La información medioambiental en España: recursos y acceso a la información pública: análisis webmétrico (2ª parte)

Manuel Blázquez Ochando
Doctor en Ciencias de la Documentación, licenciado en Documentación y diplomado en Biblioteconomía y Documentación Profesor Ayudante Doctor en la Facultad de Ciencias de la Documentación de la Universidad Complutense de Madrid. Departamento de Biblioteconomía y Documentación. Facultad de Ciencias de la Documentación. Universidad Complutense de Madrid – España.
manuel.blazquez@pdi.ucm.es

L. Fernando Ramos Simón
Doctor en Ciencias de la Información y licenciado en Periodismo y Derecho. Profesor en la Facultad de Ciencias de la Documentación de la Universidad Complutense de Madrid. Departamento de Biblioteconomía y Documentación. Facultad de Ciencias de la Documentación. Universidad Complutense de Madrid – España.
lframoss@ucm.es

Rosario Arquero Avilés
Doctora en Ciencias de la Información (Rama Documentación), licenciada en Documentación y Diplomada en Biblioteconomía y Documentación. Profesora Titular en la Facultad de Ciencias de la Documentación de la Universidad Complutense de Madrid, Departamento de Biblioteconomía y Documentación. Facultad de Ciencias de la Documentación. Universidad Complutense de Madrid – España.
carquero@ucm.es

Silvia Cobo Serrano
Doctoranda en la Facultad de Ciencias de la Documentación de la Universidad Complutense de Madrid, Master en Gestión de la Documentación y Bibliotecas, Licenciada en Documentación y diplomada en Biblioteconomía y Documentación de la CM. Investigadora en formación del programa Formación del Profesorado Universitario del Ministerio de Educación, Cultura y Deporte (España). Departamento de Biblioteconomía y Documentación. Facultad de Ciencias de la Documentación. Universidad Complutense de Madrid – España
s.cobo@ucm.es

Resumen

La meta de la investigación es el análisis webmétrico de los principales sitios de la administración central española especializada en medio ambiente, con el objetivo de estudiar la topografía, estructura, interrelación y metadescripciones de los contenidos, pudiendo posteriormente realizar su comparación con la web mexicana de la misma área de conocimiento. Para lograrlo, se ha utilizado la herramienta webcrawler Mbot, que analiza la extensión y dimensiones de la web, los rankings de sitios web con más páginas, el ratio de meta-descriptores por página, el análisis de frecuencia de los términos empleados en las descripciones y una serie de tablas estadísticas que permiten valorar la muestra. A raíz de los resultados obtenidos, se han elaborado varias recomendaciones dirigidas a mejorar la capacidad de indexación de los motores de búsqueda y suprimir malas prácticas que debilitan la capacidad de recuperación y acceso a la información pública. Entre las conclusiones, destaca la importante interrelación entre la Web española de medio ambiente y su homóloga europea, descubriendo fuentes de información poco conocidas, la recopilación de canales de sindicación que permita el seguimiento de la información pública en medio ambiente y el descubrimiento del rango ideal de frecuencias de aparición de los términos usados en las metadescripciones.

Palabras clave: Webmetría, Mbot, webcrawler, medio ambiente, administración pública, información del sector público, acceso a la información, topografía web, información medioambiental, España, México.

Environmental Information in Spain: Resources and Access to the Public Information. Webometric Analysis (Part 2)

Abstract

The goal of the research is the webometric analysis of the main Spanish central Government sites concerning environmental issues. In this line, topography, structure,

Cómo citar este artículo: BLÁZQUEZ OCHANDO, Manuel, RAMOS SIMÓN, L. Fernando, ARQUERO AVILÉS, Rosario y COBO SERRANO, Silvia. La información medioambiental en España: recursos y acceso a la información pública: análisis webmétrico (2ª parte). *Revista Interamericana de Bibliotecología* 2014, vol. 37, n° 1, pp. 13-34.

Recibido: 2013-10-22 / Aceptado: 2013-11-25

interaction, and meta-description of the contents are studied so that results can be compared with the Mexican environment websites. To achieve this, the Mbot webcrawler tool - which analyzes the site extent and dimension, the website ranking, the meta-descriptors ratio by page, the frequency analysis of the description terms and some statistical tables to estimate the sample - has been used in this academic. Because of results, several recommendations to improve the search engines indexing and remove bad practices that limit the access to public information as well as the information retrieval have been developed. It can be outlined the interplay between the Spanish environment Websites and its European counterpart, the number of unpopular information sources, syndication channels that enable public environmental information monitoring and, lastly, the ideal range of frequency of occurrence for the terms used in the meta-descriptions.

Keywords: Webometrics, Mbot, webcrawler, environment, public administration, public sector information, information access, website topography, environmental information, Spain, Mexico

1. Introducción y objetivos

El acceso a la información pública depende de las políticas de información y documentación y también de la estructuración y diseño de los medios informativos y divulgativos puestos a disposición en la red que, por ende, requieren de un estudio webmétrico profundo para la cuantificación y análisis cualitativo de las meta-etiquetas y metadatos Dublin Core. En consecuencia, en esta segunda parte de la investigación relativa a los recursos y al acceso de la información pública medioambiental, el objetivo principal del estudio es el análisis del estado del arte en relación a la web pública española especializada en materia de Medio Ambiente con la intención de establecer, *a posteriori*, una comparación con la web pública mexicana en materia medioambiental. Para el presente análisis --una vez que se ha definido y delimitado el alcance de la información ambiental a partir de la normativa nacional e internacional comentada en el artículo anterior y conociendo la complejidad de la gestión de este tipo de información-, se ha requerido la utilización de la herramienta webcrawler Mbot (Blázquez Ochando, 2013a). Esta herramienta ha sido desarrollada para analizar la web a partir de dos listas de enlaces que representan las principales páginas web relativas al sector medioambiental y que son soportadas por la administración española y mexicana, utilizadas como objeto del estudio. Debido a la extensión de los resultados obtenidos en esta investigación, en el presente artículo se relacionan los datos referidos a la web medioambiental de la administración central española y se plantean los siguientes objetivos específicos:

- Identificar y analizar los enlaces de la administración central española en cada uno de los niveles de análisis establecidos a partir de la muestra de dominios seleccionada.
- Identificar los formatos que más utiliza la administración central española para disponer la información pública medioambiental en la red.
- Estudiar el ratio por página web y el tipo de meta-etiquetas y metadatos utilizados por la administración central española.

2. Metodología

El análisis de la web requiere el empleo de programas de análisis y rastreo denominados también *Web-crawlers*. Estas aplicaciones analizan la web mediante un método bien definido: en primer lugar, emplean una lista de enlaces (constituida por dominios, sitios y páginas web), a la que se denominada *semilla*. A continuación, se extraen y analizan los contenidos de cada enlace para obtener nuevos enlaces a terceras páginas dependientes del mismo dominio o sitio web, alcanzando de esta manera distintos niveles de profundidad (Thelwall, 2001, p. 323). Posteriormente, toda la información es almacenada de forma sistemática en una base de datos (tabulada y computada) para generar los distintos informes que se ofrecen en esta investigación (Blázquez Ochando, 2013b).

El primer paso metodológico del estudio consistió en la confección de una semilla (Cothey, 2004, p.1230) especializada en Medio Ambiente y sectores afines, tales como la energía, el cambio climático, la oceanografía, geología, geografía, minería, agricultura, biodiversidad e hidrografía. También se incluyen en esta lista los centros de investigación y ministerios que están relacionados directa o indirectamente con la temática principal estudiada, véase Tabla 1 en la que se presenta una agrupación de los organismos, de acuerdo a su adscripción pública, realizada en septiembre de 2013.

Tabla 1. Semilla de enlaces originales analizados con Mbot

MINISTERIOS	ORGANISMO PÚBLICO	URL
MINISTERIO DE AGRICULTURA, ALIMENTACIÓN Y MEDIO AMBIENTE	Agencia Estatal de Meteorología	http://www.aemet.es
	Confederación Hidrográfica del Cantábrico	http://www.chcantabrico.es
	Confederación Hidrográfica del Ebro	http://www.chebro.es
	Confederación Hidrográfica del Guadalquivir	http://www.chguadalquivir.es
	Confederación Hidrográfica del Guadiana	http://www.chguadiana.es
	Confederación Hidrográfica del Miño Sil	http://www.chminosil.es
	Confederación Hidrográfica del Segura	http://www.chsegura.es
	Fundación Biodiversidad	www.fundacion-biodiversidad.es
	Mancomunidad de los Canales del Taibilla	http://www.mct.es
	Ministerio de Agricultura, Alimentación y Medio Ambiente	http://www.magrama.gob.es
MINISTERIO DE ASUNTOS EXTERIORES Y DE COOPERACIÓN	Casa Mediterráneo	http://casa-mediterraneo.es
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD	Consejo Superior de Investigaciones Científicas	http://www.csic.es
	Instituto sobre el Cambio Climático de Zaragoza	http://www.i2c2.org
	Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas	http://www.ciemat.es
	Instituto Español de Oceanografía	http://www.ieo.es
	Instituto Geológico y Minero de España	http://www.igme.es
	Instituto Nacional de Estadística	http://www.ine.es
MINISTERIO DE FOMENTO	Instituto Geográfico Nacional	http://www.ign.es
	Salvamento Marítimo	http://www.salvamentomaritimo.es
MINISTERIO DE HACIENDA Y ADMINISTRACIONES PÚBLICAS	Dirección General del Catastro	http://www.catastro.meh.es
MINISTERIO DE INDUSTRIA, ENERGÍA Y TURISMO	Comisión Nacional de Energía	http://www.cne.es
	Ciudad de la Energía	http://www.ciuden.es
	Instituto para la Diversificación y el Ahorro de Energía	http://www.idae.es
	Instituto para la Reestructuración de la Minería del Carbón y Desarrollo Alternativo de las Comarcas Mineras	http://www.irmc.es
	Ministerio de Industria, Energía y Turismo	http://www.minetur.gob.es
	Oficina Española de Patentes y Marcas	http://www.oepm.es
MINISTERIO DE LA PRESIDENCIA	Boletín Oficial del Estado	http://www.boe.es
OTROS	Consejo de Seguridad Nuclear (organismo independiente)	http://www.csn.es
	Empresa Nacional de Residuos Radiactivos S.A. (empresa pública)	http://www.enresa.es
	Observatorio Ambiental Granadilla (fundación publica mixta)	http://www.oag-fundacion.org

En la semilla especificada se observa la presencia de dos ministerios, seis confederaciones hidrográficas, siete institutos públicos de temáticas afines y siete entidades directamente relacionadas con el estudio, gestión y tratamiento del sector energético español. La selección permite reflejar una gran parte del sector medioambien-

tal de la Administración Central española. No obstante, se debe reseñar un hecho importante, que es la naturaleza y tamaño de los sitios web de los ministerios y agencias estatales con respecto al resto de instituciones y el criterio de selección. Por ejemplo, el Boletín Oficial del Estado (BOE) y el Instituto Nacional de Estadística (INE) han sido seleccionados por contener, respectivamente, información legal y estadística relativa a la materia medioambiental. Añadido a este factor, comparten una importante jerarquía de niveles de enlazamiento junto con los Ministerios, requiriendo, por tanto, un análisis en mayor profundidad dada su densidad de contenidos, ya que pueden seguirse encontrando páginas web en el 6º y 7º nivel de análisis. Debido a tales circunstancias y con el objetivo de uniformar la investigación, se ha tenido en cuenta en la configuración del webcrawler que el análisis se efectuará a tres niveles, o lo que es lo mismo, un nivel por cada página enlazada desde el enlace o dominio raíz especificado en la semilla, tal como describen en sus contribuciones (Chakrabarti; Joshi; Punera; Pennock, 2002, p.509) (Bergmark; Lagoze; Sbityakov, 2002, pp.91-106). Esta decisión ha sido tomada con el objetivo de reducir el tiempo de rastreo del webcrawler y obtener una base de conocimiento o muestra suficiente como para observar unos patrones claros durante el análisis de los datos resultantes.

El siguiente paso para el análisis de la semilla es la configuración del rastreo en el webcrawler. En este sentido, se estableció la extracción de metadatos, meta-etiquetas, canales de sindicación, imágenes, documentos, archivos multimedia, correos electrónicos y texto completo (Blázquez Ochando, 2011, p.2-3). De todos los elementos analizados, el estudio aborda con especial interés los términos y textos empleados en los metadatos y meta-etiquetas, ya que son empleados de forma directa en la indexación de las páginas web por parte de los principales buscadores (Berners Lee, 1995. p.22-23). Dicho de otra forma, el éxito en la recuperación de un contenido, con independencia de su temática, estriba en gran medida en su meta-descripción e identificación mediante metadatos y meta-etiquetas apropiados. Las meta-etiquetas son etiquetas html introducidas en el encabezado de las páginas web que describen, según las especificaciones oficiales (W3C, 1999), el título, autor, derechos, palabras clave y descripción del sitio web; es decir, campos de descripción mínimos para la identificación del sitio web. Por otra parte, los metadatos Dublin Core cualifican con 54 elementos (DCMI, 2012) aspectos más variados y útiles desde el punto de vista documental como, por ejemplo: resumen, derechos de acceso, recursos web alternativos, público objetivo, cita bibliográfica, colaborador, cobertura, fecha de creación del recurso, fecha de aceptación, fecha de *copyright*, fecha de envío, extensión, formato, partes del recurso, versiones del recurso, identificador normalizado, referenciación de terceros, idioma, licencia, soporte, procedencia, editor, relación con terceros recursos, fuente de información original, sumario de contenidos o tipo de contenido, entre otros. Junto a los elementos sujetos al análisis, también se establecen restricciones de análisis de dominio que tienen como objetivo el análisis pormenorizado de los dominios especificados en la semilla, bloqueando el análisis de páginas externas con un dominio diferente. De esta manera, se permite enfocar el análisis sólo a los enlaces especificados, obteniendo una mayor precisión en los resultados así como un menor tiempo de ejecución.

3. Resultados

Con la intención de que los resultados obtenidos en la presente investigación académica tengan una mejor comprensión, estos han sido divididos en tres grandes subepígrafes: enlaces, formatos y, por último, meta-etiquetas y metadatos.

3.1. Enlaces

La ejecución del análisis dio como resultados más de 1,5 millones de enlaces, de los cuales 679.000 son únicos y sin repetición (Henzinger, 2003)[1], véase Tabla 2. Este dato indica que más de 843.000 enlaces son empleados para la redirección y navegación de contenidos en los distintos sitios web analizados, puesto que ocupan los lugares comunes de los menús generales y contextuales, lo que supone el 55,38% del total. Esto significa que existe una gran interconexión de contenidos.

1 Los enlaces duplicados a páginas web ocurren cuando un mismo documento se encuentra en más de una localización del servidor, o bien cuando se crea redundancia para facilitar la navegación a dichos contenidos. De esta forma en cada página del sitio web se repiten los enlaces de navegación a las principales categorías y secciones del mismo. Es muy importante su detección con la finalidad de evitar distorsionar el análisis de la web, ya que no representaría adecuadamente el contenido real de la misma.

En cuanto al recuento de enlaces para cada nivel de análisis, se observa una progresión exponencial tanto en las cifras de enlaces obtenidos como de enlaces únicos, especialmente entre el nivel 2 y 3, donde se observa que de 37.177 enlaces únicos se alcanzan los 649.736, véase Tabla 3 y Figura 1 en la que se muestra la línea exponencial que describe la tendencia.

En cuanto a la topografía de la web, se observa que los dominios genéricos más reiterados son los españoles ".es", superando los 675.000 enlaces únicos y los gubernamentales en lengua española ".gob.es" con 4.724, véase Tabla 4[2]. Con cifras comprendidas entre los 500 y los 1.200 enlaces únicos se encuentran los dominios de organizaciones y comerciales ".org", ".com" .

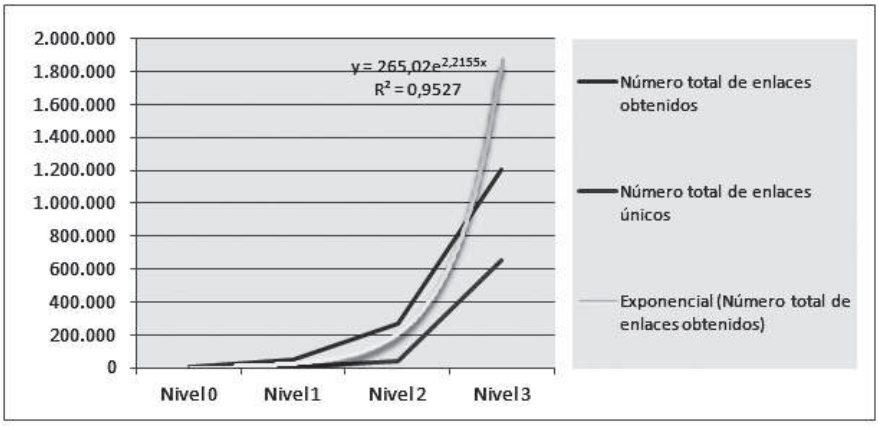
Tabla 2. Cifras generales del análisis de la web institucional relativa a medio ambiente en España

Número total de enlaces obtenidos (incl. duplicados)	1.522.881
Número total de enlaces únicos	679.439
Número total de sitios web únicos	4.030
Número total de páginas web únicas	675.409

Tabla 3. Cifras parciales para cada nivel de análisis[3]

	Nivel 0	Nivel 1	Nivel 2	Nivel 3
Número total de enlaces obtenidos (incl. duplicados)	1.330	47.801	269.855	1.203.895
Número total de enlaces únicos	1.190	7.551	37.177	649.736
Número total de sitios web únicos	200	1.061	2.277	2.290
Número total de páginas web únicas	990	6.490	34.900	647.446

Figura 1. Tendencia a un crecimiento exponencial en cada nivel de análisis



2 Tabla ordenada de forma decreciente según el número de dominios genéricos y geográficos TLD (Top-level domain).

3 La Tabla 3 sólo muestra el recuento parcial del número de enlaces únicos para cada nivel. Esto implica que existan enlaces repetidos entre varios niveles, por ejemplo entre los enlaces del nivel 1 y 2. Por ello, cuando se suman las cifras de enlaces únicos de la tabla 3 no coinciden con los valores absolutos de la tabla 2, sin que por ello la información proporcionada deje de ser exacta.

Tabla 4. Principales dominios genéricos y geográficos TLD según el número de páginas vinculadas

.es	.gob.es	.org	.com	.eu	.int	.net	.uk
676.449	4.761	1.129	693	460	102	94	36
.gov	.cat	.fr	.edu	.ch	.de	.org.es	.com.es
32	32	29	28	19	18	17	16
.info	.pt	.jp	.au	.dk	.ca	.it	.be
15	12	11	11	10	10	9	8

Si bien las cifras mencionadas son elevadas y representativas de los resultados obtenidos, resulta muy significativo profundizar en los datos recopilados sobre enlaces salientes con dominio ".eu" de forma que pueda ponerse de relieve la relación que existe entre la web de la administración central española de medio ambiente con respecto a la Unión Europea. Los resultados obtenidos demuestran que el portal de Derecho y legislación *Eurlex* es el recurso más vinculado por los sitios web de la semilla, ya que al menos un 50% de los enlaces salientes están relacionados con legislación medioambiental, gestión de recursos hídricos, mercado de energía y sostenibilidad, véase Tabla 5[4]. Otros datos de interés son los dominios enlazados, dado su alto nivel de especialización; por ejemplo, el Instituto Laue Langevin de investigación de física de neutrones, el sistema de alerta meteorológica Meteoalarm de la Unión Europea, diversas instituciones de estudios oceanográficos (PLOCAN, EMODnet, EUR-OCEANS, TPA-maritime) así como las instituciones y proyectos europeos relativos al estudio de nuevas energías como Windplatform, SNETP, FCH-JU, IGDTP, European-Biofuels. También es enlazada la Agencia Europea de Medio Ambiente, que obtiene 26 enlaces repartidos entre el Ministerio de Agricultura, Alimentación y Medio Ambiente (MAGRAMA); la Fundación Biodiversidad; la Fundación OAG y la Confederación Hidrográfica del Segura -fuente que, en resumen, es menos enlazada que otras más genéricas como el portal de la Unión Europea, la Agencia Europea de Estadística Eurostat o la Comisión Europea-. A falta de comprobar el motivo, puede introducirse la hipótesis de un posible desconocimiento de los recursos de esta institución europea, ya que armoniza en gran medida los datos agrarios y medioambientales de los países miembros.

Tabla 5. Muestra de enlaces más representativos con dominio ".eu" (I)

Semilla	Dominios enlazados	Categoría	Nº total de enlaces salientes (incl. duplicados)
http://www.ine.es/ (20) http://www.idae.es/ (15) http://www.magrama.gob.es/ (28) http://www.boe.es/ (7) http://www.minetur.gob.es/ (1) http://www.csic.es/ (1) http://www.chguadiana.es/ (10) http://www.oepm.es/ (1) http://www.i2c2.org/ (1)	*Eurlex* *http://eur-lex.europa.eu*	Derecho, legislación	246

4 Tabla ordenada de forma decreciente según el número total de enlaces salientes.

Semilla	Dominios enlazados	Categoría	Nº total de enlaces salientes (incl. duplicados)
http://www.csic.es/ (48) http://www.magrama.gob.es/ (43) http://www.ciemat.es/ (15) http://www.oepm.es/ (8) http://www.minetur.gob.es/ (7) http://www.oag-fundacion.org/ (3) http://www.igme.es/ (1) http://www.idae.es/ (3) http://www.ine.es/ (5) http://www.chsegura.es/ (3) http://www.chminosil.es/ (1) http://www.i2c2.org/ (1) http://www.fundacion-biodiversidad.es/ (4)	*Comisión Europea* *http://ec.europa.eu/*	Gobierno, Legislación	196
http://www.ine.es/ (110) http://www.magrama.gob.es/ (6)	*Eurostat* *http://epp.eurostat.ec.europa.eu*	Estadística, base de datos	121

Tabla 5. Muestra de enlaces más representativos con dominio ".eu" (II)

Semilla	Dominios enlazados	Categoría	Nº total de enlaces salientes (incl. duplicados)
http://www.magrama.gob.es/ (13) http://www.boe.es/ (7) http://www.minetur.gob.es/ (4) http://www.chguadiana.es/ (5) http://www.chsegura.es/ (2) http://www.csn.es/ (1) http://www.fundacion-biodiversidad.es/ (1) http://www.oag-fundacion.org/ (1) http://www.i2c2.org/ (1)	*Unión Europea* *http://europa.eu*	*Gobierno, portal institucional*	63
http://www.csic.es/ (43), http://www.ciemat.es/ (15)	*Instituto Laue Langevin* *http://www.ill.eu/*	*Investigación en física de neutrones*	61
http://www.csic.es/ (55)	*Keytonature* *http://www.keytonature.eu/*	*Educación medioambiental*	55
http://www.aemet.es/ (49) http://www.oag-fundacion.org/ (1)	*Meteoalarm* *http://www.meteoalarm.eu/*	*Meteorología, alertas meteorológicas*	50
http://www.oepm.es/ (22)	*OAMI Oficina de Armonización del Mercado Interior* *http://oami.europa.eu/*	*Propiedad intelectual, patentes y marcas*	37
http://www.magrama.gob.es/ (11) http://www.fundacion-biodiversidad.es/ (3) http://www.chsegura.es/ (1) http://www.oag-fundacion.org/ (1)	*Agencia Europea de Medio Ambiente* *http://www.eea.europa.eu/*	*Política medioambiental*	26

Tabla 5. Muestra de enlaces más representativos con dominio ".eu" (III)

Semilla	Dominios enlazados	Categoría	Nº total de enlaces salientes (incl. duplicados)
http://www.ciemat.es/ (15)	*European Grid Infraestructure* http://www.egi.eu/	*Telecomunicaciones, computación*	15
http://www.ciemat.es/ (15)	*ESRF European Synchrotron Radiation Facility* http://www.esrf.eu/	*Investigación, instrumentos científicos, synchrotron, rayos-x, microscopio electrónico, análisis molecular, análisis de partículas*	15
http://www.ciemat.es/ (15)	*European XFEL* http://www.xfel.eu/	*Investigación, instrumentos científicos, investigación con electrones, rayos-x, análisis del átomo*	15
http://www.ciemat.es/ (15)	*Fusion for Energy* http://fusionforenergy.europa.eu/	*Energía, fusión nuclear, ITER*	15
http://www.ciemat.es/ (15)	*PRACE Partnership for Advanced Computing in Europe* http://www.prace-project.eu	*Computación científica*	15
http://www.boe.es/ (7), http://www.magrama.gob.es/ (3), http://www.fundacion-biodiversidad.es/ (1)	*Parlamento Europeo* http://www.europarl.europa.eu	*Gobierno, Legislación*	11

Tabla 5. Muestra de enlaces más representativos con dominio ".eu" (IV)

Semilla	Dominios enlazados	Categoría	Nº total de enlaces salientes (incl. duplicados)
http://www.boe.es/ (7), http://www.magrama.gob.es/ (1)	*Oficina de Publicaciones de la Unión Europea* http://publications.europa.eu/	*Documentación europea*	8
http://www.csic.es/ (2)	*Plants day* http://www.plantday12.eu/	*Educación medioambiental*	6
http://www.magrama.gob.es/ (4)	*ECHA European Chemicals Agency* http://echa.europa.eu/	*Agencia de seguridad de productos químicos, legislación sobre sustancias químicas,*	5
http://www.magrama.gob.es/ (1)	*ENRD European Network for Rural Development* http://enrd.ec.europa.eu/	*Desarrollo rural*	5
http://www.oepm.es/ (2)	*Innovaccess* http://www.innovaccess.eu/	*Propiedad intelectual, patentes y marcas*	4

Semilla	Dominios enlazados	Categoría	Nº total de enlaces salientes (incl. duplicados)
http://www.oepm.es/ (4)	China IPR http://www.china-iprhelpdesk.eu/	*Protección de los derechos de propiedad intelectual*	4
http://www.magrama.gob.es/ (4)	*TPA maritime* http://www.tpeamaritime.eu/	*Oceanografía*	4
http://www.magrama.gob.es/ (2)	*EIONET European Environment Information and Observation Network (1)* http://www.eionet.europa.eu/ *European Centre Sustainable Consumption Production (2)* http://scp.eionet.europa.eu/	*Sostenibilidad, Medio Ambiente*	3

Tabla 5. Muestra de enlaces más representativos con dominio ".eu" (V)

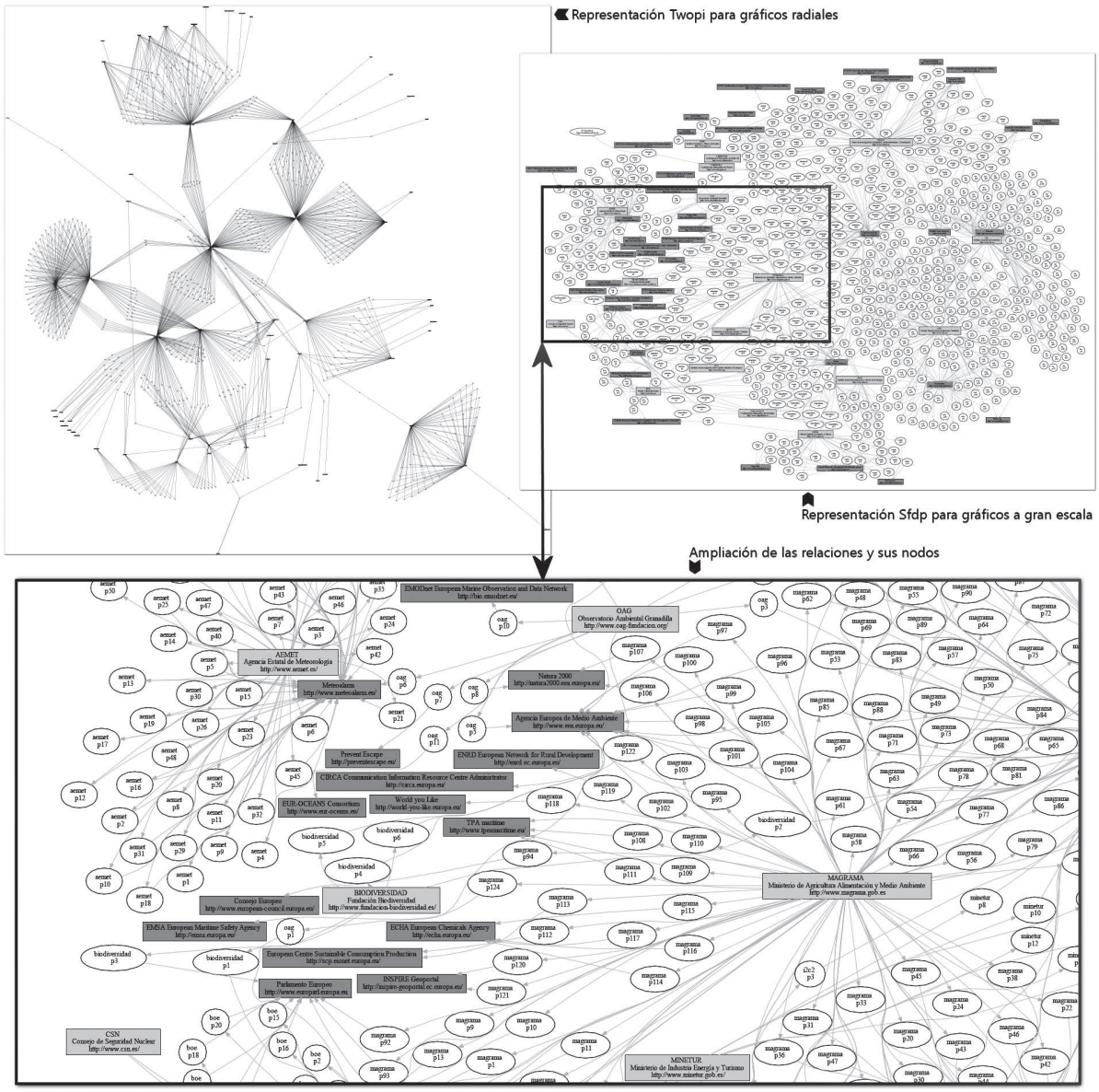
Semilla	Dominios enlazados	Categoría	Nº total de enlaces salientes (incl. duplicados)
http://www.magrama.gob.es/ (2) http://www.oag-fundacion.org/ (1)	*Natura 2000* http://natura2000.eea.europa.eu/	*Base de datos medioambientales*	3
http://www.magrama.gob.es/ (1)	*World you Like* http://world-you-like.europa.eu/	*Ecología, Educación medioambiental*	2
http://www.magrama.gob.es/ (2)	*INSPIRE Geoportal* http://inspire-geoportal.ec.europa.eu/	*Base de datos geoespaciales*	2
http://www.ciemat.es/ (1)	*Windplatform* http://www.windplatform.eu/	*Energía eólica*	2
http://www.csn.es/ (1) http://www.oepm.es/ (1)	*CORDIS Servicio de Información Comunitario sobre Investigación y Desarrollo* http://cordis.europa.eu/	*Documentación científica*	2
http://www.magrama.gob.es/ (2)	*Life at 20* http://www.life20.eu/	*Educación medioambiental*	2
http://www.aemet.es/ (2)	*MACC Monitoring Atmospheric Composition and Climate* http://www.gmes-atmosphere.eu/	*Meteorología, observación climática*	2
http://www.magrama.gob.es/ (1)	*Consejo Europeo* http://www.european-council.europa.eu/	*Gobierno, jefatura europea*	1

Tabla 5. Muestra de enlaces más representativos con dominio ".eu" (VI)

Semilla	Dominios enlazados	Categoría	Nº total de enlaces salientes (incl. duplicados)
http://www.oag-fundacion.org/ (1)	*Prevent Escape* http://preventescape.eu/	*Acuicultura*	1
http://www.oag-fundacion.org/ (1)	*PLOCAN Plataforma Oceánica de Canarias* http://www.plocan.eu/	*Oceanografía*	1
http://www.oag-fundacion.org/ (1)	*EMODnet European Marine Observation and Data Network* http://bio.emodnet.eu/	*Oceanografía*	1
http://www.oag-fundacion.org/ (1)	*EUR-OCEANS Consortium* http://www.eur-oceans.eu/	*Oceanografía*	1
http://www.magrama.gob.es/ (1)	*CIRCA Communication Information Resource Centre Administrator* http://circa.europa.eu/	*Directorio de proyectos europeos*	1
http://www.igme.es/ (1)	*Green Mines* http://www.green-mines.eu/	*Minería, sostenibilidad de la actividad minera*	1
http://www.igme.es/ (1)	*European Geosciences Union* http://www.egu.eu/	*Ciencias de la tierra, sociedad geofísica europea*	1
http://www.ciemat.es/ (1)	*IGDTP Implementing Geological Disposal of Radioactive Waste Technology Platform* http://www.igdtp.eu/	*Energía nuclear, residuos radiactivos, almacenamiento de residuos*	1
http://www.ciemat.es/ (1)	*SNETP Sustainable Nuclear Energy Technology Platform* http://www.snetp.eu/	*Energía nuclear, sostenibilidad*	1
http://www.ciemat.es/ (1)	*FCH-JU Fuel Cells and Hydrogen Joint Undertaking* http://www.fch-ju.eu/	*Energía, combustible, células de combustible, células de hidrógeno*	1
http://www.ciemat.es/ *(1)*	*European Biofuels* http://www.biofuelstp.eu/	*Energía, combustible, biofuel*	1

El análisis de los enlaces con dominio ".eu" ha sido sometido a un proceso de representación gráfica. De hecho, el webcrawler Mbot permite la generación de archivos en formato DOT legibles por la herramienta de visualización gráfica *Graphviz*, desarrollada por investigadores de la empresa AT&T, véase Figura 2.

Figura 2. Mapa en detalle de los enlaces de la web de medio ambiente de la administración pública española con la Unión Europea[5].



En la figura pueden observarse dos gráficas distintas. Por un lado, la correspondiente a la representación *Twopi* (Gansner, 2012, p.22) que es utilizada habitualmente para representar grafos de forma radial y, por otra parte, la representación *Sfdp* (Gansner, 2012, p.21) para representar gráficos a gran escala con ahorro de espacio. En ambos casos, se puede comprobar cómo existen unos pocos recuadros azules (Dominios de la Unión Europea) que reciben la mayoría de los enlaces procedentes de los recuadros verdes (Dominios de la Administración

5 Disponible en: http://www.mblazquez.es/wp-content/uploads/papiit2-mapa-web1.pdf, http://www.mblazquez.es/wp-content/uploads/papiit2-mapa-web2.pdf y http://www.mblazquez.es/wp-content/uploads/papiit2-mapa-web-grafico.png

Española) y que coinciden con los expresados en la Tabla 5. También se denota que hay más recuadros azules que verdes, lo que indica que un importante componente OUT en la muestra (Graells; Baeza Yates, 2007).

3.2. Formatos y tamaño de la web

Los resultados obtenidos en relación a los formatos de archivos detectados en los enlaces únicos permiten afirmar que la web analizada es dinámica, ya que posee un 70% de páginas editadas en PHP y ASP, lenguajes de programación del lado del servidor que hacen posible la interacción del usuario, el envío de peticiones y sus respuestas en el modelo cliente-servidor. Esto es la capacidad de realizar operaciones con bases de datos, realizar comunicaciones básicas, organizar contenidos, realizar búsquedas, con una interfaz HTML autogenerada para cada caso y petición del usuario, produciendo un cambio constante que resulta dinámico para el usuario.

Otro aspecto reseñable es la proporcionalidad en el número de páginas PHP y ASP, que son casi parejas con 4.524 y 4.235 enlaces respectivamente, véase Tabla 6. Esto significa que no hay una tecnología predominante, encontrando un equilibrio entre el número de servidores que soportan cada lenguaje, ya sea Apache Web Server en el caso de PHP y Microsoft Windows Server en el caso de ASP.

En cuanto al 30% de los enlaces restantes, estos corresponden a las páginas web en formato HTML, considerándose la parte estática de la web analizada. Los 3.662 enlaces estáticos pueden corresponder a contenidos publicados en la red para facilitar los procesos de indexación o páginas web básicas que aún no han sido actualizadas al estándar dinámico. No obstante, este resultado demuestra que la web analizada está altamente actualizada desde el punto de vista técnico.

Tabla 6. Análisis general de formatos para la totalidad de la muestra analizada

Páginas web			Sindicación		Documentos ofimáticos			Datos	Compresión		Email
html	php	asp	rss	atom	doc	xls	pdf	xml	zip	rar	email
3.662	4.524	4.235	497	240	68	101	2.149	75	125	15	1.910
Archivos de imágenes				Archivos de video						Archivos de audio	
jpg	png	gif	bmp	flv	swf	avi	mpg	mp4	wmv	mp3	wma
8.449	1.120	2.266	39	3	2	5	1	2	7	8	1

Otro aspecto destacable es la gran cantidad de canales de sindicación en formato RSS y Atom que en conjunto suman 737 enlaces. La recopilación de estos recursos tiene un gran valor desde el punto de vista de la reutilización de la información del sector público en España y más concretamente del sector medioambiental, puesto que mediante sistemas de agregadores o buscadores especialmente diseñados para el caso es posible recopilar una gran parte de la información que las administraciones y entidades analizadas publican en la red y con ello elaborar un seguimiento permanente de sus contenidos. También se observa que el principal formato de documento ofimático es PDF con más de 2.000 documentos accesibles, la gran mayoría corresponden al MAGRAMA, al Instituto Nacional de Estadística y al Boletín Oficial del Estado. Finalmente, se puede apreciar que existen más de 1.900 correos electrónicos no protegidos en toda la web analizada. Esta cifra puede indicar el número de vías de comunicación que tiene el ciudadano para establecer contacto con la Administración, ejercer sus derechos y obligaciones legales.

En cuanto al tamaño de los sitios web según su número total de enlaces, se ha comprobado que el CIEMAT es el que aglutina la mayor cantidad de enlaces (650.000), lo que representa una gran mayoría del número total de enlaces del análisis. Esto se debe a que su estructura jerárquica de contenidos es menos vertical y más horizontal; es decir, las páginas de contenidos están enlazadas directamente en niveles de enlazamiento cercanos a la página de portada o inicio del sitio web, lo cual hace que su información sea mucho más visible que, por ejemplo, la correspondiente a la Oficina de Patentes y Marcas, véase Tabla 7.

Tabla 7. Ranking de los 10 dominios con más páginas web

Rango	Dominio	Nº total de enlaces (incluyendo páginas, documentos, imágenes)	Nº enlaces analizados	Nº de dominios y subdominios	Nº de páginas
1	http://www.ciemat.es	656.524	18.848	55	637.687
2	http://www.ine.es	12.085	2.130	779	9.296
3	http://www.csic.es	11.538	3.465	186	8.069
4	http://www.magrama.gob.es	6.886	1.622	833	4.460
5	http://www.aemet.es	3.644	766	60	2.828
6	http://www.enresa.es	2.777	1.013	14	1.769
7	http://www.boe.es	2.317	502	144	1.691
8	http://www.csn.es	2.112	783	39	1.382
9	http://www.igme.es	2.065	399	692	1.063
10	http://www.oepm.es	1.849	680	150	1.256

3.3. Meta-etiquetas y metadatos

En cuanto al empleo de meta-etiquetas y metadatos, se han elaborado dos tablas que muestran tanto los dominios que las incluyen en el código fuente de sus páginas como los recuentos totales y el ratio de meta-etiquetas o metadatos por página web, que indican la densidad de uso de las mismas, véase Tabla 8 y 9. Las meta-etiquetas son las más utilizadas en la muestra analizada, ya que 26 de los 30 dominios analizados las emplean habitualmente, frente a 5 dominios que sí emplean metadatos Dublin Core. Ello no es sinónimo de mayor densidad, ya que los dominios con metadatos Dublin Core poseen un mayor ratio de metadescripciones por página web. Este es el caso de la Comisión Nacional de Energía con un total de 9 metadatos por página. Por ejemplo, el dominio con más meta-etiquetas es el CIEMAT con 56.349 meta-etiquetas, que resultan suficientes para superar la densidad que tiene en su categoría la web de Salvamento Marítimo con un ratio de 3,9 meta-etiquetas por página.

Tabla 8. Estadística del uso de meta-etiquetas para la descripción de los contenidos (I)

Dominio	Nº total de meta-etiquetas	Meta-etiquetas / Página web	Tipo de meta-etiqueta utilizada
http://www.ciemat.es	56.349	2,99	description (18.783), author (18.783), rights (18.783)
http://www.aemet.es	2.782	3,632	title (756), description (514), language (756), author (756)
http://www.enresa.es	2.404	2,373	description (418), language (993), author (993)
http://www.ine.es	2.143	1,006	description (2.143)
http://www.oepm.es	1.929	2,837	description (646), author (644). rights (639)
http://www.magrama.gob.es	1.644	1,014	description (1.593), rights (51)
http://www.chsegura.es	1.607	3,661	title (400), description (407), author (400), rights (400)
http://www.salvamentomaritimo.es	1.128	3,917	title (282), description (282), language (282), author (282)

Dominio	Nº total de meta-etiquetas	Meta-etiquetas / Página web	Tipo de meta-etiqueta utilizada
http://www.chcantabrico.es	658	2,246	title (3), description (278), language (281), author (96)
http://www.csn.es	622	0,794	description (621), author (1)
http://www.fundacion-biodiversidad.es	599	1,585	title (119), description (371), author (109)
http://www.mct.es	438	2,168	title (144), description (154), author (140)
http://www.boe.es	433	0,863	description (433)

Tabla 8. Estadística del uso de meta-etiquetas para la descripción de los contenidos (II)

Dominio	Nº total de meta-etiquetas	Meta-etiquetas / Página web	Tipo de meta-etiqueta utilizada
http://www.chguadiana.es	329	0,979	description (329)
http://casa-mediterraneo.es	228	0,987	description (228)
http://www.ciuden.es	224	0,982	description (1), author (223)
http://www.csic.es	194	0,056	description (193), author (1)
http://www.chminosil.es	174	1,794	description (97), author (77)
http://www.minetur.gob.es	173	0,41	description (173)
http://www.oag-fundacion.org	133	1,511	description (73), author (47), date (13)
http://www.catastro.meh.es	130	1,94	title (65), description (65)
http://www.idae.es	113	0,375	title (3), description (110)
http://www.chguadalquivir.es	113	0,934	description (112), date (1)
http://www.ieo.es	86	0,945	description (86)
http://www.igme.es	33	0,083	description (18), author (15)
http://www.i2c2.org	8	1	description (8)

Desde el punto de vista cuantitativo, los datos muestran una información muy clara y tendente, pero cuando se analizan en profundidad y cualitativamente, se observan rasgos que cuestionan algunos de los datos obtenidos. Este es el caso de los siguientes organismos: Salvamento Marítimo, la Oficina de Patentes y Marcas, CIEMAT, el Portal del Catastro y el Instituto Español de Oceanografía. Sucede que para todas las páginas web se repiten siempre los mismos términos para las mismas meta-etiquetas, de tal forma que no se realiza ningún tipo de descripción sobre el contenido de la web. Por ejemplo, las 1.128 meta-etiquetas detectadas para Salvamento Marítimo siempre contienen las mismas y únicas palabras (web, oficial, salvamento, marítimo), la Oficina de Patentes y Marcas contiene 1.929 repeticiones de las palabras (patentes, oficina, marcas, España), el Portal del Catastro 130 repeticiones de las palabras (portal, catastro) y así sucesivamente. Este hecho hace que la recuperación de la información en los principales buscadores dependa única y exclusivamente del texto indexado en cada página sin tener en consideración los metadatos y meta-etiquetas, lo cual penaliza la búsqueda y acceso a la información.

Tabla 9. Estadística del uso de metadatos Dublin Core para la descripción de los contenidos

Domínio	Nº total de metadatos	Metadatos / Página web	Tipo de metadato Dublin Core utilizado
http://www.cne.es	3.068	9,158	dc:title (332), dc:subject (331), dc:description (83), dc:language (331), dc:creator (332), dc:-publisher (332), dc:contributor (332), dc:rights (332), dc:date (332), dc:identifier (331)
http://www.idae.es	1.412	4,691	dc:title (292), dc:subject (135), dc:description (109), dc:publisher (292), dc:contributor (292), dc:date (292)
http://www.aemet.es	756	0,987	dc:title (756)
http://www.salvamentomaritimo.es	564	1,958	dc:title (282), dc:language (282)
http://www.irmc.es	312	2,382	dc:title (54), dc:creator (129), dc:publisher (129),

A pesar de estas malas prácticas detectadas, también es cierto que existen casos de sitios web que realizan un profuso y correcto empleo de los metadatos y meta-etiquetas. Este es el caso de la Agencia Española de Meteorología, Casa Mediterráneo, Ministerio de Industria, Energía y Turismo, MAGRAMA, IDAE, Fundación Biodiversidad y ENRESA. Se aprecia en las instituciones citadas que su sitio web realiza un uso adecuado de los instrumentos de meta-descripción, con descriptores que representan el contenido de sus principales secciones y apartados. Esto significa que una mayor densidad de meta-descriptores no siempre es garantía de una correcta representación de los contenidos. Este hecho puede ser contrastado a fondo a raíz de los datos presentados en Tabla 10. En ellas se representa el número total de términos empleados en las meta-etiquetas y metadatos Dublin Core, el número de términos relevantes contabilizados y analizados manualmente de acuerdo a la validez documental de los mismos (Gil Leiva, 2008) y sucesivamente los rangos de frecuencias de aparición de los términos desde 1.000 hasta 1 iteración, considerados como Hapax, o términos con la frecuencia de aparición más baja y el mayor poder discriminatorio. Se observa que entre los dominios que emplean meta-etiquetas, aquellos que contienen más de 30 palabras no llegan a superar el 30% de descriptores relevantes especializados en medio ambiente o con la institución a la que representan. Ello significa que el vocabulario, aún siendo correctamente empleado, no termina de describir con suficiente profundidad los contenidos y objetos de la web más especializada.

Tabla 10. Análisis cuantitativo de meta-etiquetas, frecuencias y términos relevantes (I)

Meta-etiquetas								
Dominio	Nº total términos	Términos relevantes	› 1001	101-1000	51-100	26-50	2-25	Hapax
http://casa-mediterraneo.es	1156	33	0	1	1	5	594	555
http://www.aemet.es	788	135	3	16	20	22	481	246
http://www.magrama.gob.es	680	135	4	1	0	4	410	261
http://www.minetur.gob.es	513	35	0	0	0	6	190	317
http://www.enresa.es	486	43	0	6	6	9	318	147
http://www.mct.es	323	68	0	3	1	2	104	213
http://www.fundacion-biodiversidad.es	256	69	0	19	0	2	95	140

Meta-etiquetas								
Dominio	Nº total términos	Términos relevantes	› 1001	101-1000	51-100	26-50	2-25	Hapax
http://www.chsegura.es	173	47	1	5	2	0	66	99
http://www.boe.es	161	0	0	0	1	3	75	82
http://www.csn.es	112	31	0	3	0	0	36	73
http://www.idae.es	93	15	0	0	0	0	56	37
http://www.ine.es	62	6	22	13	0	1	13	13
http://www.igme.es	58	17	0	0	0	0	48	10
http://www.chminosil.es	36	9	0	1	10	0	4	21
http://www.csic.es	34	6	0	2	5	8	8	11
http://www.ciuden.es	16	0	0	0	1	3	10	2
http://www.oag-fundacion.org	15	1	0	0	2	1	7	5
http://www.oepm.es	11	0	4	1	0	0	5	1
http://www.chguadiana.es	10	1	0	10	0	0	0	0
http://www.chcantabrico.es	7	0	4	1	0	2	0	7
http://www.ieo.es	5	1	0	0	3	2	0	0
http://www.i2c2.org	4	2	0	0	0	0	4	0
http://www.salvamentomaritimo.es	4	2	0	4	0	0	0	0
http://www.chguadalquivir.es	3	0	0	2	0	0	0	1
http://www.catastro.meh.es	2	0	0	2	0	0	0	0
http://www.chcantabrico.es	7	0	4	1	0	2	0	7

Tabla 10. Análisis cuantitativo de meta-etiquetas, frecuencias y términos relevantes (II)

Metadatos Dublin Core								
Dominio	Nº total términos	Términos relevantes	› 1001	101-1000	51-100	26-50	2-25	Hapax
http://www.idae.es	338	58	0	5	1	6	217	109
http://www.irmc.es	51	5	0	4	0	0	24	23
http://www.salvamentomaritimo.es	5	2	0	5	0	0	0	0
http://www.aemet.es	619	119	0	12	5	15	259	328
http://www.cne.es	635	50	0	8	5	7	257	358

Otro resultado de importancia es la constatación de un gran grupo de términos relevantes cuyo rango de frecuencia es medio (espectro de 2-25 repeticiones), siendo en consecuencia ligeramente más concurrido que el rango de términos con una iteración. Ello viene a corroborar la afirmación de Hans Peter Luhn en relación a que los términos con frecuencias de aparición medias tienen mejor factor de representatividad y exhaustividad a la hora de recuperar los documentos (Luhn, 1958). No obstante, debe observarse también en la Tabla 11, donde se introduce una muestra de los términos y descriptores en cuestión, que los términos Hápax con rango 1 se encuentran muy parejos al rango de frecuencias medias.

Tabla 11. Selección de términos más relevantes con rango de frecuencias medias de aparición (de 2 a 25 iteraciones y Hápax) (I)

Meta-etiquetas		
Dominio	**Términos relevantes con meta-etiquetas - Rango 2-25**	**Términos relevantes con meta-etiquetas - Rango 1 Hápax**
http://casa-mediterraneo.es	(21 términos) Cultura/l, sociedad, turismo, comercio, exposición, diplomacia, comunidad árabe, instituto francés, desarrollo sostenible, desarrollo económico, mediterráneo, seguridad social, cámara de comercio, asuntos exteriores, biología, urbana/o, turístico, geografía, geográfico, Ebro, agua	(9 términos) Playa, pesquero, pesquerías, pesca, orillas, marítima, industria/s, desierto,
http://www.aemet.es	(32 términos) Umbrales, clima, balance, variables, periodos, modelos, atmosféricos, suelo, mínimas, cambio, viento, superficie, aerónica, predicción, humedad, climáticos, aire, sequía, playas, estaciones, observación, temporal, primavera, meteoros, meteorología, climáticas, medio ambiente, mediterráneo, meteomet, Maspalomas, eumetsat, ciclogénesis explosivas	(37 términos) Verano, vegetación, troposfera, tropopausa, tormentoso, temporada, surface, sondeos, salud, reactivos, reactive, rayos, radiros, radiativo, radiation, radiacisan, radiacia, protección, prevención, nubes, nordeste, monte, medicina, litoral, isla, frontera, emission, emisión, depuración, climate, azufre, atmosféricas, archipiélagos, alergias, aerosoles, aguas, aerónico
http://www.magrama.gob.es	(26 términos) Recursos naturales, cambio climático, pesca, parques nacionales, espacios protegidos, mapa rural, espacios naturales, ganaderos, ganadería, rural, recursos marinos, marinos, pesqueros, recursos pesqueros, especies marinas, sector pesquero, miguel arias cañete, agrícolas, residuos, habitats, ecoturismo, sostenible, presas, patrimonio, natural, magrama	(64 términos) Yogur, vinos, vías, verduras, vegetal, vaca, turístico, turísticas, transporte, tortugas, tembladera, tejido, sumideros, sidra, selección de semillas, salmonela, scrapie, reservas, prevención, polvo, pesquero, pecuarias, peces, pastelería, panadería, oliva, nata, montes, mieles, mamíferos, legumbres, jamones, INM, humano, huella, hortalizas, incendios, habitat, frutas, frontera, forestales, forestal, especias, embutidos, embalajes, cultivo, cuajada, clima, caza, carnes, campo, bovina, bebidas, atún, ASECIC, arroces, arenas, animales, alimentos, ahorro, agrario, agrarias, AEMET, aceites
http://www.minetur.gob.es	(10 términos) Energía, líneas de apoyo financiero, energy, MITYC, energética, operadores, espectro eléctrico, eléctrico/s, residuos, renovables	(14 términos) Vigilancia, terminales, térmicas, seguridad, sanitaria, químicas, PIVE2, minería, materias primas, licuados, gasolina, gas, gases, explosivos, emisiones
http://www.enresa.es	(19 términos) Medio ambiente, uranio, MINER, ATC, reactores, ciclo, EURATOM, radiación alfa, emisor, radiación, vigilancia, terrenos, radiología, retirada, desmontaje, natural, medio/s, instalación, arranque nuclear	(4 términos) Zorita, Vandellos, tecnología, radiactivo

Meta-etiquetas		
Dominio	**Términos relevantes con meta-etiquetas - Rango 2-25**	**Términos relevantes con meta-etiquetas - Rango 1 Hápax**
http://www.mct.es	(20 términos) Canal, murcial, molina de segura, ETAP, agua/s, Cartagena, San Pedro del Pinatar, La Pedrera, control del consumo, depósitos, Alicante, tramo, Lorca, embalse, presa, Torrealta, suministro de agua, hídrica, estaciones, emergencia	(30 términos) Transporte, tendido, Tajo, seguridad, residuos, reparación, renovación, reforma, ramal, presas, impacto, hidrológico, hídricas, filtración, geosintéticos, ETI, embalses, eléctricas, disipación, depósito, demolición, cuenca, convenio, conservación, cloro, CHS, CETENMA, canteras, arquetas, aliviadero
http://www.fundacion-biodiversidad.es	(23 términos) Biodiversidad, verdes, ecología, custodia, sierra, protegidas, naturaleza, naturales, natura, life, infonatur, climático, carbono, ambiental, aire, verde, turismo, territorio, terrestre, sostenible, salud, litoral, huella	(22 términos) Tortugas, refugio, reconversión, playas, pesca, pecuarias, panaderos, ozono, olivar, oasis, marinas, mar, madera, lobo, jardines, humedales, ferrovial, energético/a, diversidad, culebra, bosque, agronomy
http://www.chsegura.es	(17 términos) SAIH, licitación, proceso, muestreo, campañas, vegas, sectorial, Tajo-Segura, rambla, montañas, ambiental, agua/s, ríos, costeras, Taibilla, sequía/s, embalses	(21 términos) Trasvase, sondeos, riesgo, regadíos, marino, inundación, hidrología, hídricos, geología, FEDER, embalse, edafología, ecológicos, desagües, climáticas, cenajo, caudales, biótico, altiplano, alerta, agraria
http://www.boe.es	- [No se encuentran términos en el rango especificado] -	- [No se encuentran términos en el rango especificado] -
http://www.csn.es	(16 términos) Muestras, CSN, alfa, agua/s, ICE, safety, potable/s, vigilancia, suelo/s, sedimento/s, radioyodos, radio, muestreo, medida, conservación, ambiental, aerosoles	(14 términos) Yodo, vapor, tritio, sedimento, retención, resto, recolecta, gamma, filtros, emisores, desecación, cesio, centelleo, carboactivo, ambiental
http://www.idae.es	(9 términos) Energética, IDAE, plan de eficiencia energética, ahorro energético, solar, energías renovables, energía/s, consumo energético, térmica	(2 términos) Eléctrico, domótica
http://www.ine.es	(6 términos) Revista digital, estadísticas, encuesta, métodos estándares, censo electoral, índice	- [No se encuentran términos en el rango especificado] -
http://www.igme.es	(16 términos) Topoiberia, topografía, iberia, minero, instituto minero, geológico, topography, geología, geofísica, geodesia, fenómenos atmosféricos, surficial, geophysics, geology, geodesy, igme	- [No se encuentran términos en el rango especificado] -
http://www.chminosil.es	(1 término) CHMS	(3 términos) MAGRAMA, inundación, hídrico
http://www.csic.es	(3 términos) Fomento de la cultura, RJB, PCO, científica/o	(3 términos) MIdA, education, astronomy
http://www.ciuden.es	- [No se encuentran términos en el rango especificado] -	- [No se encuentran términos en el rango especificado] -

Meta-etiquetas		
Dominio	Términos relevantes con meta-etiquetas - Rango 2-25	Términos relevantes con meta-etiquetas - Rango 1 Hápax
http://www.oag-fundacion.org	(1 término) Canarias	- [No se encuentran términos en el rango especificado] -
http://www.oepm.es	- [No se encuentran términos en el rango especificado] -	- [No se encuentran términos en el rango especificado] -
http://www.chguadiana.es	- [No se encuentran términos en el rango especificado] -	- [No se encuentran términos en el rango especificado] -
http://www.chcantabrico.es	- [No se encuentran términos en el rango especificado] -	- [No se encuentran términos en el rango especificado] -
http://www.ieo.es	- [No se encuentran términos en el rango especificado] -	- [No se encuentran términos en el rango especificado] -
http://www.i2c2.org	(2 términos) I2C2, cambio climático	- [No se encuentran términos en el rango especificado] -
http://www.salvamentomaritimo.es	(2 términos) Salvamento, marítimo	- [No se encuentran términos en el rango especificado] -
http://www.ciemat.es	- [No se encuentran términos en el rango especificado] -	- [No se encuentran términos en el rango especificado] -
http://www.chguadalquivir.es	- [No se encuentran términos en el rango especificado] -	- [No se encuentran términos en el rango especificado] -
http://www.catastro.meh.es	- [No se encuentran términos en el rango especificado] -	- [No se encuentran términos en el rango especificado] -

Tabla 11. Selección de términos más relevantes con rango de frecuencias medias de aparición (de 2 a 25 iteraciones y Hápax) (II)

Metadatos Dublin Core		
Dominio	Términos relevantes con metadatos Dublin Core	Términos relevantes con metadatos Dublin Core - Rango Hápax
http://www.idae.es	(11 términos) Energía/s, renovable/s, solar, térmica, eólica, sostenible, biomasa, mareas, marinas, osmótica, domótica	(13 términos) Turismo, primas, OFE, protocolo, prevención, molino, montaje, limpieza, industria, IFEMA, factura eléctrica, firma digital, biodiesel
http://www.irmc.es	(2 términos) Comarcas mineras, minería	(2 términos) Cuencas, carbón
http://www.salvamentomaritimo.es	- [No se encuentran términos en el rango especificado] -	- [No se encuentran términos en el rango especificado] -
http://www.aemet.es	(18 términos) Vigilancia meteorológica, balance hídrico, modelo climático, climático, temperaturas máximas, temperaturas mínimas, playas, calidad del aire, clima, agua, sequía, índice ultravioleta, humedad, medio ambiente, vapor de agua, seco, meteorología, meteomet	(49 términos) Viento, verano, vegetación, troposfera, ultramar, tropopausa, tormentoso, tormentas, terrestre, temporada, surface, sondeos, sinóptica, riesgos, río, regionales, residuales, rayos, radiactivo, radiation, quart, protección, prevención, presión, peninsular, oxidación, nubes, nordeste, monte/s, marítimas, mar, litoral, laguna, isla, fotómetros, fenómenos físicos, exposición, energía, energético, emisión, emission, eléctrica, depuración, climate, alergias, aguas, aerosoles, aeródromos, acumulada

Metadatos Dublin Core		
Dominio	Términos relevantes con metadatos Dublin Core	Términos relevantes con metadatos Dublin Core - Rango Hápax
http://www.cne.es	(14 términos) Gas, gas natural, energía, energía eléctrica, industria, reforma eléctrica, regulación energética, suministro energético, glp, gasista, petróleo, MIBEL, ley del sector eléctrico, propano	(12 términos) Vigilancia, turismo, sostenible, renovables, nuclear, minas, marina, logístico, gasistas, electra, CNE, CLH

Al analizar en profundidad la columna de términos Hápax, se observan resultados variables en el número de términos relevantes. Los casos más notables corresponden al MAGRAMA y a la Agencia Estatal de Meteorología, ya que se detecta un mayor número de términos relevantes Hápax que términos relevantes de frecuencias medias (con una diferencia mínima de 30 términos). Esto es debido al vocabulario específico del sector de la meteorología, de la agricultura y alimentación, tratado en páginas y contenidos muy concretos del sitio web, lo que explica su bajo número de repeticiones y por ello una frecuencia de aparición baja. No obstante y a pesar de que no tienen un factor representativo del conjunto de contenidos del sitio web, sí se observa un claro poder discriminatorio que facilita la recuperación de información temática más especializada como por ejemplo: enfermedades del ganado bovino, regulación legal de diversos alimentos, el indicador de impacto ambiental de la huella ecológica, los índices de radiación ultravioleta, las estaciones de fotometría y pluviometría, entre otros. En esta línea, se detectan diferencias significativas en cuanto al número de Hápax relevantes para meta-etiquetas y metadatos Dublin Core en la Agencia Estatal de Meteorología. Si bien el número de términos relevantes empleados para la meta-descripción mediante meta-etiquetas es muy parejo (32 - 37 términos), con metadatos Dublin Core se obtienen aproximadamente (18 - 49 términos). Esto indica que los textos de meta-descripción son originales y distintos para cada caso, aunque compartan algunos términos. Además, ello proporciona puntos de acceso diferentes, lo cual multiplica las posibilidades de localizar la información en un motor de búsqueda.

Finalmente, se observa que los dominios que mejor uso hacen de los metadatos Dublin Core y meta-etiquetas son precisamente los que más términos relevantes tienen. Por el contrario, los sitios web que repiten cientos de veces el mismo término a lo largo de todas las meta-etiquetas son los que menos términos relevantes tienen, o bien no se encuentran en el rango especificado de frecuencias medias.

4. Recomendaciones y conclusiones

1. Se recomienda el empleo de metadatos Dublin Core para la meta-descripción exhaustiva de los contenidos de cada sitio web. Si bien las meta-etiquetas resultan más comunes, éstas no proporcionan una descripción completa al carecer de suficientes campos. Dublin Core en su estándar básico consta de 15 campos, pero en su versión extendida alcanza los 57, lo que significa que muchos más aspectos pueden ser sometidos al análisis documental.
2. Se requiere un vocabulario más completo y exhaustivo, evitando repetir sistemáticamente los mismos términos más de 25 veces si no es necesario o está justificado. De esta forma, se evitará la penalización en la recuperación de información y se mejorará la precisión de las búsquedas realizadas por los usuarios en los principales motores de búsqueda.
3. Diseñar páginas demasiado jerarquizadas dificulta la labor de penetración y análisis de los webcrawler de cualquier buscador, ya que tardarán más tiempo en encontrar la base de contenidos que sustenta el sitio web. Esto significa que deben rediseñarse para obtener una estructura más horizontal, cuyos contenidos estén como máximo a 3 niveles de enlazamiento. Esta recomendación, lejos de ser aleatoria, define un modelo de web a seguir en la que su portada representa el primer nivel de análisis, las páginas de las secciones constituyen el

segundo nivel y de éstas dependen las páginas de contenidos en el tercer nivel.

4. Después de analizar la web de la administración central española medioambiental, se concluye que contiene 1,5 millones de enlaces (de los que 675.000 son únicos) y, por ende, el crecimiento de la muestra, nivel por nivel, es exponencial.

5. Los tres dominios más frecuentes entre todos los enlaces analizados son ".es" (676.449 enlaces), ".gob.es" (4.761 enlaces) y ".com" (1.129 enlaces). Sin embargo, realizando un análisis en profundidad sobre los dominios ".eu" se observa que, aun siendo cuantitativamente inferiores (460 enlaces), son de gran relevancia debido a la relación intrínseca de las normativas, directivas y el derecho europeo en materia de medio ambiente, así como en investigación de nuevas energías y centros de investigación con respecto a las principales instituciones españolas en la materia, quedando desentrañadas todas sus vinculaciones, tal como queda reflejado en los datos de la Tabla 5 y en el mapa de relaciones expuesto en la Figura 2.

6. La web medioambiental de la administración central española tiene un alto nivel de dinamismo, ya que sólo un 30% de la web es estática. Por otra parte, derivado del análisis de los formatos de archivo, se han obtenido más de 700 canales de sindicación que podrán ser utilizados para el seguimiento de la información publicada por la administración en relación a temas medioambientales.

7. Las páginas web con más meta-etiquetas y metadatos no siempre son las que constan de términos más relevantes, incluso detectándose malas prácticas al repetir cientos de veces un número limitado de términos. Lejos de describir los contenidos, esto dificulta la recuperación y hace inservibles las metadescripciones durante los procesos de indexación.

8. Se comprueba que los términos cuya frecuencia de aparición media se encuentran en un rango de entre 2 y 25 repeticiones, suelen tener una mayor representatividad y relevancia en la colección de documentos a la que pertenecen, demostrando nuevamente las afirmaciones de Hans Peter Luhn (1958).

9. Puede afirmarse que los dos sitios web que mejor cumplen el paradigma de meta-descripción de los contenidos son la Agencia Estatal de Meteorología (AEMET) y el Instituto para la Diversificación y Ahorro de la Energía (IDAE), ya que incorporan simultáneamente no sólo meta-etiquetas sino también metadatos Dublin Core.

5. Agradecimientos

Este trabajo es parte de un proyecto de investigación titulado "Organización del acceso, uso y reutilización de la información del sector público en España. Hacia la consolidación de una industria de la información", dirigido por Dr. L. Fernando Ramos Simón (Universidad Complutense de Madrid) y financiado por Plan Nacional de I+D en España (Ref.: CSO2010-17451).

Así mismo, este trabajo forma parte de las actividades del proyecto "Buenas prácticas en el acceso a la información gubernamental" (PAPIIT IN403113-3. Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica), dirigido por Dr. Egbert Sánchez Vanderkast (Instituto de Investigaciones Bibliotecológicas y de la Información. Universidad Nacional Autónoma de México (UNAM)) y apoyado por la Dirección General de Apoyo Académico de la UNAM.

Por último, nos gustaría agradecer el apoyo y la beca concedida en el programa de "Formación del Profesorado Universitario" por el Ministerio de Educación, Cultura y Deporte (España).

6. Referencias

1. BERGMARK, D.; LAGOZE, C.; SBITYAKOV, A. (2002). Focused crawls, tunneling, and digital libraries. En: Proceedings of the Sixth European Conference on Digital Libraries, (Rome, September 16–18), 91–106. Disponible en: http://link.springer.com/content/pdf/10.1007%2F3-540-45747-X.pdf [Consulta: 21 de octubre de 2013].

2. BERNERS-LEE, T. (1995). Hypertext Markup Language - 2.0, RFC 1866, Network Working Group. Disponible en: http://tools.ietf.org/html/rfc1866 [Consulta: 21 de octubre de 2013].

3. BLÁZQUEZ OCHANDO, M.; SERRANO MASCARAQUE, E. (2011). Análisis de la web y usabilidad:

prueba de funcionamiento de Mbot webcrawler. En: X Congreso del Capítulo español de ISKO (La Coruña, 30 junio – 1 julio). Disponible en: http://eprints.rclis.org/19104/ [Consulta: 21 de octubre de 2013].

4. BLÁZQUEZ OCHANDO, M. (2013a). "Desarrollo tecnológico y documental del webcrawler Mbot: prueba de análisis web de la universidad española". En: *XIII Jornadas Españolas de Documentación*, Fesabid, (Toledo, 21-24 mayo).

5. BLÁZQUEZ OCHANDO, M. (2013b). Mbot – Webcrawler multipropósito. Disponible en: http://mblazquez.es/mbot/ [Consulta: 21 de octubre de 2013].

6. CHAKRABARTI, S.; JOSHI, M.M.; PUNEA, K.; PENNOCK, D.M. (2002). The structure of broad topics on the Web. En: *Proceedings of the 11th World Wide Web Conference*, (Honolulu, Hawaii, May 7–11). 508-516. Disponible en: http://www.cse.iitb.ac.in/soumen/doc/www2002t/p338-chakrabarti.pdf [Consulta: 21 de octubre de 2013].

7. COTHEY, V. (2004). Web-crawling reliability. Journal of the American Society for Information Science and Technology. 55(14), 1228–1238. Disponible en: http://onlinelibrary.wiley.com/doi/10.1002/asi.20078/pdf [doi:10.1002/asi.20078] [Consulta: 21 de octubre de 2013].

8. DCMI. (2012). DCMI Metadata Terms. Disponible en: http://dublincore.org/documents/dcmi-terms/ [Consulta: 21 de octubre de 2013].

9. GANSNER, E.R. (2012). Drawing graphs with Graphviz. Disponible en: http://www.graphviz.org/doc/oldlibguide.pdf [Consulta: 21 de octubre de 2013].

10. GIL LEIVA, I. (2008). *Manual de Indización: teoría y práctica*. Gijón: Trea. 67-69.

11. GRAELLS, E.; BAEZA YATES, R. (2007). Características de la Web Chilena.

12. HENZINGER, M.R. (2003). Algorithmic challenges in Web search engines. *Internet Mathematics*, 1(1), 115–126. Disponible en: http://www.internetmathematics.org/volumes/1/1/pp115_123.pdf [Consulta: 21 de octubre de 2013].

13. LUHN, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*, 2(2), 159-165.

14. THELWALL, M. (2001). A web crawler design for data mining. *Journal of Information Science*, 27(5), 319-325. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.830&rep=rep1&type=pdf [doi: 10.1177/016555150102700503] [Consulta: 21 de octubre de 2013].

15. W3C. (1999). HTML 4.01 Specification: The global structure of an HTML document. Meta data. Disponible en: http://www.w3.org/TR/REC-html40/struct/global.html#h-7.4.4 [Consulta: 21 de octubre de 2013].