WEBWATCHING O LA MEDIDA CUANTITATIVA DE LA AUTOORGANIZACION DE LOS SISTEMAS DE CONOCIMIENTO HIPERTEXTUAL

Autores: Noelia Rodríguez
Isidro F. Aguillo
Víctor Pareja
Manuela Vázquez
CINDOC-CSIC. Madrid
nrodrigu@bib.uc3m.es

Resumen: Descripción de la presencia de instituciones del sector público de Investigación y Desarrollo de la Comunidad de Madrid en el año 2000, con el objetivo de dar a conocer una nueva técnica para la descripción del Web, o parte de él, denominada "WebWatching". Se desarrollan una serie de indicadores cualitativos y cuantitativos que permiten describirlo estructuralmente y dinámicamente para poder hacer una evaluación de calidad del mismo. Los resultados obtenidos, incluyendo el análisis comparado con muestreos efectuados en el pasado, muestran un incremento importante en el número de sedes, aunque matizado por tratarse de un sector con fuerte implantación en Internet. Sin embargo, se ha incrementado notablemente el tamaño de las mismas y tanto este valor como el de la visibilidad nacional e internacional (impacto) de las sedes parece correlacionado con la importancia y relevancia de las correspondientes instituciones, lo que indica que la autoorganización de los recursos Web en el entorno académico sigue los patrones del sistema de ciencia y tecnología.

Palabras clave: ciencia, tecnología, Internet, World Wide Web, técnicas de descripción, organización del conocimiento, cienciometría, bibliometría, Madrid, política científica, evaluación.

Abstract: This paper intends to describe the presence of the R&D public bodies in the Autonomous Region of Madrid during 2000 using a set of techniques globally known as “WebWatching". The aim is to develop and calculate quantitative and qualitative indicators for each website, so a description of the structure and dynamics of that websites could be applied for quality evaluation. The results show an important growth of the number of web sites, although such increase is expected in a sector with a long Internet tradition. However, the size and the national and international visibility (impact) of websites show also a notable increase. There is a strong correlation between the institutional relevance and the cited impact measurements, so it is possible to indicate that the self-organization of the Web resources in the academic world agrees with the pattern of the science and technology system.

Keywords: Science, technology, Internet, World Wide Web, description techniques, knowledge organization, scienciometrics, bibliometrics, Madrid, science politics, evaluation.

1. Introducción.

La Internet se ha consolidado como el sistema de información más desarrollado a escala mundial, de forma que hoy en día la WWW es una de las principales herramientas de difusión de la existencia y actividades de empresas, organismos y otras entidades tanto públicas como privadas. La presencia institucional en la Web garantiza un mejor y mayor conocimiento, aunque ello no significa que exista necesariamente una relación directa entre la importancia o prestigio de la institución y las características de su sede Web.

Con el fin de evaluar este aspecto en unas condiciones relativamente controladas, Vázquez et al. (1) desarrollaron en 1998 un estudio que relacionara la estructura lógica y volumen de información de un conjunto de sedes Web con la tipología institucional de las entidades responsables de dicha presencia en Internet. Por razones prácticas se decidió utilizar el sistema público de I+D de la Comunidad de Madrid, del que se conocía en detalle su estructura y composición y que, además, resultaba especialmente idóneo por su tamaño y accesibilidad.

Los sistemas de I+D han sido objeto de numerosos estudios, incluyendo aquellos encaminados a describir sus mecanismos de autoorganización (2). De hecho, muchos trabajos sobre la emergente Sociedad de la Información y sus consecuencias están utilizando dichos sistemas como principal objeto de análisis ya que resultan especialmente adecuados a tal fin.

Teniendo en cuenta el crecimiento explosivo del Webespacio, que a finales de 2000 superaba los 3200 millones de páginas, el estudio de sedes Web de instituciones de I+D resulta una muestra aconsejable por tamaño y por el conocimiento previo que ya se posee de las mismas. La descripción estructural y dinámica de sedes Web puede ilustrar el comportamiento de sistemas de conocimiento complejo.

Las técnicas disponibles para dicha descripción reciben, de manera informal, la denominación general de *"WebWatching"*. Estas técnicas, sobre todo aquellas de carácter cuantitativo (cibermétrico), pueden resultar especialmente útiles para estudiar la autoorganización de sistemas de cierto nivel de complejidad.

Se ha retomado el caso citado (1), revisando y actualizando la descripción del sistema público madrileño de Ciencia y Tecnología a partir de su presencia en Internet. Los datos de 1998, de carácter descriptivo se han analizado de acuerdo con criterios cibermétricos más modernos para demostrar la validez y posibilidades de esta técnica.

Se han identificado los diferentes componentes del sistema, procediéndose a una clasificación institucional de acuerdo a sus relaciones administrativas con el

fin de no realizar ninguna presunción de partida que pudiera afectar a los resultados finales. Dichos grupos son: Universidades, Consejo Superior de Investigaciones Científicas, Organismos Públicos de Investigación, unidades de investigación de Hospitales, Sociedades Científicas, Reales Academias, Bibliotecas y Museos. Sin embargo, se ha aceptado desde el principio que la propia presencia en el Web, y toda ella, es reflejo de las aportaciones académico científicas de la institución, de forma que asumimos que Internet se está convirtiendo en el medio de publicación y difusión que complementa e incluso amplia los mecanismos tradicionales de comunicación a través de la publicación en revistas científicas.

2. ¿Qué es *"WEBWATCHING"*?

Ante la necesidad de desarrollar una descripción estructural y dinámica del Web, se presenta una técnica que ya ha sido utilizada en estudios anteriores (3) para la descripción de las características del Web del Reino Unido. Dicho proyecto denominado *WebWatch* permitió desarrollar un conjunto de técnicas que en sentido amplio han conservado el nombre de *"WebWatching"* que es el que utilizamos en el presente estudio. El grupo de Kelly ha seguido aplicando estos métodos para el análisis de diferentes conjuntos del Webespacio académico británico (4) o el proyecto eLib de la British Library (5).

Distintas variantes de estas técnicas han sido utilizadas en otros estudios sobre diferentes tipos de Web, pero el conjunto no ha sido definido formalmente. Literalmente Webwatching es la observación de las sedes Web, labor que por el propio diseño de la red admite un elevado nivel de automatización. La disponibilidad de programas especializados permite describir (fundamentalmente de forma cuantitativa, pero no exclusivamente) las características estructurales de dichas sedes. Con las herramientas adecuadas se puede determinar el tamaño (número de páginas o de objetos hipermedia), organización y profundidad, o relaciones entre páginas propias y externas de las sedes, proporcionando un excelente retrato no solo de cada una de ellas en solitario sino, incluso, de conjuntos considerables de las mismas.

Por lo tanto, el "*WebWatching"* requiere de la aplicación de una amplia serie de herramientas para la recogida de datos (mapeadores, motores de búsqueda, linkcheckers, ...) y el cálculo de indicadores (6, 7). En último término, dichas estadísticas, de carácter cibermétrico, permitirán conocer cual es la estructura, contenido y dinámica de sedes Web e incluso realizar evaluaciones de calidad. (8).

Ya que se trata de una metodología en desarrollo no existe todavía una clasificación universal de los indicadores cibermétricos, aunque se han publicado diversas propuestas. Según la bibliografía recopilada existen dos posturas en el

planteamiento de indicadores calculados. Por un lado aquellos que desarrollan una clasificación teórica de los que van a analizar y por otro aquellos que calculan indicadores sin ningún tipo de clasificación. Entre los del primer grupo destacan Baeza-Yates y Castillo (9) quienes los clasifican bajo cuatro puntos de vista:

- Punto de vista de la Colección: además de obtener indicadores globales se hace un estudio del vocabulario (palabras más frecuentes).
- Punto de vista de la Página: indicadores como el tamaño, tipo de documento e idioma.
- Punto de vista del Sitio: indicadores como la profundidad de las páginas, número de páginas por sitio; cantidad de archivos de texto que existen por sitio.
- Punto de vista del Dominio: número de referencias que hay desde y hacia un dominio, representación de la estructura global de hipervínculos entre dominios, evaluación de las preferencias de los usuarios.

Estos mismos autores en un artículo posterior (10) aplican esta clasificación adaptándola al estudio del Web chileno. Los indicadores utilizados son: número total de páginas, tamaño total del texto, profundidad media de las páginas, enlaces internos (grado interno), enlaces externos (grado externo), establecimiento de una clasificación de páginas, y comprueban su mayor o menor acomodo a distribuciones informétricas tales como la ley de Zipf.

También deben ser mencionadas las dos clasificaciones desarrolladas por Aguillo (8). La primera, según el espacio a evaluar se ordenan en indicadores locales (describen cada sede) e indicadores globales (describen escenarios generales). La segunda, según el método empleado para su obtención, se clasifican en indicadores absolutos o exactos (evaluaciones exhaustivas) e indicadores relativos (evaluaciones desarrolladas a partir de motores de búsqueda).

Otros autores utilizan diferentes conjuntos de indicadores sin aplicar ninguna taxonomía explícita, tales como Veloso et al. (11), con el análisis del Web brasileño, o Leydesdorff y Curran (12), con el estudio de las relaciones dentro de la triple hélice (sistema universidad-industria-gobierno) en Holanda y Brasil. Entre los indicadores más calculados destacan, número de páginas web, cantidad de servidores y dominios, documentos por servidor, número de documentos, tamaño que ocupan, distribución de los documentos por el número de páginas, idiomas, palabras más representadas, dominios con mayor número de documentos, distribución del número de documentos por servidor, número de documentos por nivel de profundidad, porcentaje de errores por nivel, distribución de los documentos según el número de enlaces, distribución de las páginas según el número de referencias, distribución de páginas HTML según el número de imágenes referenciadas.

Para la obtención de los datos se emplean diferentes métodos, siendo el más utilizado el que consiste en la utilización de motores de búsqueda: Estos constituyen una muestra supuestamente representativa y razonablemente amplia (aunque ningún motor por sí solo indiza más de un tercio del Webespacio visible (13)), y resultan especialmente flexibles para la obtención de los datos gracias a los operadores booleanos y a los delimitadores que soportan. Con ciertas precauciones existe además la posibilidad de aplicar técnicas bibliométricas como el análisis de citas, ya que un enlace hipertextual puede asimilarse en contextos cerrados a una referencia bibliográfica.

3. Metodología.

El desarrollo de la evaluación consta de tres pasos consecutivos y complementarios:

3.1. Descripción Cualitativa.

- Comprobación de las distintas sedes institucionales del sector público de I+D de la Comunidad de Madrid del anterior estudio con la finalidad de conocer si hay nuevas incorporaciones o las ya existentes han cambiado de dirección o han desaparecido. Para la identificación de nuevas sedes institucionales se han utilizado índices nacionales (14), mientras que para el seguimiento de las sedes que han cambiado de dirección se utilizaron motores de búsqueda (Alta Vista). Esta información se ha complementado con la visita a los principales dominios, de los que por navegación interna se han extraído nuevas instituciones con presencia en el Web.
  Existen ciertos casos especiales, como las sedes correspondientes a las Reales Academias españolas, que puesto que están agrupadas todas ellas en la actual Academia de España (www.insde.es), no se han tenido en cuenta individualmente. De hecho, aunque todas ellas tienen presencia en el Web, no constituyen sedes propias.

- Se ha aplicado de forma estricta el concepto de sede Web según Aguillo (8) *" un conjunto de páginas Web ligadas jerárquicamente a una página principal, representables por la URL de esta y que forman una unidad documental, distinguible de otras, y una unidad institucional, en la que es posible identificar la responsabilidad de su autoría"*.

- El navegador Netscape ofrece la posibilidad de gestionar y anotar registros a través de los Marcadores (Bookmark), de forma que se pueden recopilar las sedes indizadas y las diferentes características constatadas tras su análisis documental y cibermétrico. Aunque no se trata de un sistema de gestión de bases de datos propiamente dicho, se pueden utilizar diferentes técnicas para

separar los campos que componen la descripción. Se procede posteriormente a la explotación de los datos a través de paquetes ofimáticos, utilizando especialmente la hoja de cálculo MS Excel. Para conocer más sobre este método de descripción téngase en cuenta el artículo de Pareja y Aguillo (15).

- Se han evaluado las siguientes instituciones:

    - Universidades[1]. Por razones prácticas sólo se han estudiado las sedes centrales, pues se reserva para un estudio posterior el análisis de las sedes de departamentos, facultades y escuelas técnicas.
    - Consejo Superior de Investigaciones Científicas (CSIC)
    - Organismos Públicos de Investigación (OPIs), en sentido amplio.
    - Sociedades Científicas[2]
    - Unidades de investigación de los Hospitales
    - Otros. Grupo del que forman parte las Reales Academias, Bibliotecas y Museos.

3.2. Descripción Cuantitativa.

Las sedes recopiladas se describieron cuantitativamente con la ayuda de programas mapeadores, en concreto Microsoft Content Analyzer, que forma parte de la suite MS Site Server Express 3.0. Este programa permite el recuento automático de los objetos, enlaces y otras características de una página web, "mapeándola" de forma completa en un intervalo de tiempo más o menos amplio según el tamaño de la misma. Tras finalizar el análisis el programa ofrece una serie de informes de los que se extraen los datos que siguen:

- Número de páginas de la sede.
- Número de enlaces hipertextuales, tanto internos (navegación) como externos (citas), presentes en dichas páginas (emitidos). Se entiende por enlaces según Aguillo (8) *"un nexo establecido entre objetos de una misma o diferentes páginas a través de la etiqueta <a> del lenguaje HTML"*
- Número de objetos totales, incluyendo los de carácter multimedia (gráficos, imágenes, audio, vídeo) y transaccionales (pasarelas). Se entiende por objetos según Aguillo (8) *"cualquiera de los ficheros informáticos que pueden formar parte de una página. Específicamente son objetos multimedia los ficheros con formatos gráficos (imágenes fijas o en movimiento) y los de audio o sonido"*.

[1] No se ha tenido en cuenta el campus en Madrid de la Universidad Pontificia de Salamanca por tratarse de una universidad que aunque situada en Madrid su campus principal se encuentra en Salamanca.

[2] Se han tenido en cuenta aquellas sedes de sociedades, asociaciones y fundaciones que tienen dominio.es. No se han evaluado aquellas sedes que tienen dominios .com o .org.

Para completar los datos obtenidos por el mapeador se emplearon motores de búsqueda con los que se pudo estimar la cantidad de enlaces dirigidos a las sedes. Teniendo en cuenta que ninguno de los motores que existen en la actualidad se puede considerar como el mejor candidato para este desarrollo, se han tenido en cuenta distintas aportaciones bibliográficas y se decidió utilizar Alta Vista (www.altavista.com). En este caso se han calculado los siguientes valores:

- Número de páginas (host/url:"URL de la sede").
- Número de enlaces recibidos (link:"URL de la sede").
- Número de enlaces recibidos externos (link:"URL de la sede" AND NOT host:"URL de la sede").
- Número de enlaces recibidos del propio país ((link:"URL de la sede" AND domain:"es") AND NOT host:"URL de la sede")).

3.3. *Cálculo de Indicadores.*

Para la elección de los indicadores estadísticos, se tiene en cuenta la clasificación dada por Aguillo (8). Fueron seleccionados teniendo en cuenta no sólo el objetivo de esta evaluación sino intentando reflejar el carácter global e hipertextual de Internet (Véase la tabla I).

Tabla I. Indicadores cibermétricos empleados.

INDICADORES	*CÁLCULO*	*DEFINICIÓN*	*HERRAMIENTA DE OBTENCION*
Tamaño Documental (S)	Número de páginas	Número de páginas de la sede.	Content Analyzer
Riqueza (R)	Número total de objetos.	Número total de objetos de la sede, incluyendo páginas, enlaces y ficheros multimedia, tanto internos como externos.	Content Analyzer
Densidad Absoluta (DA)	R/S	Relación entre número de objetos y tamaño.	Content Analyzer
Densidad Hipertextual (DH)	Enlaces emitidos totales/S	Número de enlaces por página.	Content Analyzer
Densidad Multimedia (DM)	Objetos multimedia/S	Número de objetos multimedia (sin tener en cuenta número de páginas) por página.	Content Analyzer
Endogamia (I%)	Enlaces propios de navegación/Enlaces totales	Porcentaje que representan los enlaces internos respecto del total de la sede.	Content Analyzer
Endogamia (E%)	(Objetos diferentes internos enlazados)/Objetos totales enlazados	Idem utilizando objetos diferentes como unidad	Content Analyzer.
Conectividad (C%)	Enlaces externos recibidos/Total de enlaces recibidos	Número de enlaces propios recibidos respecto al total de enlaces.	Altavista.
Factor de Impacto Web (WIF/WebIF)	Enlaces recibidos no propios/Páginas (Altavista)	Número de enlaces externos recibidos respecto del total de páginas	Altavista.
Visibilidad nacional (V%)	Enlaces externos recibidos del propio país /Páginas (Altavista)	Número de citas recibidas del propio país (.es).	Altavista.
Diversidad (H')	Fórmula de Shannon sobre los dominios con los que la sede enlaza.	Número de enlaces emitidos a diferentes dominios	Content Analyzer
Profundidad (P)	Número de niveles	Número de niveles por cada sede	Content Analyzer.

4. Resultados.

4.1. Análisis Descriptivo.

En total se han encontrado 563 sedes web dedicadas a informar de actividades de I+D en la Comunidad de Madrid. Como se explicó en apartados anteriores los tipos institucionales se agrupan en 6 categorías: Universidades tanto públicas como privadas, Consejo Superior de Investigaciones Científicas (CSIC), Organismos Públicos de Investigación (OPI´s), Sociedades Científicas, unidades de investigación de Hospitales y Reales Academias, Bibliotecas y Museos.

Puesto que existían datos correspondientes a 1996 y 1998, estos últimos actualizados con la revisión realizada en el año 2000 (60 nuevas sedes), ha sido posible estimar el crecimiento a lo largo de dicho periodo que se resume en la tabla II, donde se constata un notable incremente (media del 266%) que indica que el número total de sedes prácticamente se ha triplicado.

Tabla II. Evolución temporal de las sedes Web en la CAM.

INSTITUCIÓN	1996	1998	INCREMENTO DEL PERIODO (%)
CSIC	25	101	304
Universidad	103	373	262
OPIs	12	38	217
Otras	14	51	264

Se han organizado según el campo temático teniendo en cuenta la clasificación temática de la UNESCO (16). En la mayoría de los casos el campo temático que corresponde a una institución está expresada como subdisciplina, pero para facilitar la clasificación se ha optado por encuadrarla dentro de la disciplina superior. Como muestra la tabla III se han encontrado 689 materias agrupadas en 20 campos diferentes. En algunos casos a una misma institución se le ha dado más de un campo temático. De todas las materias encontradas se puede observar que las mejor representadas (66%), son las disciplinas relacionadas con las Ciencias Experimentales (Tecnología, seguida de Matemáticas, Física, Química, Ciencias de la Vida y Ciencias Médicas). En menor proporción están presentadas las Ciencias Sociales (33%) donde destaca la Lingüística, Política y Economía. No existen sedes clasificadas bajo Lógica; Antropología; Demografía y Ética.

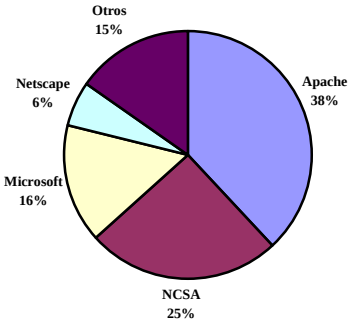
Tabla III. Grandes categorías institucionales distribuidas por campos temáticos.

CAMPO CIENTÍFICO	UNIV.	CSIC	OPIs	OTROS	TOTAL
12. Matemáticas	80	6	1	4	91
21. Astronomía Astrofísica	5	2	5	-	12
22. Física	50	29	3	8	90
23. Química	15	12	-	4	31
24. Ciencias de la Vida	20	15	5	3	43
25. Ciencias de la Tierra	8	4	5	2	19
31. Ciencias Agrarias	10	-	2	1	13
32. Ciencias Médicas	10	4	5	16	35
33. Ciencias Tecnológicas	93	20	13	1	127
53. Ciencias Económicas	32	-	3	2	37
54. Geografía	5		1	-	6
55. Historia	17	5	-	3	25
56. Ciencias Jurídicas	10	-	1	-	11
57. Lingüística	38	14	-	2	54
58. Pedagogía	7	-	-	1	8
59. Ciencia Política	34	3	3	1	41
61. Psicología	5	-	-	6	11
62. Artes y Letras	8	3	-	1	12
63. Sociología	5	2	1	1	9
72. Filosofía	7	6	-	1	14

Uno de los indicadores empleados en el estudio de Kelly and Peacock (2) es el tipo de servidor empleado por cada una de las sedes, el cual demuestra las tendencias tecnológicas en este sector. Como muestra la figura 1, los servidores más utilizados son los de dominio público como Apache (38%) y NCSA (25%), seguidos de los comerciales, donde destaca Microsoft (16%).

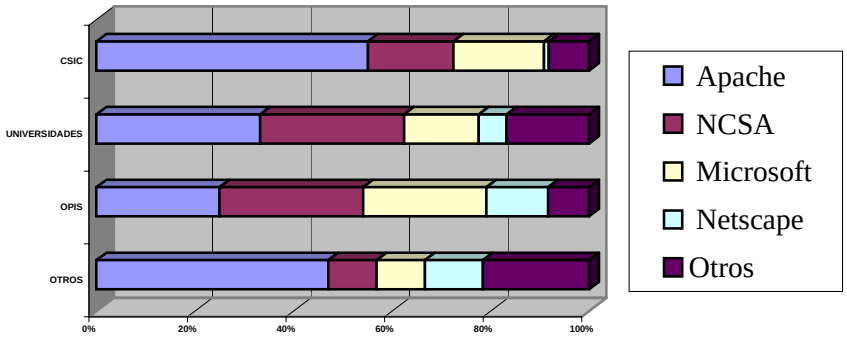
Figura 1. Tipos de servidores empleados por todas las sedes institucionales de la CAM[3].

[3] La categoría "OTROS" incluye servidores que no tienen suficiente entidad para poder ser representados aisladamente, por ejemplo servidores como CERN, Dominio, MacHTTP, Oracle_Web_Listener, Spry Safety Web, WebSTAR, http, NetPresenz, SWS, OmniHTTP, WebSite, WinHttpd, y varios.



Por categorías institucionales (figura 2) se observa la mayor homogeneidad institucional del CSIC y el mayor peso de los programas comerciales en los OPIs.

Figura 2. Distribución de tipos de servidores por grandes categorías institucionales[4].

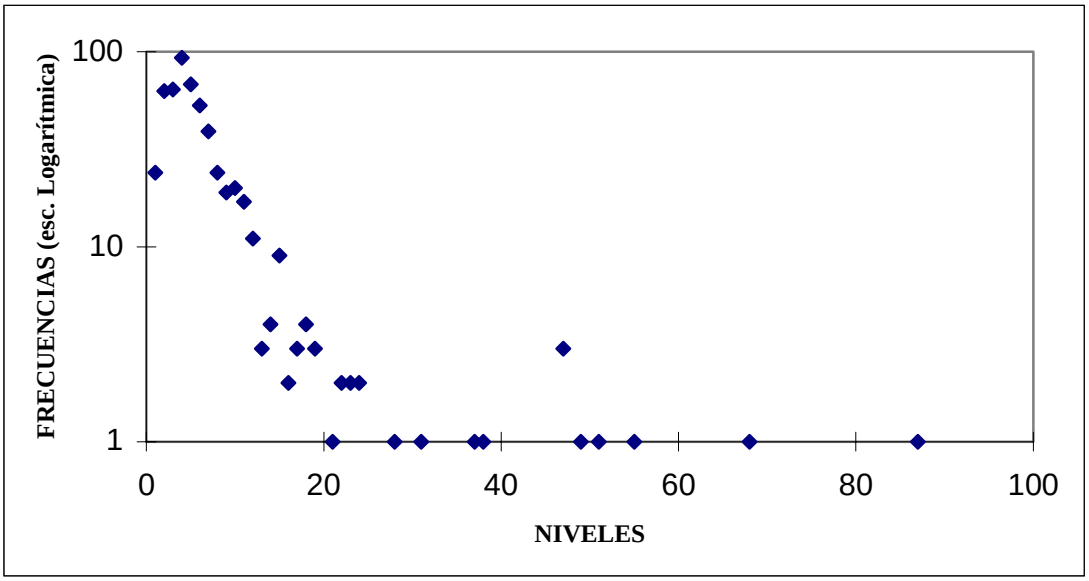


Teniendo en cuenta nuevamente el estudio de Kelly and Peacock (2) otra característica analizada es la profundidad, es decir, el recuento de niveles de directorios medidos desde la página principal (*home page*) de la sede. Desde un punto de vista práctico una gran profundidad plantea indudables desventajas, pues aquellos recursos presentes en los directorios del "fondo" resultan más opacos a los agentes de los motores de búsqueda y son por tanto menos visibles. Contrariamente, una sede que no esté razonablemente organizada en directorios y

[4] *Idem*.

subdirectorios presenta una estructura de contenidos pobre, escasamente vertebrada y posiblemente desorganizada. Aunque es evidente que hay que matizar estos datos con los valores referentes al tamaño y tipología de la sede, se trata de una medida de indudable interés cualitativo. Como muestra la figura 3 el comportamiento de la profundidad se ajusta a una distribución tipo Zipf, con un gran número de sedes con menos de 20 niveles aunque existen sorprendentemente sedes con varias docenas de niveles. Una posible causa que explique este comportamiento es la falta de madurez de algunas sedes creadas por creación de otras realmente anidadas. Con un buen planteamiento de diseño, en un futuro próximo, podría darse una drástica reducción de niveles en ciertas sedes universitarias. Asimismo, la aparición de dominios virtuales puede disminuir el número de niveles, lo que nuevamente es un buen indicador de progreso técnico, al asumir cada grupo o departamento mayores responsabilidades sobre sus servidores.

Figura 3. Distribución de las sedes institucionales de la CAM según la profundidad.



4.2. Indicadores Cibermétricos.

Como se expuso anteriormente se sigue la clasificación de Aguillo (7) explicada en la tabla I en el apartado 3.3 "Cálculo de indicadores".

4.2.1. Universidades.

En total se han evaluado las 13 sedes principales de universidades que tienen presencia en Internet. Se han tenido en cuenta tanto las universidades públicas como las privadas con la finalidad de realizar comparaciones entre ambas. Como muestra la tabla IV, la presencia (tamaño y riqueza) de las universidades públicas es mayor que las privadas. Como se esperaba la Universidad Complutense es la que se encuentra en primer lugar, lo que es coherente con su antigüedad, tamaño e importancia relativa. La Universidad Carlos III, mucho más reciente es la tercera mayor sede universitaria, fruto del reconocimiento pionero de esta institución de la importancia de Internet y de una apuesta por la calidad de la docencia y la investigación. La reciente puesta en marcha de las universidades Rey Juan Carlos I y Camilo José Cela y el menor número de disciplinas que imparten pueden justificar el hecho de que se trate de las dos sedes menores. En general esos mismos criterios pueden aplicarse al resto de las privadas, mientras que hay que indicar que la de Alcalá ha sufrido una intensa remodelación. La Politécnica incluye en su dominio un elevado número de sedes de mediano o gran tamaño, pero su sede central es relativamente modesta.

Tabla IV. Distribución de las universidades clasificadas en públicas y privadas según el tamaño.

CATEGORÍA	UNIVERSIDAD	S	R
PÚBLICAS	COMPLUTENSE	102436	164676
	AUTÓNOMA	5337	14817
	CARLOS III	3871	7904
	UNED	1533	2948
	POLITÉCNICA	1519	2565
	ALCALÁ	1234	4491
	JUAN CARLOS I	35	170
PRIVADAS	CEU	1372	1619
	NEBRIJA	832	2231
	ALFONSO X	653	2013
	PONTIFICIA	556	1181
	EUROPEA	139	393
	CAMILO JOSÉ CELA	51	191

La tabla V muestra un resumen de las medias de los indicadores cibermétricos más relevantes. Tal como se observa la endogamia de las instituciones, es decir, el número de enlaces que contribuyen a la navegabilidad interna de la sede, parecen estar dentro de unos límites razonables ligeramente

por encima del 16 %. Sin embargo, la muy baja diversidad nos indica que los enlaces supuestamente externos lo son en realidad a páginas del mismo dominio. Ello confirmaría cierto carácter cerrado sobre si mismas de las universidades, que ya ha sido señalado con cierta asiduidad en ciertos foros. En este sentido, los datos de Internet parecen confirmarlo en el aspecto académico (aunque seguramente no se puedan generalizar a la actividad investigadora).

Tabla V. Resumen de los indicadores cibermétricos de las universidades de la CAM.

	S	R	I	E	H'
n	13				
suma	119568	205199			
media	9197,5	15784,5	16,3	6,7	0,1

*4.2.2. Consejo Superior de Investigaciones Científicas***.**

El CSIC es el Organismo Público de Investigación más importante de España, con una gran concentración de institutos y centros en la Comunidad de Madrid. Dentro de ella ha habido un incremento de siete nuevas sedes con respecto a 1998, en total existen y se han analizado 101 (en el total de España, Aguillo et al. (inédito) sitúan dicha cifra en 278 sedes web en enero de 2001).

Según la tabla VII es el organismo con mayor tamaño y riqueza, siendo la media dos veces mayor que el del resto de los OPIs. Comparado con los resultados del estudio anterior se observa una importante mejora de las sedes del CSIC, puesto aunque no han aumentado de forma significativa en número, sí lo ha hecho en número de páginas. Estas, además, son más ricas en contenidos a la vista de los resultados de la densidad absoluta, que confirman una elevada relación de objetos por página. En cuanto al número de enlaces emitidos se observa una alta endogamia. La diversidad de impacto muestra cuales son los dominios a los que se han emitido enlaces, siendo los dominios más enlazados España (.es, alrededor del 90%), Alemania (.de) y Reino Unido (.uk).

El alto factor de impacto (enlaces externos recibidos respecto del total de páginas), la alta visibilidad nacional (citas recibidas del propio país) y la alta conectividad (enlaces propios recibidos respecto al total de enlaces) indican que se trata de una institución que ofrece información valorada por la comunidad científica tanto nacional como (en menor medida) internacional. El hecho de tener actividad fundamentalmente investigadora justifica también esta visibilidad, que no queda empañada por la alta endogamia que se puede justificar por la fuerte ligazón administrativa entre centros.

Tabla VII . Medias de los indicadores más representativos del CSIC en la CAM.

	S	R	DA	I	E	H'	WebIF	C	V
n	101								
suma	47667	173144							
media	476,67	1731,4	4,7	8,9	0,4	0,2	13,6	1,1	9,5

En las tablas VIII y IX se ofrece un listado de las diez mayores sedes por las que está compuesto el CSIC con los indicadores cibermétricos más representativos, organizadas por tamaño y factor de impacto. Llamamos la atención de la importante presencia del CINDOC y la Coordinación de Bibliotecas, cuyos servicios de información se han hecho indispensables incluso fuera del ámbito del CSIC y que cuentan con un gran prestigio en toda la comunidad académica.

Tabla VIII. Diez mayores sedes del CSIC en la CAM según su tamaño.

INSTITUCIÓN	S	R	DA	I	E
Inst. Química Física Rocasolano. Dpto. Cristalografía	15245	67934	4,46	136,40	0,02
C. Comunicaciones CSIC-REDIRIS (REDIRIS)	6642	14543	2,19	9,26	0,35
C. Nac. Biotec.. G. Diseño Proteínas	3786	9121	2,41	12,00	0,33
Consejo Superior Invest. Científicas (CSIC)	2997	11391	3,80	4,07	0,56
C. Invest. Biológicas (CIB)	1648	2910	1,77	7,44	0,33
Inst. Ciencia Materiales Madrid (ICMM)	1534	6068	3,96	5,23	0,36
C. Inform. Doc. Científica (CINDOC)	818	3877	4,74	4,72	0,70
Unidad Coordinación Bibliotecas	726	5916	8,15	2,70	0,81
Inst. Química Física Rocasolano. Dpto. Estructura y Dinámica Molecular. G. Teoría Fluidos	719	3945	5,49	23,75	0,12
Clío. RedIRIS	716	2407	3,36	5,99	0,47

Tabla IX. Diez mayores sedes del CSIC en la CAM según el factor de impacto.

INSTITUCIÓN	WebIF	C	V	S
Unidad Coordinación Bibliotecas	686	1	594	726
C. Comunicaciones CSIC-REDIRIS (REDIRIS)	18,88	1,03	10,35	6642
Consejo Superior Invest. Científicas (CSIC)	3,07	1,14	2,03	2997
C. Inf. y Doc. Científica (CINDOC)	1,14	1,17	0,44	818
Clío. RedIRIS	0,90	1,09	0,29	716
Inst. Ciencia Materiales Madrid (ICMM)	0,82	3,24	0,11	1534
C. Invest. Biológicas (CIB)	0,18	1,16	0,07	1648
Inst. Química Física Rocasolano. Dpto. Estructura y Dinámica Molecular. G. Teoría Fluidos	0,18	1,60	0,06	719
Inst. Química Física Rocasolano. Dpto. Cristalografía	0,03	2,30	0,01	15245
C. Nac. Biotec.. G. Diseño Proteínas	0	0	0	3786

4.2.3. Organismos Públicos de Investigación (OPIS).

El resto de los Organismos Públicos de Investigación cuentan con 38 sedes presentes en Internet, existiendo un crecimiento del 217% con respecto al año 1996 (incluyendo 8 sedes nuevas más).

Se caracterizan por su alto nivel de endogamia, desde el punto de vista de los enlaces y de los objetos emitidos por página. Aunque la diversidad no se muestra en la tabla X, más del 90% de los enlaces emitidos son a dominios españoles (.es), con la excepción de Madri+d con un 88,7% de sus enlaces a dominios de organizaciones (.org). En ambos casos hay que destacar la sede de la Agencia Espacial Europea de Villafranca del Castillo, que no sólo tiene una alta endogamia sino también una gran diversidad (0,66 bytes), lo que está justificado por tratarse de la agencia de carácter europeo con fuertes relaciones internacionales, además de una relevante fuente de información.

Tabla X. Medias de los indicadores cibermétricos de los Organismos Públicos de Investigación de la CAM.

	S	E%	I%
n	38		
suma	8343		
media	225,5	80	74

En la tabla XI se presentan los mayores Organismos Públicos de Investigación de la CAM con sus indicadores cibermétricos más representativos, organizados por tamaño.

Tabla XI. Diez mayores Organismos Públicos de Investigación de la CAM según su tamaño.

INSTITUCIÓN	S	R	I%	E%
C. Investigaciones Sociológicas	2072	777	97,82	83,78
Inst. Salud Carlos III	1512	3296	96,88	86,68
Inst. Nac. Meteorología	821	1682	96,12	95,66
Agencia Espacial Europea Villafranca del Castilla	781	4992	56,33	32,13
C. Investigaciones Energéticas, Medioambientales y Tecnológicas	451	352	68,42	51,7
Laboratorio Astrofísica Espacial y Física Fundamental	447	1547	83,61	61,93
C. Nac. Investigaciones Oncológicas Carlos III (CNIO)	311	891	96,54	93,27
Inst. Investigación Básica. Grupo Física Partículas	255	1445	86,91	73,08
Inst. Toxicología	252	644	74,67	74,69
Inst. Nac. Investigación y Tecnología Agraria y Alimentaria	150	215	89,73	81,86

4.2.4. Sociedades Científicas, Asociaciones y Fundaciones.

Hemos identificado 41 sedes de sociedades, asociaciones y fundaciones. Tienen una alta densidad absoluta indicando que existe una buena proporción de objetos y páginas. El porcentaje de enlaces emitidos está relacionado con la alta diversidad observándose que los dominios más enlazados son los españoles (.es). Sin embargo, la Sociedad Española de Medicina de Urgencia cuenta con un 74,1% de sus enlaces a dominios de organizaciones (.org). También es destacable el alto nivel de enlaces por página que reciben de instituciones externas nacionales y extranjeras (factor de impacto), de instituciones nacionales exclusivamente (visibilidad) y de sí mismo (conectividad), como muestra la tabla XII.

Tabla XII. Medias de los indicadores cibermétricos de las Sociedades Científicas, Asociaciones y Fundaciones de la CAM.

	S	R	DA	I%	E%	WebIF	C	V
N	41					10		
Suma	3981	12209						
Media	99,5	305,2	10,2	70	60	3,9	1,4	2,4

Se trata de instituciones menos conocidas que las anteriores, haciendo que esto tenga influencia en la presencia de sus sedes en Internet. En las tablas XIII y XIV se muestran cuales son las cinco mayores sociedades, asociaciones y fundaciones de la CAM según el tamaño y el factor de impacto.

Tabla XIII. Cinco mayores Sociedades Científicas, Asociaciones y Fundaciones de la CAM según su tamaño, e indicadores relacionados.

INSTITUCIÓN	S	R	DA	I%	E%
As. Hispanoamericana de C. de Investigación y Empresas de TELECOM	1051	1445	1,37	25	12
S. Esp. de Cardiología	654	1663	2,54	70	84
As. Esp. de Contabilidad y Administración de Empresas	302	490	1,62	91	76
As. Int. de Sociología	255	459	1,80	70	28
S. Esp. de Medicina Nuclear	214	425	1,99	95	64

Tabla XIV. Cinco mayores Sociedades Científicas, Asociaciones y Fundaciones de la CAM según su factor de impacto e indicadores relacionados.

INSTITUCIÓN	WebIF	C	V	S
S. Esp. de Medicina Nuclear	5	1	0,5	214
As. Hispanoamericana de C. de Investigación y Empresas de TELECOM	1,88	0,93	0,97	1051
S. Esp. de Cardiología	1,01	0,89	0,12	654
As. Esp. de Contabilidad y Administración de Empresas	0	0	0	302
As. Int. de Sociología	0	26	0	255

4.2.5. Hospitales.

A pesar de que las instituciones médicas están experimentando en la actualidad una mejora en su presencia en Internet, sin embargo, no se han encontrado resultados diferentes con respecto al estudio anterior. Se han tenido en cuenta sólo las unidades de investigación de los hospitales de la CAM, identificándose las mismas 4 sedes. Según el número de páginas, la que mayor presencia tiene es la unidad de investigación del Hospital Universitario de San Carlos (115 páginas).

En las tablas XV y XVI se puede ver un resumen de las medias de los indicadores de las sedes según el tamaño y el factor de impacto.

Tabla XV. Unidades de investigación de los Hospitales de la CAM con sus indicadores cibermétricos según el tamaño.

INSTITUCIÓN	S	R	DA	DH	DM	I%	E%	H'
Hospital Universitario de San Carlos. Unidad de Investigación	115	6	6	8	5	0	20	0,7
Hospital Ramón y Cajal. Unidad de Investigación	41	371	3,23	8,49	1,22	82	70	0
Clínica Puerta de Hierro. Unidad Mixta de Investigación	1	1	1	0	0	0	100	0
Hospital Universitario de la Princesa. Unidad de Epidemiología Clínica	1	546	13,32	24,56	2,12	49	10	0,6

Tabla XVI. Unidades de investigación de los Hospitales de la CAM con sus indicadores cibermétricos según el factor de impacto.

INSTITUCIÓN	WebIF	C	V	S
Hospital Universitario de la Princesa. Unidad de Epidemiología Clínica	25	1	0	1
Clínica Puerta de Hierro. Unidad Mixta de Investigación	1,79	0,44	0,2	1
Hospital Universitario de San Carlos. Unidad de Investigación	0	0	0	115
Hospital Universitario Ramón y Cajal. Unidad de Investigación	0	1	0	41

4.2.6. Reales Academias, Bibliotecas y Museos.

Este grupo es muy heterogéneo y no se pueden sacar opiniones concluyentes sobre la presencia de las mismas en Internet. Sin embargo, hay que destacar que, al igual que en el caso de los hospitales, no han variado y los nuevos indicadores no ofrecen ninguna información relevante que las caracterice.

5. Conclusiones.

A la vista de los resultados obtenidos se observa que el número de instituciones nuevas aparecidas en el sector público de Investigación y Desarrollo de la Comunidad de Madrid en el año 2000 es de 60 nuevas instituciones, por lo que ha habido un incremento con respecto al año 1998 (266%).

En lo que respecta a las características cualitativas los campos científicos con mayor presencia son las Ciencias Experimentales, principalmente las Ciencias Tecnológicas, y las menos representadas son las Ciencias Sociales.

En cuanto a las características cualitativas muestran las actuales tendencias en cuanto a tecnologías de servidores, y aunque siguen siendo los servidores de dominio público las más empleadas, existe una gran variedad de otros programas utilizados. La profundidad de las sedes muestra una situación coyuntural, con estructuras no plenamente consolidadas, que pueden dificultar la visibilidad y navegabilidad de las mismas.

En cuanto a las características cuantitativas se ha observado una mejora en la presencia de las instituciones del CSIC ya que si en 1998 el tamaño medio de las instituciones era menor que el del resto de los OPIs, actualmente ha mejorado llegando a ser dos veces mayor. De esta manera se muestra el esfuerzo realizado por esta institución en mejorar su presencia en la red ya que el aumento en el tamaño se debe no a la presencia de nuevas instituciones sino a la mejora de las ya existentes.

En el resto de los casos se demuestra que ya existe una relación directa entre el peso académico y/o científico de una institución y su presencia en Internet. En el caso de las universidades, el CSIC y los OPIs, sus sedes Web están creciendo en tamaño y a la vez son las instituciones que mayor cantidad de enlaces reciben de otras instituciones debido a la importancia de las mismas dentro del sistema de investigación y desarrollo de la Comunidad de Madrid.

Por último, con respecto a las técnicas de *"WebWatching"*, para describir y caracterizar sedes Web, resultan especialmente cómodas y sencillas de aplicar, sus resultados admiten interpretaciones y comparaciones útiles a distintos niveles y, con matices, pueden ser utilizadas para realizar evaluaciones de calidad.

6. Bibliografía.

1. Vázquez, Manuela; Aguillo, Isidro, and Pareja, Víctor (1998). "Observatorio Madrileño de la Ciencia y la Tecnología en Internet: Presencia de Instituciones y Centros de investigación y desarrollo madrileños en el World Wide Web". Informe final.

2. Leydesdorff, Loet (2001)A Sociological Theory of Communication: The Self-Organization of the Knowledge-Based Society. Amsterdam, Universal Publishers/uPUBLISH.com. 359 pags.

3. Kelly, Brian and Peacock, Ian (1999). "WebWatching UK Web Comunities: Final Report for the WebWatch Project". British Library Research and Innovation Report 146. British Library Research and Innovation Centre.

4. Kelly, Brian (1997) “”WebWatching UK Universities and Colleges”. *Ariadne*. 12 <http://www.ariadne.ac.uk/issue12/web-focus/intro.html/. [Consultado el 26 de septiembre del 2000].

5. Kelly, Brian.(2000) “WebWatch:WebWatching eLib project Web Sites”. *Ariadne*. 26.<http://www.ariadne.ac.uk/issue26/web-watch/> [Consultado el 15 de enero del 2001].

6. Aguillo, I. F. (1998). Herramientas de Segunda Generación. Anuario SOCADI 1998. Socadi, Barcelona.

7. Aguillo, I. F (1999). Del multibuscador al metabuscador: Los agentes trazadores en Internet. Actas del IV Congreso ISKO-España, EOCONSID’99. Granada, 22-24 de abril de 1999.

8. Aguillo, Isidro (2000). "Indicadores: Hacia una evaluación objetiva (cuantitativa) de sedes web". *VII Jornadas Españolas de Documentación. Bilbao* .(Bilbao 19-21 de octubre del 2000). Bilbao: FESABID.

9. Baeza-Yates, Ricardo and Castillo, Carlos (2000). “Relating Web Characteristics”. <http://www.todocl.com/stats/rbaeza.pdf>. [Consultado el 8 de enero del 2001].

10. Baeza-Yates, Ricardo and Castillo, Carlos (2000). "Caracterizando la Web chilena".<http://www.todocl.com/stats/jun2000/wcl2000.html>. [Consultado el 8 de enero del 2001].

11. Veloso, Eveline A.; Moura, Edleno S. de; Golgher, Paulo B.; Silva, Altigran S. da; Almeida, Rodrigo B.; Laender, Alberto H. F.; Ribeiro-Neto, Berthier, and Ziviani, Nivio. (2000) "Um Retrato da Web Brasileira". <http://www.ppgia.pucpr.br/sbc2000/eventos/semish.htm>. [Consultado el 8 de enero del 2001].

12. Leydesdorff, Loet and Curran, Michael (2000) "Mapping University-Industry-Goverment Relations on the Internet: The Construction of Indicators for a Knowledge-Based Economy". *Cybermetrics*.; 4.

13. Sullivan, Danny (2000). Search Engine Sizes. SearchEngineWatch.com

<http://www.searchenginewatch.com/reports/sizes.html> [Consultado el 8 Noviembre 2000].

14. RedIris (2000). Catálogo de recursos. Sistemas de información existentes en la Red Académica española <http://www.rediris.es/recursos> [Consultado el 14 de Septiembre 2000].

15. Pareja, Víctor Manuel & Aguillo, Isidro (1996). “Experiencias de indización rápida para recursos de información en Internet: El caso del directorio de recursos europeos para PYMEs”. Actas de las *V Jornadas Españolas de Documentación Automatizada* (Cáceres, 17, 18 y 19 de octubre de 1996). 601-608. Cáceres: FESABID, ABADMEX, Universidad de Extremadura.

16. CICYT (1996). Plan Nacional de I+D. Instituciones y Centros de Investigación y Desarrollo. Madrid. Comisión Interministerial de Ciencia y Tecnología.