Etiquetado social y *blog-scraping* como alternativa para la actualización de vocabularios controlados

Aplicación práctica a un tesauro de Biblioteconomía y Documentación

Gonzalo Mochón-Bezares, Eva Méndez-Rodríguez y Ángela Sorli-Rojo

Universidad Carlos III. Facultad de Humanidades, Comunicación y Documentación. Departamento de Biblioteconomía y Documentación / Consejo Superior de Investigaciones Científicas. Departamento de Postgrado y Especialización, España / gmochonb@gmail.com; emendez@bib.uc3m.es; angela.sorli@csic.es

Resumen

El objetivo de este artículo es comparar las etiquetas en lenguaje libre, tomadas en nuestro caso de blogs especializados en ciencias de la información (information sciences), frente al lenguaje controlado no estructurado de las listas de palabras clave, con el fin de comprobar cuál de estos dos es una mejor fuente de nueva terminología para el Tesauro de Biblioteconomía y Documentación. Para ello, se extrajeron las etiquetas de autor de 127 blogs sobre biblioteconomía y documentación mediante técnicas de *web scraping*, y se compararon con los listados de descriptores e identificadores de la base de datos ISOC Biblioteconomía y Documentación (ISOC-BD). El análisis de las etiquetas de autor de blogs ha aportado 186 nuevos términos, mientras que los listados de la base de datos han proporcionado 130 términos. Se concluye que las etiquetas en lenguaje libre pueden ser una mejor y más rápida vía de aporte de nueva terminología a los vocabularios controlados que los listados de lenguaje controlado no estructurado.

Palabras clave

Etiquetado social
Mantenimiento de tesauros
Blogs
Biblioteconomía y documentación
Extracción terminológica

Abstract

Social tagging and blog-scraping as an alternative for updating controlled vocabularies: Practical application to a library and information science thesaurus. The aim of this paper is to compare the use of free language tags, taken in our case from specialized blogs on information sciences, against the unstructured controlled language of keywords lists, for verifying which of them is the best source of new terminology for the Librarianship Thesaurus and Documentation. To do this, authors' labels were extracted from 127 blogs on librarianship and information science using web scraping techniques, and were compared with descriptors and identifiers lists of the ISOC library and documentation database (ISOC-BD). The results of the analysis of authors' tags in blogs contribute with 186 new terms, while the database lists only 130 terms. It is concluded that free language tags could be a better and faster way for contributing new terminology to controlled vocabularies than unstructured controlled language lists.

Keywords

Social tagging
Thesauri maintenance
Blogs
Library and information science
Terminological extraction

Artículo recibido: 13-07-2017. Aceptado: 20-09-2017

1. Introducción

La actualización de los tesauros y otros sistemas de organización del conocimiento (SOC) es una tarea que presenta problemas en el mantenimiento que debe seguir a toda creación de vocabularios. Uno de estos problemas es que el alcance y uso de algunos descriptores o conceptos cambia con el paso del tiempo, a la vez que se incorporan términos nuevos y otros caen en desuso. En disciplinas que tienen un alto componente tecnológico, como ocurre en Biblioteconomía y Documentación (ByD), el cambio terminológico sucede con rapidez y resulta bastante difícil reflejarlo en los esquemas de contenido o vocabularios semánticos que se utilizan para representar y recuperar la información.

En la literatura especializada se pueden encontrar diferentes formas utilizadas en la actualización de los vocabularios controlados, principalmente sobre tesauros y ontologías. En estas se distinguen métodos automáticos o semiautomáticos destinados a la extracción de terminología, tanto de los textos de fuentes originales como de la información asociada a estas (metadatos). Si nos fijamos en la extracción de datos de las fuentes, podemos encontrar estudios en los que se aplican métodos del procesamiento del lenguaje natural que tienen como objetivo identificar nuevos términos destinados a tesauros, como el de L. Araujo y J. R. Pérez-Aguera (2006), quienes proponen un sistema de coincidencia de patrones en diccionarios para identificar relaciones jerárquicas entre los términos, y de esta manera poder usarlos en la generación automática de tesauros.

Otro trabajo sobre dicha temática es el realizado por P. Arnold y E. Rahm (2014), que analizan la Wikipedia buscando relaciones semánticas entre términos que puedan servir para complementar tesauros y diccionarios. En lo que se refiere a los metadatos como origen de los términos, se puede hacer referencia al trabajo de J. Wang (2006) sobre la asociación entre palabras de los títulos de obras y los descriptores asignados a las mismas para dotar a un tesauro de nueva terminología. Otro modelo de método semiautomático de actualización de vocabularios controlados es el elaborado por M. Vállez, et. al. (2015), que comparan los descriptores usados en la representación de un corpus de texto y las palabras clave empleadas por los usuarios durante sus consultas, con vistas a obtener términos candidatos para actualizar un vocabulario.

Otro tipo de sistema de organización del conocimiento, habitual en la bibliografía especializada, es la folksonomía. Las investigaciones sobre este tipo de SOC se han centrado en el etiquetado en lenguaje natural para extraer ontologías empleando diversos métodos, recogidos en distintos estados de la cuestión (Limpens, Gandon y Buffa, 2009; García-Silva et al., 2012). Sin embargo, los conjuntos de etiquetas no controladas apenas han sido utilizados como base para el desarrollo de tesauros. Solamente se han encontrado tres ejemplos sobre el empleo de folksonomías como fuente de nueva terminología para los tesauros: uno referido a un tesauro sobre la Guerra Civil Española (SIDBRINT) y los otros dos dedicados a la crítica y mejora del contenido del Tesauro de Biblioteconomía y Documentación (TByD).

En el trabajo sobre el tesauro SIDBRINT se analiza la interrelación entre un sistema tradicional de recuperación de información basado en un tesauro y otro fundamentado en el comportamiento de los usuarios (folksonomías), con el fin de comprobar la integración de ambos en un sistema de información digital. Los autores señalan que la actividad de los usuarios, como consumidores de información que asignan términos a los registros bibliográficos, si puede aportar conceptos que ayuden a mejorar las consultas a realizar en el sistema SIDBRINT junto con el mencionado tesauro (Masó-Marema y Sebastià-Salat, 2013).

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 37 (diciembre 2017)

En lo que se refiere a las propuestas de mejora para el TByD, el estudio más antiguo es el de Luis Rodríguez Yunta (2009), en el que se analizan las etiquetas de tres tipos diferentes de medios sociales (un servicio de promoción social, cuatro blogs mantenidos por profesionales de la documentación y un servicio de marcadores sociales), comparando los resultados obtenidos con el contenido del TByD. La finalidad es comprobar si el etiquetado libre puede aportar nueva terminología a dicho vocabulario controlado. Se concluye que los lenguajes documentales tradicionales pueden resultar inútiles si no se transforman e incorporan mayor capacidad de actualización y enriquecimiento de su contenido, como sí ocurre con las folksonomías. Otro estudio sobre este tema es el trabajo académico realizado por Luis Alonso Soriano (2013), en el que se extraen etiquetas de autor en lenguaje libre de diez blogs en español elaborados por profesionales de biblioteconomía y documentación entre octubre de 2010 y junio de 2012. Alonso Soriano identifica 58 nuevos términos candidatos para el TByD, la mayoría de ellos relativos a las tecnologías de la información y la comunicación y al desarrollo de la web 2.0. Este autor concluye que el etiquetado social en blogs puede ser una fuente terminológica de gran valor para la actualización de los vocabularios controlados.

Los resultados obtenidos en los dos estudios antes citados constatan la necesidad que tiene el TByD de actualizar su contenido para disponer de un nivel operativo adecuado dentro de las materias que abarca y más acorde con el ámbito de la web. Además, demuestran la importancia del etiquetado de autor en blogs como fuente de terminología para la ciencia de la información y la documentación.

En sus comienzos, el TByD fue un vocabulario controlado creado para ayudar en la indización y recuperación documental de la base de datos ISOC Biblioteconomía y Documentación (ISOC-BD) del Consejo Superior de Investigaciones Científicas (CSIC). Su primera versión, editada en forma de libro, (Mochón Bezares y Sorli Rojo, 2002) constaba de 913 términos preferidos distribuidos en doce grupos jerárquicos y 239 términos no preferidos. En 2005 se hizo una segunda versión del TByD con motivo de su aplicación al portal de revistas digitales Temaria para facilitar el acceso por materias a los artículos (Abadal et al. 2005). Esta segunda versión reunía un total de 1.113 descriptores y 394 términos no preferidos, e incrementó de forma considerable las relaciones entre sus términos. Ambas versiones del tesauro fueron elaboradas de acuerdo con la norma ISO 2788 de construcción de tesauros (ISO, 1986), y tuvieron como principal fuente de terminología el listado de palabras clave resultado de la indización manual de los artículos incluidos en ISOC-BD.

Teniendo en cuenta lo expuesto más arriba, tanto en lo relativo a la nueva actualización del TByD como en lo que respecta al análisis de alternativas para la extracción de nuevos términos con los que actualizar tesauros, el objetivo de este trabajo es doble:

» Por un lado, se investiga la representación de conceptos en las etiquetas en lenguaje libre, agrupadas en folksonomías, y en el lenguaje controlado no estructurado (listas de palabras clave o descriptores) de una base de datos con el fin de comprobar su grado de idoneidad en el enriquecimiento de tesauros.
» Por otro, se busca conocer la terminología más novedosa para cada una de las categorías que integran el TByD, con el fin de incorporarla al mismo. Se exceptúa el área de museología ya que en los blogs empleados no hay ninguno especializado en dicha materia.

Para ello hemos utilizado un conjunto de blogs seleccionados por especialistas con el objetivo de extraer sus etiquetas de autor o materias a través de herramientas de extracción de contenido de páginas web, el denominado *web scraping*, y lo hemos

comparado con el conjunto de resultados obtenidos de las palabras clave presentes en los campos de descriptores e identificadores recogidos en la base de datos ISOC-BD.

2. Metodología

La metodología empleada combina, por un lado, la extracción terminológica de las listas de términos procedentes de la indización en la base de datos ISOC-BD, de la que ya nos hemos servido en las dos versiones anteriores del TByD, y por otro, la extracción de términos a partir de blogs, utilizando tecnologías de web scraping y text mining, lo que nosotros denominamos *blog scraping*.

En lo que se refiere a los listados de términos de la base ISOC-BD, conviene señalar que los campos de interés para nuestro propósito son los denominados "descriptores" e "identificadores". La lista de descriptores es fruto de la indización humana y puede contener tanto los conceptos que el indizador considere que están representados en la literatura, como los términos que aparezcan en los textos de los artículos o en las palabras clave de autor. Los identificadores son los nombres propios de personas u organizaciones que aparecen en los documentos indizados. De estos últimos, interesan especialmente nombres de servicios de redes sociales, sistemas o programas informáticos (buscadores, sistemas operativos, navegadores, gestores de referencias bibliográficas o gestores de sistemas de gestión de contenidos), lenguajes de programación, esquemas de metadatos, siglas de instituciones e índices de citas. Para conseguir los listados completos de descriptores e identificadores de la base de datos ISOC-BD se realizaron consultas por año de publicación de los documentos recogidos en la misma, seleccionando de los resultados solamente la información de interés para nuestro caso, como se aprecia en el gráfico 1.

El otro modo de adquisición de términos empleado se basa en el web scraping, que supone la extracción y creación de una representación estructurada de datos de un sitio web de forma automática usando un screen scraper o un motor de extracción de datos, y que almacena el contenido en una base de datos local o una hoja de cálculo (Cording, 2011). En nuestro caso se busca extraer en formato .csv los datos sobre la asignación de materias o etiquetas en un conjunto de blogs, lo que supone una parte muy pequeña en comparación con la información contenida en las páginas consultadas.

Como fuente terminológica única en el apartado de los blogs, se ha optado por el conjunto recogido en el directorio Biblogsfera, creado por M. A. Vera Baceta. Este directorio comenzó siendo un grupo de bitácoras seleccionadas en una encuesta a diversos profesionales españoles vinculados con la biblioteconomía y documentación en 2013, aunque posteriormente se le han ido añadiendo más bitácoras sobre estos temas (Vera Baceta, 2013; 2015). Biblogsfera recopila a mayo de 2017 un total de 200 blogs, de los cuales 127 disponen en sus *posts* de asignación de materias o etiquetas que identifican su contenido, y además permiten el análisis del contenido por medio de robots, ambas condiciones consideradas indispensables para llevar a cabo nuestro análisis.

La temática principal tratada en estas 127 publicaciones alcanza a todas las áreas recogidas en el TByD, excepto a los apartados Lenguajes y Lingüística y Museología, que quedan sin representación directa. De todos los blogs, seis se consideran generales pues sus contenidos abarcan varias materias tratadas en el tesauro, siendo el objeto principal de los restantes el siguiente: bibliotecas y biblioteconomía (47), archivos y archivística (20), temática relacionada con la web (acceso abierto, arquitectura de la información, curación de contenidos, gestión de comunidades, usabilidad) (15),

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 37 (diciembre 2017)

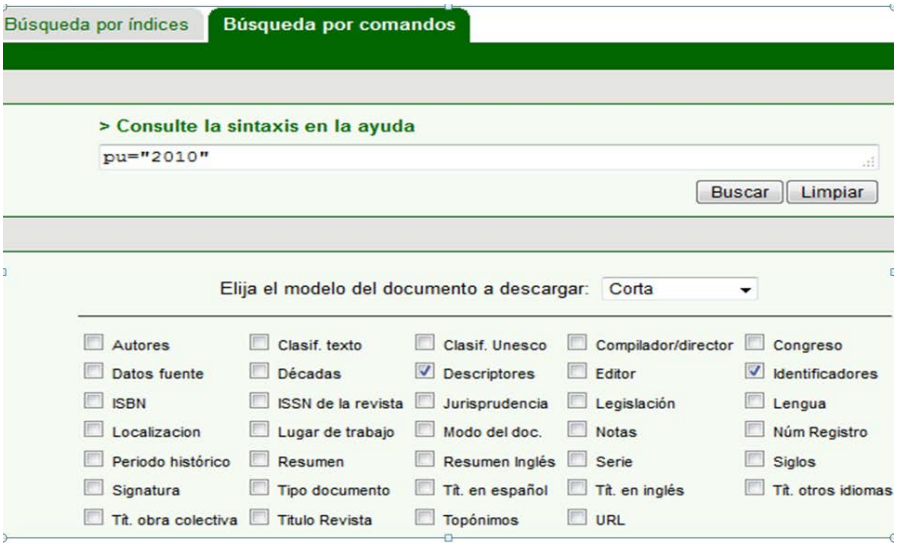

Gráfico 1. Detalle de consulta y selección de información en la base de datos ISOC-BD

centros de documentación (14), gestión de información en organizaciones (10), profesionales y usuarios de la información (6), estudios métricos de información (5) e industria de la información (4).

Para las labores de *blog-scraping* se ha optado por la herramienta *OutWit Hub* (Version Pro 5.0.1.57), específicamente diseñada para labores de *web scraping*, de la que interesa especificar las funcionalidades *scraper*s y *macro*. La primera permite recuperar el contenido que se encuentra entre los marcadores iniciales y finales establecidos por el usuario, que pueden ser etiquetas completas de elementos HTML o bien partes de estas. Como los datos que nos interesan están incluidos en todas las páginas como contenido de dos únicos atributos, las consultas realizadas a través de la funcionalidad *scrapers* no resultaron complicadas, como puede observarse en el Gráfico 2.

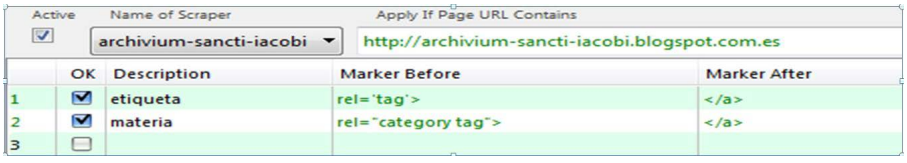

Gráfico 2. Marcadores usados en la consulta mediante la función scrapers

Los marcadores de las consultas son muy sencillos. El inicio se especifica la parte que interesa del atributo *rel* y el valor que este toma en cada uno de los casos, mientras que al final se pone la etiqueta de cierre del elemento. El objetivo es recuperar el contenido de la etiqueta o la categoría que en la página aparecen en forma de enlace.

Esta vía facilita la extracción de la información en la página de inicio del blog, pero no en las páginas que desciendan a un segundo o inferior nivel, que es donde suelen presentarse los posts más antiguos. Para llevar a cabo esta acción se debe crear una *macro,* diferente para cada blog, en la que se definan entre corchetes los rangos mínimos y máximos que puedan tomar los valores cambiantes de las URLs, que en muchos casos son los números de las páginas (Gráfico 3).

Sin embargo, este tipo de macro no es recomendable para la consulta de blogs que incluyan datos de variabilidad compleja -como son horas o fechas- en las URLs de sus posts. Pues, al buscar por cada uno de los valores que se encuentren

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 37 (diciembre 2017)

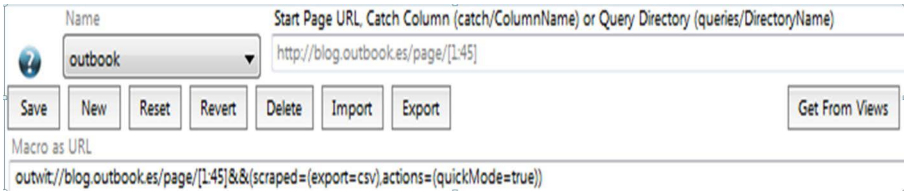

Gráfico 3. Detalle de consulta con macro

entre el rango mínimo y máximo asignado para cada fecha u hora, la acción se ralentiza excesivamente y el nivel operativo de la herramienta resulta escaso. Para evitar este problema se empleó la opción *Dig Within Domain* disponible dentro de la *macro*, que explora de forma sistemática todos los enlaces internos encontrados dentro del sitio de un blog, y limitando nuestra búsqueda de contenido a las etiquetas o categorías.

Para que los términos obtenidos, tanto de los listados de la base ISOC-BD como de los blogs del directorio Bibliogsfera, puedan ser considerados descriptores deberán cumplir las siguientes condiciones:

» que no estén incluidos en la edición del año 2005 del TByD. Si se comprueba que un nuevo término, que tenga la consideración de término no preferido en la versión de 2005, haya sido más utilizado en la literatura que su descriptor equivalente, ambos términos intercambiarán sus estatus (de término no preferido a descriptor y viceversa).
» que pertenezcan a una de las áreas temáticas del Tesauro.
» que tengan una frecuencia mínima de 3 en los listados obtenidos de la base de datos ISOC-BD, o que hayan sido usados en tres blogs distintos. En los casos de sinonimia entre términos candidatos, se considera descriptor el que tenga mayor frecuencia de uso, quedando el de menor frecuencia como término no preferido.
» la lengua preferente es el castellano. Sin embargo, la lengua inglesa es también admitida en aquellos casos en que los conceptos estén más representados en los documentos por términos en inglés que por sus equivalentes en castellano.

En el resto de los aspectos relacionados con la forma gramatical de los términos, se han seguido las recomendaciones de las normas ISO 2788 (1986) y 25964-1 (2011) sobre construcción de tesauros.

3. Resultados y discusión

Dado que la consulta a la base de datos ISOC-BD y la extracción de etiquetas de los blogs son procesos diferentes, se detalla primero lo obtenido en cada caso de forma separada y, después, se analizan los resultados finales de ambos estableciendo comparaciones parciales y totales con los contenidos de las versiones del TByD realizadas en los años 2002 y 2005.

3.1. Resultados de extracción de la base de datos ISOC-BD

A finales del mes de mayo de 2017 la base de datos ISOC-BD disponía de un total de 17.859 referencias, en las cuales se contabilizaron 18.111 términos de indización, que divididos por su categoría en la base de datos resultaban 11.536 descriptores y 6.575 identificadores. De estas cifras se eliminaron en primer lugar los términos que no alcanzaban una frecuencia de tres, posteriormente aquellos que perteneciesen a campos del conocimiento ajenos al TByD, y finalmente los

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 37 (diciembre 2017)

que ya estuvieran presentes en la segunda versión del tesauro. Tras aplicar dichos criterios y eliminar identificadores tales como nombres propios y de organizaciones, el conjunto quedó reducido a 130 términos, como se puede apreciar en la tabla 1. La mayor parte de dichos términos están relacionados con las tecnologías de la información, actividades de la sociedad de la información relacionadas con dichas tecnologías, los profesionales de la información y los usuarios. Las cantidades de palabras clave insertadas en las otras áreas resultan escasas, sorprendiendo la total ausencia de términos candidatos sobre archivística, dado el gran número de documentos sobre esa materia indizados en ISOC-BD.

3.2. Resultados de extracción de los blogs

Los 127 blogs de los que se han extraído etiquetas o materias han resultado una importante fuente terminológica, aunque también se han constatado grandes diferencias entre ellos. Una de ellas es la periodicidad irregular en la publicación, que en el conjunto manejado va desde los 1.633 posts publicados en *Deakialli Documental* durante sus once años de actividad frente a los escasos 27 incluidos en *Archivos de la Administración: entre el Poder y la Memoria* durante más de cinco años. Otra diferencia está en la variedad de etiquetas usadas para enlazar los posts, que queda a la voluntad de cada autor. Como ejemplo de esta diferencia está los blogs *Social Media Strategies*, con 185 etiquetas diferentes, y *GrancanariaPA,* donde se emplean solamente 2. Dadas las diferencias entre los blogs no se han obtenido medias de productividad (posts por blogs), ni de representación de contenidos (etiquetas empleadas en cada blog) por no ser de utilidad para nuestro análisis.

Categorías del Tesauro	**Base ISOC**		**Blogs**	
		%		%
Archivística	0		0	
Biblioteconomía	3	2,31	3	1,61
Ciencias y técnicas auxiliares	6	4,62	7	3,76
Estudios métricos de la información	7	5,38	7	3,76
Fuentes de información	2	1,54	5	2,69
Lenguajes y lingüística	2	1,54	5	2,69
Museología	0	0	0	0
Proceso documental	6	4,62	8	4,30
Profesionales de la información y usuarios	22	16,92	33	17,74
Sociedad de la información	23	17,69	26	13,98
Tecnologías de la información y las comunicaciones	58	44,62	87	46,77
Unidades de información	1	0,77	5	2,69

Tabla 1. Porcentajes obtenidos por categoría del TByD y fuente de terminología

Para evitar una acumulación innecesaria de términos, se decidió acometer una selección preliminar del vocabulario obtenido en cada uno de los blogs tras la extracción terminológica. En esta primera selección individual se eliminaron los identificadores (nombres de personas o entidades, marcas comerciales y topónimos), así como las etiquetas personales que solo tienen sentido para el autor y aquellos términos que no estuviesen vinculados con las áreas temáticas del TByD. Tras esta primera revisión quedaron un total de 4.733 etiquetas diferentes procedentes de los 127 blogs examinados.

En una segunda selección se eliminaron todas las etiquetas que no estuviesen recogidas en al menos tres blogs y que ya estuviesen presentes en la segunda versión del TByD. También se corrigieron las variantes ortográficas de acuerdo a las de normas ISO 2788 (1986) e ISO 25954 (2011), dejando una sola forma para cada concepto. Tras este nuevo proceso selectivo, la cifra final de términos procedentes del etiquetado de blogs fue 186, con una importante presencia de vocabulario de Tecnologías de la información y la comunicación, Sociedad de la información y de los colectivos de Profesionales de la información y usuarios, tal y como se aprecia en la tabla 1.

Si se analizan los grupos de temática principal de los blogs junto con los resultados obtenidos en estos por cada área del tesauro, se puede constatar que cada publicación no limita el contenido de sus posts exclusivamente a su objeto central sino que tratan diferentes temas. De esta manera, podemos comprobar que las áreas de Archivística y Biblioteconomía proporcionan muy pocas etiquetas, 0 y 3 respectivamente, cuando los blogs sobre estos temas son los más numerosos dentro del conjunto que hemos manejado. En el caso de la archivística ocurre algo similar a lo visto en los descriptores de la base ISOC-BD, y es que, aunque está suficientemente representada con 20 blogs que proporcionan un considerable número de etiquetas, no ha aportado ningún candidato a descriptor.

3.3. Resultados totales

La cantidad total de términos candidatos resulta considerablemente superior en el caso de las etiquetas de los blogs (186) frente a la obtenida (130) con los descriptores e identificadores de la base ISOC-BD. En lo que a la distribución porcentual de términos por categorías se refiere, existe cierta similitud entre los dos conjuntos de términos, pues en ambos casos más de tres cuartas partes del total pertenecen a las áreas de Tecnologías de la información y las comunicaciones, Sociedad de la información y Profesionales de la información y usuarios, quedando el resto de grupos terminológicos con escasos porcentajes o incluso nulos.

Una vez unidos los dos conjuntos se procedió a la revisión final de ambos listados dejando un descriptor para cada concepto, eliminando las sinonimias o cuasi-sinonimias y estableciendo las equivalencias entre los mismos. Además, se adaptó la forma gramatical de los términos adecuándola a los requisitos recogidos en la norma ISO 25964-1 (2011), poniendo en plural las etiquetas de los términos contables (que los autores de los blogs suelen incluir en singular) y cuidando que las siglas se muestren desarrolladas.

Con esta última revisión, se consiguió un resultado total de 259 términos (205 descriptores y 54 términos no preferentes) que se asignaron al grupo temático que les correspondía dentro del TByD. También se aprovechó el manejo del elevado conjunto de términos que proporcionaron las fuentes para revisar el contenido de la segunda versión del tesauro. Como resultado de dicha revisión 37 términos variaron su posición en el índice jerárquico del tesauro, de entre los cuales 11 cambiaron su estatus de términos no preferidos a descriptores y 9 perdieron su condición de descriptores pasando a ser términos equivalentes. Estos cambios de status de los términos se hicieron en base a la frecuencia de uso que estos presentan en las fuentes empleadas, mientras que los cambios de ubicación se debieron a una mayor adecuación de cada concepto a un ámbito de aplicación concreto.

La tabla 2 muestra los porcentajes de descriptores y no descriptores por materia. En el caso de los términos no preferidos la agrupación se hace de acuerdo a las categorías en las que están incluidos los descriptores de los que son sinónimos.

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 37 (diciembre 2017)

Categorías del Tesauro	**Descriptores**		**No descriptores**	
		%		%
Archivística	0	0	0	0
Biblioteconomía	5	2,44	0	0
Ciencias y técnicas auxiliares	9	4,39	2	3,70
Estudios métricos de la información	10	4,87	0	0
Fuentes de información	11	5,36	3	5,55
Lenguajes y lingüística	3	1,46	2	3,70
Museología	0	0	0	0
Proceso documental	5	2,44	3	5,55
Profesionales de la información y usuarios	19	9,27	12	22,22
Sociedad de la información	25	12,19	9	16,66
Tecnologías de la información y las comunicaciones	113	55,12	22	40,74
Unidades de información	5	2,44	1	1,85

Tabla 2. Porcentajes de descriptores a ingresar en el Tesauro por categoría

Las cifras de descriptores presentadas en la tabla 2 señalan que las categorías de Tecnologías de la información y las comunicaciones (TIC), Profesionales de la información y usuarios y Sociedad de la información acumulan la mayoría de los nuevos términos, aunque con unos niveles algo inferiores a los que estas áreas han obtenido cuando se han manejado por separado los listados obtenidos de las fuentes. Los porcentajes de no descriptores están algo más distribuidos entre las mencionadas categorías que los correspondientes a los descriptores, pero también con un alto grado de concentración frente al resto de grupos.

A continuación se detallan ejemplos de nuevos descriptores del TByD agrupados por sub-áreas dentro de cada grupo jerárquico, si bien se exceptúan los casos de nombres propios o servicios web para no alargar innecesariamente la relación:

» En Biblioteconomía se añaden formas de servicios de préstamo (autopréstamo, préstamo digital), formas de promoción cultural (talleres de lectura) y de vías de acceso al documento bibliotecario (acceso libre a los fondos).
» En Ciencias y técnicas auxiliares destacan términos sobre acceso y uso de información (acceso a la información pública, datos abiertos), el derecho sobre información y la conservación de documentos (copyleft, gestión de derechos digitales, preservación digital).
» En Estudios métricos de la información se incorporan términos sobre análisis de datos en la web (altmetría, analítica web), colaboración científica (análisis de redes sociales, data sharing), análisis de citas (índice G, índice H).
» En Fuentes de información se incluyen términos sobre documentos sonoros (audiolibros), formatos de libros (libros en braille), tipología de la literatura gris (trabajos académicos) y diversos índices de citas (Journal Citation Reports).
» Lenguajes y lingüística se incrementa con nuevos términos relativos a la indización y tipos de vocabularios (etiquetas, folksonomías).
» En Proceso documental se incorporan estándares de catalogación (Resource Description and Access), formatos de registros bibliográficos (MARC21) y nuevas formas de catalogación (catalogación social).
» En Profesionales de la información se incorporan novedades sobre perfiles adaptados al entorno laboral (community managers, curadores de contenido),

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 37 (diciembre 2017)

vías de formación profesional y universitaria (educación abierta, e-learning, campus virtuales) y formas de consumo de información en el entorno electrónico (lectura digital, reutilización de información).

» En Sociedad de la información se incorporan términos sobre la administración electrónica (gobierno abierto, transparencia de la información), la imagen en la red (reputación online, identidad digital) y distintas formas de colaboración (micromecenazgo). También se recoge terminología sobre la actividad científica y la industria de la información (ciencia abierta, e-ciencia).

» En TIC se incluyen sistemas operativos (Ubuntu, Android), lenguajes de programación (PHP, Python), buscadores (buscadores académicos, buscadores especializados), esquemas de metadatos (METS, MODS) y sistemas de gestión de contenidos (Drupal). Se identifican también elementos de web social (social media, marcadores sociales), gestión de contenidos (curación de contenidos, gestión de comunidades), acceso abierto (ruta verde, ruta dorada), y arquitectura de la información (microformatos, esquemas de metadatos).

» Las Unidades de información tienen escasas novedades sobre archivos y bibliotecas (bibliotecas sin libros, archivos parlamentarios), además de ejemplos de bibliotecas digitales (Google books).

Categorías del Tesauro	**2ª ver.**	**3ª ver.**	**diferencia**
Archivística	71	68	-3
Biblioteconomía	50	56	6
Ciencias y técnicas auxiliares	74	83	9
Estudios métricos de la información	43	49	6
Fuentes de información	162	165	3
Lenguajes y lingüística	61	62	1
Museología	40	39	-1
Proceso documental	98	100	2
Profesionales de la información y usuarios	95	117	22
Sociedad de la información	32	58	26
Tecnologías de la información y las comunicaciones	267	337	70
Unidades de información	122	125	3

Tabla 3. Número de descriptores por categoría en segunda y tercera versión del TByD

En la tabla 3 se registra variación en el número de descriptores por categoría entre la versión de 2017 y la de 2005. En la mayoría de los grupos, las diferencias no alcanzan la cantidad de términos nuevos aceptados como descriptores recogidos en la tabla 2, e incluso en dos de estas disminuye. Ello es debido a los cambios que se realizaron sobre 37 términos, bien por una reubicación dentro de la estructura jerárquica o bien por haberse modificado el status de los términos.

	1ª ver. (2002)	**2ª ver. (2005)**	**3ª ver. (2017)**
Términos	1.152	1.514	1.773
Descriptores	913	1.117	1.322
No descriptores	239	397	451
Notas de alcance	36	96	115
Relaciones asociativas	1.594	2.403	2.784

Tabla 4. Comparativa de cantidades de las tres versiones del TByD.

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 37 (diciembre 2017)

A modo de resumen se muestra en la tabla 4 los datos sobre el crecimiento que ha experimentado los principales aspectos del TByD en su última actualización.

El contenido completo de la tercera versión del TByD se encuentra disponible para su consulta en http://vocabularyserver.com/tbyd/index.php.

4. Conclusiones

Este estudio refleja una aportación metodológica novedosa: la extracción, con técnicas de web scraping, de términos especializados procedentes de blogs para actualizar la terminología de tesauros, y de forma más concreta, para poner al día el contenido del Tesauro de Biblioteconomía y Documentación. De la misma manera, se realiza una comparación entre las palabras clave empleadas en los campos de descriptores e identificadores de la base de datos ISOC-BD y las etiquetas o materias asignadas a los posts de 127 blogs sobre Biblioteconomía y Documentación elaborados por especialistas, con el fin de comprobar cuál es mejor fuente de terminología para actualizar el contenido del TByD.

Todo esto implica una nueva aportación a la tradicional dicotomía entre el lenguaje libre y los lenguajes controlados, pues se realiza una comparación entre conjuntos de etiquetas en lenguaje libre (folksonomías) y listados de palabras clave en lenguajes controlados, a fin de analizar la verdadera capacidad de actualización y renovación que puede tener cada uno de estos tipos de vocabulario en el tesauro que nos ocupa.

Los resultados obtenidos en la comparación realizada demuestran que las etiquetas asignadas por autores al contenido de los blogs (categoría de etiquetado social) aportan una terminología más variada y actualizada que las palabras clave asignadas a artículos científicos por documentalistas en la base de datos ISOC-BD, aun a pesar de que, en nuestro caso, el conjunto original de etiquetas manejado era menor que el de descriptores e identificadores de dicha base de datos (4.377 frente a 11.536 descriptores y 6.575 identificadores). Si se evalúan ambas formas de representación del conocimiento desde el punto de vista económico (costes de personal, elaboración y mantenimiento del software) y de acuerdo a las labores de obtención de términos, el etiquetado en blogs también da un mejor resultado que la indización realizada por especialistas en la mencionada base de datos. Las tareas de indización en bases de datos y su mantenimiento suponen un elevado coste económico, mientras que en la asignación de etiquetas por los autores este es prácticamente nulo. Si bien, en este último caso, habría que añadir el gasto que supone el software empleado para la extracción de etiquetas.

Si nos fijamos en la distribución por ámbito de los nuevos conceptos del TByD, se comprueba que el mayor incremento terminológico se ha dado en las categorías relacionadas con la web y su desarrollo tecnológico: la web 2.0 o web social. Dicho desarrollo ha producido una nueva forma de uso de la Red, así como una adaptación de los entornos laborales y formación de profesionales de la información, junto con un cambio de paradigma del acceso, uso y comercialización de información.

En cuanto a la capacidad de aportación de las novedades terminológicas para cada una de las categorías del TByD en base al origen de los vocablos manejados, la conclusión es que las etiquetas de los blogs solo destacan en cinco de las doce áreas del tesauro (Fuentes de la información, Lenguajes y Lingüística, Profesionales de la información y Usuarios, TIC y Unidades de información), ya que en las otras las cantidades aportadas son muy similares a las conseguidas en la base ISOC-BD.

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 37 (diciembre 2017)

La menor presencia de descriptores nuevos en la base ISOC-BD, frente a la gran diversidad de etiquetas aportadas por los blogs, no le resta valor como fuente de terminología para actualizar el Tesauro de Biblioteconomía y Documentación. Por lo tanto, creemos que tanto los conjuntos de etiquetas de autores en lenguaje libre como los listados de descriptores e identificadores de la base de datos ISOC-BD son buenas fuentes de nuevos términos, y por ello deben emplearse de forma complementaria para las labores de actualización y mantenimiento del TByD.

ISSN 1514-8327 (impresa) / ISSN 1851-1740 (en línea)
Información, cultura y sociedad / 37 (diciembre 2017)

Referencias bibliográficas

» Abadal, Ernest; A. Estivill; J. Franganillo; J. Gascón y J. M. Rodríguez Gairín. 2005. L'accés multilingüe per matèries a articles de revista. En *La dimensión humana de la organización del conocimiento*. Congreso del capítulo español de ISKO (5: 2005: Barcelona). Barcelona: Universidad de Barcelona. p. 33-50.

» Alonso Soriano, Luis. 2013. *Etiquetado social como fuente terminológica para el mantenimiento de vocabularios: Análisis aplicado al Tesauro de Biblioteconomía y Documentación del CINDOC*. TFM presentado en el Máster Universitario en Bibliotecas y Servicios de Información Digital. Universidad Carlos III. MS. 70 p.

» Araujo, Lourdes y J. R. Pérez-Agüera. 2006. Enriching thesauri with hierarchical relationships by pattern matching in dictionaries. En *FinTAL: International Conference on Natural Language Processing*. (5th: 2006: Turku). p. 268-279. < http://eprints.rclis.org/8351/> [Consulta: 19 Junio 2017].

» Arnold, Patrick y E. Rahm. 2014. Extracting Semantic Concept Relations from Wikipedia. En *WIMS'14. Proceedings of the International Conference on Web Intelligence, Mining and Semantics*. (4th: 2014: Thessaloniki). <http://dbs.uni-leipzig.de/file/wims2014-final.pdf> [Consulta: 19 Junio 2017].

» Cording, Patrick Hagge. 2011. *Algorithms for Web Scraping*. Lyngby: Technical University of Denmark. <http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/6183/pdf/imm6183.pdf> [Consulta: 14 Junio 2017].

» García-Silva, Andrés; O. Corcho; H. Alani y A. Gómez-Pérez. 2012. *Review of the state of the art: Discovering and Associating Semantics to Tags in Folksonomies*. En *The Knowledge Engineering Review*. Vol. 27, no. 1, 57-85. <http://oa.upm.es/6376/1/Review_of_the_state.pdf> [Consulta: 19 Junio 2017].

» International Standard Office (ISO). 1986. *ISO 2788:1986. Guidelines for the establishment and development of monolingual thesauri*. Ginebra: International Standard Office.

» International Standard Office (ISO). 2011. *ISO 25964-1:2011. Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval*. Ginebra: International Standard Office.

» Limpens, Freddy; F. Gandon y M. Buffa. 2009. *Linking Folksonomies and Ontologies for Supporting Knowledge Sharing: a State of the Art*. <https://hal.inria.fr/hal-00530371/document> [Consulta: 14 Junio 2017].

» Masó-Marema, Gemma y M. Sebastià-Salat. 2013. The integration of folksonomies within a thesaurus in a social science Web portal: SIDBRINT. En *Information research*. Vol. 18, no. 3. <http://InformationR.net/ir/18-2/paperS04.html> [Consulta: 19 Junio 2017].

» Mochón Bezares, Gonzalo y A. Sorli Rojo. 2002. *Tesauro de biblioteconomía y documentación*. Madrid: Consejo Superior de Investigaciones Científicas.

» Rodríguez Yunta, Luis. 2009. Etiquetado libre frente a lenguajes documentales. Aportaciones en el ámbito de biblioteconomía y documentación. En *Nuevas perspectivas para la difusión y organización del conocimiento*. Congreso ISKO-España. (9: 2009: Valencia). Valencia: Universidad Politécnica. p. 832-845. <http://eprints.rclis.org/15836/1/Comunicacion_Luis_RYunta_ISKO2009.pdf> [Consulta: 19 Junio 2017].

» Vállez, María; R. Pedraja-Jiménez; L. Codina; S. Blanco y C. Rovira. 2015. Updating controlled vocabularies by analysing query logs. En *Online Information Review*. Vol. 39, no. 7, 870-884.

» Vera Baceta, Miguel Ángel. 2013. *Aproximación a la BIBLOGSFERA española: Composición, autoría, estructura, contenidos y definición*. Trabajo académico presentado en la Universidad de Murcia. MS. 80 p. <https://digitum.um.es/xmlui/handle/10201/36967> [Consulta: 19 Junio 2017].

» Vera Baceta, Miguel Ángel. 2015. *Biblogsfera: Comunidad de Blogs relacionados con la Biblioteconomía y la Documentación*. <http://biblogsfera.com/> [Consulta: 19 Junio 2017].

» Wang, Jin. 2006. Automatic thesaurus development: Term extraction from title metadata. En *Journal of the American Society for Information Society and Technology*. Vol. 57, no. 7, 907-920.