Revista Española de Documentación Científica,
33, 4, octubre-diciembre, 643-663, 2010
ISSN: 0210-0614. doi: 10.3989/redc.2010.4.763

NOTAS Y EXPERIENCIAS / *NOTES AND EXPERIENCES*

Tesauros en acceso abierto en Internet. Un análisis cuantitativo

Gonzalo Mochón Bezares*, Angela Sorli Rojo**

Resumen: Se realiza un análisis cuantitativo de tesauros de acceso abierto en diversas lenguas europeas para comprobar las tendencias en su edición, así como su grado de adaptación a las directrices de construcción de tesauros. Los parámetros analizados en cada tesauro son el nombre de la institución o persona editora y su ubicación geográfica, el número de términos, las principales materias tratadas, los modos de consulta, la última fecha de actualización, el nivel de contenido de la introducción y la ayuda en línea al usuario. Los resultados indican un predominio de los tesauros editados en Europa Occidental; que el inglés es la lengua más usada; que las ciencias sociales tienen el mayor número de tesauros entre las áreas de conocimiento observadas; que sólo una tercera parte de los tesauros dispone de un motor de búsqueda; que hay un nivel medio de contenido muy bajo en las introducciones; y ausencia de ayuda al usuario en numerosas páginas. En conclusión, gran cantidad de tesauros en acceso abierto no aprovechan las ventajas que ofrece Internet para la consulta de sus contenidos, y muchos de ellos deberían mejorar sus interfaces.

Palabras clave: Tesauros abiertos en línea, Internet, evaluación.

Publicly available Thesauri on the Internet. A quantitative analysis

Abstract: *This work offers a quantitative analysis of publicly available thesauri in different European languages in order to check publication trends and their degree of adherence to thesauri construction guidelines. The following parameters were analysed: publisher's name and location, number of terms, subject coverage, query modes, latest update, level of content in introduction and on-line help. The results reveal the prevalence of thesauri published in Western Europe; English is the most frequently used language; social sciences have the highest number of thesauri among the knowledge areas observed; only one third of the thesauri has a search engine; on average, there is a poor level of content in the introductions and in the on-line help. In conclusion, many publicly available thesauri do not take advantage offered by the Internet for querying their content and should improve their interfaces.*

Keywords: *publicly available thesauri, Internet, evaluation.*

* Aspy System, Madrid, España. Correo-e: gomobez@yahoo.es.
** IEDCYT, CSIC, Madrid, España. Correo-e: angela.sorli@cchs.csic.es.
Recibido: 16-11-2009; 2.ª versión: 12-2-2010; aceptado: 16-2-2010.

1. Introducción

Desde su desarrollo en las décadas de los cincuenta y los sesenta, los tesauros han sido herramientas útiles para el almacenamiento y recuperación de la información, y han gozado de cierto prestigio entre los profesionales de la documentación de todo el mundo. La rápida expansión de la World Wide Web ofreció, y continua ofreciendo, una estupenda oportunidad para la difusión de los tesauros en formato electrónico. La adaptación de los tesauros al entorno web supone una simplificación de su manejo gracias los enlaces hipertextuales, un abaratamiento en los costes de actualización, permite la integración del usuario en los procesos de creación y gestión; y ofrece la posibilidad de reutilización e interoperabilidad entre distintos formatos, como la que ofrece SKOS-Core (De la Cueva, 1999; Shiri y Revie, 2000; Arano, 2005). Otros autores señalan una serie de inconvenientes o limitaciones como son el alto coste de mantenimiento y de actualización, la dificultad de adición de relaciones entre los términos, y la falta de mecanismos que relacionen los términos con objetos (Moreiro, 2007; García-Jiménez, 2002).

Frente a la limitación del uso de tesauros en entornos muy concretos como las bases de datos temáticas y los sitos web, hay autores que creen necesaria una mayor flexibilidad de estas herramientas para su adaptación a la recuperación documental en Internet (García-Marco, 2008; Rodríguez, 2009), y para su manejo por usuarios sin grandes conocimientos de lenguajes documentales (Dalbin, 2007).

En distintas bases de datos (LISA, ISOC, ISTA, INSPEC y Francis) y conjuntos de repositorios como Open DOAR (Directory of Open Access Repositories) y Registry of Open Access Repositories (ROAR), se ha observado una carencia de trabajos de carácter cuantitativo sobre tesauros en formato electrónico en la literatura científica. La falta de un estudio de este tipo ha motivado la realización de este trabajo sobre tesauros de distintas materias en acceso abierto en Internet, que tiene como fin conocer las principales características de los mismos en lo que se refiere a los distintos formatos de presentación, las categorías temáticas, su ubicación geográfica, el nivel de actualización y la forma de edición de los tesauros. Los datos obtenidos para cada una de las características se han cruzado con los pertenecientes a otras variables para poder comprender mejor la situación actual de los tesauros de acceso gratuito en Internet y comparar las tendencias que marcan los datos obtenidos. Por el contrario, el objetivo de este artículo no es evaluar el contenido semántico de estos tesauros, ni observar el grado de eficacia de los mismos en la realización de consultas en las bases de datos para los que han sido creados.

2. Metodología

La obtención de datos para este trabajo se ha basado en la descripción de ciento setenta y dos tesauros en acceso abierto en Internet en lengua inglesa, francesa, española, alemana, italiana y portuguesa, realizada por los autores, y

publicada en diversos artículos de la *Revista Española de Documentación Científica* entre los años 2007 y 2009 (Mochón, 2007a; 2007b; 2008a; 2008b; 2008c; 2009). Estos tesauros se recuperaron de los resultados de las consultas realizadas con cada uno de los siguientes términos: «tesauro», «tesauros», «thesaurus» y «thesauri», en las versiones avanzadas de los motores de búsqueda Google y Yahoo, y que fueron filtradas por cada uno de los países en los que se utilizan alguna de las lenguas señaladas más arriba. Además, para poder completar la consulta se utilizaron distintos directorios de tesauros disponibles en Internet. A partir del conjunto de lenguajes controlados obtenidos en su día, se ha procedido a una revisión minuciosa de los mismos, eliminando todos aquellos tesauros desaparecidos o que no ha sido posible localizar, y actualizando los enlaces rotos de tesauros todavía existentes. Tras suprimir doce títulos del listado original por no poder ser localizados, el conjunto final de tesauros quedó reducido a ciento sesenta y cuatro.

Sobre este conjunto de tesauros se ha elaborado un análisis cuantitativo de los tesauros en acceso abierto en Internet atendiendo a diversos indicadores como son la ubicación geográfica de los sitios web en los que se encuentran los tesauros; las instituciones encargadas de la edición y/o el mantenimiento; la cobertura temática; la disponibilidad del contenido por idiomas; fecha de última actualización del contenido, el número total de términos declarado, los objetivos para los que fueron concebidos los tesauros, las modalidades de consulta del contenido, y la existencia de introducción y ayuda al usuario. Dichos indicadores han sido tomados en parte de trabajos sobre evaluación de tesauros (Álvaro y otros, 1989b), de directrices sobre construcción de tesauros y vocabularios controlados (ANSI, 2005; AENOR, 1990), y de obras teóricas sobre la materia (Naumis, 2007), junto con otros de elaboración propia que se consideraron importantes a durante el análisis realizado a las páginas web en las que se presentan los tesauros.

Todos los datos que se han utilizado en este estudio han sido extraídos de las propias páginas web en las que están albergados los tesauros y han sido contrastados entre los meses de julio y septiembre de 2009.

3. Resultados

3.1. Distribución por países

Con el fin de conocer la distribución geográfica de los tesauros, estos se presentan agrupados por grandes áreas geográficas y detallando en cada una de ellas el número elaborado en cada país junto con información procedentes de otras variables examinadas:

1. El *continente europeo* es el área geográfica con mayor número de tesauros: sesenta y ocho. La temática de los mismos es variada, y los organis-

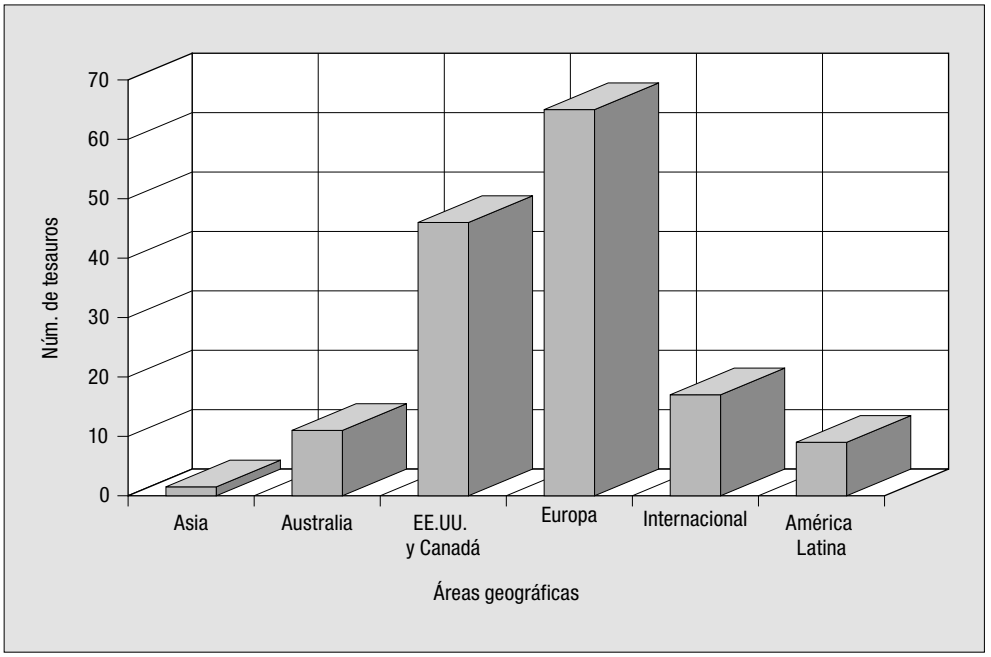
mos encargados de su edición y mantenimiento son, en su mayoría, organismos de la administración pública y organismos de investigación. La relación de países, atendiendo al número de tesauros producidos, es como sigue:

- *España*. Presenta un elevado número de tesauros en acceso abierto: veintidós. Entre los mismos destacan los diez tesauros editados por el CINDOC, del CSIC, que tratan principalmente sobre Humanidades y Ciencias Sociales. También hay que recordar aquellos tesauros editados por universidades españolas (seis) y distintos organismos de la Administración Publica (cinco), todos ellos de temática variada.
- *Alemania*. Sorprende el escaso número de tesauros disponibles, seis, en las páginas web de instituciones alemanas. En su mayoría han sido editados por facultades y escuelas universitarias. En cuanto a su ámbito temático, hay que destacar los referidos a Documentación.
- *Francia*. Este país ofrece un alto número de tesauros en línea de acceso abierto, dieciocho. Entre las materias de los tesauros disponibles destacan las Ciencias de la Salud, con ocho vocabularios controlados, elaborados principalmente por organismos ministeriales e institutos de salud publica. También se pueden encontrar, aunque en menor número, tesauros sobre Humanidades y Ciencias Sociales editados por organismos ministeriales e institutos de investigación.
- *Reino Unido*. De los doce tesauros en acceso abierto existentes en el Reino Unido, es importante destacar la presencia de los relativos a las Humanidades (ocho). Los museos y las organizaciones dedicadas a la recuperación del patrimonio cultural, además de las universidades, son las principales instituciones editoras de tesauros en este país.
- *Otros países*. Los diez tesauros restantes de la zona europea se localizan en los siguientes países: Países Bajos con tres tesauros; Italia y Suiza, con dos cada uno; y Finlandia, Portugal y Noruega, con uno, cada uno. Entre estos países se observa una clara vocación por los tesauros sobre Ciencias Sociales, y la edición de los mismos se debe en su mayoría a Centros de Investigación y Universidades.

2. *Asia*. El Tesauro Irandoc, sobre Ciencias Naturales, editado por el Iranian Research Institute for Scientific Information and Documentation, es el único tesauro elaborado en el continente asiático que ha sido incluido en el presente estudio. Presenta una versión en persa y otra en inglés.
3. *Australia*. Los catorce tesauros editados en este país tratan fundamentalmente sobre Humanidades, Ciencias Sociales y Ciencias de la Salud. Estos han sido elaborados por entidades de la administración pública, tanto de ámbito estatal como local, y por la Biblioteca Nacional.
4. *Internacionales*. Los diecinueve tesauros que componen este conjunto han sido elaborados por organismos supranacionales, en especial por

FIGURA 1

Número de tesauros por grandes áreas geográficas



organismos dependientes de Naciones Unidas y de la Unión Europea. Nueve de estos tesauros tratan sobre distintas Ciencias Sociales, dos están dedicados a las Ciencias de la Salud o el resto son multidisciplinares.

5. *América del Norte*. Esta zona del continente americano tiene una nutrida representación de tesauros en acceso abierto (cuarenta y ocho), pero no llega a superar al conjunto de países europeos. La distribución de tesauros por países es la siguiente:

   - *Estados Unidos*. Es el país del mundo que dispone de mas tesauros en acceso libre en Internet, treinta y ocho. Entre estos se encuentran representadas todas las áreas del conocimiento, destacando en especial las Ciencias de la Salud con once tesauros y las Ciencias Sociales con diez. La elaboración y/o edición de estos lenguajes controlados se debe, entre otros tipos de instituciones, a Universidades (once), Bibliotecas Nacionales (siete), Institutos y Agencias Nacionales (diez) y Asociaciones privadas (cinco).

     Asimismo, se puede señalar que en esta nación se encuentran los tesauros de mayor extensión del mundo: el Art and Architecture Thesaurus y el Thesaurus of Geographic Terms, ambos elaborados por la Fundación Paul Getty.

- *Canadá*. Los diez tesauros canadienses de acceso libre por Internet, elaborados en su práctica totalidad por organismos de la administración pública, se incluyen dentro de las áreas de Ciencias de la Salud y Ciencias Sociales.

6. *América Latina*. Esta zona geográfica es una de las que menos tesauros en acceso abierto recoge: doce. De entre los países de esta área destaca Brasil.
   - *Brasil*. Es el país de América Latina que presenta un mayor número de tesauro, cinco. La responsabilidad editorial recae sobre distintos Institutos Nacionales y Universidades. La temática preponderante entre estos tesauros es Ciencias Sociales.
   - *Otros países*. Varios son los países latinoamericanos que tienen tesauros en Internet: Argentina y Costa Rica, con dos tesauros cada uno, y Chile, México y Colombia con un tesauro cada uno. Distintos organismos de la Administración Publica y Universidades son los editores de estos vocabularios. Las áreas del conocimiento tratadas son las Ciencias Sociales y las Humanidades.
7. *Países desconocidos*. En dos tesauros analizados, uno sobre Humanidades y otro sobre Ciencias de la Salud, ha resultado imposible averiguar en que países se ubican sus entidades editoras.

3.2. Entidades

Conocer la autoría de un tesauro puede resultar algunas veces una tarea bastante difícil. Sin embargo, no ocurre lo mismo cuando se trata de conocer la entidad responsable de su edición. En numerosas ocasiones, las entidades editoras figuran como responsables únicos de los tesauros, siendo también los principales, y a veces únicos, destinatarios de los mismos.

TABLA I

Números de tesauros por tipo de institución editora

Tipos de institución	Número
Bibliotecas	16
Centros de documentación	8
Institución desconocida	4
Entidades supranacionales	21
Organismos de la Administración Pública	47
Organismos de investigación	23
Universidades	33
Varios	12

Dada la dificultad mencionada, se opta por recuperar las entidades editoras de los tesauros con el fin de agruparlas por tipos de institución. Los tipos de entidad obtenidos son los siguientes: bibliotecas, centros de documentación, entidades supranacionales, organismos de la administración pública, organismos de investigación, universidades, varios y desconocidos.

Los resultados obtenidos, incluidos en la tabla I, se han comparado con los de otras variables contempladas en el estudio para intentar comprender mejor la información obtenida:

- *Bibliotecas*. Se han contabilizado dieciséis bibliotecas como responsables de la edición de tesauros en acceso abierto. La mayor parte de los cuales son lenguajes controlados elaborados por distintas bibliotecas sectoriales nacionales de Estados Unidos y Australia, con el fin de normalizar los vocabularios de las disciplinas que les son propias, y que en muchos casos tienen también como fin las labores de indización de bibliotecas y centros de documentación sectoriales de menor entidad.
- *Centros de Documentación*. Es un tipo de adscripción con baja representatividad, solamente ocho de los tesauros analizados han sido creados o editados por centros de documentación. La presencia de centros de documentación como responsables de la edición/elaboración de tesauros es mayor en las áreas de Ciencias Sociales y Humanidades.
- *Entidades Supranacionales*. Bajo este epígrafe se reúnen veintiún tesauros en línea, la mayor parte de los cuales son de gran extensión, elaborados o editados por organizaciones como la ONU, OIT, FAO, UNESCO, o la Comisión Europea, entre otras. Las áreas temáticas representadas en los tesauros de estas instituciones son variadas, aunque hay una mayoría de carácter multidisciplinar.
- *Organismos de la Administración Pública*. Bajo esta denominación un tanto abstracta se agrupan todas las entidades de las administraciones públicas de ámbito estatal, regional y local, responsables de la elaboración y/o edición de tesauros en línea de acceso abierto. Se trata del grupo más numeroso de entidades editoras (cuarenta y siete) observado. Este tipo de entidad tiene una presencia más destacada en países como Estados Unidos, Francia, Australia o Canadá.
- *Organismos de Investigación*. Son entidades de carácter no universitario dedicadas a la investigación, en su mayoría de carácter público aunque también se han encontrado algunas de carácter privado. Se han recogido veintitrés tesauros elaborados por organismos de investigación, entre los que destacan los diez tesauros elaborados por el antiguo Centro de Información y Documentación Científica (CINDOC), del CSIC, y mantenidos por su sucesor, el Instituto de Estudios Documentales sobre Ciencia y Tecnología (IEDCYT).
- *Universidades*. Este epígrafe engloba a treinta y tres tesauros editados por distintos departamentos universitarios, y también aquellos realizados por

alumnos o doctorandos con el fin de presentarlos como trabajos de curso, o para la obtención de un grado académico. Se observa un alto número de tesauros editados y/o elaborados por universidades norteamericanas (nueve), alemanas (cinco), españolas (seis) y británicas (cinco). Por el contrario, se aprecia un número escaso de tesauros editados por universidades francesas (dos).

- *Varios*. Bajo este epígrafe se recogen doce tesauros elaborados y editados bien por particulares no vinculados a una institución universitaria o de investigación, bien por empresas privadas, por fundaciones privadas y organizaciones no gubernamentales. Estos tesauros son de temática variada y la mayoría de los mismos se encuentran en páginas web de los Estados Unidos.
- *Desconocidos*. En este apartado se incluyen todos aquellos tesauros que no proporcionan información suficiente sobre la institución responsable de su edición en internet. Solamente cuatro tesauros no informan sobre la entidad que es responsable de su edición electrónica.

3.3. Materias

La temática de los tesauros analizados se ha agrupado mediante una clasificación «ad hoc» en las siguientes áreas de conocimiento: Humanidades, Ciencias Naturales, Ciencias Sociales, Ciencias de la Salud, Ciencias Experimentales y Tecnologías de la Información y las Comunicaciones. La asignación a un área de conocimiento concreta se ha realizado en base a la materia o materias tratadas en los conjuntos de términos de cada tesauro. En el caso de aquellos tesauros que tratan sobre más de una materia, se ha procedido de la siguiente manera: en los casos de los tesauros que tratan sobre materias dispares se ha optado por incluirlos en la categoría de «Multidisciplinares», mientras que los tesauros que tratan sobre dos o más materias afines se han incluido en la categoría correspondiente a las materias tratadas.

Al contrastar los datos obtenidos correspondientes a las áreas del conocimiento con otros parámetros observados en el análisis de los tesauros estudiados, se obtiene la siguiente información sobre los mismos:

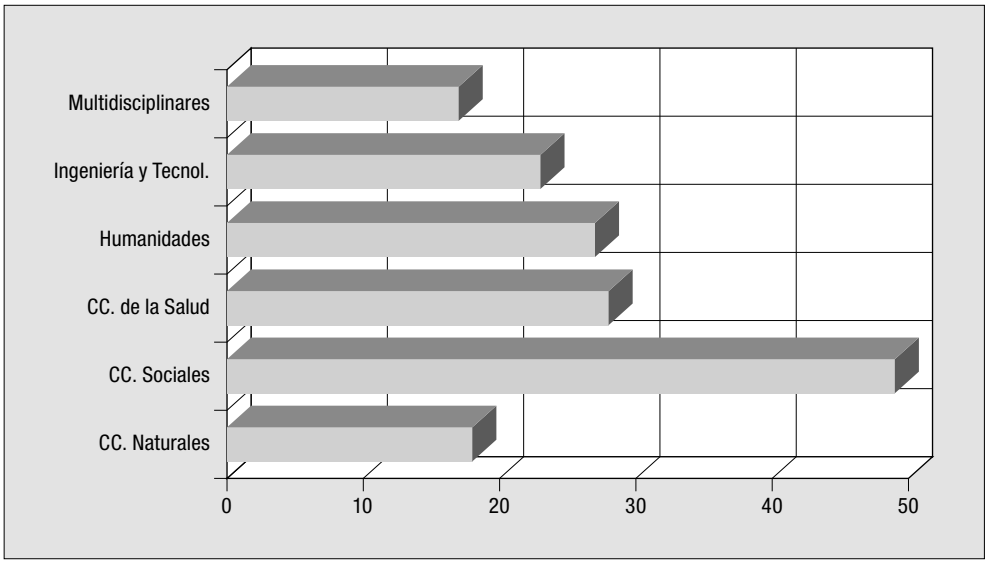
- *Ciencias Naturales*. Este grupo incluye diecinueve tesauros que tratan, entre otras disciplinas, de Física, Química, Biología, Geología y Medio Ambiente. Este conjunto de vocabularios controlados destaca por el número de obras editadas por organismos de investigación, entidades privadas y universidades de Estados Unidos.
- *Ciencias Experimentales y Tecnologías de la Información y las Comunicaciones*. Se trata de un grupo heterogéneo en el que se incluyen veinticuatro tesauros sobre Matemáticas, Ciencias de la Información (Information Science), Informática, Ingeniería, Astronomía y Defensa. Dentro de este conjun-

to destaca el número de tesauros elaborados en España y Canadá, con cinco y seis vocabularios respectivamente. Al cruzar los datos de este área del conocimiento con los relativos a los del tipo de institución, se observa un claro predominio de los elaborados o editados por universidades.

- *Ciencias de la Salud*. Incluye veintiocho tesauros sobre distintas especialidades de Medicina, Farmacia y Ética de la Salud, en su mayoría en lengua francesa o inglesa. Dentro de esta área temática, se pueden destacar dos subgrupos de tesauros: uno de lenguajes elaborados o editados por organismos públicos de Francia y Canadá (7 tesauros en francés), y otro formado por lenguajes debidos a universidades e instituciones públicas norteamericanas (diez tesauros en inglés). Se puede señalar también la importancia de la obra «Medical Subject Headings» (MeSH) de la National Library of Medicine de Estados Unidos, de la cual el Institut National de la Santé et de la Recherche Médicale (INSERM) ha elaborado una versión en francés, y el Centro Latinoamericano y del Caribe de Información en Ciencias de la Salud (BIREME), una versión en portugués. El tesauro MeSH también ha servido como fuente de la cual se han extraído términos para numerosos tesauros sobre ciencias de la salud.
- *Ciencias Sociales*. De entre todas las áreas temáticas recogidas en esta clasificación, la de Ciencias Sociales es la que reúne mayor número de tesauros (cuarenta y nueve), dado el amplio rango de materias que abarca. Dentro de este grupo destaca el número de obras editadas en países como

FIGURA 2

Número de tesauros agrupados por grandes áreas de conocimiento



Estados Unidos (diez), España (ocho), especialmente las elaboradas por organismos de investigación, y las elaboradas por instituciones supracionales (nueve).

- *Humanidades*. Bajo esta denominación se incluyen veintisiete tesauros sobre materias tales como Historia, Arte, Arqueología, Música, Cine, Museología, Archivística y Literatura. Los países con mayor número de lenguajes controlados relativos a las materias de este grupo temático son España, Francia y Reino Unido, cada uno con cinco tesauros. Por otra parte, sorprende el bajo número de tesauros editados por instituciones norteamericanas (dos).
- *Multidisciplinares*. Bajo este epígrafe se recogen, como ya se ha señalado más arriba, los tesauros que tratan disciplinas heterogéneas. En este grupo se incluyen diecisiete tesauros, entre los que destacan aquellos elaborados o editados por instituciones supranacionales (cuatro) e instituciones de la administración pública de Estados Unidos (cuatro) y Australia (cuatro).

3.4. Idiomas

A pesar de la limitación de la búsqueda, ya señalada en la introducción, a las lenguas alemana, española, francesa, inglesa, italiana y portuguesa, se han podido contabilizar hasta treinta y siete lenguas de edición diferentes en los tesauros recuperados, tal y como se puede observar en la tabla II. Este elevado número de idiomas se debe a la presencia de treinta y cinco tesauros que tienen su contenido disponible en más de una lengua (dieciséis bilingües y diecinueve multilingües). Al haber contabilizado cada una de las versiones en las que están disponibles estos tesauros plurilingües como si fueran tesauros independientes, se puede crear una falsa imagen sobre la existencia de tesauros en acceso abierto en determinados países asiáticos y europeos.

TABLA II

Número de tesauros por idioma

Idiomas	Número de tesauros
Alemán	17
Español	47
Francés	47
Inglés	106
Italiano	12
Portugués	14
Otros idiomas	71

Teniendo presente el problema que supone la contabilidad de las versiones idiomáticas de los tesauros plurilingües, en el análisis realizado se obtienen los siguientes resultados:

- De los ciento sesenta y cuatro tesauros recuperados en este trabajo, ciento seis tienen disponible su contenido en lengua inglesa, de los cuales setenta y tres han sido editados por instituciones de países anglófonos.
- De los cuarenta y siete tesauros recuperados en lengua francesa, algo más de la mitad han sido elaborados por instituciones de países francófonos: Francia y Canadá, con dieciocho y diez tesauros, respectivamente. El resto de tesauros contabilizados en esta lengua se debe a versiones en francés de tesauros multilingües.
- De los cuarenta y siete tesauros en lengua española, veintidós han sido elaborados por instituciones españolas y siete por instituciones de determinados países de América Latina: Argentina, Chile, Costa Rica, Colombia y México. El resto de los tesauros disponibles en español proceden de versiones de tesauros multilingües elaborados principalmente por organismos supranacionales.
- Los tesauros disponibles en lengua alemana suman diecisiete, aunque solamente seis de los mismos han sido editados por instituciones alemanas. El resto son versiones en alemán de tesauros multilingües elaborados o editados por instituciones internacionales.
- En el caso de los tesauros disponibles en lengua italiana sucede lo mismo que con los tesauros en alemán. Se han contabilizado un total de doce tesauros en italiano, pero solamente dos han sido editados por instituciones italianas, siendo el resto versiones de tesauros multilingües.
- Los tesauros en lengua portuguesa ascienden a catorce, de los cuales cinco han sido editados por instituciones brasileñas y solamente uno se debe a una institución de Portugal. El resto de los tesauros contabilizados en esta lengua son versiones en portugués de tesauros plurilingües.

El resto de los idiomas recogidos se debe, como se ha señalado más arriba, a las versiones de los tesauros plurilingües, aunque también se han recuperado dos tesauros monolingües en lengua catalana.

En determinados tesauros multilingües observados, como el Tesauro de la UNESCO o el AGROVOC, varía el número total de términos dependiendo de la lengua en la que esté escrito el índice que se maneje. Esta situación da lugar a los llamados «tesauros no-simétricos», y es debida a la diferencia conceptual de un mismo término en distintas lenguas (IFLA, 2009).

3.5. Fecha de publicación y/o actualización

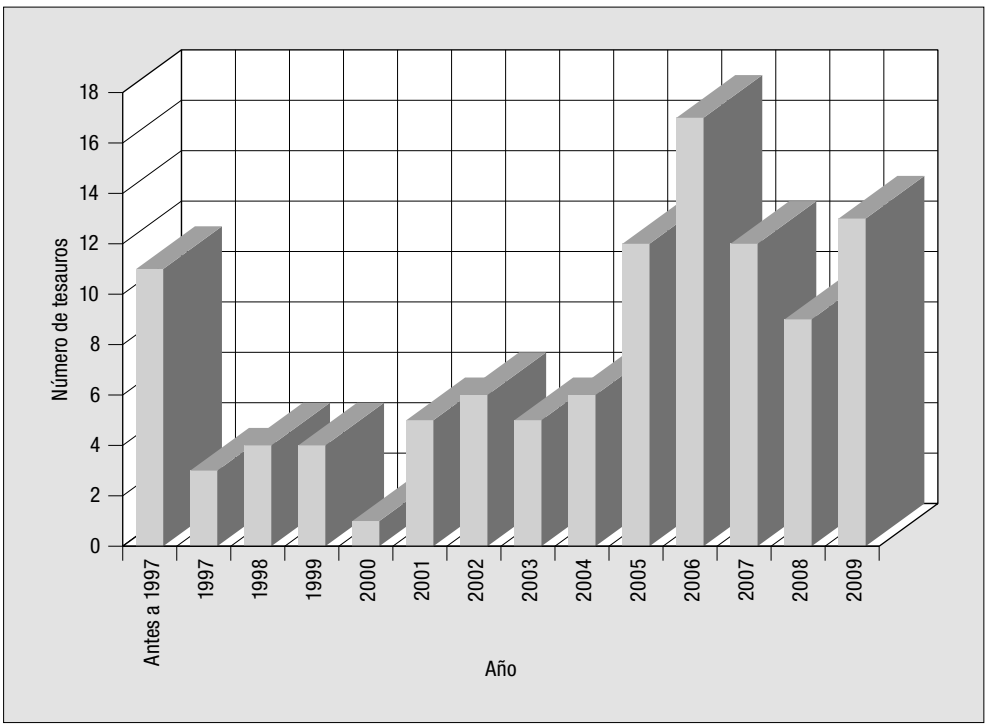
La fecha de última actualización es un elemento importante a consignar en la introducción de un tesauro, según recomiendan las directrices de construcción

de tesauros monolingües elaboradas por distintos organismos de normalización (ANSI, 2005; AENOR, 1990). La fecha de incorporación del último término o de la última actualización de un tesauro es un dato importante para conocer el grado de obsolescencia del mismo, dado que en algunas materias esto puede significar la perdida de vigencia de parte o todos los términos que contiene.

A pesar de la importancia de consignar la fecha de última actualización, cincuenta y cinco de los ciento sesenta y cuatro examinados no sólo no incluyen la citada fecha sino que ni siquiera dan información sobre la fecha de edición del tesauro (figura 3b). El número de tesauros con datación conocida se presentan en la figura 3a agrupados por tramos de anualidades.

FIGURA 3a

Número de tesauros por año de última edición o actualización

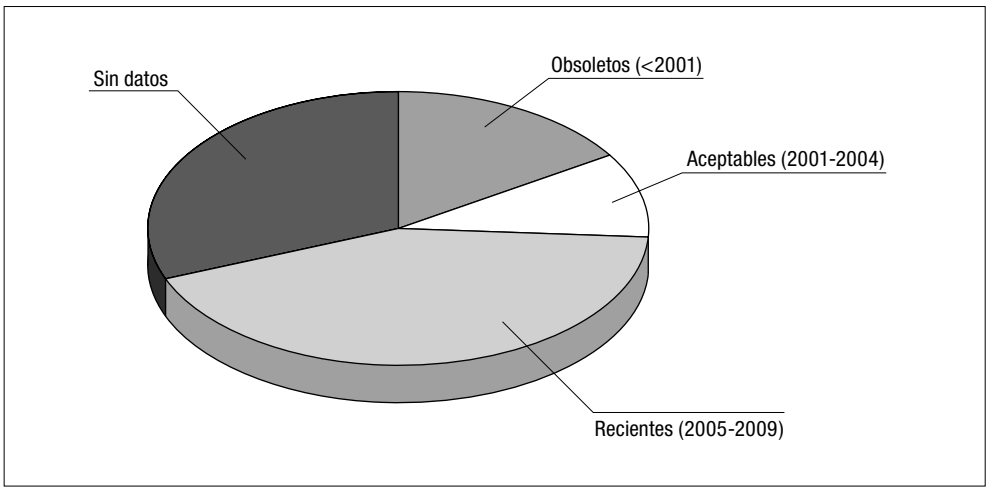


En la figura 3a se aprecia un elevado número de tesauros con una antigüedad que a primera vista se antoja excesiva. La primera barra del gráfico representa doce tesauros editados entre los años 1984 y 1996, y que en muchos casos son índices en pdf de sus versiones impresas, cuando no son directamente imágenes escaneadas de sus ediciones en papel, como es el caso del Population Multilingual Thesaurus. En las anualidades comprendidas entre los años 1997 y 2000, se

observa un bajo número de lenguajes controlados actualizados. Estos dos grupos de tesauros, cuyo contenido no ha sido actualizado en los últimos diez años, han quedado obsoletos (figura 3b), con lo que su operatividad queda muy limitada. Algunos de estos lenguajes son trabajos de alumnos universitarios o proyectos elaborados por departamentos universitarios que no han tenido continuidad, por lo que han quedado obsoletos. Los últimos cinco años representados en la figura 3a (2005-2009) recogen un total de sesenta y tres, considerados como recientes (figura 3b), de los ciento nueve tesauros con fecha de actualización conocida y aquellos que declaran una actualización constante (57,7% del total), entre los que hay una cantidad considerable de tesauros sobre Educación y Ciencias de la Salud. Además de los datos referidos a las materias, se puede comprobar que la mitad de los tesauros elaborados en Australia y Canadá han sido actualizados en el período indicado.

FIGURA 3b

Porcentaje de tesauros por nivel de actualización



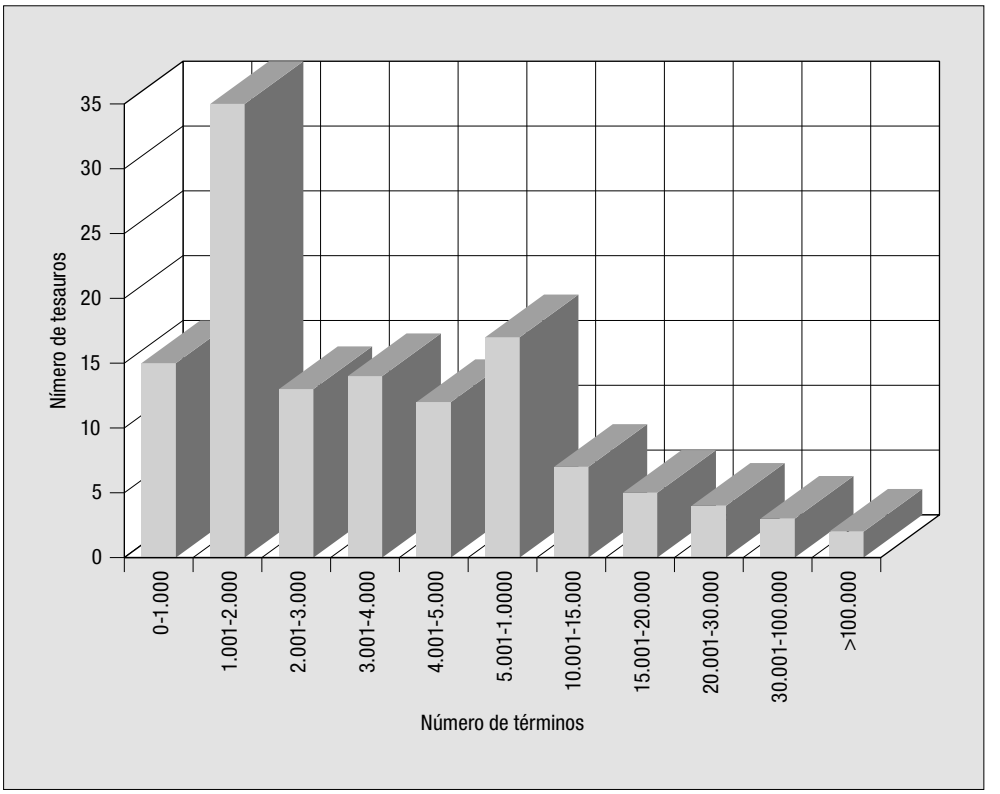
3.6. Número de términos

Las diferentes directrices de construcción de tesauros recomiendan que los autores indiquen la cifra de términos incluidos en los mismos, a ser posible desglosando la cantidad de términos preferentes y no preferentes (ANSI, 2005; AENOR, 1990). En el caso que nos ocupa, se han contabilizado treinta tesauros que no aportan ninguna información sobre la cifra de términos incluidos, y de los ciento treinta y cuatro restantes que si lo hacen, muchos no indican cual es el porcentaje de términos preferentes y no preferentes.

En la figura 4 se muestra el número de total de términos de cada tesauro agrupados en una serie no uniforme de datos: por millares hasta los 5.000 términos, por tramos de 5.000 desde los 5.001 hasta los 20.000 términos, por decenas de millar desde los 20.001 hasta los 30.000 términos; en un tramo desde 30.001 hasta 100.000 términos, y otro para los tesauros con una cantidad de términos superior a 100.000.

FIGURA 4

Número de tesauros agrupados por número de términos



Al analizar los distintos tramos por cantidades de términos junto con otras variables contempladas, se ha podido constatar lo siguiente:

- Algo menos de la mitad (sesenta y cuatro) de los tesauros cuya extensión es conocida contienen menos de 3.000 términos, siendo la banda entre 1.001 y 2.000 términos la que mayor número de tesauros incluye: treinta y cinco. En este intervalo de 1 a 3.000 términos se encuentran la mayoría de

tesauros sobre Ingeniería, Tecnologías de la Información y las Comunicaciones, Humanidades y Ciencias Naturales.

- Entre los tesauros de Ciencias Sociales hay una mayoría que supera los 4.000 términos de extensión. La presencia de tesauros de grandes dimensiones en este área del conocimiento, puede deberse a que la misma abarca ciencias con un vocabulario muy extenso como pueden ser la Economía, el Derecho o la Educación.
- La mayoría de los tesauros multidisciplinares se incluyen en la banda entre los 5.001 y los 10.000 términos. La gran extensión de este tipo de tesauros podría explicarse por su carácter holístico.
- Los tesauros sobre Ciencias de la Salud presentan una gran disparidad. Por un lado, se encuentran los tesauros de menor extensión (hasta 2.000 términos), la mayoría de ellos de carácter sectorial dentro de su área de conocimiento. Por otro, están los de gran extensión (superiores a 20.000 términos) como los Medical Subject Headings, con sus versiones en Francia y Brasil, y el Crisp Thesaurus, que son de carácter médico más general.
- Llama la atención el elevado número de tesauros, veintisiete, que superan los 10.000 términos, destacando los de Ciencias de la Salud, Educación y Geografía. En esta última disciplina encontramos el Getty Thesaurus of Geographic Terms, un tesauro con una extensión de un millón de términos, lo que le convierte en una herramienta de difícil manejo para el usuario novel.

3.7. Objetivos

El objetivo por el cual se confecciona un tesauro es la razón de ser del mismo y, por tanto, es necesario que quede claramente explicitado en su introducción (ANSI, 2005; AENOR, 1990). No obstante esta necesidad, en treinta y nueve de los tesauros analizados en este estudio no se recoge o explica el propósito de su realización en las páginas web que presentan su contenido, aunque se debe suponer que la confección de un tesauro siempre debe tener un motivo (Naumis, 2007).

La razón o propósito de la creación de la mayoría de los tesauros de acceso abierto es servir de ayuda en las labores de indización y recuperación documental, tanto en bases de datos concretas (noventa y uno tesauros) como en bases de datos no específicas (doce tesauros). Otro objetivo observado es la asignación de materias tanto en catálogos de bibliotecas o como en los distintos formatos de metadatos utilizados para la recuperación de información en los sitios web por buscadores. Con esta finalidad se han creado quince tesauros, la mayoría de ellos en lengua inglesa. Estos objetivos descritos, que en nuestra opinión son prácticamente uno, son los únicos que son citados como fines por distintos autores y organizaciones que han escrito este tema (Naumis, 2007; Murakami, 2005; Soergel, 2003, ANSI, 2005).

TABLA III

Número de tesauros por objetivo

Objetivo de los tesauros	Núm. tesauros
Indización, asignación de materias y ayuda a la recuperación documental en bases de datos o catálogos concretos.	91
Indización y ayuda a la recuperación de documentos en bases de datos no concretas.	12
Asignación de materias en formatos de mandatos para recuperación de páginas web.	15
Trabajos de curso, tesinas o tesis doctorales.	4
Otras.	3
Finalidad no declarada o desconocida.	39

Otras finalidades presentes en los tesauros analizados, aunque con un número reducido de ejemplos en cada caso, son las siguientes: presentarlos como trabajos de investigación en la universidad, ya sea como trabajos de curso, memorias de licenciatura o tesis doctorales (cuatro tesauros); utilizarlos como herramientas en las clases prácticas sobre indización documental de distintas universidades (dos tesauros); y que sirvan como herramientas de intercambio de información entre distintos repositorios en Internet (un tesauro).

Un caso aparte son los tesauros elaborados por organizaciones supranacionales, que tienen como objetivo principal servir como herramientas de lenguaje controlado a unidades de información con recursos limitados (Álvaro y otros 1989a).

3.8. Consulta del contenido

Los tesauros analizados permiten el acceso a su contenido de dos formas no excluyentes: consultando los distintos índices disponibles o utilizando los buscadores.

3.8.1. Buscadores

De los ciento sesenta y cuatro tesauros estudiados, se han recuperado noventa y uno (55,4% del total) que disponen de una herramienta de búsqueda como ayuda a la consulta de sus contenidos. Todos estos buscadores proporcionan acceso a todos los términos del tesauro, tal y como recomienda la norma ANSI/NISO Z39.19 de 2005 (ANSI, 2005). Aunque en algunas ocasiones también recuperan términos incluidos en el texto de las notas de alcance o de las notas históricas, lo que puede dar lugar a resultados erróneos.

Los principales elementos de ayuda para la consulta a través de los buscadores, son los siguientes:

- Los símbolos de truncamiento, sobre todo al final de los términos, cuyo uso está permitido en sesenta y nueve tesauros. En algunos casos el truncamiento impide la consulta por términos compuestos.
- En lo que se refiere al uso de operadores booleanos, se permite el uso del operador AND en veinticuatro de los tesauros consultados, mientras que el operador OR sólo puede utilizarse en seis de los mismos, y el operador AND NOT, únicamente en dos.
- La búsqueda exacta o búsqueda por frase se encuentra disponible en dieciséis de los tesauros, y la forma más usual es mediante el uso de las comillas.

En casi todos los tesauros que disponen de un buscador, los resultados de las consultas suelen presentarse en forma de listado alfabético en el que los términos aparecen como hiperenlaces.

Se ha observado también que seis tesauros, entre los que están el World Bank Thesaurus y el National Public Health Language, incluyen enlaces para la búsqueda en Google o Yahoo a través de sus descriptores. Operación que resulta muy poco efectiva, dada la naturaleza desestructurada de los datos presentes en la World Wide Web.

3.8.2. Índices

La presentación del contenido de los tesauros editados en papel ha venido realizándose en índices alfabéticos y sistemáticos, ignorándose en la mayoría de los casos otras formas recogidas en distintas versiones de directrices de tesauros, como es la representación grafica (ANSI, 2005; AENOR, 1990). En la tabla IV se recogen los índices por su forma de visualización y por su tipo de presentación del contenido, contabilizando en muchos casos más de un índice por tesauro.

TABLA IV

Tipos de índices por forma de visualización

Forma de visualización / Tipos de índices	En línea (HTML, XHTML, XML)	PDF	Otros formatos de text (word, rtf, txt)
Sistemático	74	27	4
Alfabético	98	38	9
Temático	6	2	—
Permutado KWIC	3	12	—
Permutado KWOC	5	7	—

Al analizar la forma de visualización de los índices hipertextuales se puede observar un alto numero de índices de tipo alfabético (noventa y ocho), considerado el mas útil para los usuarios (Naumis, 2007), y un número algo menor de tipo sistemático (setenta y cuatro). Otros formatos minoritarios son los de tipo temático (seis) y tipo permutado KWIC y KWOC (ocho).

Al considerar los índices en formato de texto, se aprecia una importante presencia del formato pdf (ochenta y seis) en todos los tipos de índices frente a la escasa presencia de otros formatos como doc, txt o rtf (trece). En muchos casos, estos índices son solamente un complemento de los índices hipertextuales.

Existen otras opciones de visualización de índices en XML (ocho) o SKOS-Core (seis), y Z-Thes (cuatro).

3.9. Ayuda al usuario e introducción

3.9.1. Ayuda al usuario

En un entorno web, lo ideal es diseñar páginas que sean intuitivas o, en su caso, proporcionar a los usuarios una explicación de aquellos elementos cuya funcionalidad no esté claramente explicita.

En el caso de los tesauros en línea se debería incluir una página de ayuda en la que se explique la función de los índices, así como de los distintos elementos que puedan servir de ayuda en la consulta de su contenido y de las distintas abreviaturas usadas.

De todos los tesauros analizados, solamente sesenta y cuatro tienen una ayuda al usuario que puede considerarse completa para un correcto manejo de los mismos. Los otros cien tesauros o bien carecen por completo de ayuda para los usuarios (sesenta y tres tesauros) o bien se ha considerado incompleta por no incluir todos los aspectos para que el usuario pueda realizar una consulta satisfactoria (treinta y siete tesauros).

3.9.2. Introducción

En diferentes obras y directrices sobre construcción de tesauros (Naumis, 2007; ANSI, 2005; AENOR, 1990), se recomienda la inclusión de una introducción en la que se explique con claridad distintos detalles del tesauro y su manejo. A pesar de las recomendaciones, sesenta y cinco de los tesauros analizados carecen de introducción, es decir, restan una información fundamental sobre sus características y el proceso de su confección. Treinta y dos de las introducciones examinadas apenas dan información sobre características como el número de términos, la fecha de última actualización o el objetivo del tesauro, por lo que son consideradas de escaso contenido informativo. Otras treinta y cuatro aportan mayor información que las anteriores pero sin llegar a cubrir lo recomendado por las directrices. Solamente treinta y tres tesauros disponen de una introducción en la que se recoja información sobre todos los puntos recomendados en las normativas.

4. Conclusiones

El análisis de los tesauros en acceso abierto en Internet nos demuestra que todavía son consideradas herramientas útiles, como lo demuestra el alto porcentaje de tesauros analizados (38,41%) que han sido elaborados o actualizados en los últimos cinco años, para las labores de almacenamiento y recuperación de información, especialmente en bases de datos y catálogos de bibliotecas, y la asignación de materias en formatos de metadatos. Sin embargo, en numerosos casos se ha podido observar que los autores de tesauros no han sabido aprovechar las ventajas que ofrece la World Wide Web.

En el caso de la presentación de los contenidos en formato electrónico, se puede comprobar que ésta ha avanzado poco respecto al formato de edición en papel, como ya han señalado otros autores (Arano, 2004). De todo lo observado apenas se han encontrado tesauros cuyo contenido no sea presentado en algunos de los *índices clásicos*: alfabético, sistemático o permutado. Estos índices se muestran en formatos de hipertexto orientados a la visualización (html y xhtml), y que suelen ir acompañados en numerosos casos por sus versiones en formato texto para su descarga. Solamente en catorce tesauros se encuentra disponible el contenido en otros formatos, como xml o SKOS-Core, que favorecen la interoperabilidad.

Otros aspectos revisados han sido las áreas geográficas y las lenguas en los que se ha editado los tesauros. En el primer caso, queda patente la hegemonía de los países europeos en la elaboración de tesauros con más del 41% de los casos analizados. En estos países se han elaborado diversas obras como el National Agricultural Library Thesaurus, el Medical Subject Headings y el ERIC Thesaurus, que han servido de referencia para la construcción de otros tesauros de sus respectivas áreas de conocimiento. Dos países con un alto número de tesauros son España y Francia, en los cuales diversas instituciones de la administración pública y organismos de investigación han realizado una excelente labor en el control terminológico de diversas materias.

En lo que a los idiomas de publicación de refiere, hay un claro predominio del inglés, como ya señaló Michelle Hudon (Hudon, 2003). Más del 64% de los tesauros estudiados están escritos en inglés o tienen al menos disponible una versión del contenido en lengua inglesa, en el caso de los tesauros multilingües. El elevado uso de este idioma se explica por su carácter de lengua vehicular dentro de las distintas disciplinas.

Del análisis de los datos sobre las áreas de conocimiento de los tesauros, se puede concluir una mayor presencia de tesauros sobre Ciencias Sociales, Humanidades y Ciencias de la Salud, y una presencia muy reducida de lenguajes sobre Ciencias Naturales y de carácter multidisciplinar. El elevado número de vocabularios controlados sobre Ciencias Sociales y Humanidades puede explicarse por la falta de normalización terminológica de las disciplinas contempladas, y el bajo número de tesauros de las Ciencias Naturales, por la elevada estandarización de los lenguajes empleados (Álvaro y otros, 1989b).

En lo que se refiere a la usabilidad de las interfaces, se aprecia una ausencia de ayuda al usuario en numerosos tesauros, lo que puede traducirse en un bajo porcentaje de uso de esas herramientas.

Por último, se hace necesario señalar la falta de profesionalidad observada en algunos tesauros, los cuales carecen de un conjunto mínimo de datos descriptivos (autoría o editor, fecha de actualización, número de términos u objetivos) que los haga más manejables. Esta ausencia sitúa a estos tesauros en formato electrónico un paso más atrás respecto a la edición impresa, cuando debería ser al contrario.

5. Bibliografía

Álvaro Bermejo, C.; Villagrá Rubio, A., y Sorli Rojo, A. (1989a). Desarrollo de lenguajes documentales formalizados en lengua española: una evaluación. I. Vigencia teórica y práctica de lenguajes controlados. *Revista Española de Documentación Científica*, vol. 12 (2), 154-159.

Álvaro Bermejo, C.; Villagrá Rubio, A., y Sorli Rojo, A. (1989b). Desarrollo de lenguajes documentales formalizados en lengua española. II. Evaluación de los tesauros en lengua española. *Revista Española de Documentación Científica*, vol. 12 (3), 283-305.

American National Standars Institute (ANSI) (2005). *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*. . Bethesda: NISO Press. p. 172. Disponible en http://www.niso.org/kst/reports/standards [consulta: 12 de octubre de 2009].

Arano, S. (2005). Los tesauros y las ontologías en la Biblioteconomía y Documentación. *Hipertext.net*. n.º 3. Disponible en http://www.hipertext.net/web/pag260.htm [consulta: 6 de octubre de 2009].

Arano, S., y Codina, L. (2004). La estructura conceptual de los tesauros en el entorno digital: ¿nuevas esperanzas para viejos problemas?. En *9es jornades catalanes d'informació i documentació: un espai de reunió, de diàleg, de participació, Barcelona, 25 i 26 de novembre de 2004*. Barcelona: Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya, 41-58.

Asociación Española de Normalización y Certificación (AENOR) (1990). *Directrices para el establecimiento y desarrollo de tesauros monolingües: UNE 50-196.90. Equivalente a ISO 2788-1986*. Madrid: AENOR.

Cueva Martín, A. de la (1999). Acceso y utilización de tesauros en Internet. *Revista Española de Documentación Científica*, vol. 22 (4), 531-540.

Dalbin, S. (2007). Thésaurus et informatique documentaires: partenaires de toujours. *Documentaliste-Sciences de l'Information*. vol. 44 (1). Disponible en http://www.cairn.info/resume.php?ID_ARTICLE=DOCSI_441_0042 [Consulta: 6 de octubre de 2009].

García Jiménez, A. (2002). *Organización y gestión del conocimiento en la comunicación*. Gijón: Trea., 203.

García Marco, F. J. (2008). Las normas de tesauros se ponen al día. Vocabularios estructurados para la recuperación de información en el entorno digital. En *Anuario ThinkEPI*, 57-62.

Hudon, M. (2003). True and tested products: thesauri on the web. *The Indexer*, vol. 23, n.º 3, 115-119.

International Federation of Library Associations and Institutions (IFLA) (2009). Working Group on Guidelines for Multilingual Thesauri. *Guidelines for Multilingual Thesauri*. La Haya: IFLA. p. 30. Disponible en http://archive.ifla.org/VII/s29/pubs/Profrep115.pdf [Consulta: 12 de octubre de 2009].

Mochón Bezares, G., y Sorli Rojo, A. (2007a). Tesauros de ciencias de la salud en Internet. *Revista Española de Documentación Científica*, vol. 30 (1), 107-124.

Mochón Bezares, G., y Sorli Rojo, A. (2007b). Tesauros de ciencias sociales en Internet. *Revista Española de Documentación Científica*, vol. 30 (3), 395-419.

Mochón Bezares, G., y Sorli Rojo, A. (2008a). Tesauros multidisciplinares en Internet. *Revista Española de Documentación Científica*, vol. 31 (1), 129-139.

Mochón Bezares, G., y Sorli Rojo, A. (2008b). Tesauros de humanidades en Internet. *Revista Española de Documentación Científica*, vol. 31 (3), 437-452.

Mochón Bezares, G., y Sorli Rojo, A. (2008c). Tesauros de ciencias naturales en Internet. *Revista Española de Documentación Científica*, vol. 31 (4), 647-658.

Mochón Bezares, G., y Sorli Rojo, A. (2009). Tesauros de ciencias experimentales y tecnologías de la información y la comunicación en Internet. *Revista Española de Documentación Científica*, vol. 32 (2), 115-127.

Moreiro, J. A. (2007). La representación de los contenidos digitales: de los tesauros automáticos a las folksonomías. En *VI Workshop CALSI: Información digital: nuevas perspectivas en la sociedad del conocimiento, Valencia 14, 15 y 16 de noviembre de 2007*. Disponible en: http://www.calsi.org/2007/wp-content/uploads/2007/11/jamoreiro.pdf [consulta: 6 de octubre de 2009].

Murakami, T. R. M. (2005). *Tesauros e a World Wide Web*. Sao Paulo: T. R. M. Murakami, p. 92. Disponible en http://eprints.rclis.org/10432/1/murakami-tesauros.pdf [Consulta: 12 de octubre de 2009].

Naumis Peña, C. (2007). Los tesauros documentales y su aplicación en la información impresa, digital y multimedia. México: Universidad nacional Autónoma de México. Centro Universitario de Investigaciones Bibliotecológicas, 284.

Rodríguez Yunta, L. (2009). Etiquetado libre frente a lenguajes documentales. Aportaciones en el ámbito de Biblioteconomía y Documentación. En *IX Congreso ISKO-España, Valencia (Spain), 11-13 March 2009*. Valencia: UPV. p. 832-845. Disponible en http://eprints.rclis.org/15836/1/Comunicacion_Luis_RYunta_ISKO2009.pdf [Consulta: 6 de octubre de 2009].

Shiri, A. A., y Revie, C. (2000). Thesauri on the web: current developments and trends. *Online Information Review*, vol. 24, (4), 273-279.

Soergel, D. (2002) Thesauri and ontologies in digital libraries: tutorial. En *6th European Conference on Research and Advanced Technology for Digital Libraries*. September 16-18. Pontifical Gregorian University, Rome, Italy. Disponible en http://www.dsoergel.com/cv/B63_rome.pdf [Consulta: 7 de octubre de 2009].