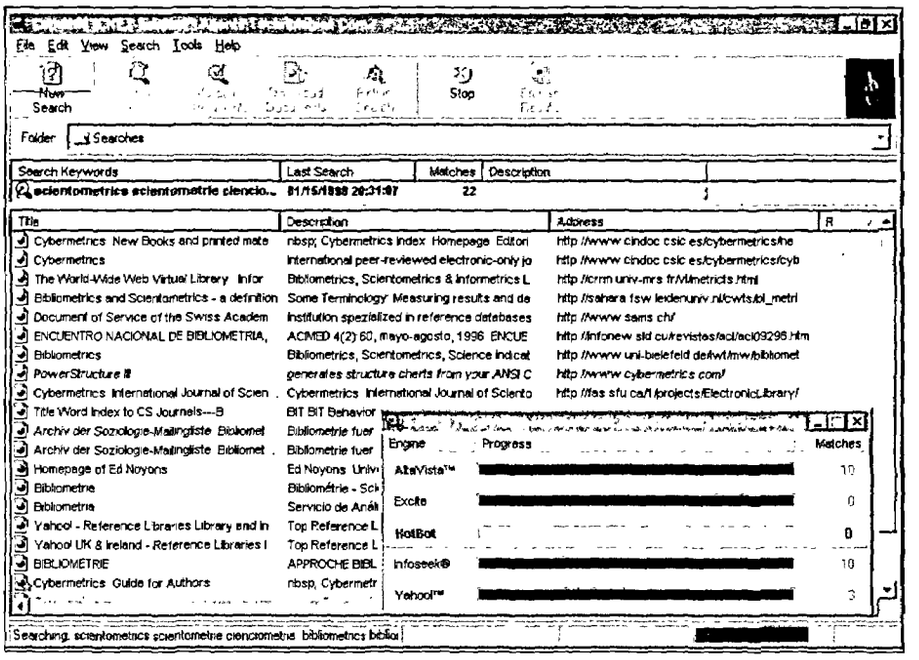


En una categoría muy superior nos encontramos con *ZurfRider 1.0*, un producto nuevo <www.zurf.com>, que incorpora un sistema de refinado múltiple atendiendo a distintos criterios (palabras, direcciones, región geográfica), cuyos términos se pueden activar o desactivar según las necesidades. Las opciones de exportación son excelentes, aunque se dejan notar diversos *bugs* que creemos se resolverán en próximas versiones.

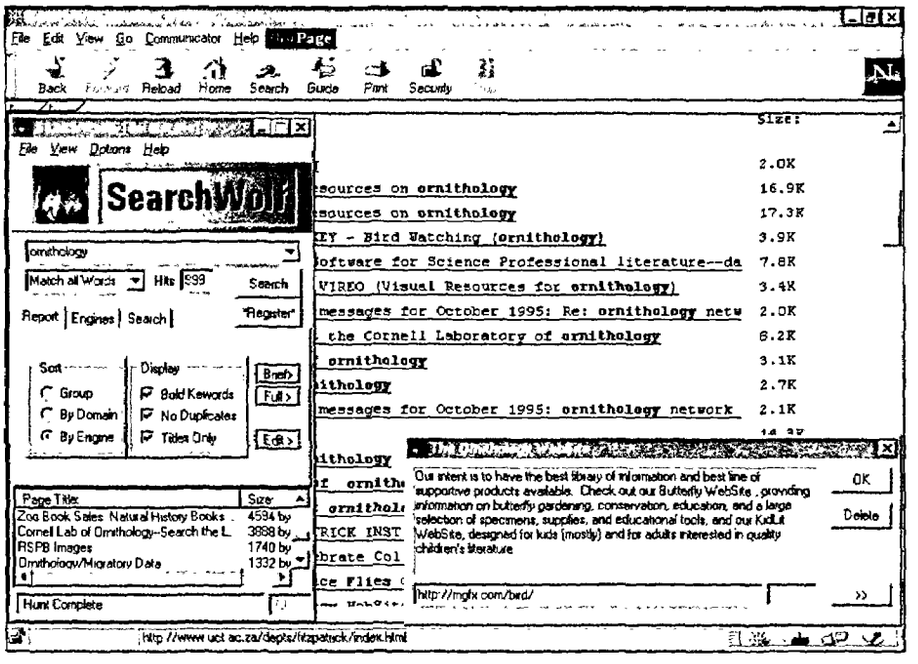


Como se indica en su propio nombre, *Quest 98* de Inforia <www.inforia.com> es otro producto nuevo que avanza en las posibilidades que ofrecían los anteriores, prestando especial cuidado a los mecanismo de configuración y filtro.

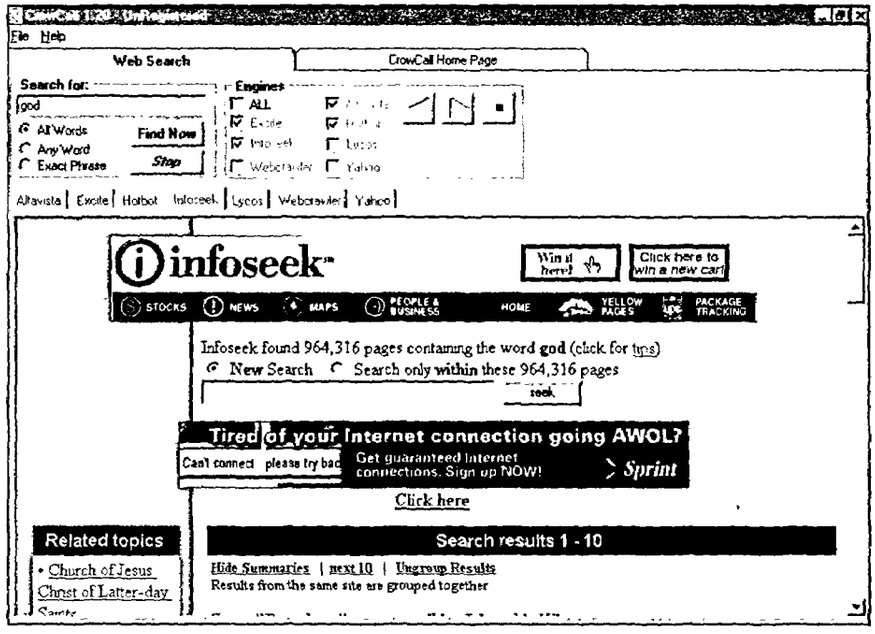
Symantec <www.symantec. com> ha retirado *WebFind FastFind* 1.0, que formaba parte de una interesante *suite* comercial.



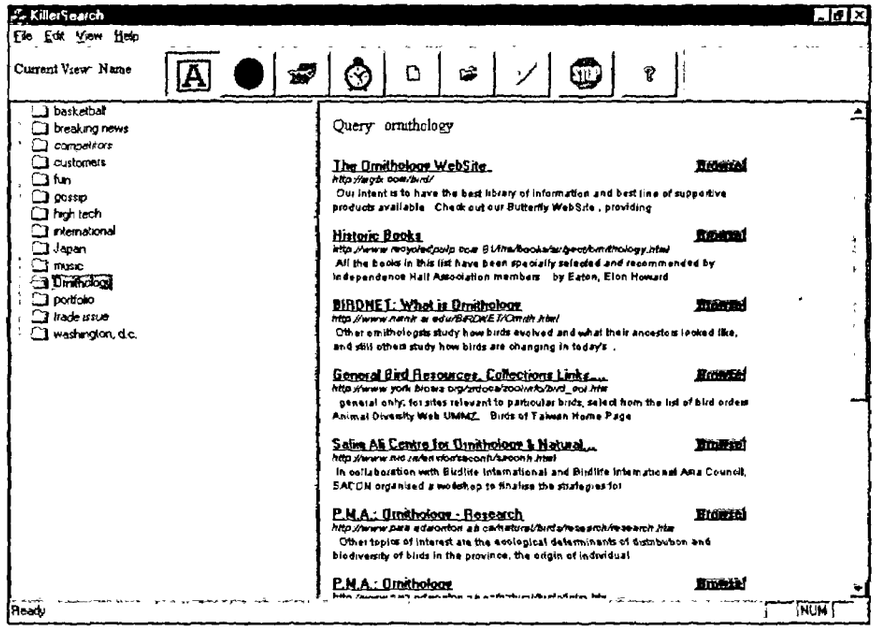
Copernic 1.2 es el nuevo nombre del multibuscador de Agents Technologies <www. agents-tech.com>, que posee uno de los sistemas de exportación más interesantes, tanto por el formato final (html) cuanto por la capacidad de eliminar duplicados y organizar los resultados.



La empresa MSW <www. msw.com.au> edita una amplia serie de herramientas de segunda generación que, a medida que van alcanzando versiones superiores, resultan más robustas e interesantes. *SearchWolf 2.02* es el multibuscador, que ofrece las respuestas organizadas como páginas Web. Sin embargo, el nivel de control es muy limitado y los resultados no resultan fáciles de utilizar.

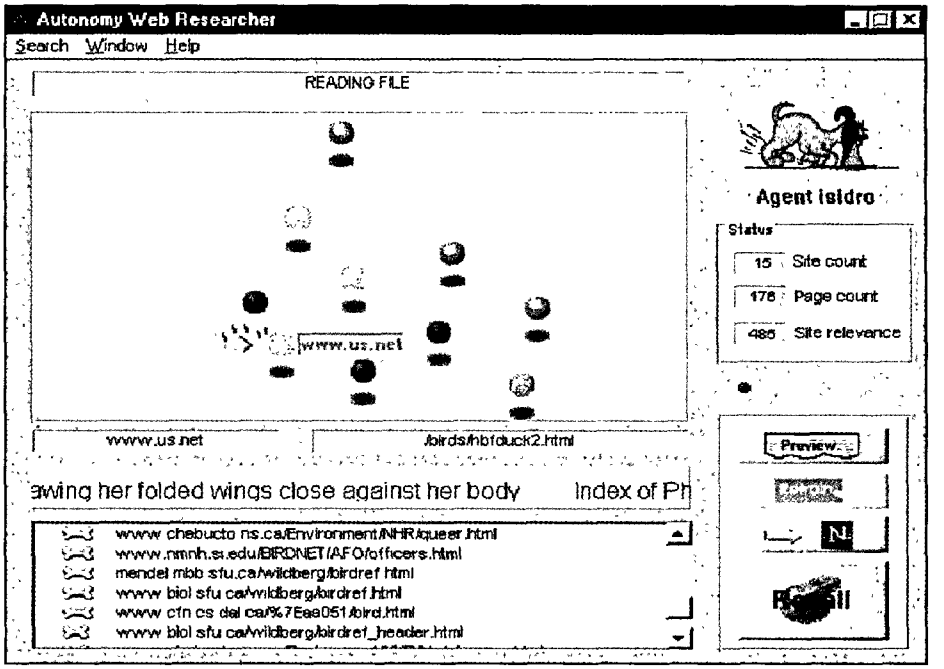


CrowCall 1.20 <www.alphalink.com.au/~pbrooks/CrowCall/index.htm> es un buscador pluriventana, capaz de mostrar los resultados de búsqueda en la propia página de cada motor, ofreciendo posibilidades de navegación como cualquier *browser*. No resulta especialmente recomendable sobre otros productos.

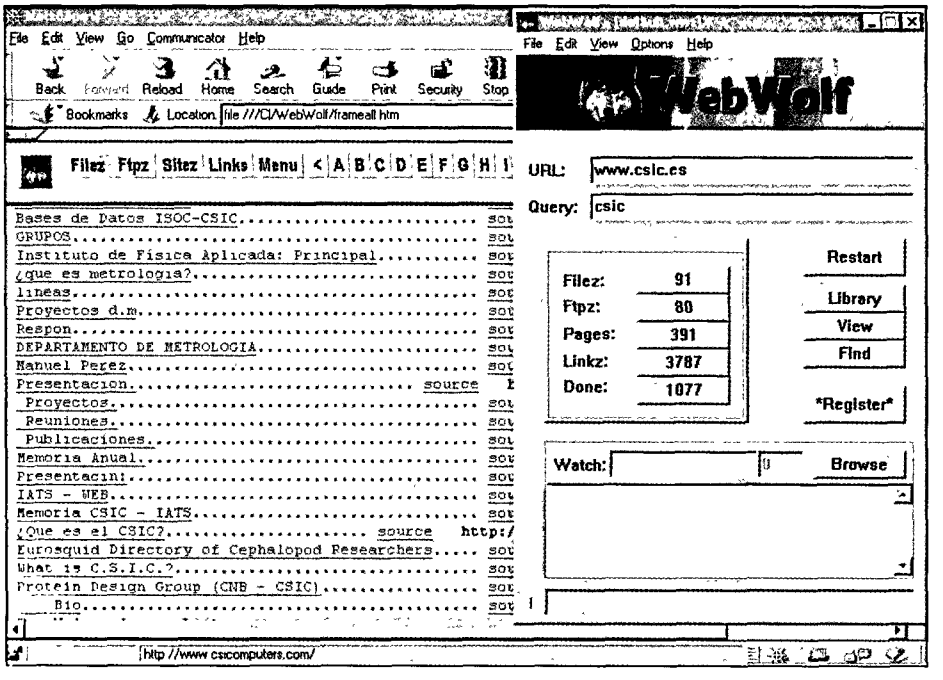


Killer Search 1.12 <www.killersearch.com> presenta un sencillo interfaz que evita el tener que entrar a los buscadores de uno en uno. Sin embargo, la disposición de los resultados es tal como la realizan los diferentes motores. Algunas opciones para programar las sesiones en horarios no habituales pueden resultar interesantes.

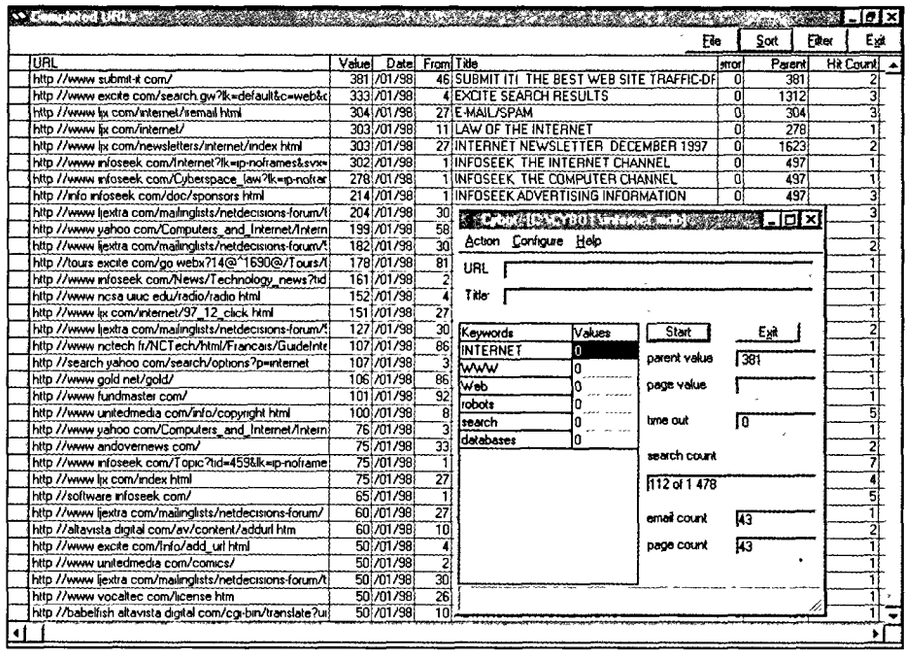
Los tres siguientes son auténticos metabuscadores y utilizan agentes inteligentes en el desarrollo de su tarea. El mecanismo resulta extremadamente simple pero los resultados son espectaculares. Desde una serie de páginas semilla (índices o motores) el agente localiza direcciones que puedan cumplir los criterios de la estrategia definida. La búsqueda continua explorando los enlaces y comprobando que las páginas alcanzadas son pertinentes. Si ocurre así se genera una trayectoria que une los recursos pertinentes a través de los enlaces cruzados. Existe la posibilidad de interrumpir el proceso y aleccionar al agente perfilando la estrategia, bien directamente (añadiendo o excluyendo más términos), bien indirectamente validando manualmente las sedes propuestas. El agente interpreta esas instrucciones y prosigue su labor, incluso de forma *offline* o con apoyo de librerías externas creadas por otros usuarios o los productores del programa.



El pionero es *Agentware* 2.0 de Autonomy Systems <www. agentware.com>, que sigue ofreciendo uno de los sistemas más pintorescos, aunque muy interesante. La utilidad final queda algo mermada por la lentitud del agente y las limitaciones de exportación, aunque merece la pena probarlo aunque sólo sea por conocer el concepto en el que se basa (destacar el sistema de aleccionamiento).



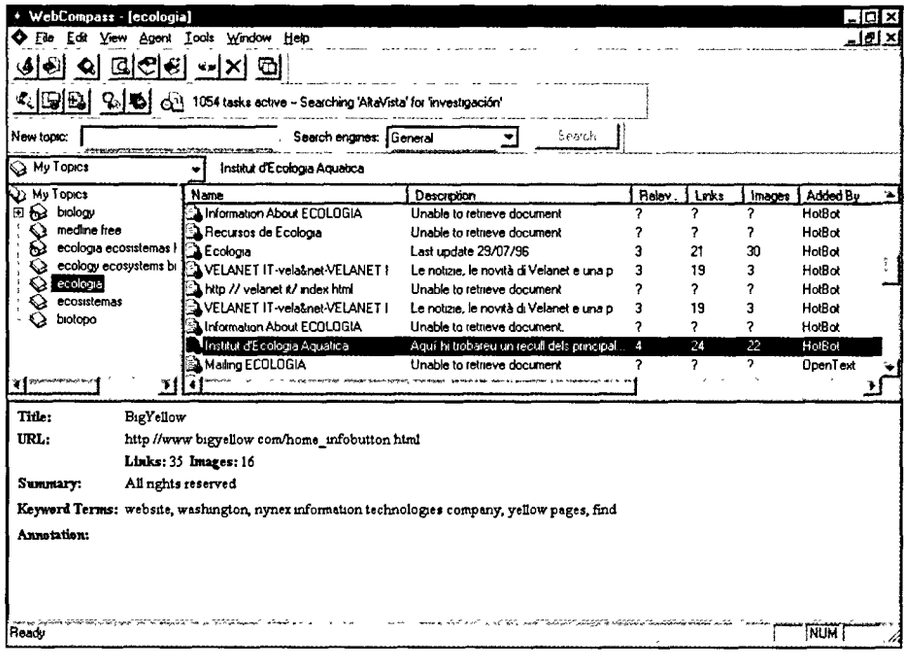
WebWolf 2.02 (W3x y W95) es el metabuscador de MSW. Destaca por su presentación final de los resultados, que pueden ser utilizados inmediatamente. Asimismo presenta la particularidad de poder restringir el proceso a una única dirección, lo que resulta especialmente útil en sedes muy grandes.



Cybot 2.42 de Virtual Gallery <www.the ArtMachine.com/ Cybot.htm> es un potente metabuscador que permite además utilizar los motores de búsqueda como semillas de inicio. Se pueden añadir o editar nuevos motores u otras direcciones, indicando mediante un ingenioso sistema de pesos las prioridades en los procesos de búsqueda. El programa resulta muy potente porque a través de la descripción de los resultados es posible perfilar la asignación de pesos. Destacar, por último, un excelente mecanismo de exportación.

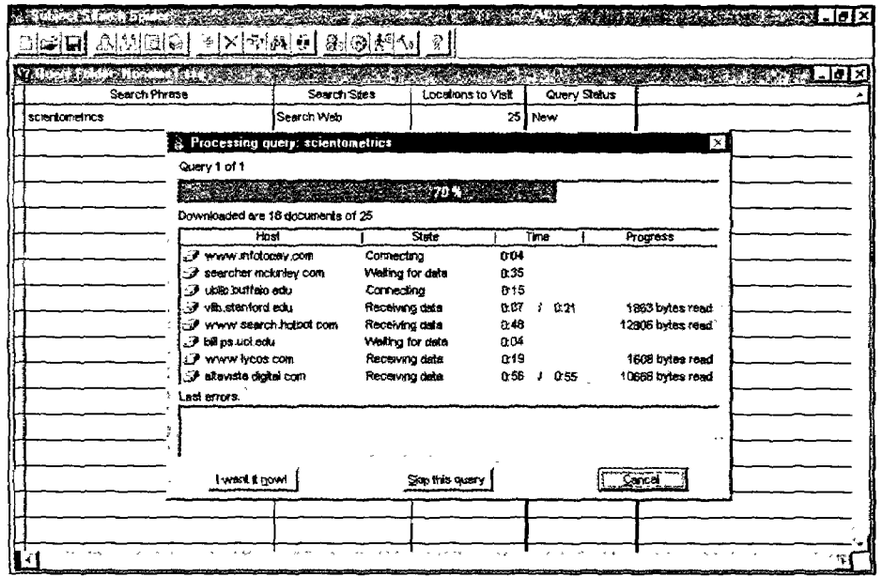
2.4. Multibuscadores que generan índices y/o extraen resúmenes

Un conjunto de programas que suponen una evolución natural de los anteriores, en la medida que son capaces de extraer palabras clave o resúmenes de los contenidos de páginas Web identificadas tras la realización de una multibúsqueda con la estrategia deseada. El resultado es un perfil anotado, que además se puede editar. Realizando este último proceso de forma manual (estrategia global semiautomática) se puede generar un magnífico documento de síntesis.



Quaterdeck <www.qdeck.com> distribuye comercialmente *WebCompass* 3.0, uno de los mejores programas de los que figuran en esta revisión. Totalmente configurable y muy visual, tanto los índices como los resúmenes que proporciona son espectaculares.

WebSleuth 1.52 de Prompt Software <www.promptsoftware.com> profundiza en la descripción de las sedes, extrayendo no solo las palabras clave y los conceptos principales, sino que además los organiza por orden alfabético. Con la ayuda de un visualizador de las páginas siempre se tiene acceso al documento fuente.

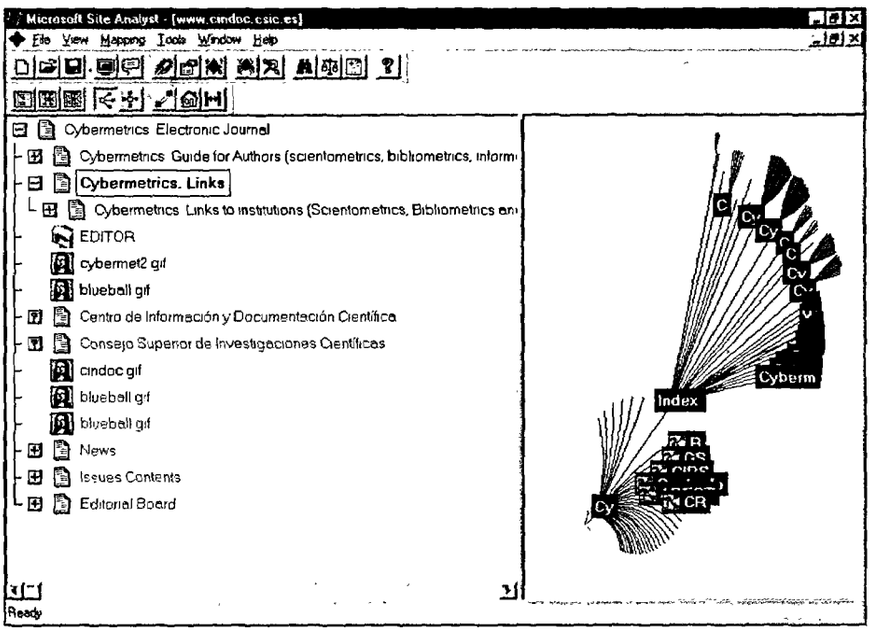


Subject Search Spider 1.03 <www.kryltech.com> es un multibuscador muy sofisticado que ofrece una presentación completa con registros listos para su utilización. La indización es poco sofisticada, pero adecuada para la preparación de pequeños informes o directorios anotados.

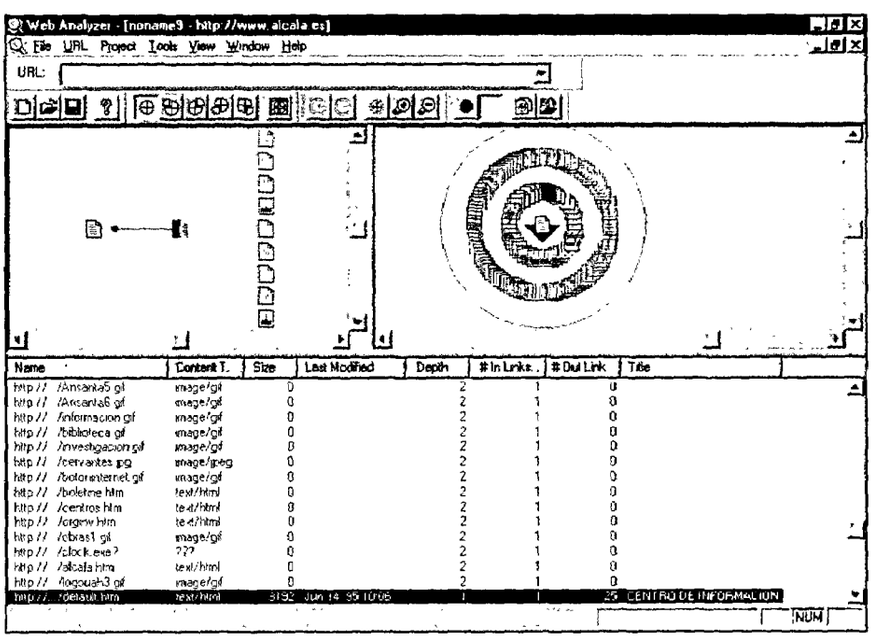
Por último, citaremos que otro programa destacable, *EchoSearch* 2.01, ha sido retirado, lamentablemente de forma definitiva, por Iconovex <www.iconovex.com>.

2.5. Programas mapeadores

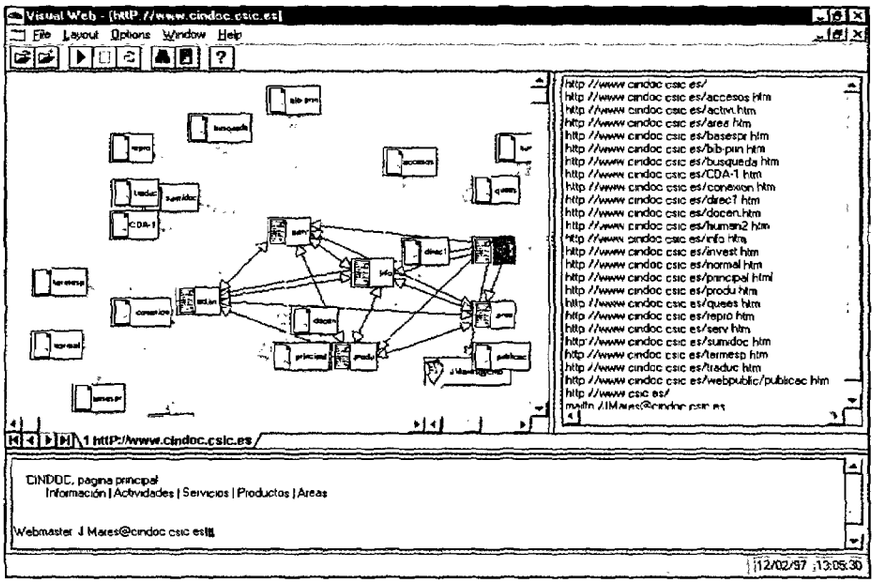
Se trata de herramientas diseñadas para ayudar a los *Webmasters* en el mantenimiento de sus sedes, pero su capacidad para describirlas tanto de forma gráfica como numérica puede tener indudables aplicaciones documentales.



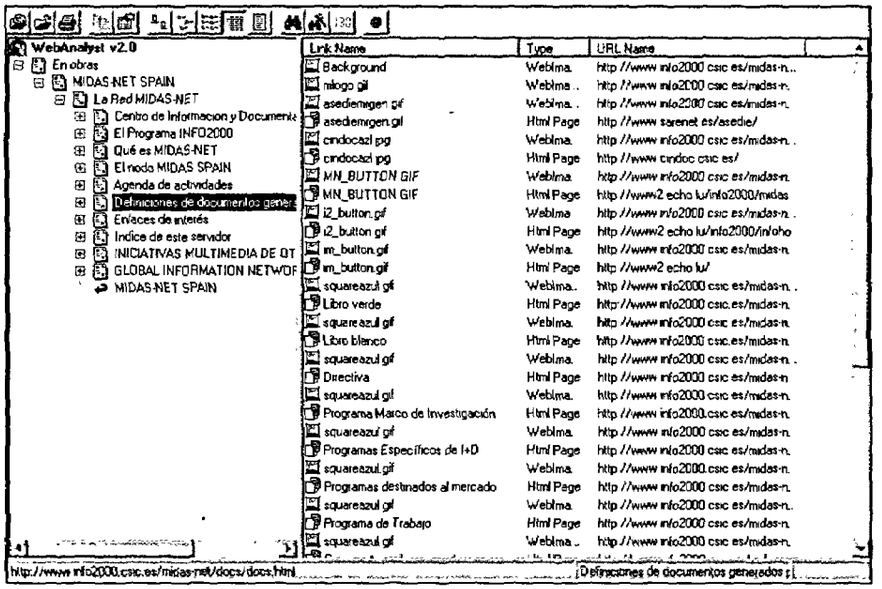
Microsoft <backoffice.microsoft.com> no podía faltar de esta revisión. El *Site Analyst* (v. 2.0) no es un producto realmente independiente, pues forma parte del servidor inter/intranet Backoffice. Aunque Backoffice está diseñado para Windows NT, el programa Site Analyst corre bajo W95, eso sí, con notables dificultades. Es un programa muy gráfico, llamativo en la presentación de los árboles de contenidos, aunque destaca por los informes detallados que suministra.



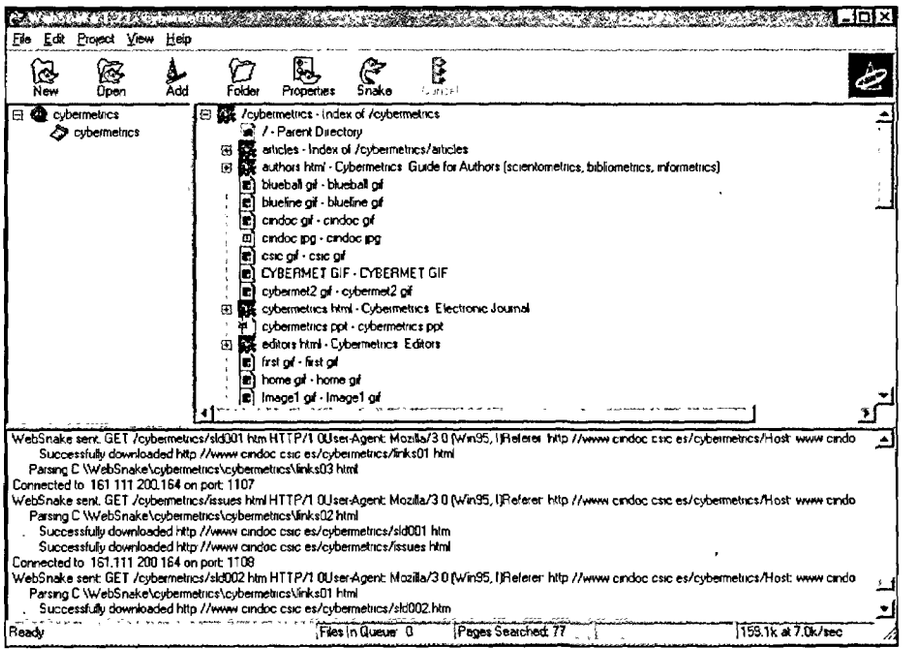
Aunque no alcanza las prestaciones del anterior, *WebAnalyzer* 2.0 <www.everyware.com>, es un programa rápido y eficaz y, desde luego, consume muchos menos recursos. Muy recomendable para descripciones generales e imprescindible si la sede a analizar es muy grande.



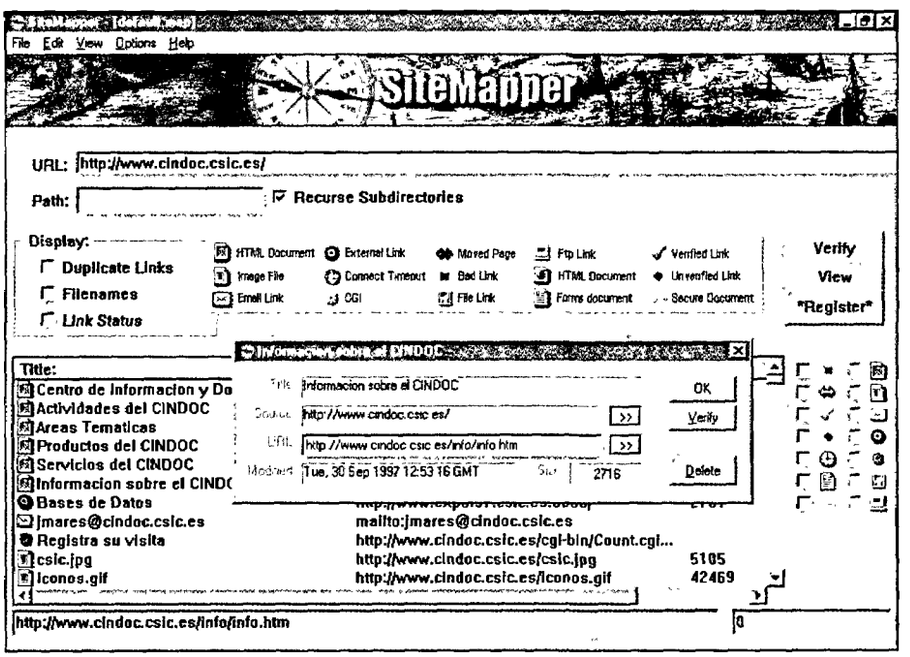
Visual Web 1.07 producido por Innovative Software (www.isg.de> resulta un poco confuso, no aportando opciones novedosas.



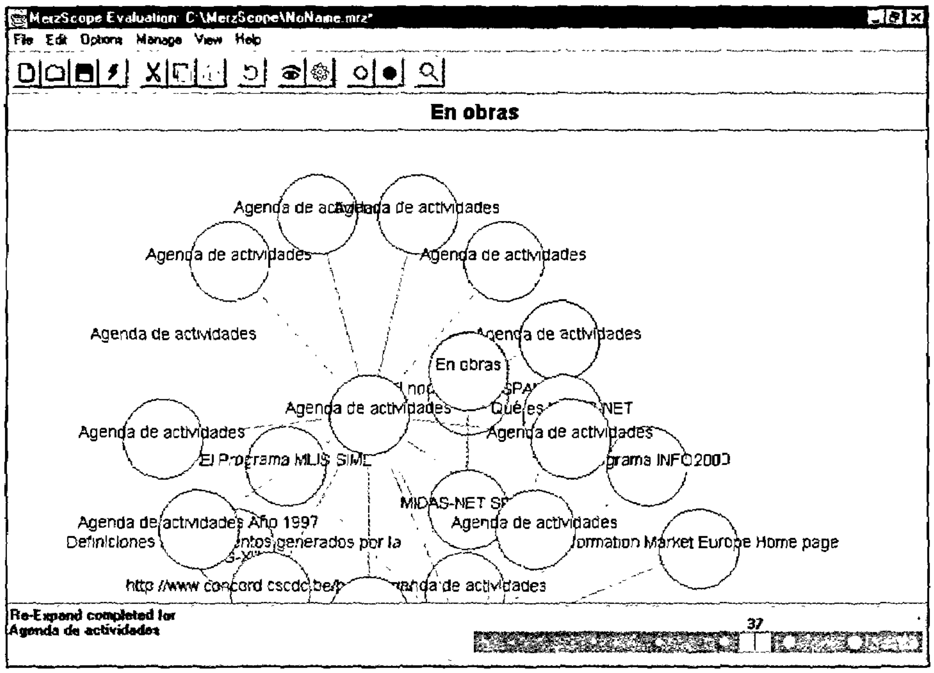
WebAnalyst 2.0 de AdageUS <www.adageus.com> ya solo está disponible como producto comercial, y aunque con notable apoyo gráfico y posibilidades de configurar temporalmente los análisis no creemos alcance las prestaciones de otros programas.



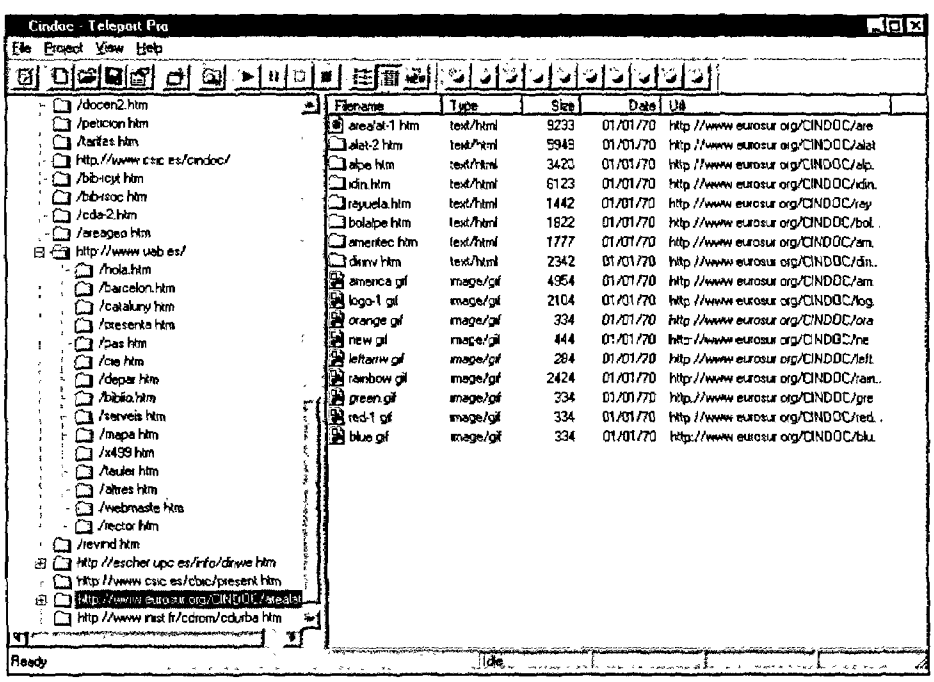
WebSnake 1.23 de Anawave <www.anawave.com/websnake> es el programa recomendable para descripciones simples, ya que ofrece la información básica con celeridad y buena presentación. Al igual que alguno de los otros programas se puede utilizar como un volcador fiable y sofisticado.



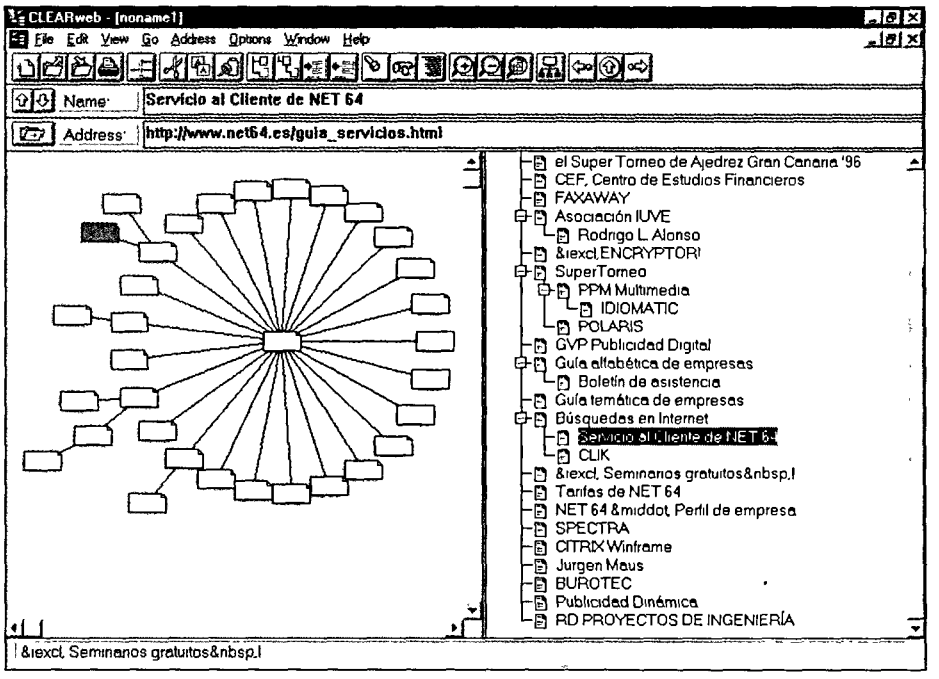
El recién llegado *Site Mapper* 1.00 <www.msw.com.au/ mapper> es un programa incipiente, cuya auténtica potencialidad es muy difícil de evaluar en este momento. El hecho que forme parte de la serie Wolf augura un interesante desarrollo que hay que vigilar.



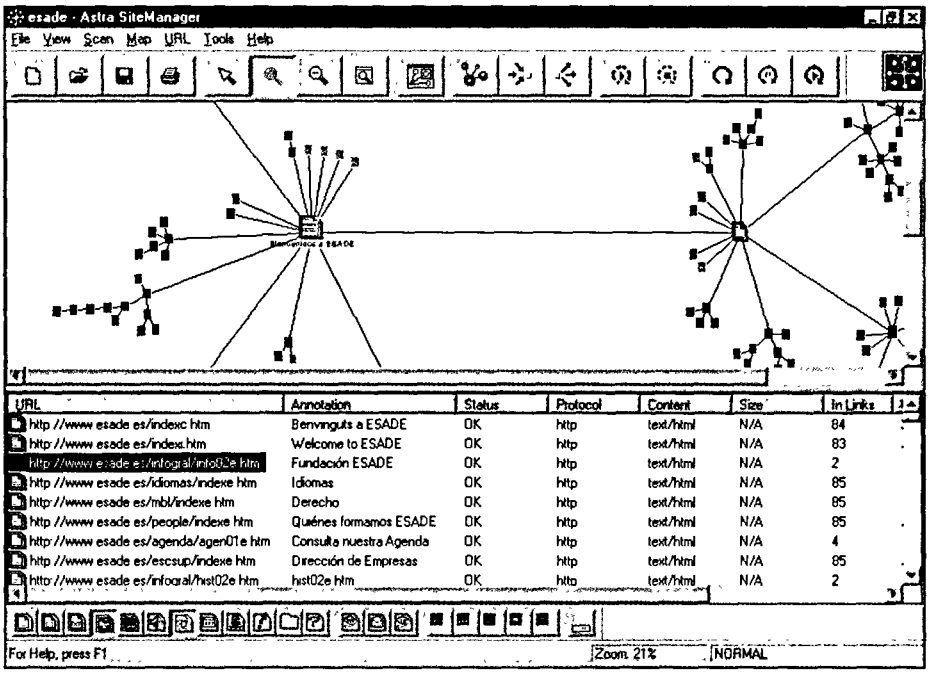
MerzScope <www.merzcom. com> no resulta especialmente útil, aunque sus gráficos son grandes y legibles y están bien anotados. Estos coloridos mapas pueden acompañar a los informes generados por otros productos.



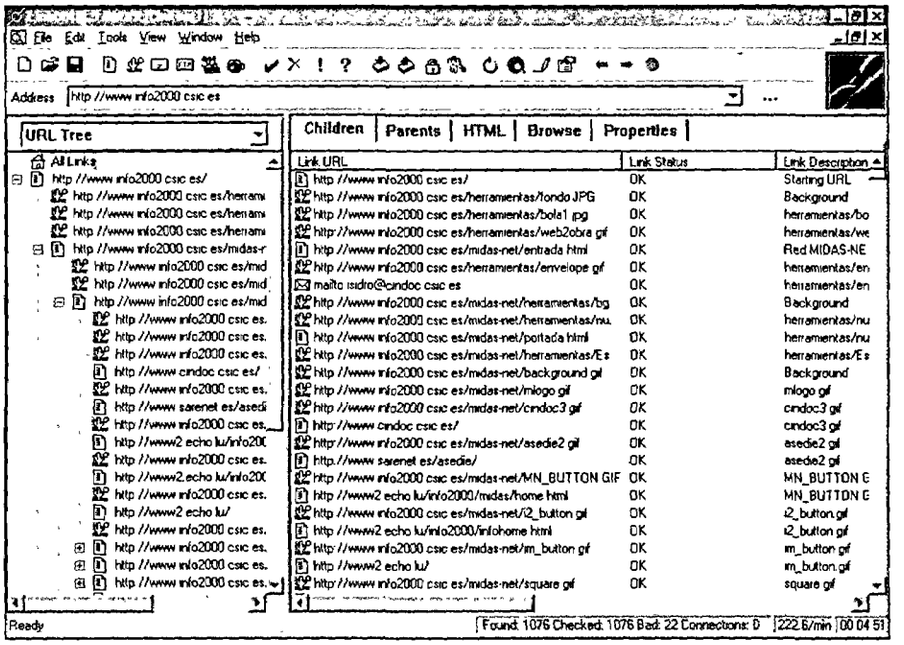
El más rápido de todos los mapeadores es *Teleport Pro* 1.28 de Tennyson Maxwell <www.tenmax.com>, aunque los informes que ofrece son modestos. No es de extrañar porque es básicamente un volcador sofisticado, lo que le hace recomendable si además de describir una sede se pretende archivarla. En este aspecto es muy superior, por celeridad, a Analyst.



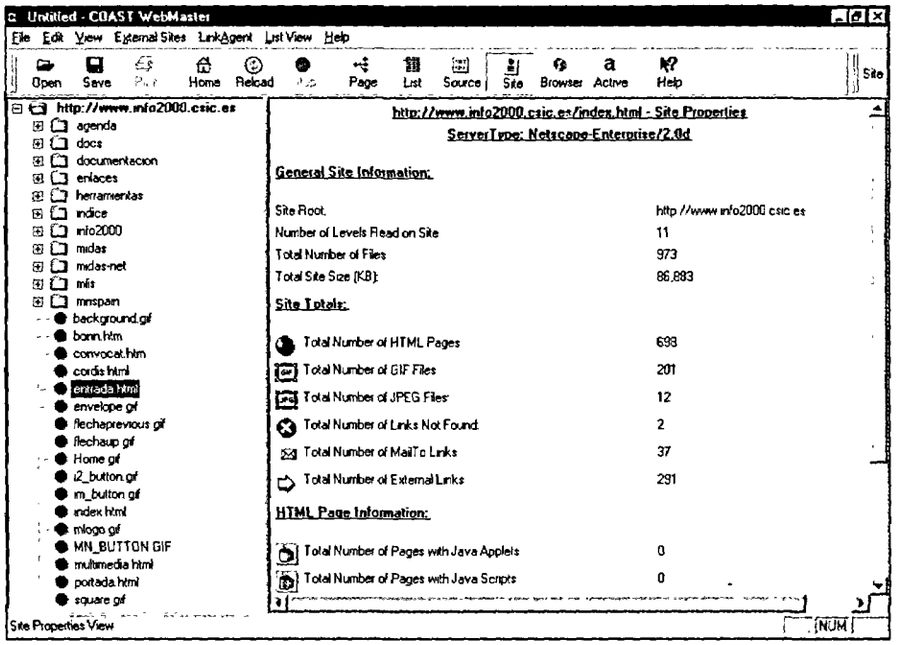
Los dos siguientes programas destacan por su visualización gráfica de las sedes web (mapas), aunque *CLEAR Web* 1.02 <www.clearweb. com> es claramente inferior en todos los aspectos a Astra Site Manager.



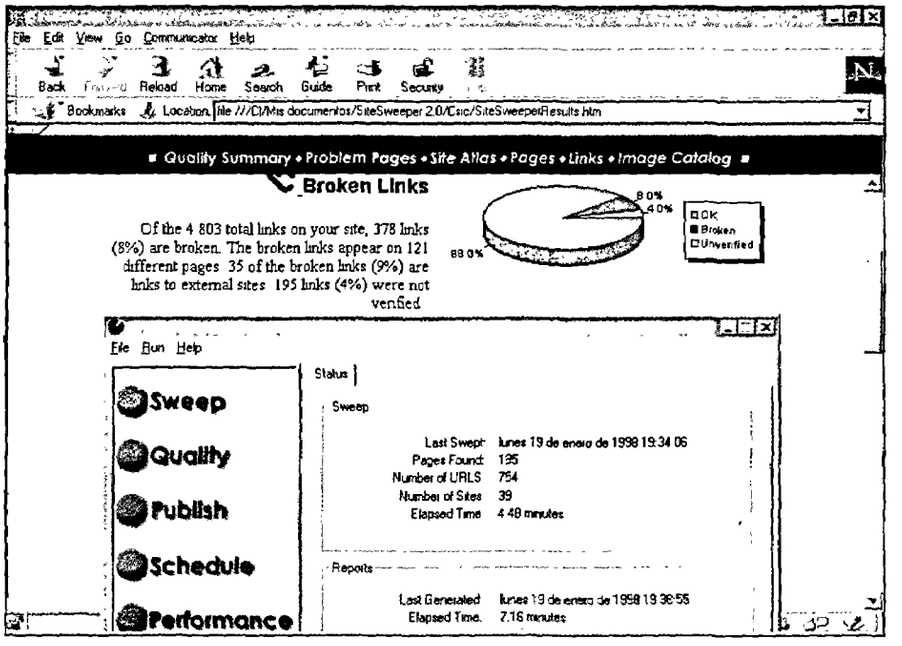
A igualdad de prestaciones con otros programas, *Astra Site Manager* 1.03 <www.merc-int.com> ofrece uno de los mejores mapas (si no el mejor), que se puede visualizar de diferentes formas y a distintos niveles de detalle.



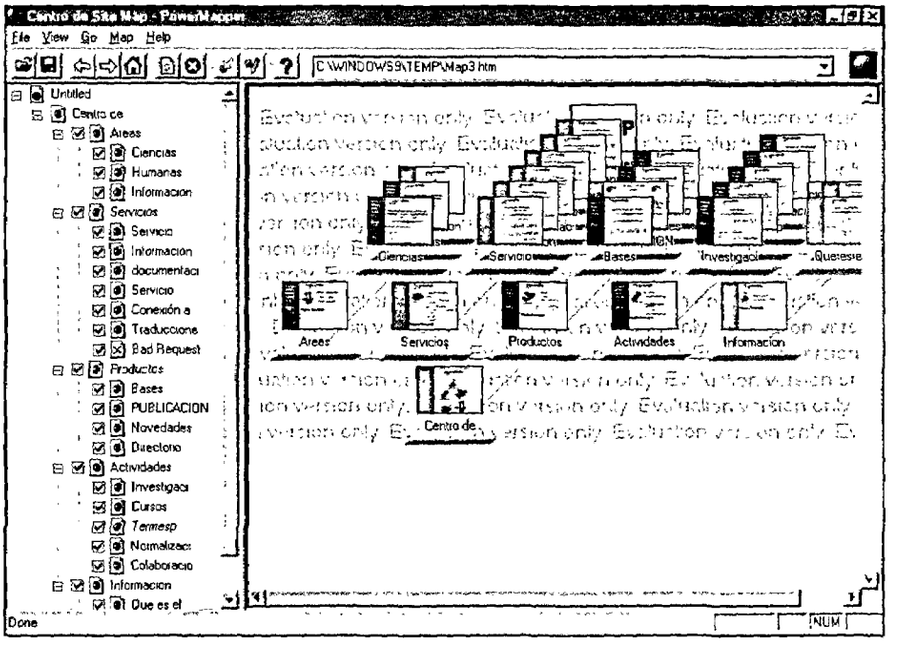
LinkBot Pro 3.5 de Tetranet Software <tetranetsoftware. com> ofrece buenas prestaciones, la mayoría de ellas presentes en programas anteriores.



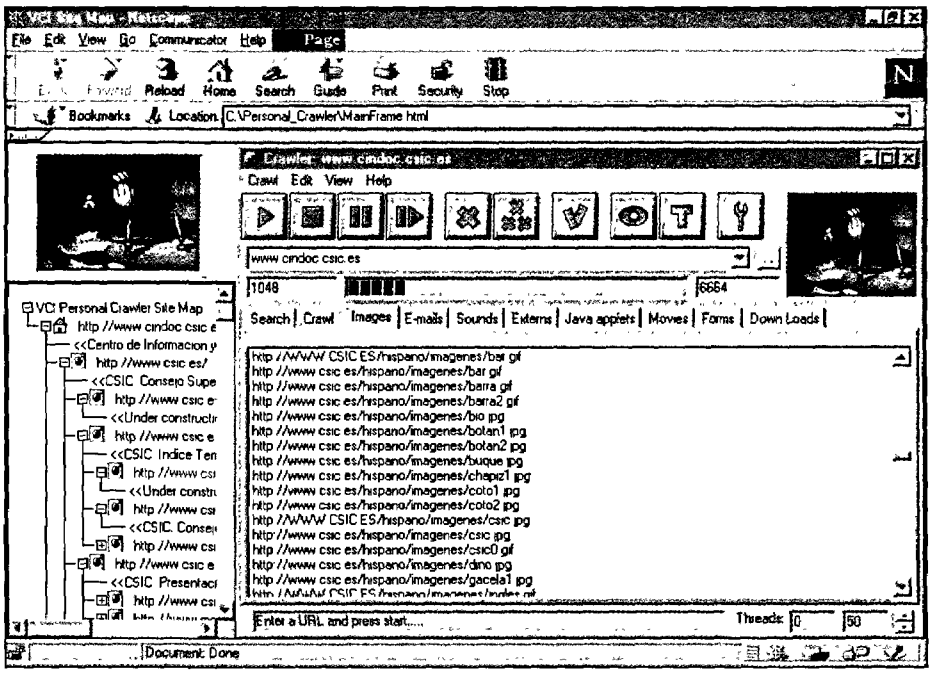
WebMaster 2.02 de Coast Software <www.coast.com> suministra un informe relativamente amplio, aunque lo suficientemente básico como para que el proceso no consuma demasiado tiempo.



SiteSweeper 2.0 de Site Technologies <www.sitetech.com> es un programa modesto que describe el estado de los enlaces de una sede Web. Aunque formalmente forma parte de los programas que comprueban enlaces (validadores), el hecho de que pueda ser utilizado para analizar direcciones externas nos ha llevado a su inclusión en este apartado.



Visualmente *Power Mapper* resulta uno de los programas más interesantes gracias a su presentación 3D, aunque sus potencialidades son inferiores a la de otros programas. No hemos podido evaluar la versión 2.0 que distribuye Electrum Multimedia <www. electrum.co.uk>.



Personal Crawler 1.01 <www.vci.co.il> no es un mapeador en sentido estricto, pero tiene un gran capacidad para analizar una dirección y extraer los diferentes objetos y enlaces de la misma.

Conclusiones

Las herramientas descritas tienen numerosas aplicaciones documentales y pueden ayudar definitivamente a la integración de la Internet en la labor diaria de cualquier profesional de la información. No sólo hay que contar con un previsible incremento de la productividad, sino que se abre la puerta a la prestación de nuevos servicios. Por citar algunos de las más relevantes:

a. Creación de bases de datos bibliográficas y de recursos Web de forma automática, recuperación la información de múltiples fuentes simultáneamente.

b. La indización automática y semi-automática de los recursos Internet como ayuda a la descripción documental en proyectos horizontales o sectoriales (verticales).

c. La constitución de archivos periódicos de determinados recursos, programando la frecuencia y el alcance de los volcados.

d. La edición de catálogos de recursos, con posibilidades de consulta *offline*.

e. La descripción cualitativa, gráfica y cuantitativa de las sedes lo que podría, mediante el análisis de los enlaces, constituir una excelente herramienta de aplicación en políticas de información e investigación.