IX CONGRESO ISKO-ESPAÑA (Valencia, 11-13 de marzo de 2009)
Nuevas perspectivas para la difusión y organización del conocimiento
Actas del Congreso. Valencia: UPV, vol 2, pp. 832-845.

ETIQUETADO LIBRE FRENTE A LENGUAJES DOCUMENTALES. APORTACIONES EN EL ÁMBITO DE BIBLIOTECONOMÍA Y DOCUMENTACIÓN.

Luis Rodríguez Yunta
CSIC, Centro de Ciencias Humanas y Sociales, Unidad de Análisis Documental y Producción de Bases de Datos ISOC, Madrid, España. E-mail: luis.ryunta@cchs.csic.es

Resumen
Se analiza el uso de etiquetas o tags en los blogs, servicios de promoción social de noticias y marcadores sociales, por parte de profesionales del campo de la Biblioteconomía y la Documentación. Los datos obtenidos en varios ejemplos de este tipo de recursos se comparan con los términos incluidos en un tesauro especializado de este mismo ámbito. A través de esta comparación se sistematizan cuáles son las aportaciones del etiquetado libre frente a las herramientas tradicionales. Los tesauros pueden aprovechar las folksonomías como una fuente de gran calidad para actualizar y ampliar su cobertura terminológica. Este objetivo debe considerarse prioritario si se quiere que los tesauros puedan representar un papel en el futuro inmediato en los sistemas de recuperación sobre texto completo.

Abstract
This communication analyzes the use of tagging on sites such as blogs, digging services and social bookmarkings, by library and information science professionals. The analysis results of this type of resources are compared with the terms included into a specialized thesaurus in this area. By this comparative study, the author systematizes which are the contributions of the free tagging in front of the traditional skills on content analysis. The thesauri can make good use of the folksonomies like a high quality source for updating and expanding the terminological scope. This aim must be considered a priority if a central role for thesauri in the full-text search systems is pretended in an immediate future.

Palabras clave
Etiquetado libre, folksonomías, tesauros, indización, web social.

1. Introducción

El tagging o etiquetado libre se está imponiendo como una nueva modalidad de indización en lenguaje natural, especialmente a través de las herramientas y recursos de la web social. Se utiliza de forma habitual, en especial en las denominadas herramientas de la web social. Así, se pueden localizar fácilmente ejemplos del uso del tagging en sitios web muy diversos, como las herramientas para compartir fotografías (*Flickr*, http://www.flickr.com/), bitácoras o blogs (*Comunicación cultural*, http://www.comunicacion-cultural.com/), servicios de promoción de noticias (*Meneame*, http://meneame.net/) o sistemas colaborativos de marcadores (*Delicious*, http://delicious.com/).

Folksonomía, tagging o etiquetado son conceptos ya recogidos en la literatura profesional. Generalmente se presentan como alternativas novedosas para la organización y clasificación de la información, en el contexto de estas nuevas herramientas colaborativas de uso creciente en la web. La práctica del tagging es una de las características que definen el concepto de web 2.0 o web social, en la que el usuario se ha transformado de consumidor pasivo en un activo "prosumidor" de información. Este neologismo, "prosumidor", aún cuando no está aún aceptado por la RAE, parece necesario para hacer referencia a las personas que son productoras y consumidoras de un mismo producto. Y por ello se identifica este fenómeno con una real democratización de la información y el conocimiento (Rodríguez Palchevich, 2008).

Yusef Hassan distingue entre dos posibles usos del tagging, uno de carácter social, otro de sentido personal. En su opinión sólo puede hablarse de un modelo novedoso de indización cuando se práctica la "indización social agregada", es decir, cuando varios usuarios indizan un mismo recurso (Hassan, 2006). Otros autores utilizan el concepto de "etiquetado colaborativo" (McGregor y McCulloch, 2006), o reclaman una diferenciación entre el tagging de autor y el tagging realizado por los lectores (Seoane, 2007a).

Así como el concepto de etiquetado se relaciona con la indización, el de folksonomía se presenta como una alternativa entre los lenguajes documentales. Puede definirse como una clasificación social que se genera por consenso a través de las aportaciones de los usuarios (Wright, 2004). Pero su naturaleza es radicalmente distinta a los lenguajes documentales tradicionales. Clasificaciones, taxonomías o tesauros son lenguajes creados por expertos, para ser utilizados posteriormente en un sistema de información. Por el contrario, una folksonomía se crea por agregación de información sin ningún punto de partida previo, y por ello se puede interpretar como reflejo de un poder popular (Quintarelli, 2005).

Este renacimiento de los lenguajes no controlados también ha recibido críticas. Se cuestiona incluso el interés de la denominada web 2.0:

> "Amateurismo y charlatanería conviven en la escritura colaborativa de la Web 2.0. Si bien se trata de herramientas de alta productividad para formar comunidades, en muchos casos no aportan calidad a nivel de contenidos, sólo experiencias de producción no-profesional, poco fiables". (Pardo, 2007)

El etiquetado hereda todos los problemas tradicionales de los vocabularios no controlados (Ros, 2008). Es una forma desestructurada de aplicar metadatos para describir recursos o documentos en la web. Este sistema implica grandes limitaciones para la recuperación de información: carencias de precisión, sin control de sinónimos, ausencia de estructura jerárquica, baja tasa de recuperabilidad,… (Quintarelli, 2005). En definitiva, supone la apuesta por un sistema de recuperación basado en la serendipia (serindipity), muy lejos del intento de construcción de sistemas que aseguren cierto equilibrio entre

exhaustividad y pertinencia. El éxito de la serendipia es un hecho constatable en la generalización del uso de buscadores tipo Google frente a los directorios tipo Yahoo! (Seoane, 2007b).

Pero además existe una limitación propia que sugiere cierta improvisación en su sistema de recuperación: la sintaxis de las etiquetas o tags se limita en muchos casos a unitérminos, sin que exista una forma única para presentar entradas compuestas por más de una palabra. Frecuentemente los sistemas de búsqueda a través de etiquetas utilizan la nube de etiquetas o "tags cloud", como un recurso visual, en el que el tamaño de letra se relaciona con la frecuencia de utilización en la asignación de entradas. Por ello, la falta de normas para la construcción de términos compuestos tiene un efecto negativo para la fiabilidad de esta modalidad de recuperación, un posible tema de búsqueda que haya recibido varias "expresiones" diferentes, no obtendrá un lugar destacado en la nube de etiquetas. Otro elemento que puede sorprender a los profesionales, acostumbrados a la seriedad de los léxicos documentales, es la elevada presencia en las redes sociales de etiquetas de tipo afectivo o subjetivo (Kipp, 2008), la tendencia a utilizar entradas dirigidas a uno mismo o a amigos, e incluso el uso de spam.

Entre las ventajas, el etiquetado libre tendría las propias de todo sistema de indización en lenguaje natural: simplicidad, transparencia, establecimiento de pesos por popularidad y aparición inmediata de nuevos términos. También se señala su atractivo visual, su sentido lúdico y por supuesto, el añadido de utilizar una economía de escala (Serrano, 2007). Pero un claro argumento para utilizar folksonomías es que resultan "mejor que nada", puesto que introducen un elemento para mejorar la capacidad de recuperación en un contexto en el que no es viable la aplicación de lenguajes controlados (Shirky, 2005).

Como soluciones, Yusef Hassan (2006) aboga por la aplicación de soluciones invisibles para el usuario final, como el empleo de "modelos propios de la indización automática sobre la indización social: ponderación mediante el empleo de las frecuencias de uso del tag, ponderación de los taggers por autoridad, desambiguación del significado en función del contexto, etc."

Los estudios relativos al uso del tagging dentro de la web semántica, se basan igualmente en herramientas que limiten la dispersión del vocabulario. Las propuestas se dirigen hacia las agrupaciones de etiquetas, con conceptos como key-tags (Catarino y Baptista, 2008) o la introducción de fórmulas de clustering en la visualización de tag-clouds (Hassan y Herrero, 2006). El etiquetado parece pues una opción con capacidad para consolidarse y puede contribuir incluso al desarrollo de la web semántica, puesto que facilita una red de términos que se autoalimenta de forma continua (Seoane, 2005 y 2007b).

Sin embargo, cabe reflexionar sobre si este modelo de indización mediante términos libres debe ser utilizado por los profesionales de la documentación, defensores tradicionales de los lenguajes controlados. ¿Significa una renuncia a las herramientas básicas de control de la terminología? ¿Son sus aportaciones a la recuperación de información también de interés en medios profesionales?

2. Objetivos

Con esta comunicación se pretende analizar cuáles son las aportaciones del etiquetado libre como herramienta de indización libre utilizada por profesionales de la documentación. Esta práctica se compara con las herramientas tradicionales de control de vocabulario.

Desde las Ciencias de la Documentación se ha mantenido una polémica tradicional entre el empleo de lenguajes controlados y las ventajas de los lenguajes libres. Los profesionales de la Biblioteconomía y Documentación han defendido generalmente la normalización de la terminología empleada en los campos de materias o descriptores. Control y normalización son señas de identidad que hasta el momento parecían irrenunciables. En la elaboración de productos documentales, la práctica de la asignación libre de entradas de materias, generalmente denominadas palabras clave, era realizada por los autores de los documentos. Las bases de datos bibliográficas habitualmente han venido trasformando estas palabras clave en términos controlados, tomando las elecciones de los propios autores como meras sugerencias, como un punto de partida que podía mantenerse o modificarse. Así, por ejemplo, un estudio comparativo entre las palabras clave de autor en revistas españolas y los descriptores asignados por los documentalistas de las bases de datos del CSIC, obtuvo unos resultados de similitud en torno al 60% (Gil y Alonso, 2005).

Con la traducción de los conceptos tratados por un documento a entradas extraídas de un tesauro o listado previo se pretende dar mayor consistencia y eficacia a las bases de datos. En este contexto, Lancaster (2002) marca un doble objetivo para el control del vocabulario:

- Facilitar la representación consistente de las materias por parte de indizadores y usuarios que recuperan, evitando la dispersión de los elementos relacionados.
- Facilitar la realización de una búsqueda amplia sobre una materia enlazando los términos con relaciones paradigmáticas o sintagmáticas.

No obstante, los sistemas con lenguaje natural ofrecen una ventaja sobre los sistemas que utilizan lenguaje controlado. El uso de un vocabulario ilimitado permite una gran especificidad en la recuperación (Lancaster, 2002). Con la generalización de la documentación en formato electrónico, su uso puede ser predominante:

> "Parece evidente que el lenguaje natural será la norma en la recuperación de información y que el uso de los vocabularios controlados convencionales disminuirá. Existen numerosas razones para ello, como los elevados costes del proceso intelectual humano, la rápida disminución de los costes de almacenamiento automatizado, el creciente volumen de texto que se encuentra accesible por ordenador (incluyendo el correo electrónico y el texto completo de revistas y periódicos), y la reducción gradual de la dependencia de intermediarios cualificados en la búsqueda online." (Lancaster, 2002, p. 188).

Moreiro utiliza el concepto de indización libre en contraposición a indización controlada, para referirse al modelo de asignación de términos de materias sin la existencia previa de un vocabulario de referencia que determine la forma unívoca de las entradas que pueden emplearse en un sistema (Moreiro, 2004). Se trata de una modalidad dentro de la "indización humana". En la indización libre se asignan entradas por extracción de conceptos explícitos presentes en el texto o por asignación de conceptos implícitos. Se caracteriza por la ausencia de control semántico, por lo que se produce una lista ilimitada de entradas. Pero el lenguaje natural de partida tiene ambigüedades, redundancias y presencia de conceptos implícitos que hacen difícil manejar sus expresiones. Su eficacia puede variar según el contexto; se argumenta que su rendimiento depende de la aplicación de "lenguajes muy exactos, los propios de las ciencias aplicadas y la tecnología, ya que su terminología es muy estable" (Moreiro, 2004, p. 145)

Como ventajas de la indización libre, Moreiro (2004) señala las siguientes:

- No se precisa inversión para construir lenguajes documentales.
- Son lenguajes evolutivos.
- Ofrecen una enorme riqueza de vocabulario.
- Son fácilmente automatizables, al trabajar sobre todo con unitérminos.
- Se obtienen resultados satisfactorios cuando se combinan con los términos propios de un entorno científico-técnico específico.

Los profesionales de la Documentación han sido tradicionalmente valedores del empleo de lenguajes controlados, por su consistencia en la representación sistemática del análisis documental de contenido y su capacidad para combinar búsquedas genéricas y específicas. Sin embargo, a partir de las herramientas de la web social, el etiquetado también está siendo aplicado por los propios documentalistas y bibliotecarios. Parece oportuno reflexionar sobre este hecho ¿Se trata de una renuncia a valores tradicionales de la disciplina? ¿O es una adaptación a las demandas de nuevas generaciones de usuarios? ¿Hay un cambio de paradigma en la percepción de las herramientas de recuperación?

La terminología empleada en el etiquetado social es muy versátil ya que puede referirse a la descripción del contenido pero también a aspectos subjetivos, atributos o elementos del contexto. Pero esta característica no la diferencia en realidad del modelo de indización aplicado en bases de datos documentales como los servicios de recuperación de fotografías comerciales. En este sentido, la práctica del etiquetado social, como el empleo de descriptores, no constituye un modelo único de indización. ¿Cabe entonces hablar de un modelo de uso profesional dentro de los sistemas que emplean etiquetas libres?

3. Metodología

Para poder analizar el uso del etiquetado social entre profesionales de la documentación en España se ha realizado un análisis del tagging utilizado en diferentes recursos. La búsqueda se ha realizado en la primera semana del mes de noviembre de 2008, acumulando las entradas encontradas en los seis meses anteriores (de mayo a octubre de 2008). Estas listas de términos se han comparado con la terminología controlada admitida en el Tesauro de Biblioteconomía y Documentación editado por el CINDOC.

Se han utilizado tres tipos de fuentes:

A) *DocuMenea* (http://www.documenea.com/), un servicio de promoción social de noticias entre profesionales. Se trata de un ejemplo modélico de herramienta de la web social, por su carácter colaborativo y abierto a la participación de los usuarios. Está dedicado a noticias en español especializadas en Biblioteconomía y la Documentación. Su funcionamiento se basa en la participación de los usuarios, que pueden seleccionar aquellas noticias que les hayan parecido más relevantes, a través de la votación directa. En la portada de *DocuMenea* sólo permanecen aquellas noticias que han recibido un número suficiente de votos. Los redactores del servicio añaden etiquetas. La web muestra una nube de las etiquetas más frecuentes, de las últimas 48 horas, última semana, último mes, último año o todas.

B) Blogs mantenidos por profesionales. Se han seleccionado algunos ejemplos de bitácoras personales que emplean etiquetas y son medios con cierto prestigio y difusión:
   - *Deakialli DocuMental* (http://www.deakialli.com/), blog gestionado por Catuxa Seoane y

Vanesa Barrero, desde marzo de 2003.

- *El Documentalista Enredado* (http://www.documentalistaenredado.net/), blog en el que colaboran Marcos Ros, María Elena Mateo y esporádicamente Julio Ruiz. Ofrece contenidos que se remontan a mayo de 2004.
- *Documentación, biblioteconomía e información*, el blog de Álvaro Cabezas (http://www.lacoctelera.com/documentacion), cuyos posts se iniciaron en julio de 2005.
- *Bibliotecarios 2.0* (http://bibliotecarios2-0.blogspot.com), la bitácora de Nieves González, que comenzó en septiembre de 2006.

Los datos de todos estos blogs se han tratado de forma conjunta, suponiendo que uno de los posibles usos de estos recursos sería la sindicación conjunta de este grupo de fuentes.

C) *Delicious* (http://delicious.com/, antes del.icio.us), un servicio de marcadores sociales. En este recurso se han seleccionado las etiquetas empleadas en los bookmarks de Dídac Margaix (http://delicious.com/didacmargaix) y Nieves González (http://delicious.com/nievesglez). En ambos casos se trata de dos profesionales que se han significado en la divulgación del uso profesional de las herramientas de la web social. El número de recursos que tienen seleccionados en *Delicious* hace suponer que se trata de dos usuarios experimentados en esta herramienta, por lo que pueden tomarse como modelo.

Los resultados obtenidos a partir de estas tres fuentes se han comparado con el *Tesauro de Biblioteconomía y Documentación* realizado por Gonzalo Mochón y Ángela Sorli, publicado por el CINDOC en 2002 y actualizado en línea en 2005 (cuya versión puede consultarse en la web en http://thes.cindoc.csic.es/index_BIBLIO_esp.html).

4. Resultados

4.1 Análisis de las etiquetas presentes en las noticias de *DocuMenea*

En el periodo de mayo a octubre de 2008 se han consultado un total de 506 noticias distribuidas y promocionadas en *DocuMenea*. Estas noticias tienen asignadas etiquetas, que pueden ser palabras únicas o grupos nominales, generalmente separados por coma. Sin embargo, se ha constatado la presencia de registros en los que no se ha respetado este criterio y se utilizan otros separadores (punto y coma o punto), o se incluyen varias palabras sin relación semántica clara, pero sin emplear ningún separador. Si se considera la posibilidad de crear un índice automático a partir de estas etiquetas pueden tenerse en cuenta varios separadores alternativos, pero no el simple espacio, ya que con frecuencia une palabras que constituyen una etiqueta única. Siguiendo este criterio se ha localizado un total de 1263 etiquetas con algún separador entre ellas, con una media de 2,5 por noticia.

De estas etiquetas, eliminando duplicidades, se encuentran 653 entradas diferentes. La mayor parte se han empleado con una frecuencia muy baja. Tan sólo once etiquetas alcanzan cierta visibilidad en este periodo: google (53), internet (46), bibliotecas (29), biblioteca (25), web 2.0 (24), buscadores (23), redes sociales (23), blogs (18), libros (17), revistas científicas (14) y digitalización (13).

En esta primera lista ya puede verse una primera consecuencia del control de vocabulario: la dispersión de entradas por el uso indiscriminado de singulares y plurales: biblioteca/s. A ello se añade la utilización de sinónimos que afectan a la visibilidad de algunos conceptos. Por ejemplo, la web 2.0

(utilizado 24 veces) aparece también como web social (6) y Web2.0 (sin espacio de separación, 2 veces). Pero la mayor dispersión se presenta en el concepto de los libros electrónicos que figuran como: "e-books" (escrito con comillas, 1vez), "libros electrónicos" (idem con las comillas, 1 vez), e-books (2), ebooks (8), ebooks libros electrónicos (tal cuál sin separador, 1 vez), e-libro (3), e-libros (1), elibros (1), libro digital (1), libro electrónico (8), libro electrónico digital (1), libro-e (1), libros electrónicos (1), libros-e (1) y libros-electrónicos (1). Un total de 15 formas diferentes que suman 32 entradas, a las que podrían añadirse dos más específicas, libro electrónico celular (1) y libro online (1). Sin duda, un tema estrella que puede quedar difuminado en una nube de etiquetas.

De la comparación del total de 653 términos con el *Tesauro de Biblioteconomía y Documentación* se han encontrado los siguientes casos:

- 165 entradas (25’3%) de nombres propios, personas, empresas, productos y lugares geográficos. Se pueden asimilar con identificadores, que no son recogidos en el listado de descriptores del tesauro.
- 77 etiquetas (11’8%) que se corresponden de forma exacta con entradas admitidas en el tesauro.
- 21 términos (3’2%) que se corresponden de forma exacta con entradas de términos equivalentes en el tesauro y que por tanto pueden remitirse de forma automática a un término admitido.
- 59 entradas (9%) con pequeñas diferencias: singular-plural, empleo de partículas. Estas variantes podrían agruparse con términos del tesauro con herramientas de búsqueda como el stemming o lematización. También se incluyen en este grupo las etiquetas formadas por unitérminos que en el tesauro se corresponden con un único término aunque esté en forma compuesta.
- 44 etiquetas (6’7%) que presentan diferentes variantes de especificidad en el tesauro. Su recuperación no podría asignarse de forma automática a una entrada del tesauro, ya que precisaría un juicio de desambiguación.
- 19 entradas (2’9%) con erratas: erratas tipográficas y construcciones de términos sin relación semántica pero sin separadores que los discriminen. Como consecuencia se dificulta su asignación automática a un término admitido de un tesauro.
- 268 términos (41%) no presentes en el tesauro: nuevos conceptos o posibles formas equivalentes no recogidas de forma expresa.

Así pues, el tesauro permitiría reconocer de forma automática tan sólo el 24% de las diferentes entradas de un hipotético índice de etiquetas acumuladas en este periodo (sumando equivalencias exactas con las de gran similitud). Por el contrario, el porcentaje de posibles términos candidatos que no se encuentran reflejados en el tesauro es muy elevado: un 41% de las diferentes entradas utilizadas.

4.2 Análisis de las etiquetas utilizadas en los blogs seleccionados

En el periodo de mayo a octubre de 2008 se han localizado un total de 179 artículos con etiquetas publicados en los cuatro blogs seleccionados. Las etiquetas van separadas con comas, y pueden ser palabras únicas o grupos nominales. En el caso de sintagmas nominales las palabras suelen ir separadas por espacios, aunque aparecen también casos en los que se emplean guiones. En ocasiones dentro de un mismo artículo se emplean ambos criterios, sin ninguna normalización. En la suma de estas fuentes se ha localizado un total de 654 etiquetas con algún separador entre ellas, con una media de 3,65 por artículo.

De estas etiquetas, eliminando duplicidades, se encuentran 241 entradas diferentes. Como en el caso de *DocuMenea*, la mayor parte se han empleado con una frecuencia muy baja, pero se aprecia una mayor concentración en las entradas más utilizadas, pues el número de etiquetas con más de diez usos se eleva a quince, a pesar de proceder de menos registros. Esta mayor consistencia puede deberse a que los casos utilizados representan a un grupo reducido de autores de blogs, mientras que en *DocuMenea* intervienen un mayor número de redactores. Las etiquetas más frecuentes en el periodo de estudio fueron: biblioteca 2.0 (utilizado 20 veces), congresos (20), web_2.0 (19), blogs (17), citas (14), alfin (14), bibliotecas (14), Internet (13), IVBP (etiqueta críptica que se refiere al IV Congreso de Bibliotecas Públicas, 13 veces), libros (13), web 2.0 (13), bibliotecarios (12), bibliotecas_universitarias (11), comunicacion cientifica (11) y web-2.0 (10).

En esta lista salta a la vista la presencia de entradas duplicadas relativas a la web 2.0 (13 veces) utilizada también con guión como web_2.0 (19) y web-2.0 (10). Pero a estas entradas habría que sumar las de web 2 0 (9), web social (2) y web-social (5). Y a ello se añade la presencia de otras entradas muy relacionadas con la aplicación profesional de esta tendencia: alfin_2.0 (3 veces), biblioteca 2.0 (20), biblioteca 2 0 (sin punto, con espacio, 1 vez), biblioteca_2.0 (con un guión, 9 veces), bibliotecarios 2.0 (1), bibliotecas web social web_2.0 herramientas aplicaciones (sin separadores, 1 vez), comunidad-2.0 (1), docente2.0 (1) y opac social (3).

De la comparación de estos 241 términos con el *Tesauro de Biblioteconomía y Documentación* se han encontrado los siguientes casos:

- 45 entradas (18’7%) de nombres propios, personas, empresas, productos y lugares geográficos que se pueden asimilar con identificadores, que no son recogidos en el listado de descriptores del tesauro.
- 40 etiquetas (16’6%) que se corresponden de forma exacta con entradas admitidas en el tesauro.
- 8 términos (3’3%) que se corresponden de forma exacta con entradas de términos equivalentes en el tesauro y que pueden remitirse de forma automática a un término admitido.
- 32 entradas (13’3%) con pequeñas diferencias (en los casos ya especificados en el análisis de las etiquetas de *DocuMenea*).
- 11 etiquetas (4’6%) que presentan diferentes variantes de especificidad en el tesauro. Su recuperación no podría asignarse de forma automática a una entrada del tesauro, ya que precisaría un juicio de desambiguación.
- 1 entrada (0’4%) que no utiliza separadores que permitan discriminar los diferentes términos, por lo cuál se dificulta su asignación automática a un término admitido de un tesauro.
- 104 términos (43’2%) no presentes en el tesauro: nuevos conceptos o posibles formas equivalentes no recogidas de forma expresa.

Así pues, si se constituyese una base de datos alimentada por los metadatos de los artículos de estas cuatro fuentes, el tesauro permitiría reconocer de forma automática el 33% de las diferentes entradas de un hipotético índice de etiquetas acumuladas en el periodo analizado. Por el contrario, el porcentaje de posibles términos candidatos que podrían incorporarse al tesauro sería bastante superior: un 43% de las diferentes entradas utilizadas.

4.3 Análisis de las etiquetas utilizadas en *Delicious*

En esta fuente se han analizado solamente dos casos por que no se han encontrado otros ejemplos que pudieran compararse. Se buscaron marcadores unidos por cierta similitud de intereses unida a la

presencia de una cantidad suficiente de recursos seleccionados. En el bookmark de Dídac Margaix se presentan 575 recursos, etiquetados con 269 entradas. En el de Nieves González se han utilizado 527 etiquetas para describir 800 recursos. En total 796 etiquetas, de las que 728 son entradas diferentes. Al tratarse de dos profesionales especializados en la docencia sobre herramientas de la web social, cabría esperar una gran similitud en el uso de entradas. Sin embargo, sólo existe coincidencia en 64 casos, un 8% de las etiquetas analizadas.

En este recurso, no se ha establecido una distribución de cada uno de los casos de posible comparación con el tesauro, por que la dispersión de formas de las entradas en varios idiomas dificultaba el análisis. En los recursos marcados en los bookmarks de los dos profesionales seleccionados, figuran etiquetas en español y en inglés, pero muchas de estas pueden proceder de los registros realizados por terceras personas, que los usuarios de *Delicious* pueden incorporar.

No obstante, a partir de esta fuente también pueden localizarse ejemplos de nuevos términos no presentes en el *Tesauro de Biblioteconomía y Documentación*, que se analizan conjuntamente con los procedentes de las otras dos fuentes descritas anteriormente.

4.4 Análisis de las etiquetas que no se corresponden con entradas del tesauro

El principal efecto práctico de la comparación entre las folksonomías y un tesauro radica en la detección de nuevas entradas que pueden enriquecer el lenguaje controlado, en especial de cara a una ampliación de su ámbito de utilización. Para que un tesauro pueda utilizarse para la recuperación en el texto completo o sobre resúmenes, resulta indispensable que abarque de la forma más explícita posible, todas las formas en las que pueden expresarse los temas de interés. A partir de las folksonomías analizadas se han localizado entre un 41% y un 43 % de posibles candidatos presentes en los índices de etiquetas, sin tener en cuenta nombres propios o formas con alto grado de similitud.

Entre estos términos no contemplados en el tesauro se pueden distinguir diferentes casos que hay que valorar de forma independiente:

a) Conceptos genéricos de uso común que sin embargo no fueron recogidos por el tesauro, posiblemente por no considerarlos necesarios para la descripción de los temas tratados en este área temática. Se trata de entradas como por ejemplo: Biología, Biomedicina, Cultura, Demografía, Ecología, Economía, I+D, Medio ambiente, Música o Política científica. La aparición de estas etiquetas muestra que no es posible limitar el léxico necesario para el análisis de contenido a los términos específicos de una disciplina. Estos conceptos probablemente no deban entrar en un tesauro especializado, pero sería de gran utilidad contar con una lista auxiliar o un macrotesauro de referencia.

b) Otros términos de uso común, de significado más específico: contenidos, contenidos digitales, crisis, datos, docencia, ocio, premios,… Este grupo de entradas plantea otro tipo de dificultad, por su carácter más específico resulta más complicado plantear la confección de listas auxiliares. Y algunos de estos conceptos resultan de gran ambigüedad. No puede establecerse una solución sencilla para este conjunto.

c) Entradas que deberían figurar como términos equivalentes de un descriptor presente en el tesauro, pero que han sido incluidas. El caso más llamativo es el de la entrada “blog”. Mientras el tesauro analizado, en su actualización de 2005, admite la forma “weblog” y como término equivalente

recoge “cuadernos de bitácora”, en el etiquetado utilizado por los profesionales domina el uso de “blog” y se recogen incluso varios derivados directos: blogging, blogosfera, bloguers. En este caso además de introducir estas entradas en el tesauro, debería incluso replantearse el término preferente a la vista de cuál es el uso mayoritario. Otro caso llamativo es el empleo de numerosos sinónimos y anglicismos para el caso de los “libros electrónicos”. Esta entrada figura en el tesauro pero no sus equivalentes empleados por los profesionales: ebooks, e-books, libros-e,… Igualmente, el tesauro recoge el término “minería de datos”, mientras en las etiquetas se emplea el anglicismo “data mining”.

d) Nuevos conceptos necesarios, términos que definen ámbitos de interés que han eclosionado en pocos años. Se incluyen aquí las entradas referidas al propio objeto de interés de esta comunicación (folksonomías, serendipia, tags o tagging), pero también otros muchos conceptos: alfabetización digital, alfin, AI, buscadores semánticos, CRAI, GIO, GPS, mashups, promoción web, redes sociales, RFID, SEO, web 2.0, web 3.0, wikis,.. A estos conceptos centrados en aspectos tecnológicos se añaden los debates más sociológicos que también van interesando a los profesionales y que se manifiestan en conceptos como “nativos digitales” o “generación google”. Esta afluencia de términos plantea la necesidad de actualización constante de una herramienta terminológica como un tesauro. Parte de esta terminología puede ser efímera, pero se debe recoger y reflejar la evolución en el uso científico de los términos.

e) Términos que definen aspectos que no son novedosos, pero que no han sido recogidos por el tesauro. Se trata de temas que atañen a la profesión pero sobre los que hay escasa bibliografía académica, lo cuál puede ser la causa principal que ha retrasado su inclusión en el tesauro. Ejemplos de este tipo sería entradas como escuelas de biblioteconomía, datasets, facsímiles, gestores de citas, hosts, software libre, tutoriales,… Estos casos plantean la necesidad de ampliar las fuentes terminológicas que se tienen en cuenta al elaborar un tesauro.

f) Conceptos más específicos que suponen variantes de los existentes, como por ejemplo: microblogging, blogs profesionales, blogs personales, estadísticas web, libro electrónico celular, libros acuáticos, televisión en Internet. Se trata de términos candidatos cuya necesidad debe evaluarse.

g) Por último, cabe constatar que en las folksonomías también aparecen otros términos de dudosa utilidad, por su carácter anecdótico o subjetivo, como por ejemplo: frikadas, lecturas pendientes, blogs de alumnos,… Se trata de una de las características que dificultan la generalización del etiquetado social, pero su peso porcentual es reducido en el ámbito de los recursos realizados por profesionales.

5. Conclusiones

Como resultado del análisis efectuado cabe preguntarse qué aportan las folksonomías a los profesionales de la documentación. Ante todo abren un campo para la innovación y la experimentación, constituyen una fuente terminológica de indudable valor. Su incorporación en los recursos documentales es una tendencia con futuro. Su utilización en las bibliotecas virtuales y los repositorios ha sido señalada por algunos autores (McGregor y McCulloch, 2006) como una oportunidad que permite conectar a los productores de servicios documentales con los usuarios y creadores de contenidos. Cabe esperar que su utilización se expanda y es necesario prestarles atención.

Los tesauros representan un modelo de lenguaje documental propio de otro momento histórico, en el que la apuesta de mayor calidad en la recuperación de información se correspondía con sistemas de carácter referencial. Su adaptación a la recuperación de información a texto completo en las fuentes electrónicas disponibles en Internet es uno de los principales retos actuales, que permitiría un avance considerable para la construcción de herramientas de la web semántica (Pérez Agüera, 2004). Pero la mayor parte de los tesauros que pueden consultarse actualmente, han sido realizados como meras herramientas de apoyo en el proceso de indización humana para la alimentación de bases de datos documentales. Los términos que los constituyen se corresponden con los temas tratados por la literatura científica en revistas académicas. Con su diseño actual, su aplicación directa a la búsqueda a texto libre sobre otro tipo de documentos, arrojaría resultados muy pobres, por que no hay una suficiente explicitación de las relaciones semánticas que un analista humano pone en juego en el análisis documental.

Si bien los tesauros tradicionales no han sido diseñados para su uso sobre texto libre, cabría esperar mejores resultados al aplicarse sobre un folksonomía, que constituye un sistema terminológico ya filtrado y limitado a posibles términos de búsqueda. El análisis realizado sobre los ejemplos seleccionados para este trabajo muestra que aún aplicando herramientas de reconocimiento de variantes básicas (singular/plural y diferentes separaciones en la sintaxis de un término), la comparación automática entre las etiquetas y el tesauro apenas reconocería entre un 24% (*DocuMenea*) y un 33% de los términos (suma de blogs seleccionados). En parte este mal resultado es producto del diferente tipo de documentación que ha servido de base para la construcción de los sistemas analizados: artículos académicos en el caso del tesauro, noticias y recursos web en las folksonomías. En segundo lugar, por el desfase temporal: el tesauro fue publicado en 2002 y actualizado en 2005, las etiquetas analizadas se corresponden con temas de actualidad en 2008.

Por tanto, los porcentajes deben tomarse con cierta precaución, carecen de valor estadístico, pero si resultan de utilidad para resaltar las carencias de un tesauro creado a partir de una documentación académica, para enfrentarse a la búsqueda sobre documentos de actualidad que cuentan con etiquetas libres como principal recurso para favorecer la recuperación. Si proliferan los recursos de información en los que predomine este modelo de “indización libre” los tesauros tradicionales corren un claro riesgo de resultar inservibles, si no se transforman. En este contexto, las folksonomías pueden servir como fuente preferente para enriquecer las entradas de un tesauro, ayudando a detectar nuevos descriptores y también términos equivalentes de los ya incorporados. El enriquecimiento de los tesauros si permitiría plantear su aplicación a la búsqueda en la web.

Como indicaba Spiteri (2007) para la incorporación de las folksonomías en los catálogos de bibliotecas, sería de interés contar con unas recomendaciones básicas sobre su redacción. Debe superarse el uso masivo de unitérminos, que a menudo resultan insuficientes, así como buscar soluciones para la desambiguación de entradas polisémicas.

En la actualidad hay claras diferencias entre tesauros y folksonomías. Como ya se ha señalado, el tesauro analizado en esta comunicación se elaboró a partir de los descriptores utilizados en el análisis de artículos de revistas científicas de Biblioteconomía y Documentación. Las folksonomías utilizadas parten sobre todo de la categorización de noticias, en las cuáles la atención a lo novedoso es prioritaria y se puede constatar con mucha mayor antelación que en la bibliografía más académica. En segundo lugar, las folksonomías son creadas por los propios autores, sin el filtro o la censura que introduce el documentalista. Cuando se elabora un tesauro se busca habitualmente la mayor corrección posible en la selección de los términos preferentes. Esto implica evitar los anglicismos innecesarios, que sin embargo pueden ser utilizados con mucha frecuencia por los autores. Pero todas las variantes de un

concepto deberían reflejarse en una herramienta de control del vocabulario técnico.

Estos dos factores, agilidad para captar las novedades y cercanía al uso social real, otorgan cierta ventaja a las folksonomías frente a las herramientas tradicionales de control de vocabulario. Los tesauros podrían aprovechar estas mismas ventajas si utilizan el etiquetado social como fuente de interés en su mantenimiento y actualización. Las etiquetas utilizadas por los autores también pueden servir como indicadores para replantear el término preferente frente a los equivalentes.

Como reflexión final cabe preguntarse también qué pueden aportar los tesauros a las folksonomías en un contexto profesional. Pese al uso creciente de etiquetas, su presencia no permite ofrecer un sistema de recuperación eficaz, por su variabilidad y dispersión de entradas. En la línea que indicaba Yusef Hassan (2006) será necesario aplicar soluciones invisibles para el usuario final, para lo cuál sería muy importante contar con herramientas de control del vocabulario. Sólo si se adaptan e incorporan nuevos términos y variantes léxicas con agilidad, los tesauros estarán en condiciones de representar un rol en los sistemas de búsqueda en el texto libre en un futuro inmediato. Este objetivo debe considerarse prioritario si se quiere que los tesauros puedan representar este papel.

Sin embargo, la simple comparación no resuelve todos los problemas que pueden limitar la posible aplicación de los tesauros para la búsqueda en texto libre. Un aspecto relevante a resolver radica en las diferencias que existen en la construcción semántica entre etiquetas y descriptores. En las folksonomías aparecen entradas que precisan una mayor concreción para romper su ambigüedad, un factor que no puede solucionarse fácilmente. La práctica de la "indización libre" no permite establecer cuando un término precisa una mayor concreción para ser admitido como etiqueta en un documento. Esta característica, propia de los lenguajes controlados, no es aplicable a una folksonomía.

Un segundo aspecto que queda sin resolver es la presencia de etiquetas en varios idiomas, que se ha detectado en los recursos marcados en *Delicious.* En una biblioteca digital o en una plataforma de revistas electrónicas se puede identificar el idioma que se emplea en las palabras clave de los autores, por ejemplo mediante metaetiquetas Dublín Core que lo especifiquen. Por el contrario, en los recursos que emplean el etiquetado no está prevista esta opción.

6. Bibliografía

CATARINO, Maria Elisabete; BAPTISTA, Ana Alice. "Social Tagging and Dublin Core: A Preliminary Proposal for an Application Profile for DC Social Tagging". En: *Proceedings ELPUB 2008 Conference on Electronic Publishing, Toronto*. Disponible en: http://elpub.scix.net/cgi-bin/works/Show?100_elpub2008 [consulta 15-10-2008].

GIL LEIVA, Isidoro; ALONSO ARROYO, Adolfo. "La relación entre las palabras clave aportadas por autores de artículos de revista y su indización en las bases de datos ISOC, IME e ICYT." *Revista Española de Documentación Científica*, 2005, vol. 28, n. 1, pp. 62-79. Disponible en: http://redc.revistas.csic.es/index.php/redc/article/view/165/219 [consulta 18-10-2008].

HASSAN MONTERO, Yusuf. "Indización social y recuperación de información". *No Solo Usabilidad*, 2006, n.5. Disponible en: http://www.nosolousabilidad.com/articulos/indizacion_social.htm [consulta 15-10-2008].

HASSAN MONTERO, Yusuf; HERRERO SOLANA, Víctor. "Improving Tag-Clouds as Visual

Information Retrieval Interfaces". En: *I Internacional Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006*, Mérida, 2006. Disponible en: http://www.nosolousabilidad.com/hassan/improving_tagclouds.pdf [consulta 14-10-2008].

IDUMM. "Investigación online (2): Etiquetado útil de enlaces en del.cio.us y similares". *Idumm Blog*, 10 de Diciembre de 2006. Disponible en: http://www.idumm.org/blog/?p=49 [consulta 26-10-2008].

KIPP, Margaret E.I. "@toread and Cool : Subjective, Affective and Associative Factors in Tagging". En: *Proceedings of the Annual Conference of the Canadian Association for Information Sience (CAIS)*, Vancouver, 2008. Disponible en: http://eprints.rclis.org/archive/00013788/ [consulta 18-10-2008].

LANCASTER, F.W. *El control del vocabulario en la recuperación de información*. Valencia: Universitat, 2002.

McGREGOR, George; McCULLOCH, Emma. "Collaborative Tagging as a Knowledge Organisation and Resource Discovery Tool". *Library Review*, 2006, vol. 55, n. 5-6, pp. 291-300. Pre-print disponible en: http://eprints.rclis.org/archive/00005703/ [consulta 14-10-2008].

MOCHÓN, Gonzalo; SORLI, Ángela. *Tesauro de Biblioteconomía y Documentación.* Madrid: CINDOC, 2002. La versión actualizada en 2005 está disponible en: http://thes.cindoc.csic.es/index_BIBLIO_esp.html [consulta 09-11-2008].

MOREIRO GONZÁLEZ, José Antonio. *El contenido de los documentos textuales: su análisis y representación mediante el lenguaje natural*. Gijón: Trea, 2004.

PARDO KUKLINSKI, Hugo. "Un esbozo de ideas críticas sobre la Web 2.0". En: Cobo Romaní, Cristóbal; Pardo Kuklinski, Hugo. *Planeta Web 2.0. Inteligencia colectiva o medios fast food*. Grup de Recerca d'Interaccions Digitals, Universitat de Vic. Flacso México. Barcelona / México DF, 2007.

PÉREZ AGÜERA, José Ramón. "Automatización de tesauros y su utilización en la web semántica". *BiD: textos universitaris de biblioteconomia i documentació*, 2004, n. 13. Disponible en http://www2.ub.es/bid/consulta_articulos.php?fichero=13perez2.htm y http://eprints.rclis.org/archive/00004176/ [consulta 09-11-2008].

QUINTARELLI, Emanuele. "Folksonomies: power to the people". En: *ISKO Italy – UniMIB Meeting*, 2005, Milán. Disponible en: http://www.iskoi.org/doc/folksonomies.htm [consulta 16-10-2008].

RODRÍGUEZ PALCHEVICH, Diana. *Nuevas tecnologías Web 2.0: Hacia una real democratización de la información y el conocimiento*. 2008. Disponible en: http://eprints.rclis.org/archive/00013897/ [consulta 20-10-2008].

ROS MARTÍN, Marcos. "Folksonomías, marcado social y filtrado social de noticias". En: *Comunidad de prácticas: Web social para profesionales de la información*, SEDIC, 2008. Disponible en: http://comunidad20.sedic.es/?m=2008&w=23 y http://eprints.rclis.org/archive/00013709/ [consulta 26-10-2008].

SEOANE, Catuxa. "El éxito de las folksonomías". *Deakialli DocuMental*, 28 de julio de 2005. Disponible en: http://www.deakialli.com/2005/07/28/el-exito-de-las-folksonomias/ [consulta 26-10-2008].

SEOANE, Catuxa. “Tagging de autor versus tagging de lector: profesionales y consumidores”. *Deakialli DocuMental*, 5 de junio de 2007. Disponible en: http://www.deakialli.com/2007/06/05/tain-de-autor-versus-tain-de-lector/ [consulta 26-10-2008].

SEOANE, Catuxa. “Flexibilidad de las folksonomías”. En: *Anuario ThinkEPI*, 2007, pp. 74-75. Disponible en http://eprints.rclis.org/archive/00011558/ [consulta 28-10-2008].

SERRANO COBOS, Jorge. “Tags, folksonomies y bibliotecas”. En: *Anuario ThinkEPI*, 2007, pp. 71-73. Disponible en: http://www.thinkepi.net/notas/2007_16.pdf [consulta 28-10-2008].

SHIRKY, Clay. “Folksonomies + controlled vocabularies”. *Many 2 many Weblog*, January 7, 2005. Disponible en: http://many.corante.com/archives/2005/01/07/folksonomies_controlled_vocabularies.php [consulta, 25-10-2008].

SPITERI, Louise F. “The structure and form of folksonomy tags: the road to the public library catalogue”. En: *La interdisciplinariedad y la transdisciplinariedad en la organización del conocimiento científico. Actas del VIII Congreso ISKO-España*. León: Universidad, 2007, pp. 459-467.

WRIGHT, Alex. “Folksonomy”. *Alex Wright’s Blog*, August 23 2004. Disponible en: http://www.alexwright.org/blog/archives/000900.html [consulta 25-10-2008].