B.8. Las normas de tesauros se ponen al día: vocabularios estructurados para la recuperación de información en el entorno digital

Por **Francisco-Javier García-Marco**

García-Marco, Francisco-Javier. "Las normas de tesauros se ponen al día: vocabularios estructurados para la recuperación de información en el entorno digital".
En: *Anuario ThinkEPI*, 2008, pp. 57-62.

Resumen: *Se presentan los proyectos internacionales de reforma y ampliación de las normas sobre tesauros. El objetivo de estas iniciativas es refundir las normas y recomendaciones existentes, adaptar estos instrumentos al nuevo entorno digital, explicitar su relación con otras herramientas de control de vocabulario y facilitar la interoperabilidad entre los diferentes lenguajes documentales. Las normas constituyen una oportunidad para hacer visible ante la comunidad internacional la gran experiencia acumulada por los profesionales de la información y la documentación en la representación y organización del conocimiento.*

Palabras clave: *Vocabularios controlados, Tesauros, Clasificaciones, Taxonomías, Internet, Interoperabilidad, Normas internacionales.*

Title: Updating thesaurus standards: Structured vocabularies for information recovery in the digital environment

Abstract: *International projects for reforming and broadening the scope of the current technical guidelines and standards on thesauri are presented. The aims of these projects are to recast the existing standards and guidelines, to adapt them to the new digital environment, and to specify their relationship with other vocabulary control tools, making integration easier. These new standards are an excellent opportunity to enhance the visibility of the great corpus of knowledge and experience that information and documentation professionals have accumulated regarding knowledge representation and organization.*

Keywords: *Controlled vocabularies, Thesauri, Classifications, Taxonomies, Internet, Interoperability, Integrated systems, International standards.*

La revolución de internet alcanza a los tesauros

NO SE PUEDE PENSAR LA INFORMACIÓN Y LA DOCUMENTACIÓN sin referencia a la nueva infraestructura informativo-documental que conforma internet y que permite a la vez –aunque con variado éxito según los distintos aspectos implicados– la edición, publicación, archivo, recuperación y difusión de los documentos en cualquier medio.

En torno a esa nueva plataforma de comunicación y acumulación de la información y el conocimiento están convergiendo y estructurándose –a un ritmo más rápido que lento– el resto de los medios: radio, prensa, televisión, publicidad, comercio de la música, el sonido y el vídeo, enseñanza, interacción con administraciones públicas y empresas, comunicación telefónica, etc. Internet ha revolucionado el mundo de la comunicación y, por tanto, la vida cotidiana y laboral. Como resultado, la comunicación social se ha acelerado, se ha multiplicado, se ha hecho más eficiente, y también se ha convertido en una avalancha.

Lógicamente, tantos cambios tenían que revolucionar también la labor del profesional de la información y la documentación que, al fin y al cabo, es un profesional de la comunicación, un mediador que ayuda a poner en contacto la información acumulada con las necesidades de los usuarios.

Muchas áreas de la información y la documentación han cambiado para ponerse a tono

con la nueva realidad. Así, el concepto de documento se ha teñido de multimedia y digital, internet se ha convertido en el ámbito por excelencia de la referencia y la recuperación documental, las bases de datos documentales y los catálogos se han transformado en inevitables secciones de los portales de sus organizaciones, la catalogación ha transmutado en asignación de metadatos, los lenguajes documentales se han convertido en ontologías para la recuperación de la información,…

La ola del cambio tenía que alcanzar también a los tesauros, un campo muy específico pero fundamental de las documentaciones especializadas, y la respuesta a este reto, después de fraguarse durante casi una década, ha empezado a concretarse en estos dos últimos años.

De la documentación técnica a internet

Los primeros tesauros se desarrollaron dentro del ámbito científico-técnico en los años cincuenta y sesenta para superar las dificultades que planteaba a la recuperación el sólo empleo de palabras clave –nótese la analogía actual con internet– y, posteriormente, de descriptores.

El primero en ser publicado en 1959 fue el de *DuPont*, y los primeros en ser utilizados por una amplia base fueron el *Chemical Engineering Thesaurus* –publicado en 1961 por el *American Institute of Chemical Engineers*– y el *Thesaurus of Astia Descriptors* –publicado por el *US Defense Documentation Center* en 1962. La herramienta alcanzó su madurez en 1967 con el *Thesaurus of Engineering and Scientific Terms* del *Engineers Joint Council* y el *US Department of Defense*.

A la práctica siguieron las normas, lo cual es también una lección interesante. La primera norma propiamente dicha sobre tesauros fue la *Ansi Z39.19-1974* titulada *Thesaurus Structure, Construction and Use*, que fue publicada por el American *National Standards Institute* (*Ansi*, 1974) en 1974 y revisada en 1980. La nueva herramienta se expandió rápidamente en el contexto internacional. Se multiplicaron las publicaciones y manuales y aparecieron las recomendaciones de la *Unesco* y las diferentes normas internacionales y nacionales. En los años setenta la iniciativa normalizadora internacional la llevó el programa *Unisist* de la *Unesco* (1973, 1980) siguiendo la estela norteamericana. En la década de los ochenta tomó el relevo –con cierta lógica– la *International Organization for Standardization* (1985, 1986).

Respecto a la situación en España, la nueva herramienta fue presentada en un libro de **Lasso de la Vega** de 1966, pionero de la amplia labor de difusión realizada posteriormente por **Currás**, de la excelente práctica realizada –notablemente por el *Cindoc*– y de la investigación realizada por un notable grupo de autores (**García**, 2002, 2006). A pesar de la pronta difusión en nuestro país, que contribuyó con numerosos tesauros propios, la aprobación de las correspondientes adaptaciones por la *Agencia Española de Normalización y Certificación (Aenor)* de las recomendaciones aprobadas por la *Sociedad Internacional para la Normalización (ISO)* se produjo en el quicio de los noventa. El borrador de las *Directrices para el establecimiento y desarrollo de tesauros monolingües –UNE 50-106-90*, equivalente a la *ISO 2788:1986*, una corrección de la *ISO 2788* de 1975, a su vez inspirada en las recomendaciones de la *Unesco* de 1970– fue publicado en la *Revista Española de Documentación Científica* en los años 1989 y 1990 (*Aenor*, 1989, 1990a). Finalmente, la norma fue efectivamente aprobada dicho año (*Aenor*, 1990b). Los borradores de las *Directrices para la creación y desarrollo de tesauros multilingües –UNE 50-125*, un trasunto de la *ISO 5964*, publicada en 1985– fueron publicados en la *Revista Española de Documentación Científica* en 1996 y 1997 (*Aenor*, 1996, 1997a), y las norma, aprobada y publicada en julio de 1997 (*Aenor*, 1997b). Ambas pueden consultarse fácilmente en el recopilatorio de normas UNE sobre documentación publicado por la *Aenor* (1999).

Las iniciativas de reforma

La necesidad de acomodar los tesauros a la realidad marcada por internet se dejó sentir muy pronto entre los expertos, y ha cristalizado recientemente. De hecho, en este momento existen ya dos normas reformadas de tesauros que abordan la problemática de su uso

en internet: se trata de la norma *Ansi Z39.19* y la norma *BS 8723*.

De forma semejante a lo ocurrido al comienzo de esta historia, norteamericanos y británicos han tomado la delantera en su reforma. No obstante, es importante señalar también la formación de un grupo de trabajo de la *International Federation of Library Associations (Ifla)* con el objeto de poner al día las recomendaciones para tesauros multilingües (*Ifla*, 2006), en estrecha coordinación con las iniciativas anglosajonas.

Los norteamericanos ya constataron la necesidad de poner al día los tesauros ante los cambios que se estaban produciendo en el marco de la información electrónica y de la *World Wide Web* en 1998, cuando la *Ansi Z39.19-1980* (segunda revisión de la primera norma estadounidense) fue revisada y confirmada. Atendiendo a este consenso, la *National Information Standards Organization (Niso)* organizó al año siguiente un taller de trabajo nacional, el *Workshop on Electronic Thesauri*, que tuvo lugar el 4 y 5 de noviembre de 1999 con el objetivo concreto y explícito de investigar la conveniencia y posibilidad de desarrollar un estándar para los tesauros electrónicos. En la reunión participaron, entre otros, la *American Psychological Association (APA)*, la *American Society of Indexers (ASI)*, y la *Association for Library Collections and Technical Services (Alcts)* de la *American Library Association (ALA)*.

A partir de las recomendaciones aprobadas en ella, se desarrolló la cuarta edición de la norma *Z39.19*, a la que se dio el título *Guidelines for the construction, format, and management of monolingual controlled vocabularies*, y que fue publicada el año pasado (*Ansi*, 2006). Se trata de un extenso documento de 172 páginas en el que, partiendo de un análisis preliminar de la necesidad del control de vocabulario, se presentan los objetivos, conceptos, principios y estructura de los tesauros; se definen las normas de tratamiento terminológico –selección, alcance y forma de los términos simples y complejos—; las relaciones; las técnicas de presentación; los aspectos relacionados con la interoperabilidad; y los aspectos relacionados con su construcción, prueba, mantenimiento y gestión.

La norma americana aporta novedades muy importantes: en primer lugar, adopta un enfoque dirigido a todo tipo de recursos de información, tanto tradicionales como electrónicos; en segundo lugar, por ampliar su alcance –anteriormente centrado exclusivamente en los tesauros– a los "vocabularios controlados", incluyendo concretamente "las listas de términos controlados, anillos e sinónimos, taxonomías y tesauros"; en tercer lugar, por abordar el problema de la interoperabilidad de vocabularios en el marco de la Red. Sin embargo, también tiene ciertas limitaciones: aborda tan sólo los vocabularios monolingües –con el criterio de que los multilingües deben ser abordados por el organismo internacional competente– y deja de lado herramientas de vocabulario controlado muy importantes.

En este momento existen ya dos normas reformadas de tesauros que abordan la problemática de su uso en internet: se trata de la norma *Ansi Z39.19* y la norma *BS 8723*

En cuanto a la nueva norma británica *BS 8723* –su historia puede ser consultada en **Gilchrist** (2007)–, se trata también un documento muy amplio que consta de cinco partes. La primera parte establece las definiciones y conceptos comunes para todos los tipos de vocabularios controlados para la recuperación de la información en sistemas de información. La parte segunda se ocupa de los tesauros propiamente dichos –es decir, de la antigua norma *BS 5723* (=*ISO 2788*)—, aunque va más allá: pues, proporciona pautas sobre su uso y gestión digital. La tercera parte trata otros vocabularios estructurados, concretamente los esquemas de clasificación, los tesauros, las listas de encabezamientos, las taxonomías y las ontologías. La cuarta parte aborda la interoperabilidad entre vocabularios en general y, muy específicamente, el mapeo entre ellos. Dentro de esta parte se trata el problema del multilingüismo como un caso especial y aborda allí los aspectos tratados por la norma sobre tesauros multilingües *BS 6723 (ISO 5964)* y, por tanto, la integra. Finalmente, la quinta parte trata de los protocolos y formatos para el intercambio de datos sobre vocabularios controlados. Las cuatro primeras partes han sido publicadas en 2006 y 2007. La parte cinco

Aenor se ha sumado con rapidez a los esfuerzos internacionales para revisar y ampliar el alcance de las normas sobre tesauros y para facilitar su adaptación al entorno español

está en fase de elaboración y trata de la gestión electrónica de los vocabularios mediante el uso de xml y otras tecnologías asociadas.

La norma británica supone varios avances importantes respecto a la norteamericana, en especial la consideración de la interoperabilidad lingüística, el análisis de facetas –un aspecto clave en la teoría moderna de los lenguajes documentales—, la inclusión de los sistemas de clasificación como herramientas complementarias de los tesauros y puntos clave de una estrategia de interoperabilidad entre los lenguajes documentales en conjunto, y el abordaje de otros lenguajes como las listas de encabezamientos de materia y las ontologías, sin olvidar las listas de descriptores y las taxonomías que ya fueron abordadas también en la última edición de la *Ansi Z39.19*.

Finalmente, este verano el grupo inglés propuso a la *Organización Internacional de Normalización* la revisión de las normas *ISO 2788* y *5964*, y la propuesta británica fue apoyada por una amplia mayoría de países con derecho a voto.

El Grupo de vocabularios controlados para la recuperación de información de *Aenor 50*

La *Asociación Española de Normalización y Certificación (Aenor)* se ha sumado con rapidez a los esfuerzos internacionales para revisar y ampliar el alcance de las normas sobre tesauros, y para facilitar su adaptación al entorno español. Así, el 23 de junio de 2006 se constituyó oficialmente el *Grupo de Vocabularios Controlados para la Recuperación de Información* del *Comité 50*[1], siguiendo la estela de las iniciativas norteamericana *Ansi/Niso Z39.19-2005* y, especialmente, de la más completa propuesta por el *British Standard Institute*.

En consonancia con ellas, el grupo comparte los objetivos de reunificar las normas relativas a tesauros monolingües y multilingües –a saber la *UNE 50106:1990*, *UNE 50106:1995 Erratum* y *UNE 50125:1997*– en una sola que incluya también el resto de los lenguajes documentales, especialmente los encabezamientos de materia, las clasificaciones y las taxonomías. Igualmente, se busca establecer los mecanismos de interconexión entre los distintos lenguajes y, finalmente, conectar éstos con los avances que se están produciendo desde la informática en el campo de las ontologías y de los mapas temáticos (*topic maps*).

El grupo ha estudiado los cambios propuestos, y ha mantenido contactos con los colegas británicos que han promovido la norma *BS 8723* colaborando en la revisión de las partes 3 y 4 de su trabajo. Las propuestas más importantes se refieren a la necesidad de tener en cuenta en la redacción de la norma los estándares sobre control de autoridades *Isar* y *Isbd (AR)*, la norma sobre *topics maps*, así como la necesidad de considerar las clasificaciones de *business archives* en el contexto más amplio de la tradición archivística. Ahora es importante ampliar el grupo y dar cabida a todas las partes españolas interesadas para colaborar en la nueva norma internacional y en la futura norma española.

Conclusiones

El proceso de digitalización y puesta en red de gran cantidad de recursos que ha supuesto internet es a la vez un gran reto y una gran oportunidad para los lenguajes documentales, en particular, y para los documentalistas, en general.

Los usuarios de la Red y de los servicios de información y documentación reclaman –y premian– un acceso integrado a los diferentes tipos de recursos cualquiera que sea su medio y forma; que supere las barreras lingüísticas y terminológicas que estorban la recuperación de la información que plagan la Web hoy; y que se complemente el brillante acceso específico que proporciona *Google* con recursos para la navegación temática (*topic maps*, etc.) y el *browsing*. También piden mayor precisión en la búsqueda temática y un grado de exhaustividad más ajustado a las diferentes necesidades.

Está claro que los tesauros son la herramienta ideal para soportar terminológicamente estas funciones. Pero, para ello, es necesario que los tesauros en internet –y en general, los lenguajes documentales– puedan conectarse entre sí en redes de recuperación, de forma que se supere el localismo y el "especialismo" que los lastra en la actualidad, pues sólo sirven para buscar dentro de sitios web y bases de datos concretas, al contrario que *Google*, por ejemplo, que permite realizar búsquedas en el conjunto de la Red.

De hecho, en un sentido amplio, *Google* es a la Web lo que el conjunto de los lenguajes documentales es a internet invisible. Conectar ambos mundos, superar la común barrera lingüística –buscar en recursos de diferentes idiomas utilizando sólo uno de ellos– y el mosaico de provincias que componen el segundo, es uno de los grandes retos de la Red.

De alguna manera y como nuevo punto de referencia del mundo de la recuperación de la información, *Google* ejemplifica muy bien los límites del paradigma actual de recuperación de la información en internet mediante buscadores, que se basa en la combinación de las búsquedas postcoordinadas sobre texto libre, los métodos de ponderación de palabras clave y el análisis de citas. La experiencia cotidiana es que *Google* resulta excelente para proporcionar resultados a búsquedas muy específicas y para ofrecer resultados interesantes en las búsquedas temáticas. Sin embargo, no discrimina suficientemente, al menos todavía, la calidad de los resultados temáticos; y, sobre todo, no ofrece un panorama sistemático del ámbito temático recuperado y no ofrece acceso a las mejores bases de datos de referencias del mundo, muchas de ellas gratuitas, que están indizadas con vocabularios específicos[2].

Lo contrario ocurre en internet invisible. Los recursos seleccionados por bases de datos indizadas de calidad suelen estar descritos con vocabularios muy diferentes, que van desde la *Clasificación Decimal Universal* y la Clasificación Decimal de Dewey, a los *Encabezamientos de Materia de la Library of Congress* y sus derivados en otras lenguas o las *Medical Subject Headings*, pasando por los diferentes tipos de tesauros. Esta situación exige programar su consulta en sesiones diferentes, y convierte a la recuperación en una tarea muy costosa en tiempo y recursos. Su consulta integrada exigiría que se asegurase la compatibilidad entre los diferentes lenguajes a través de su mapeo, y en definitiva alcanzar un concepto de moda: la interoperabilidad entre los diferentes lenguajes documentales. Lógicamente, ello supone previamente la clarificación de la estructura de los diferentes tipos de lenguajes y de los medios para interconectarlos.

Google es a la Web lo que el conjunto de los lenguajes documentales es a internet invisible

Las normas que se han presentado están precisamente orientadas a ese fin. La norma americana *Ansi/Niso Z39.19-2005* y la británica *BS 8723* siguen esa dirección al abordar los tesauros dentro de un marco más amplio, el de los vocabularios controlados para la recuperación de información. En especial, la norma británica supone un gran avance en dicho camino, pues contempla un amplio rango de modelos de vocabularios controlados, al que habría que incorporar el ámbito de las clasificaciones archivísticas.

Si se avanza en la ruta marcada por esos esfuerzos, se habrá dado un paso de gigante en la superación del provincianismo que lastra a los lenguajes documentales en la era de la globalización y, por ende, en la configuración de una teoría integrada de los lenguajes documentales, en la que a cada familia de lenguajes documentales y más ampliamente de sistemas de recuperación se le reconozca lo suyo: a los alfabéticos el acceso alfabético y la cercanía a los términos usados por los usuarios, a los sistemáticos la navegación conceptual y la posibilidad de ofrecer ordenaciones conceptuales rigurosas y predecibles, a los probabilísticos la oferta de información ponderada de forma objetiva, a la extracción de palabras clave su inmediatez, sencillez y potencia,…

En cualquier caso, como señala **Gilchrist** (2007), la publicación de estándares sólidos en el campo de la organización de los vocabularios de organización y recuperación de la información constituye una gran oportunidad para hacer visible ante el resto de los colegas técnicos y científicos la gran experiencia acu-

mulada por los profesionales de la información y la documentación en la resolución de los problemas relacionados con la representación y organización del conocimiento.

Notas

1. Está formado por **Carmen Agustín Lacruz**, **Carmen Caro Castro**, **Francisco Javier García Marco** (coordinador), **José Ángel Martínez Usero** y **Rosa San Segundo**. El autor agradece al resto de los miembros su ayuda, su colaboración y sus aportaciones.

2. Es cierto, con todo, que *Google Scholar* y el incipiente tratamiento terminológico, entre otras iniciativas, suponen pasos importantes en la solución de estos problemas.

Referencias

Aitchison, J.; **Clarke, S. D.** "The thesaurus a historical viewpoint, with a look to the future". En: *Cataloging & Classification Quarterly*, 2004, v. 37, n. 3/4, pp. 5-21.

Asociación Española de Normalización y Certificación. "Documentación: directrices para el establecimiento y desarrollo de tesauros monolingües: parte 1". En: *Revista Española de Documentación Científica*, 1989, v. 12, n. 4, pp. 463-483.

Asociación Española de Normalización y Certificación. "Documentación: directrices para el establecimiento y desarrollo de tesauros monolingües: parte 2". En: *Revista Española de Documentación Científica*, 1990, v. 13, n. 1, pp. 601-629.

Asociación Española de Normalización y Certificación. Directrices para el establecimiento y desarrollo de tesauros monolingües. Madrid: Aenor, D.L., 1990. 47 p. *UNE 50-106-90*. Equivalente a *ISO 2788-1986*.

Asociación Española de Normalización y Certificación. Documentación. Directrices para el establecimiento y desarrollo de tesauros monolingües. Madrid: Aenor, D.L., 1995. *UNE 50106:1995 Erratum.*

Asociación Española de Normalización y Certificación. "Documentación. Directrices para la creación y desarrollo de tesauros multilingües: *ISO 5964-1985*, *UNE 50-125*". En: *Revista Española de Documentación Científica*, 1996, v. 19, n. 4, pp. 439-467.

Asociación Española de Normalización y Certificación. "Documentación. Directrices para la creación y desarrollo de tesauros multilingües: *ISO 5964-1985*, *UNE 50-125*: continuación". *Revista Española de Documentación Científica,* 1997, v. 20, n. 1, pp. 63-82.

Asociación Española de Normalización y Certificación. Documentación. Directrices para la creación y desarrollo de tesauros multilingüe. Madrid: Aenor, 1997. 77 p. Normas *UNE 50125*. Norma equivalente a *ISO 5964:1985*.

Asociación Española de Normalización y Certificación. Documentación: recopilación de normas *UNE*. 3ª ed. Madrid: Aenor, 1999. 580 p.

British Standards Institute. *BS 8723*, Structured vocabularies for information retrieval. London: Bristish Standards Institute, 2006.

García-Marco, Francisco-Javier. "La literatura científica sobre lenguajes poscoordinados en España: de la divulgación del concepto a la internet". En: *Documentación de las Ciencias de la Información*, 2002, v. 25, pp. 291-319.

García-Marco, Francisco-Javier. "Ontologías, taxonomía y tesauros: manual de construcción y uso". En: *El profesional de la información*, 2006, v. 15, n. 4, pp. 317-318.

Gil-Urdiciaín, Blanca. "Origen y evolución de los tesauros en España". En: *Revista General de Información y Documentación*, 1998, v. 8, n. 1, pp. 63-110.

Gilchrist, Alan. "Revisión de las Normas Británicas BS5723 y BS6723 para el diseño y uso de tesauros: un breve informe de progreso". En: *Scire*, 2007, v. 13, n. 1.

International *Federation of Library Associations, Classification and Indexing Section, Working Group on Guidelines for Multilingual Thesauri.* Guidelines for Multilingual Thesauri. IFLA, April 2005. *http://www.ifla.org/VII/s29/pubs/Draft-multilingual thesauri.pdf*

National Information Standards Organization (Estados Unidos); American National Standards Institute. Guidelines for the construction, format, and management of monolingual controlled vocabularies: an American national standard. Bethesda, MD.: National Information Standards Organization, 2006. *Ansi/Niso Z39.19-2005*. Approved July 25, 2005 by the *American National Standards Institute*.

Sociedad Internacional de Normalización. Documentation. Guidelines for the establishment and development of multilingual thesauri. Gèneve: ISO, 1985. *(ISO 5964).*

Sociedad Internacional de Normalización. Documentation, guidelines for the establishment and development of monolingual thesauri = documentation, principes directeurs pour l'etablissement et le developpement de thesaurus monolingues. 2nd ed. Geneva: ISO, 1986. (International standard; *ISO 2788*).

Unesco, Unisist. Principes directeurs pour l'établissement et le développement de thésaurus multilingues. Paris: Unesco, mai 1980. 88 p. (*PGI/80/WS/12*).

Unesco, Unisist. Principes directeurs pour l'établissement et le développement de thésaurus monolingues. Paris: Unesco, sept. 1973. 34 p. (*SC/WS/555*).