El uso de metadatos en las webs universitarias españolas

José L. Alonso Berrocal
Carlos G. Figuerola
Ángel F. Zazo
Universidad de Salamanca (España)

Resumen

El objetivo es mostrar la utilización real de metadatos y los que se usan de todas las universidades españolas. Emplearemos el robot SACARINO (presentado en Ibersid 2005) para la recogida de todos los datos y a partir de ahí procesaremos la información para poder obtener los resultados. Se mostrarán los resultados de forma gráfica, lo que facilitará una rápida comprensión de los mismos.

Palabras clave: Recuperación de información. Cibermetría. Metadatos. Webs. Universidades. España.

Abstract

The objective is to show the real use of metadata and which ones are used by all the Spanish universities. We will use the robot SACARINO (that was presented in Ibersid 2005) for the collection of all the data and then we will process the information in order to obtain the results, that will be shown in a graphical way so that they are easy to understand.

Keywords: Information retrieval. Cybermetrics. Metadata. Webs. Universities. Spain.

1. Introducción

El trabajo planteado en este artículo es un avance de los resultados obtenidos en el proyecto de investigación del MEC del programa de Estudios y Análisis y con referencia EA2006-0080. Inicialmente se planteó como un póster a presentar en el congreso Ibersid del 2006 y finalmente se reconvirtió en una comunicación al mismo.

La recogida de información que realizamos para dicho proyecto nos ha permitido estudiar la situación y el empleo de las etiquetas META en las universidades españolas. Tratamos de analizar exactamente cuál es su utilización real, analizando además los diferentes tipos de etiquetas que se pueden emplear. También

nos centraremos en la utilización de la especificación Dublin Core, para comprobar su extensión. Mostraremos aquí los datos más generales y las conclusiones preliminares de dicho estudio.

2. Metodología

En el proceso de recogida de la información hemos empleado el bot SACARINO, presentado en Ibersid 2005, que nos permite realizar la fase de *crawling* de forma plenamente satisfactoria, empleando técnicas de recogida autónoma de la información. Hemos obtenido la información de todas las universidades españolas valiéndonos de la posibilidad de utilizar diferentes semillas e intentando así optimizar la recogida de datos.

De las 73 universidades ha sido posible obtener la información de todas excepto de la UPSA (Universidad Pontifica de Salamanca), donde no se permitió el acceso a su información por parte de nuestro bot.

Algunos trabajos han analizado también las etiquetas META (Merlo y Sorli, 2000; Vidal y Salvador, 2000), pero se centraban solamente en las páginas de las bibliotecas de diferentes centros, empleando técnicas manuales en algunos de los casos y con un tamaño de muestra de unos pocos de cientos de páginas.

En nuestro caso se ha recorrido de forma automática toda la web de cada universidad y el total de páginas ha sido de 4.200.000, de modo que los resultados a los que llegamos pueden ser más extrapolables a la situación real.

En relación con los metadatos, un buen trabajo a consultar es el de Méndez (2002), que los define como "destinados a ordenar y describir la información contenida en un documento entendido como objeto, de tal forma que se erigen como reveladores tanto de la descripción formal, como del análisis de contenido, en aras a mejorar el acceso a esos objetos de información de la Red". Lo que parece claro en la actualidad es que los metadatos tienen una función tanto identificadora como descriptiva, el contexto de la red y la posibilidad de interpretación por máquina.

Otros trabajos interesantes son los de Caplan (2003) y Hillmann y Westbrooks (2004). Algunas direcciones web que pueden aportar informaciones relevantes son DCMI (2003a, 2003b, 2003c), IETF (1999), Normaweb (2001), W3C (1999, 2006).

De los diferentes tipos de metadatos —encontramos una interesante tipología en Dempsey y Heery (1998)— que se pueden implementar —DC/DCMI, TEI, RDF, etcétera—, nosotros nos vamos a centrar exclusivamente en el empleo de las denominadas *etiquetas META*, codificadas bajo HTML. Estas etiquetas van ubicadas en la cabecera del documento HTML de la siguiente forma:

```
<HEAD>
<META http-equiv="Content-Type" content="text/html;
charset=utf-8">
<META name="AUTHOR" content="José Luis Alonso Berrocal">
<LINK rel="StyleSheet" href="../personal.css"
type="text/css" media="all">
<TITLE> Grupo REINA: Miembros: Página personal de
José Luis Alonso Berrocal</TITLE>
</HEAD>
```

Las etiquetas META requieren, en principio, de una intencionalidad por parte del editor, para poder incorporarlas al documento. Si bien esto es cierto para la mayoría de estas etiquetas, en la actualidad muchos programas editores de HTML incorporan de forma automática algunas de ellas, eliminando de esta forma la intencionalidad, si bien, en la mayoría de los casos, se trata de etiquetas relacionadas con el propio editor HTML. Las etiquetas META tienen dos atributos principales:

1. *http-equiv*: pensado para recabar información de los encabezados del protocolo http, aunque se le han añadido otros usos como el de los PICS (*Platform for Internet Content Selection*) con el fin de valorar el contenido de las páginas.
2. *name*: identifica los tipos de propiedades y aparece junto a la opción content para definir el valor de la propiedad:

   ```
   <META name="propiedad" content="valor">
   ```

La lista de propiedades para el atributo *name* es muy extensa y además la Dublin Core Metadata Initiative (http://dublincore.org/) se implementa también sobre la etiqueta META, además de hacerlo sobre modelos RDF/XML. Algunos ejemplos de propiedades son los siguientes:

- *Description*. Algunos motores de búsqueda incluyen esta información junto con los resultados de la búsqueda, por lo que para que sea realmente útil debería contener la mejor descripción posible del documento. Por ejemplo:

  ```
  <meta name="description" contents="Este documento
  trata sobre los elementos META (metatags). Explica
  la diferencia entre los elementos que especifican
  el atributo http-equiv y el atributo name, y hace
  un repaso de los más utilizados." lang="es">
  ```

- *Keywords*. Algunos buscadores utilizan este elemento para clasificar los documentos por palabras clave. Existen multitud de tutoriales que explican cómo especificar palabras clave para lograr mejores posiciones en los

resultados de los buscadores. Muchos de estos métodos constituyen una especie de "fraude", lo cual ha provocado que su uso produzca resultados inútiles desde el punto de vista del usuario. Ejemplo:

```
<meta name="keywords" contents="metatags,meta
name,meta http-equiv">
```

No está claro si las palabras deben separarse por comas, espacios, o coma y espacio. Existen diversas teorías al respecto.

- *Autor*. El autor del documento es la persona con la que querríamos ponernos en contacto por algún asunto relacionado con el documento.
- *Copyright*. Información sobre el copyright del documento. Por ejemplo:

```
<meta name="copyright" content="&copy; 2006, José
Luis Alonso Berrocal" lang="es">
```

- *Robot*. Controla la acción de los motores indizadores sobre el documento. En general puede tomar uno o más de los siguientes valores: INDEX, NOINDEX, FOLLOW, NOFOLLOW. Además, ALL equivale a INDEX, FOLLOW y NONE equivale a NOINDEX, NOFOLLOW. Por ejemplo:

```
<meta name="robots" content="NOINDEX, NOFOLLOW">
```

Los robots no pueden indexar esta página y no pueden recorrerla para buscar nuevos enlaces.

```
<meta name="robots" content="NOINDEX, FOLLOW">
```

Los robots no pueden indexar esta página pero sí recorrerla para buscar nuevos enlaces.

```
<meta name="robots" content="ALL">
```

Los robots pueden indexar esta página y recorrerla para buscar nuevos enlaces.

- *Dublin Core*. Es un estándar de metadatos que define un conjunto de propiedades recomendadas para descripciones bibliográficas electrónicas, y su objetivo es promover la interoperabilidad entre modelos descriptivos dispares. Estas propiedades incluyen, entre otras, título, autor, descripción, tema, editor, tipo de recurso, fecha de publicación, idioma, formato, documentos relacionados, derechos de autor, etcétera. Por ejemplo:

```
<meta name="DC.Title" content="Ibersid 2006 -
Metadatos">
```

```
<meta name="DC.Creator" content="José Luis ALonso
Berrocal">
<meta name="DC.Subject" content="Metadatos">
<meta name="DC.Description" content="Los
elementos META de HTML: tipos, funciones, etc.">
<meta name="DC.Publisher"
content="Berrocal.reina">
<meta name="DC.Date" content="2006-10-26">
<meta name="DC.Type" scheme="DCMIType"
content="Text">
<meta name="DC.Format.Medium"
content="text/html">
<meta name="DC.Relation.isPartOf"
content="http://reina.usal.es/">
<meta name="DC.Identifier"
content="http://reina.usal.es/articulos/metadatos">
<meta name="DC.Language" content="es">
<meta name="DC.Rights" content="(c) 2006 by JOsé
Luis Alonso Berrocal. All rights reserved.">
```

Puede encontrarse toda la información en el sitio web de Dublin Core (http://dublincore.org).

3. Resultados

Los resultados son un avance de la investigación en marcha; nos centraremos en los aspectos más relevantes. El empleo de etiquetas META en las universidades españolas es el indicado en la figura 1.

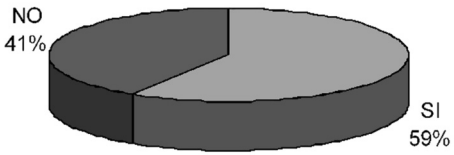

Figura 1. Porcentaje de empleo de las etiquetas META

Del porcentaje de páginas que emplean metas, en la figura 2 podemos ver el reparto que se produce, atendiendo a los grandes tipos de etiquetas meta disponibles, que como podemos ver se centra en el atributo *http-equiv*.

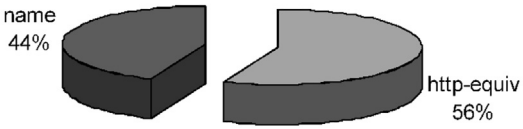

Figura 2. Reparto de atributos meta

En cuanto al número de etiquetas META existente en cada página (figuras 3 y 4), el 50% se lo reparten las páginas con 1 y 3 etiquetas. Apenas un 15% del total tienen más de 5 etiquetas por página.

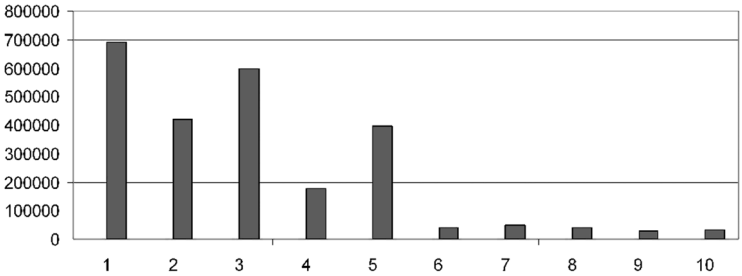

Figura 3. Número de etiquetas por página

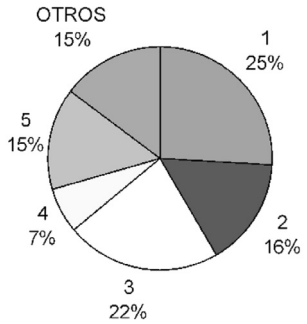

Figura 4. Porcentaje del número de etiquetas por página

En cuanto al número de etiquetas META existente en cada página para el atributo *http-equiv* (figuras 5 y 6), el mayor porcentaje es para las páginas con 1 etiqueta, destacado de forma clara.

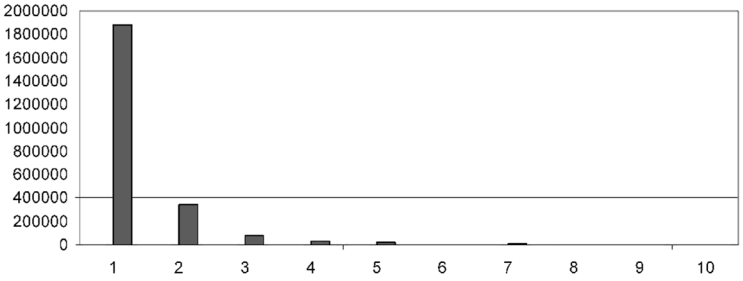

Figura 5. Número de etiquetas http-equiv

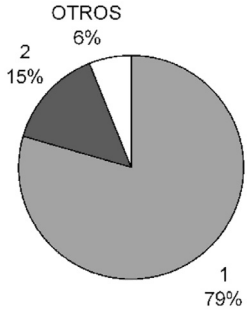

Figura 6. Porcentaje de etiquetas http-equiv

En cuanto al número de etiquetas META existente en cada página para el atributo *name* (figuras 7 y 8), el mayor porcentaje es para las páginas con 2 etiquetas, que junto a las de 1 etiqueta superan el 50%.

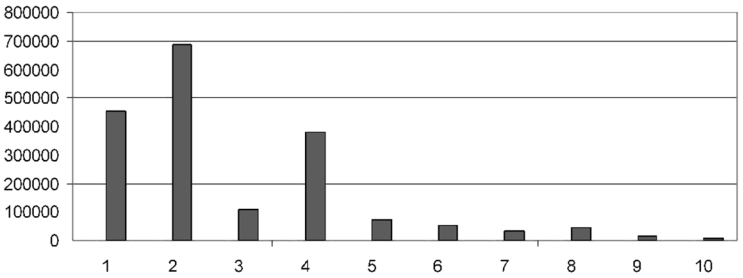

Figura 7. Número de etiquetas name

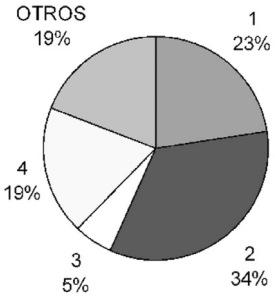

Figura 8. Porcentaje de etiquetas name

Centrándonos en las opciones más utilizadas del atributo *http-equiv* (figura 9), vemos claramente que la más utilizada es la *content-type*. Podemos observar que se emplean algunas opciones que no pertenecen al atributo *http-equiv*, posiblemente por desconocimiento del adecuado empleo. Hay que señalar, aunque en porcentajes mínimos, *DC.description*, *title*, *description*, *keyword*.

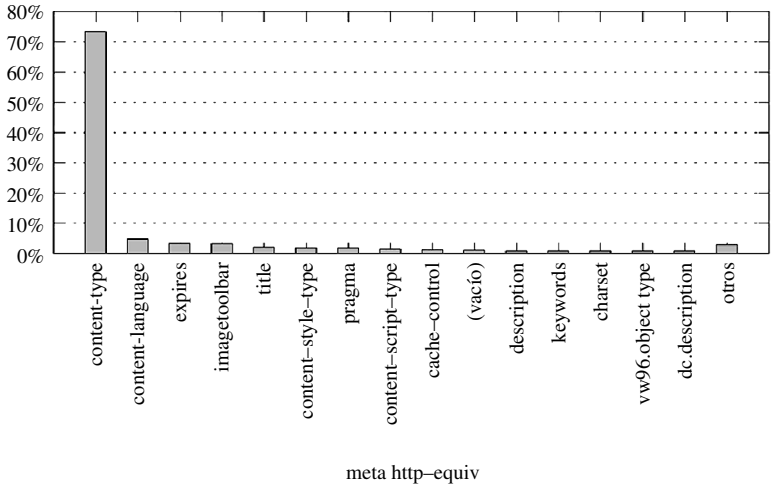

Figura 9. Opciones del atributo http-equiv

Respecto a las opciones más utilizadas del atributo *name* (figura 10), la más utilizada es la opción *generator* (que hace referencia al programa editor empleado). Debemos tener en cuenta que muchos programas editores de páginas HTML generan esta etiqueta y esta opción de forma automática, sin intervención del usuario y sin requerir de este una validación de la misma. Le siguen *keywords* y *des-*

cription. Hay que destacar el mínimo empleo de la especificación *Dublin Core*. El elemento *otros* recoge todas las posibilidades empleadas, sumando un porcentaje amplio, que indica el empleo de una gran variedad de opciones.

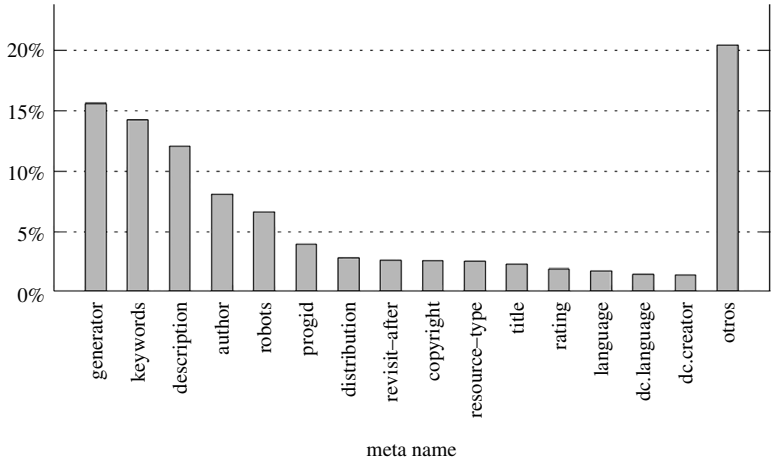

Figura 10. Opciones de la etiqueta name

Centrándonos en las opciones más utilizadas de la especificación *Dublin Core* (figura 11), dentro del escaso empleo de las mismas, destaca *DC.language* y *DC.creator*.

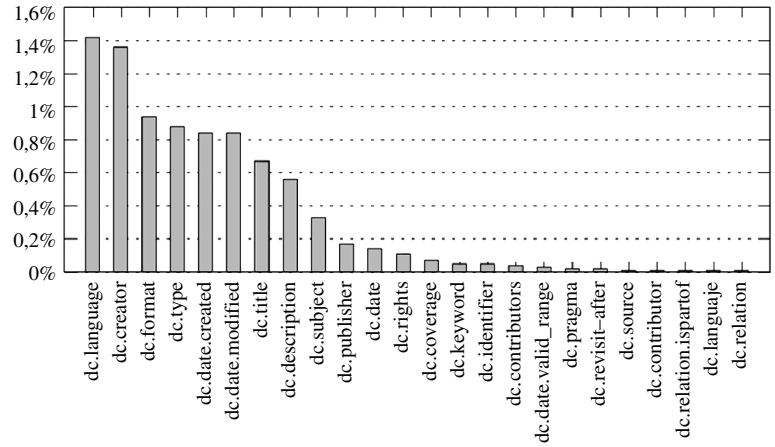

Figura 11. Opciones Dublin Core *más utilizadas*

4. Conclusiones

1. El 60% de las páginas web de las universidades españolas tienen etiquetas META y de forma mayoritaria se emplea el atributo *http-equiv*.
2. En el atributo *name* la variedad de opciones es enorme y con una distribución muy irregular.
3. Hemos podido observar la escasa implantación de la especificación DC, lo que implica una enorme dificultad para aplicar técnicas de recuperación de información basadas solamente en metadatos.
4. Las etiquetas META pueden ser un buen complemento a las técnicas de recuperación de información clásicas, basadas en el modelo vectorial; pueden aportar matices interesantes, como nuestro grupo de investigación ya ha demostrado (Figuerola *et al.*, 2006).

5. Trabajo de futuro

Muchos son los aspectos que quedan por valorar, como la distribución por universidades, distinguiendo entre universidades públicas y privadas. También debemos analizar la posible implementación de otros tipos de metadatos, destacando el modelo RDF bajo XML.

También deseamos realizar un estudio de algunas de las opciones del atributo *name*, como *title*, *keywords*, etcétera, y si se ajustan a los contenidos.

Referencias

Caplan, P. (2003). Metadata fundamentals for all librarians. Chicago: American Library Association, 2003.

DCMI (2003a). Expressing Dublin Core in HTML/XHTML meta and link elements. URL: <http://dublincore.org/documents/dcq-html/>. Consultado: 2006-10-25.

DCMI (2003b). Guidelines for implementing Dublin Core in XML. URL: <http://dublincore.org/documents/dc-xml-guidelines/>. Consultado: 2006-10-25.

DCMI (2003c). Foro español del Dublin Core Metadata Initiative. URL: <http://listserv.rediris.es/archives/dcmi-es.html>. Consultado: 2006-10-25.

Dempsey, L.; Heery, R. (1998). Metadata: A current view of practice and issues. // Journal of Documentation. 54:2 (1998) 145-172.

Figuerola, C. G.; Alonso Berrocal, J. L.; Zazo, A. F.; Rodríguez Vázquez de Aldana, E. (2006). Web page retrieval by combining evidence. // Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers. Lecture Notes in Computer Science. 4022 (2006) 880-887.

Hillmann, D. I.; Westbrooks, E. L. (2004). Metadata in practice. Chicago: American Library Association, 2004.

IETF (1999). Encoding Dublin Core Metadata in HTML, Internet RFC 2731. URL: <http://www.ietf.org/rfc/rfc2731.txt>. Consultado: 2006-10-25.

Merlo Vega, J. A.; Sorli Rojo, A (2000). El uso de metainformación en las webs de las bibliotecas españolas. // Jornadas Españolas de Documentación (7. 2000. Bilbao). La gestión del conocimiento: retos y soluciones de los profesionales de la información. Bilbao: Universidad del País Vasco, 2000. 154-164.

Méndez Rodríguez, E. (2002). Metadatos y recuperación de información: estándares, problemas y aplicabilidad en bibliotecas digitales. Gijón: TREA, 2002.

Normaweb (2001). Grupo de trabajo del SEDIC sobre Normalización para la Recuperación de Información en Internet. URL: <http://www.sedic.es/gt_normalizacion.htm>. Consultado: 2006-10-25.

Vidal Bordés, F. J.; Salvador Oliván, J. A. (2000). La implementación de metadatos y Dublin Core en sedes y páginas de bibliotecas y centros de documentación de universidades y centros de investigación de la red iris. // Jornadas Españolas de Documentación (7. 2000. Bilbao). La gestión del conocimiento: retos y soluciones de los profesionales de la información. Bilbao: Universidad del País Vasco, 2000. 197-209.

W3C (1999). Resource Description Framework (RDF). URL: <http://www.w3.org/RDF/>. Consultado: 2006-10-25.

W3C (2006). Extensible Markup Language. URL: <http://www.w3.org/TR/REC-xml/>. Consultado: 2006-10-25.