Metodología para la obtención de datos con fines cibermétricos

Natalia Arroyo Vázquez y Víctor Manuel Pareja Pérez
Laboratorio de Internet
CINDOC – CSIC

natalia@cindoc.csic.es
vmpareja@cindoc.csic.es

Resumen

Se propone una metodología, basada en el concepto de sede web como unidad documental básica de análisis, para extraer datos del web con el objetivo de que puedan ser empleados en la obtención de resultados en estudios cibermétricos. Las etapas de la metodología propuesta son las siguientes: identificación; selección y recogida de sedes web; codificación descriptiva de las sedes; cuantificación automatizada, recogida de datos y posterior volcado a la base de datos; así como el mantenimiento de la base de datos.

1. Introducción.

En el Laboratorio de Internet del CINDOC se viene trabajando, desde hace algo más de cinco años, en la obtención de indicadores cibermétricos de ciencia y tecnología en al ámbito europeo. Este trabajo atraviesa en este momento por una fase experimental de definición de conceptos y métodos debido a lo novedoso de esta subdisciplina en la que se sustenta: la Cibermetría, que trata de aplicar los métodos bibliométricos al web.

Por ello, y dada esta situación de fase experimental, el objetivo de la presente comunicación es la descripción de una metodología, así como su problemática, que se ha venido fraguando con el trabajo diario durante algo más de cinco años y que, de hecho, sigue sujeta a los cambios y modificaciones impuestos tanto por el carácter dinámico del web como por los objetivos y prioridades identificados en cada momento.

Tomando como unidad básica de análisis el concepto de sede web, nuestro trabajo consiste en extraer unos datos cuantitativos a partir de los cuales se obtendrá una serie de indicadores de ciencia y tecnología.

El método empleado para ello constaría de varias etapas:

1. Identificación, selección y recogida de sedes web.

2. Codificación descriptiva de las sedes.
3. Cuantificación automatizada de dichas sedes.
4. Recogida de datos y volcado a la base de datos
5. Mantenimiento de la base de datos.

Estas etapas no están exentas de limitaciones y problemas metodológicos propios de la idiosincrasia de Internet y de las herramientas empleadas, que pasamos a abordar a continuación.

2. Identificación, selección y recogida de sedes web.

Antes de abordar esta primera etapa conviene aclarar un concepto, el de sede web, que fue ideado por Isidro Aguillo y definido por primera vez en 1998. Posteriormente autores (Lluis Codina, Necip Facil Ayan et al., etc.) han venido trabajando sobre él. El primero de ellos emplea como sinónimos los términos "lugar web", "página web" o "*homepage*", creando así una asociación de ideas que lleva a confusión.

El segundo propone la expresión *logical domain*, entendida como un grupo de páginas que tienen una relación semántica determinada y una estructura que las relaciona entre ellas, por oposición al dominio físico, que se identifica únicamente por el nombre del dominio. Este concepto se asemejaría en gran medida a nuestra idea de sede web, con la particularidad de que esta última —la sede web— respondería a una visión más amplia, como veremos a continuación.

En definitiva, el concepto de sede web podría ser definido 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.

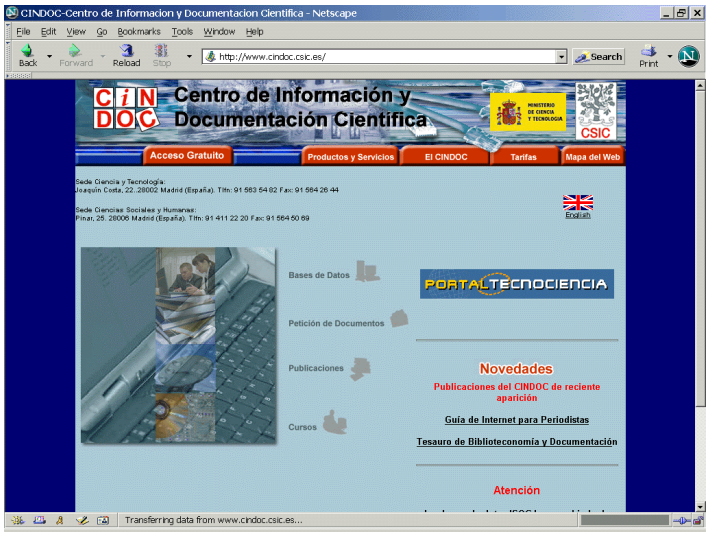


Figura 1. Ejemplos de sedes web institucionales: www.csic.es y www.cindoc.csic.es.

Algunos ejemplos de sedes web serían las del Consejo Superior de Investigaciones Científicas (CSIC) o la del Centro de Información y Documentación Científica (CINDOC), ambas de carácter institucional, y representadas por las URLs www.csic.es y www.cindoc.csic.es respectivamente, o la de la revista electrónica *Cybermetrics*, de carácter temático.

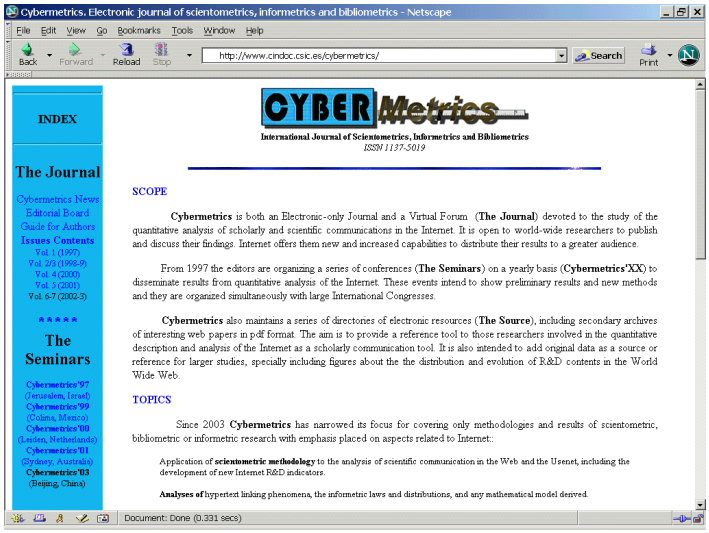


Figura 2. Ejemplo de sede web temática: www.cindoc.csic.es/cybermetrics.

En los dos primeros casos, la sede es equiparable a un dominio específico, no así el tercero, por lo que se infiere que no es condición *sine qua non* para que exista una sede web el tener dominio propio. Una sede web lo es por cuanto forma una unidad documental, si bien se aloja en un directorio o subdirectorio de un dominio que corresponda a otra sede.

El proceso de identificación de sedes web no puede ser realizado automáticamente debido a la ausencia de patrones fijos en las URLs que las representan, por lo que el método que se ha venido siguiendo no es otro que la búsqueda y navegación en el Web.

El siguiente paso consistiría en seleccionar las sedes previamente identificadas atendiendo a unos criterios fijados de antemano. En la práctica este proceso se realizaría de forma simultánea al anterior, si bien sería necesario discernir entre todo lo encontrado y aquello que nos interesa, atendiendo a unos criterios previamente establecidos.

Y por último, estas sedes se recogerían en una base de datos junto con alguna información básica, como podría ser el nombre de la sede y la institución de la que depende.

3. Codificación descriptiva.

La codificación descriptiva consiste en una breve descripción de las sedes web de tres tipos:

- Codificación institucional.
- Codificación geográfica.
- Codificación por materias.

Codificación institucional

La primera de ellas, la codificación institucional, tiene por objeto identificar de la forma más rápida y sencilla posible el tipo de institución a la que representa cada sede web recogida en la base de datos con el fin de posibilitar la extracción de resultados fiables. Para ello ha sido creada específicamente una clasificación adaptada a nuestras necesidades y que aún continúa abierta a posibles modificaciones.

La mayor dificultad consiste en entender cabalmente y homogeneizar, a efectos de clasificación, los sistemas universitarios y de investigación propios de cada país, que difieren en algunos aspectos unos de otros. En este sentido, se encuentran instituciones que, como los “colleges” británicos o los “Lehrstuhl” alemanes, son característicos de sus respectivos ámbitos geográficos. La solución adoptada ha sido integrarlos junto con otras instituciones del mismo nivel. Este sistematización se complica para el caso de insitutiones de investigación no universitarias.

Codificación geográfica

La herramienta escogida para realizar la descripción geográfica son los códigos NUTS (Nomenclature of Territorial Units for Statistics) de EUROSTAT (Oficina Europea de Estadística), ya que se adaptan perfectamente al área territorial objeto del trabajo: la Europa de los quince. Esta clasificación alfanumérica divide a cada país en

regiones que reflejan sus divisiones administrativas y territoriales, de acuerdo a tres niveles. Por ejemplo, en España, el primer nivel, o nivel 0, es el país (representado por dos letras ES), el segundo lo constituyen las Comunidades Autónomas y el tercero cada una de las provincias, codificados ambos por tres dígitos, cada uno de los cuales se corresponde con un nivel diferente.

La codificación ha sido realizada en todos los casos al nivel más específico posible, según la ubicación geográfica de la institución autora de la sede web o representada en esta.

Figura 3. Códigos NUTS. España.

Codificación por materias

El último tipo de descripción que se realiza es la descripción por materias, que tiene por objetivo conocer en qué medida quedan representadas las diferentes disciplinas y áreas del conocimiento. Para ello se emplea como herramienta la clasificación UNESCO de códigos para las áreas de ciencia y tecnología.

Esta clasificación se basa en una codificación numérica de seis dígitos. Ha sido construida a tres niveles: mediante el primero de ellos, representado por los dos primeros dígitos, podemos identificar los campos generales en que se divide la ciencia y tecnología. El segundo representa a las disciplinas científicas, que suponen una descripción general de grupos de especialidades en ciencia y tecnología, mientras que el último se refiere a subdisciplinas –las entradas más específicas de la nomenclatura–, que identifican las actividades que se realizan dentro de una disciplina.

Al tratarse de una clasificación creada exclusivamente para campos de ciencia y tecnología todo aquello que quede fuera de estas áreas es indescriptible. Ejemplos de ello son instituciones como las propias universidades o las bibliotecas, por citar los casos más claros. La solución adoptada ha sido la creación de estándares, empleando para ello los códigos más próximos; de esta forma, las universidades quedan recogidas en "Pedagogía" y "Política educativa" y las bibliotecas en "Documentación" ―en el campo de Lingüística―. Este factor deberá ser tenido muy en cuenta durante la fase de recuperación de información, ya que, de no ser así, podrían producirse sesgos importantes.

Por último, señalar también el desequilibrio entre los campos de ciencia y tecnología, muy desarrollados, y los de humanidades y sociología, escasamente representados, derivado de su especialización en los primeros.

4. Cuantificación automatizada de sedes web.

La cuantificación automatizada de sedes web consiste en extraer –como su propio nombre indica– datos cuantitativos para cada sede mediante programas denominados *mapeadores*, que no hacen otra cosa que navegar por la sede. El funcionamiento del programa se basa en suministrar una URL de partida y va entrando a través de los en los enlaces hipertextuales en los diferentes directorios y subdirectorios que encuentra a su paso, lo que posibilita contabilizar todos los recursos que contiene: páginas web, enlaces, ficheros de audio, vídeo, de texto, etc.

El programa que se emplea para realizar esta labor es *Microsoft Site Analyst*, una versión *shareware* subsimida por Microsoft en el paquete *Back Office* a partir de un antiguo programa llamado *Webmapper*. Posterioremente, aquella empresa lo actualizó a la versión 2.0, lo introdujo en el pack *Microsoft Site Server* y lo denominó *Content Analyzer*. Esta versión tenía como novedad la traducción del programa al castellano.

Tras cierto tiempo de prueba para someter a ambas versiones a un examen comparativo, se llegó a la conclusión de que la versión 1.0 resultaba más eficiente y operativa. Dicho examen ha llevado a determinar el uso generalizado de *Site Analyst*, si bien *Content* se utiliza para algunos casos puntuales donde *Site* queda limitado en sus resultados por diversos motivos.

La duración del *mapeo* de una sede web varía en función de la cuantía y extensión de los elementos que contenga dicha sede, de la potencia del equipo que esté mapeando y de la capacidad o nivel de saturación de las comunicaciones de que se

disponga en un momento concreto. Así, se pueden obtener los resultados tanto a los cinco minutos de iniciarse el proceso como pueden transcurrir varias semanas.

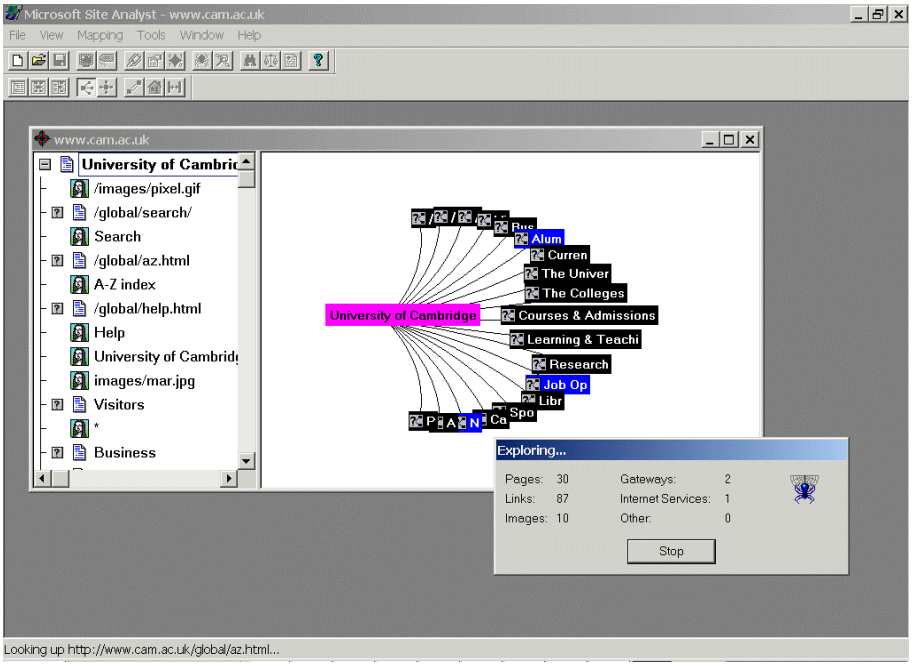


Figura 4. Funcionamiento de Microsoft Site Analyst.

El resultado de este proceso son una serie de informes, en formato *html*, en mayor o menor número dependiendo del tamaño de la sede, que contienen los datos cuantitativos objeto de la introspección de la sede. Es en el informe llamado *Summary* donde se concentran los datos generales que darán cuenta de las características de la sede web.

Microsoft Site Analyst Logo

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

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	3246	9270169
Images	737	10666631
Gateways	28	N/A
Internet	589	N/A
Java	0	0
Applications	165	5397524
Audio	0	0
Video	0	0
Text	6	3458
WebMaps	0	0
Other Media	52	1508965
Totals	**4823**	**26846747**

Status Summary

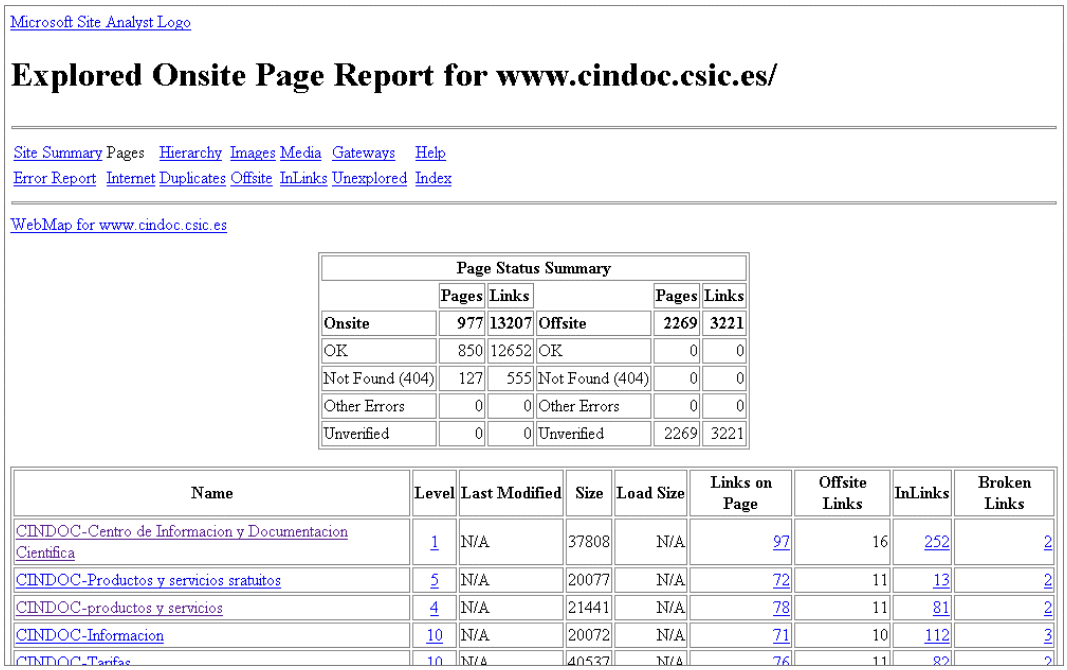
	Objects	Links
Onsite	**1792**	**24137**
OK	1597	23272
Not Found (404)	188	733
Other Errors	0	0
Unverified	7	132
Offsite	**3031**	**4314**
OK	0	0
Not Found (404)	0	0
Other Errors	0	0
Unverified	3031	4314
Totals	**4823**	**28451**

Map Statistics

Map Date	Mar 04 13:32 2003
Levels	15
Avg Links/Page	29

Server Summary

Domain:	www.cindoc.csic.es
Server Version:	Apache/1.3.23 (Unix) (Red-Hat/Linux) mod_fastcgi/2.2.12 Midgard/1.4.4/SG mod_python/2.7.6 Python/1.5.2 mod_ssl/2.8.7 OpenSSL/0.
HTTP Version:	1.1



Microsoft Site Analyst Logo

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

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	**977**	**13207**	**Offsite**	**2269**	**3221**
OK	850	12652	OK	0	0
Not Found (404)	127	555	Not Found (404)	0	0
Other Errors	0	0	Other Errors	0	0
Unverified	0	0	Unverified	2269	3221

Name	Level	Last Modified	Size	Load Size	Links on Page	Offsite Links	InLinks	Broken Links
CINDOC-Centro de Informacion y Documentacion Cientifica	1	N/A	37808	N/A	97	16	252	2
CINDOC-Productos y servicios gratuitos	5	N/A	20077	N/A	72	11	13	2
CINDOC-productos y servicios	4	N/A	21441	N/A	78	11	81	2
CINDOC-Informacion	10	N/A	20072	N/A	71	10	112	3
CINDOC-Tarifas	10	N/A	40537	N/A	76	11	82	2

Figura 5. Resultados de Site Analyst para la sede www.cindoc.csic.es.

En el informe resumen de la página de *Summary* se ofrece información concerniente a los resultados totales de la cuantificación, así como una breve descripción del servidor remoto. Dichos resultados aparecen en dos columnas con títulos diferentes: *Objects statistics* y *Status summary*. En ambas, los totales de los datos que se aportan son idénticos pero cambia la perspectiva del análisis realizado para cada columna. Así, en la primera de ellas, se resalta la tipología de recursos encontrados, así como su tamaño en bytes: páginas html, imágenes, pasarelas, servicios de internet, ficheros de audio o de vídeo, etc.; mientras que en la segunda columna la visión que se suministra es la de estos mismos recursos tomados como elementos internos o externos a la sede.

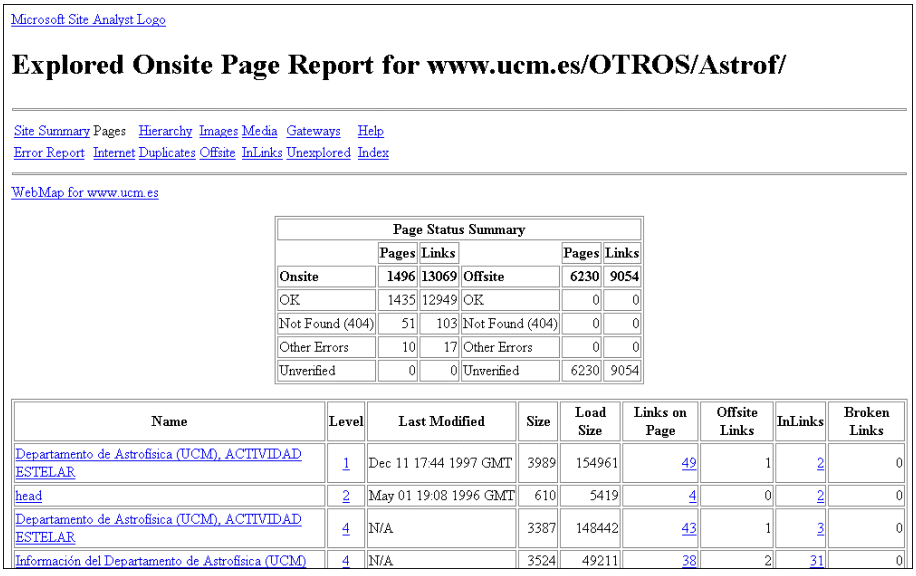
Por otra parte, es preciso mencionar que la utilización del programa *Site Analyst* (y en mayor medida su actualización *Content Analyzer*) plantea una serie de limitaciones e inconvenientes debidos a una casuística diversa, que se puede sintetizar en los siguientes aspectos: .

- Programación avanzada de las sedes web: Java, JavaScript, Flash, php, etc. Al no emplear estos tipos de programación una estructura de hipervínculos basados en *html*, los *mapeadores* no podrán cuantificar las sedes basadas en ellos, puesto que, como ya se comentó, lo que estos programas cuentan son, básicamente, enlaces. El problema que se origina es que estos tipos de programación tienden a ser cada vez más utilizados en la Red.

- Frames o marcos. Las sedes web construidas con este diseño sólo están formadas por una página dentro de la cual se pueden visualizar otras, por lo que a veces el programa sólo contabilizará la página que centraliza a las demás.

- Directorios compartidos por varias sedes. En algunas ocasiones se pueden encontrar diferentes sedes web alojadas dentro de un mismo directorio, por lo que técnicamente es imposible hacer una cuantificación diferenciada e independiente de cada sede. De esta manera, no podrán obtenerse datos fiables de cada una de ellas porque no se contabilizaría sólo esa sede, sino todo lo contenido en ese directorio en conjunto.

Precisamente, en la figura 5 se muestra el caso de dos sedes ubicadas en el mismo directorio y cuyos datos al cuantificarlas son prácticamente idénticos porque se ha mapeado el contenido completo de ficheros del directorio. El ejemplo procede de un servidor de la Universidad Complutense de Madrid, dentro de un subdirectorio dedicado a Astrofísica; en dicho subdirectorio se hallan tanto la sede del Departamento de Astrofísica, como las de los grupos adscritos al mismo, el Grupo de Actividad Estelar del Departamento (representada por la página *actividad.html*), y el Grupo de Investigación Extragaláctica (que viene identificado por la página *galaxias.html*).



Microsoft Site Analyst Logo

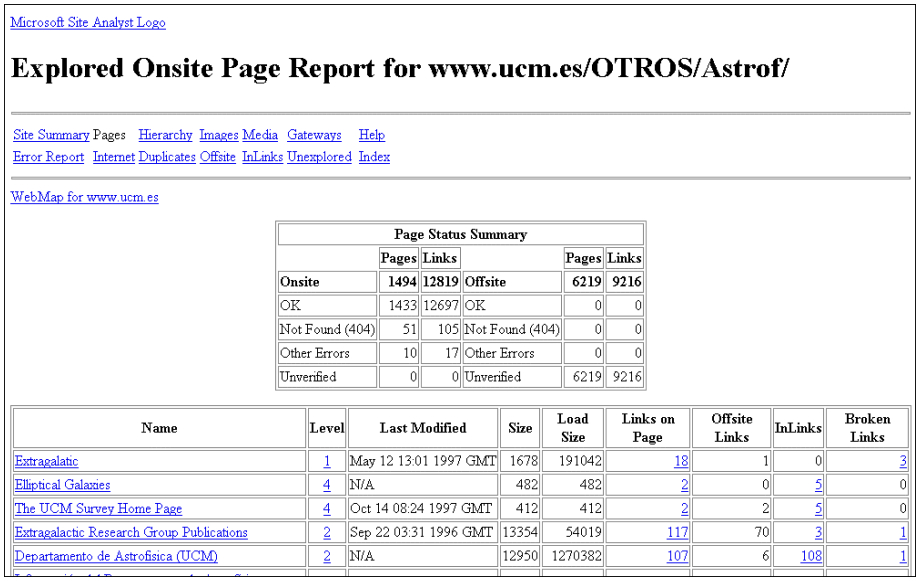
Explored Onsite Page Report for www.ucm.es/OTROS/Astrof/

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

WebMap for www.ucm.es

Page Status Summary					
	Pages	Links		Pages	Links
Onsite	**1496**	**13069**	**Offsite**	**6230**	**9054**
OK	1435	12949	OK	0	0
Not Found (404)	51	103	Not Found (404)	0	0
Other Errors	10	17	Other Errors	0	0
Unverified	0	0	Unverified	6230	9054

Name	Level	Last Modified	Size	Load Size	Links on Page	Offsite Links	InLinks	Broken Links
Departamento de Astrofisica (UCM), ACTIVIDAD ESTELAR	1	Dec 11 17:44 1997 GMT	3989	154961	49	1	2	0
head	2	May 01 19:08 1996 GMT	610	5419	4	0	2	0
Departamento de Astrofisica (UCM), ACTIVIDAD ESTELAR	4	N/A	3387	148442	43	1	3	0
Información del Departamento de Astrofisica (UCM)	4	N/A	3524	49211	38	2	31	0



Microsoft Site Analyst Logo

Explored Onsite Page Report for www.ucm.es/OTROS/Astrof/

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

WebMap for www.ucm.es

Page Status Summary					
	Pages	Links		Pages	Links
Onsite	**1494**	**12819**	**Offsite**	**6219**	**9216**
OK	1433	12697	OK	0	0
Not Found (404)	51	105	Not Found (404)	0	0
Other Errors	10	17	Other Errors	0	0
Unverified	0	0	Unverified	6219	9216

Name	Level	Last Modified	Size	Load Size	Links on Page	Offsite Links	InLinks	Broken Links
Extragalatic	1	May 12 13:01 1997 GMT	1678	191042	18	1	0	3
Elliptical Galaxies	4	N/A	482	482	2	0	5	0
The UCM Survey Home Page	4	Oct 14 08:24 1997 GMT	412	412	2	2	5	0
Extragalactic Research Group Publications	2	Sep 22 03:31 1996 GMT	13354	54019	117	70	3	1
Departamento de Astrofisica (UCM)	2	N/A	12950	1270382	107	6	108	1

Figura 6. Sedes que comparten un mismo directorio. www.ucm.es/info/Astrof/actividad.html y www.ucm.es/info/Astrof/galaxias.html.

Iniciando *Site Analyst* por cualquiera de estas entradas de sedes distintas, computa siempre los mismos o muy similares datos, por lo que no se da una discriminación particularizada de la cuantía de cada sede.

- Gran consumo de recursos mientras trabaja el programa, especialmente de memoria RAM, por lo que es necesario poder disponer de equipos informáticos potentes que soporten estas tareas de mapeado, especialmente para las sedes más grandes.

- Errores de programa y/o del servidor mapeado. En algunas ocasiones *Site Analyst* no soporta la tarea de cuantificar la sede y algún tiempo después de empezar a mapear, bien se cierra solo o bien se queda bloqueado. Estos errores se pueden solucionar en ciertas ocasiones empleando *Content Analyzer*, pero en otros este programa vuelve a repetir el error. Entonces debería intentarse localizar dónde está el fallo que lo provoca e intentar cuantificar de nuevo la sede recurriendo a una de las opciones del programa, que permite obviar las páginas problemáticas.

Un buen ejemplo de esto es la sede de la Universidad de La Rioja, que en un principio se intentó cuantificar con *Site Analyst*; al quedar bloqueado el programa, se hizo un segundo intento, esta vez con *Content Analyzer*. Entonces se observó que el programa siempre quedaba bloqueado en las mismas páginas, y se logró finalmente por medio de restringir las páginas problemáticas.

Sin embargo, en muchas otros casos se hace muy difícil detectar dónde se ha producido el error, especialmente cuando el programa se cierra por sí solo, sin generar ningún mensaje de error, por lo que hay que conformarse con esta pequeña limitación. En todo caso, como ya se ha comentado, este tipo de problemas supone un porcentaje casi insignificante.

- Las **pasarelas** son, tal y como las define Aguillo, puntos de entrada a algún otro servicio de información que requiere una inclusión de datos por iniciativa del solicitante, un término de búsqueda, password, o elegir una opción. Esta actividad genera una respuesta por parte de la base de datos que está detrás de la pasarela, proporcionando la respuesta del sistema con la correspondiente información. *Site Analyst* no soporta las pasarelas, pero sí *Content Analyzer* –en la mayor parte de las ocurrencias–, por lo que en este caso se emplea esta otra versión del programa.

- Los **bucles** se dan cuando, al cuantificar el programa una sede determinada, comienza a contar las mismas páginas una y otra vez, entrando en una secuencia circular infinita por causas desconocidas. La solución es la misma que en el caso de las pasarelas: intentar la cuantificación con *Content Analyzer*; de esta forma en la mayoría de ocasiones se obtienen resultados satisfactorios.

- Otro problema que afecta a la fiabilidad de resultados finales viene determinado por la existencia de **dos o más dominios o URL para identificar la misma sede**. Esto solo es particularmente grave de cara a la cuantificación cuando la

estructura de los hiperenlaces es absoluta[1], de manera que las páginas de la sede estén construidas alternando y combinando esos varios dominios o URL representativas de la sede. El fenómeno que provoca es que, partiendo de la URL seleccionada como punto de inicio para el mapeo, el programa no reconoce, en buena lógica, como propios o internos los enlaces que estén construidos con la sintaxis de la URL alternativa.

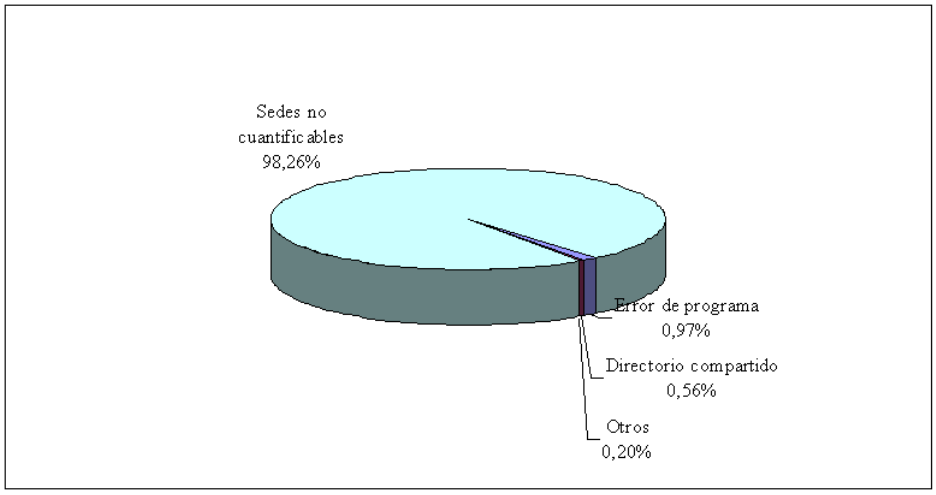


Figura 7. Sedes no cuantificables.

Todas estas limitaciones no llegan a afectar significativamente a los estudios de este tipo, puesto que atañen a una mínima parte de las sedes analizadas. En la figura 7 se puede observar cómo, de las más de 25.000 sedes recogidas en nuestra base de datos, sólo un pequeño porcentaje (el 1,73%) no ha podido ser cuantificado. Este porcentaje es la suma del 0.97% procedente de errores del programa, el 0,56% de directorios compartidos y el 0,20% de otros tipos de problemas.

5. Recogida de datos y volcado a la base de datos.

Una vez obtenidos los informes resultantes del proceso de cuantificación de sedes web es necesario seleccionar qué datos son los más adecuados para el tipo de estudio que se quiere realizar y volcarlos a la base de datos. En nuestro caso concreto se han seleccionado datos sobre páginas, enlaces y objetos internos y externos.

El proceso de extracción o recogida de los resultados se realiza de forma automatizada, mediante un programa creado especialmente para este fin que agiliza enormemente el trabajo. Ya recogidos en un formato de texto, los datos son finalmente exportados a la base de datos.

6. Mantenimiento y actualización de la base de datos.

[1] Enlace absoluto es aquel que está diseñado haciendo constar toda la url al completo desde http://www...., por oposición a enlace relativo en el que se escribe solamente el fichero al que apunta el enlace puesto que la conexión con el servidor es la opción por defecto.

La fase de mantenimiento y actualización de la base de datos es sin duda alguna la más costosa en tiempo y recursos debido a una de las características propias del web: su dinamismo. Para hacerse una idea de ello, baste con mostrar las cifras de la última revisión llevada a cabo en la base de datos durante el año 2002: la actualización de más de 25.600 registros, contando con unos recursos humanos de una media de tres personas y con unos recursos materiales de más de 20 PCs, ha supuesto más de siete meses de trabajo.

Sobre el obsolescencia del Web como sistema de información, el profesor de la Universidad de Oklahoma, W. Koehler, ha realizado varios estudios. En ellos ha venido investigando sobre los diferentes aspectos a tener en cuenta en este sentido. El que aquí se destaca es el concepto de **permanencia**, que este investigador norteamericano define como la medida de la probabilidad de que los documentos web lleven la misma URL a lo largo del tiempo (pero no necesariamente el mismo IP) o de que, si cambia de URL, se redireccione a la nueva URL.

En las sucesivas actualizaciones de nuestra base de datos no sólo hemos encontrado sedes que hayan cambiado de URL, sino también otras a las que podríamos denominar, empleando una vez más la terminología de Koehler, "intermitentes", es decir, que en algún momento de su vida, si es que pudiéramos hablar de tal, no han podido ser encontradas, no se ha podido contactar con ellas, o su acceso ha estado prohibido. En ocasiones esto supone una desaparición total de la sede, pero en otras es simplemente un estado temporal.

El último aspecto a destacar es el **crecimiento** del Web, del que todos conocemos de su magnitud, y que, cuando se trabaja con grandes volúmenes de información, supone una importante limitación a tener en cuenta.

En el gráfico de barras de la figura 8 podemos observar cómo en unos pocos meses casi un 20% de las sedes de nuestra base de datos, como media, han sufrido algún tipo de cambio, ya haya sido en la URL, ya en la posibilidad de conectar con ellas. Esto implica la necesidad de trabajar a un rápido ritmo de actualización, sólo posible cuando el volumen de la muestra es considerablemente menor.

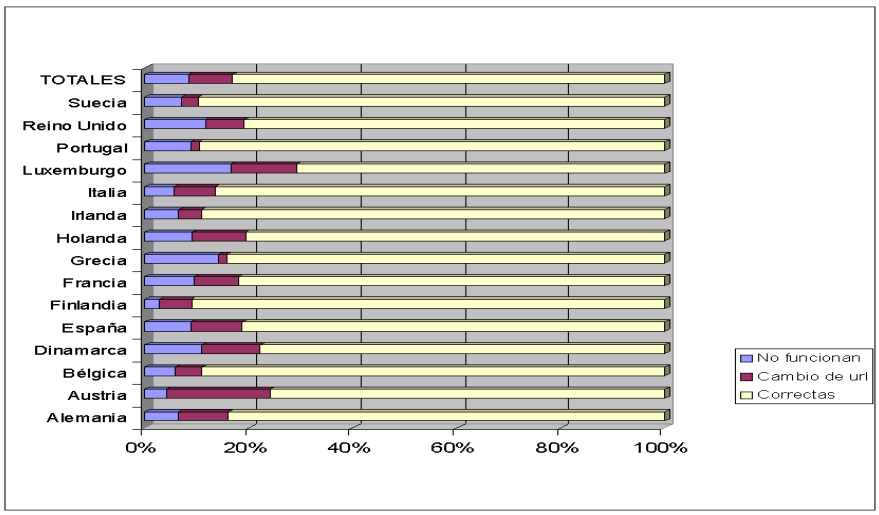


Figura 8. Sedes web que sufrieron cambios de abril a julio 2002.

7. Conclusiones.

Después de varios años de trabajo aplicando esta metodología, se concluye que pese a sus limitaciones ya expuestas, su utilidad para la obtención de datos con fines cibermétricos queda suficientemente contrastada. Esta utilidad viene determinada preferentemente por la capacidad como técnica cuantitativa para dotar de mayor objetividad y acercamiento a la realidad que las cuantificaciones de carácter estimatorio, con lo que sin ser fiables al 100% permite establecer patrones de comportamiento en la información presente en Internet. Su éxito –el de esta metodología– aumenta cuando la cantidad de las sedes de la muestra y el tamaño de estas es menor.

Por otra parte, esta metodología sirve para sistematizar la conceptualización documental de la información en la Red por cuanto se parte de la noción de "sede web", aunque es preciso perfilarla y establecer una estructuración y tipología que será objeto de otro trabajo.

En definitiva, y después de bastante tiempo de contraste para la tarea de mapeado, *Site Analyst* se convierte en la mejor herramienta para la cuantificación de las sedes, incluso comparándolo con su versión actualizada de *Content Analyzer*.

Aún así, y dadas las continuas transformaciones que a programación web se refieren, trastocando y limitando así los resultados y la capacidad de este software, se hace preciso encontrar técnicas o software que soporten la cuantificación independientemente del lenguaje en que las sedes estén construidas. De cómo sea la evolución de la construcción web dependerá el éxito de la metodología que se expone en este artículo.

8. Bibliografía

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

2. Aguillo, Isidro (2002). Web Characterization for Cybermetric Purposes: Terminology and Definitions.

3. Codina, Lluís (2000). Evaluación de recursos digitales en línea: conceptos, indicadores y métodos. Revista Española de Documentación Científica. 23(1):9-44.

4. Koehler, Wallace (1999). An analysis of Web page and Web site constancy and permanence. Journal of the American Society of Information Science. 50 (2):162-180.

5. Koehler, Wallace. Web Document Management. Disponible en: http://www.ou.edu/cas/slis/courses/LIS5990A/slis5990/.

6. Necip Facil Ayan; Wen-Syan Li, and Okan Kolak (2002). Automating extraction of logical domains in a web site. Data & Knowledge Engineering. 43:179-205.

7. Pareja, Víctor Manuel; González, Ana; Aguillo, Isidro (1999). Ciencia y tecnología españolas en Internet: valoración a través de la presencia de organismos públicos españoles y de sus revistas electrónicas. Arbor. 639: 367-390.