Herramientas de control del vocabulario científico-técnico: glosarios y tesauros del CINDOC

ÁNGELES MALDONADO MARTÍNEZ
TERESA ABEJÓN PEÑA
CINDOC-CSIC

Índice

Introducción
1 Tesauros
1.1 Contenido y estructura
1.2 Historia
1.3 Metodología de elaboración
1.4 Estructura y presentación
2 Glosarios especializados
3 Acceso a tesauros y glosarios

Introducción

El Centro de Información y Documentación Científica (CINDOC), perteneciente al Consejo Superior de Investigaciones Científicas (CSIC), cuenta en el momento actual con una colección de herramientas terminológicas, tesauros y glosarios especializados en temas científico-técnicos, que ha ido elaborando y actualizando a lo largo de su existencia. El presente texto tiene por finalidad hacer una presentación de dichas herramientas haciendo especial hincapié en la metodología de elaboración, estructura y disponibilidad actual de las mismas.

1 Tesauros

1.1 Contenido y estructura

Un tesauro trata de plasmar el sistema conceptual utilizado en una materia o disciplina determinada y puede ser definido como un conjunto de términos estructurados jerárquicamente según la generalidad o especificidad de sus significados y relacionados por vínculos asociativos y de equivalencia. En un tesauro podemos conocer de un término aquellos otros que mantienen algún tipo de relación semántica con él: sus equivalentes, los que tienen una significación más amplia (TG, términos genéricos), los que representan conceptos más específicos (TE, términos específicos) y aquellos otros que por diversos motivos se le asocian (TR, términos relacionados). Existen tesauros que incluyen además equivalencias idiomáticas.

Tradicionalmente en un tesauro no se definen los términos, sino que solamente se añade en alguno de ellos lo que se conoce como nota de alcance (NA). Una nota de alcance no constituye una definición sobre la significación del término, sino que es más bien una aclaración de cómo debe utilizarse el mismo dentro del sistema documental dado.

Ejemplos de términos extraídos de un tesauro:

Establecimientos comerciales
UP Locales comerciales
TG Equipamiento comercial
TE Centros comerciales
Comercios tradicionales
Grandes almacenes
Mercados
Supermercados
TR Actividades comerciales
Comercio
Edificios comerciales

Enfermedades de los ojos
EN Eye diseases
FR Maladies des yeux
UP Enfermedades oculares
Trastornos de la visión
TG Enfermedades
TE Ceguera
Enfermedades de la córnea
Enfermedades de la retina
Errores de refracción
Glaucoma
TR Diabetes
Lentes de contacto
Oftalmología
Ojos
Visión

Los tesauros son herramientas terminológicas utilizadas en el análisis y recuperación de información en bases de datos documentales con dos objetivos:

A Controlar el vocabulario. Controlar el vocabulario significa identificar dentro de un campo semántico todos los conceptos representados por más de un término. La identificación de términos equivalentes hace posible minimizar la pérdida de información en las búsquedas realizadas en un sistema documental automatizado.

B Conocer todos los términos relacionados con un concepto determinado, lo que ayuda a añadir más términos adecuados para enriquecer tanto los análisis de contenido de los documentos como las estrategias de búsqueda para recuperar información.

1.2 Historia

El interés del ahora CINDOC, nacido en 1992 de la unificación del ICYT (Centro de Documentación Científica y Técnica) y del ISOC (Centro de Documentación en Ciencias Sociales y Humanidades), en la elaboración de tesauros se manifiesta a principio de los años 80 en dos hechos que marcarán el comienzo de esta línea de trabajo que continúa en la actualidad. Por una parte la concesión al ISOC en 1987 por parte de la CICYT (Comisión Interministerial de Ciencia y Tecnología) de un Proyecto de Investigación para la elaboración de tesauros sectoriales en el campo de las Ciencias Sociales. El otro hecho relevante es el encargo al ICYT por parte de la UNESCO de la traducción y adaptación al español del *Tesauro Spines*, vocabulario pensado para el control y desarrollo de políticas científicas y tecnológicas.

Con ayuda del proyecto concedido por la CICYT se elaboraron los tesauros de Urbanismo, Economía y Psicología y se avanzaron los trabajos de los de Derecho, Ciencias Políticas y Sociología. Todos estos tesauros nacieron con el propósito de controlar el vocabulario en cada uno de los ficheros de la base ISOC, base de datos bibliográfica que recoge el vaciado de las revistas científicas editadas en España especializadas en Ciencias Sociales y Humanas.

Posteriormente se ha completado y publicado el *Tesauro de Derecho* (1997), gracias a la colaboración de otras instituciones del ámbito jurídico español. Dentro de las Ciencias Humanas se publicó posteriormente el *Tesauro de Historia Contemporánea de España* (1999) y más

recientemente el *Tesauro de Biblioteconomía y Documentación* (2003). En el campo de las Ciencias habremos de mencionar el *Tesauro ICYT de Biología Animal* (1995) y el *Tesauro de Geología* (2001).

Mención aparte requiere el *Tesauro ISOC de Topónimos*, cuya primera edición data de 1993 y la tercera y última de 2004 (actualmente en imprenta). Este tesauro recoge en el plano nacional los municipios españoles agrupados por provincias y comunidades autónomas, y a nivel internacional los países con sus capitales. En el caso de la Unión Europea y de América Latina se añaden ciudades de más de 100.000 habitantes, y en el resto de los países las ciudades mayores de 1.000.000 de habitantes. En el caso de los municipios catalanes, vascos, gallegos, valencianos, mallorquines y asturianos se recogen las denominaciones en español y en las lenguas de las comunidades autónomas. En el caso de topónimos no españoles se ha intentado recoger el abanico más amplio de grafías de los mismos dando preferencia a la española.

Finalmente, es interesante mencionar el *Tesauro HEREIN* que recoge la terminología utilizada en la gestión del patrimonio cultural. Este tesauro de más de 600 términos en 10 idiomas diferentes (inglés, francés, español, búlgaro, rumano, griego, húngaro, esloveno, lituano y polaco), se ha realizado dentro del proyecto europeo IST «Herein 2. European Heritage Network». En la actualidad puede ser consultado en la página web de la propia red europea del patrimonio (http://www.european-heritage.net).

1.3 Metodología de elaboración

Prácticamente todos los tesauros, salvo *HEREIN*, por su multilingüismo, y el de topónimos, por la naturaleza de los términos, fueron realizados con la misma metodología. Todos ellos fueron coordinados por un documentalista especialista en el tema tratado, participando además en la elaboración expertos en las diferentes materias, quienes en unos casos establecieron las diferentes relaciones entre los términos, y en otros dieron el visto bueno a lo anteriormente organizado. Las fases de elaboración fueron las siguientes:

A Selección de fuentes

Los términos de indización procedentes de las bases de datos ISOC e ICYT formaron el núcleo fundamental que sirvió de base para la elaboración de los diferentes tesauros. Sin embargo, este conjunto de términos no fue suficiente y hubo que completar cada una de las terminologías especializadas con términos extraídos de otras fuentes: tesauros de igual

temática o temáticas próximas existentes en otros idiomas, diccionarios y léxicos especializados.

B Extracción de términos

Una vez seleccionadas las fuentes se procedió a la extracción sistemática de los términos aparecidos en ellas.

C Normalización de términos

En cuanto a la forma gramatical, género y número de los términos, se siguieron las especificaciones contempladas en la norma ISO 2788-1986 versión española en el caso de los tesauros elaborados con anterioridad a la aprobación de la norma UNE 50106:1990 sobre Directrices para el establecimiento y desarrollo de tesauros monolingües. Una vez aprobada esta norma se siguieron sus especificaciones.

D Determinación de grandes áreas y distribución de términos

Paralelamente a la extracción de términos se pensó en la estructura jerárquica básica de cada uno de los tesauros, determinando las grandes áreas temáticas o familias semánticas en que más tarde se distribuirían los términos. A partir de ese momento se trabajó individualizadamente con cada familia en el establecimiento de relaciones semánticas.

E Establecimiento de relaciones de equivalencia

Agrupados los términos por áreas temáticas, el paso siguiente fue identificar los términos de igual significación o significación muy próxima dentro de las mismas. Detectadas las equivalencias semánticas se procedió a elegir uno de los términos como representativo del concepto, adjudicándole la categoría de descriptor o término principal. El término o términos equivalentes se remitirían al principal.

F Establecimiento de la estructura jerárquica

Una vez elegidos los descriptores se comenzó a construir la jerarquía con los mismos. Paulatinamente se establecieron los diferentes niveles jerárquicos hasta crear una estructura piramidal donde quedaron encajados todos los términos principales ocupando el nivel que les

correspondiera según el grado de especificidad-generalidad de sus significados en relación con el resto.

G Establecimiento de relaciones asociativas y notas de aplicación

Creada la estructura jerárquica, se pasó a establecer las relaciones de asociación entre términos que, aun no manteniendo una vinculación jerárquica directa, tuvieran sin embargo una relación de vecindad semántica. Finalmente se elaboraron las notas de aplicación, que introdujeron ciertas aclaraciones sobre el sentido de alguno de los términos.

H Informatización

Una vez establecidas todas las relaciones posibles entre los términos, hubo que informatizar el resultado con el fin de conseguir una presentación adecuada para la edición. Para ello se utilizó el módulo de creación de tesauros del gestor de bases de datos estándar BASIS, capaz de generar los listados básicos en los que se difunde un tesauro: alfabético, jerárquico y permutado. En algunos casos también se utilizó el programa de creación de tesauros para microordenadores conocido como CAT (Construcción Automática de Tesauros), elaborado por los informáticos del propio CINDOC.

I Edición provisional

Finalizados todos los trabajos anteriormente relacionados, se procedió a una edición provisional, para hacerla circular con el fin de recoger opiniones externas al equipo de trabajo.

J Revisión y edición definitiva

La edición provisional sirvió para poner a prueba los tesauros. Una vez editados primeramente en papel y posteriormente en formato electrónico, pudieron ser utilizados en la indización de documentos y en la elaboración de estrategias de búsqueda, comprobándose así su adecuación al tema desarrollado.

K Mantenimiento y actualización

Debido al paso del tiempo y a su uso, en la mayoría de los tesauros se han incluido nuevos términos con el objetivo de mantenerlos actualizados.

1.4 Estructura y presentación

La presentación de los tesauros mencionados responde al mismo esquema. Todos ellos presentan los términos que contienen en tres índices diferentes: alfabético, jerárquico y permutado. A la hora de utilizar un tesauro, lo aconsejable es consultar los tres índices, ya que cada uno ofrece una información diferente sobre un término dado. A continuación se presentan las características de cada uno de los índices.

A Índice alfabético

En el índice alfabético están presentes todos los términos contenidos en el tesauro, descriptores o términos principales y no descriptores o sinónimos. Ambas categorías se diferencian gráficamente, pues los primeros suelen ir en letras mayúsculas y los segundos en minúsculas. Cada uno de los términos principales se presenta con todos sus términos relacionados, mencionando la naturaleza de la relación. Los no descriptores tan sólo remiten a los términos principales.

B Índice jerárquico

En este listado aparecen representadas las cadenas jerárquicas completas. Este índice da una visión completa de cada cadena jerárquica y proporciona el encuadre de los términos dentro de ellas. La representación de la cadena jerárquica se hace a través de sangrados y puntos. Los listados alfabético y permutado remiten a este índice jerárquico a través de los códigos que aparecen en todos los términos principales o descriptores.

C Índice permutado KWIC

Índice en el que aparecen todos los términos del tesauro, descriptores y no descriptores, alfabetizados por todas las palabras significativas de que están compuestos. Este índice reúne todos los términos que tienen una misma palabra o al menos una misma raíz común.

2 Glosarios especializados

Otra línea de trabajo desarrollada en el CINDOC relacionada con la terminología es la creación de glosarios especializados. Estos trabajos tienen como principal finalidad recoger y sistematizar la terminología específica de un campo científico determinado, establecer definiciones de los términos utilizados en el mismo y establecer equivalencias con otros idiomas, todo ello para propiciar el intercambio entre profesionales y expertos.

En este campo merece la pena resaltar el trabajo del grupo de terminología del antiguo ICYT, que entre otros trabajos, contribuyó a finales de los años 80 a la actualización de EURODICAUTOM, banco de datos terminológicos de la Unión Europea, con la aportación de más de 30.000 términos de diversos campos pertenecientes a la Ciencia y a la Tecnología.

En el momento presente existen cuatro glosarios publicados de estas características antes mencionadas: *Diccionario de Alimentos*, *Terminología Multilingüe de Drogas*, *Terminología Multilingüe de Máquinas-herramienta* y *Glosario de Términos de Acuicultura*. Los tres primeros han sido realizados por los componentes del grupo TermEsp (Terminología española), que desarrolla sus actividades en el CINDOC desde 1985. El cuarto ha sido desarrollado por los componentes del Centro de Documentación en Acuicultura, todos ellos asimismo pertenecientes al CINDOC.

Estos glosarios tienen todos una estructura similar. Constan de un índice alfabético general donde los términos en castellano van acompañados de sus equivalentes en alemán, inglés y francés (en el caso del *Glosario de Términos en Acuicultura* también aparece el equivalente en italiano). En los casos en que procede también figuran los sinónimos en español. Además, algunos términos van acompañados de una breve nota aclaratoria que precisa su significado. Este índice alfabético general se acompaña de un índice permutado y otro alfabético de los términos en alemán, inglés y francés.

Como consecuencia de la pertenencia a REALITERM (Red Palatina de Terminología) el CINDOC viene participando en la elaboración de diversos glosarios multilingües. En la actualidad se acaba de finalizar uno especializado en el «genoma humano», y se está participando en la elaboración de uno sobre «geomática» y otro sobre «terminología».

Finalmente apuntemos que en estos momentos y gracias a la participación en el proyecto ESLEE (Estudio de los Lenguajes Especializados en Español (I). Elaboración y desarrollo de vocabularios

científicos y técnicos, referencia BFF2001-1506) financiado por el MICYT, se encuentran en marcha dos léxicos especializados, uno en gestión del patrimonio cultural y otro en prospectiva tecnológica. Para la elaboración de estos léxicos se está trabajando con corpus especializados creados *ad hoc* de donde se están extrayendo los términos, y las definiciones y contextos de los mismos, contrastando y completando la información con otras fuentes. Los términos finalmente seleccionados irán acompañados de sus equivalentes en lengua inglesa. El interés de estos nuevos léxicos para el equipo que trabaja en terminología en el CINDOC es la experimentación de una metodología avanzada del trabajo terminológico que incluye una mayor automatización de determinadas fases del mismo. Esta metodología podrá ser utilizada en el futuro para la elaboración de herramientas terminológicas similares aplicadas a otros campos de especialidad. Por otra parte, uno de los dos grandes grupos de destinatarios principales del proyecto ESLEE, el de los traductores e intérpretes, seguirá siendo también beneficiario principal de la creación de nuevos léxicos especializados, teniéndose en cuenta en el futuro sus necesidades específicas como usuarios de estos recursos.

3 Acceso a tesauros y glosarios

Existen varios soportes de edición de los tesauros y glosarios mencionados. De prácticamente todos hay una edición impresa que puede ser adquirida en el propio CINDOC por todo aquel que esté interesado. También pueden ser consultados a través de Internet (http://pci204.cindoc.csic.es/tesauros/index.html). Las ediciones web tanto de los tesauros como de los glosarios son una reproducción de las versiones impresas, pues se puede acceder al contenido de los mismos a través de los índices alfabético sistemático, jerárquico y permutado. La ventaja que trae consigo la consulta a través de Internet es el poder hacer uso de los hiperenlaces, lo que permite saltar automáticamente entre términos y entre índices. El uso de los tesauros y glosarios a través de Internet es gratuito y está pensado para realizar consultas puntuales.

Tesauros y Glosarios del CINDOC

Tesauro	Última edición	Idiomas	Términos			Índices			Edición	
			Total	Descrip	No Descrip	Alfab.	Jerárq.	Permut.	Papel	Internet
Tesauro Spines	1988	ES, EN, FR	10.832	7.993	2.839	SÍ	SÍ	SÍ	SÍ	SÍ
Tesauro ISOC Urbanismo	1992	ES	4.422	3.652	770	SÍ	SÍ	SÍ	SÍ	SÍ
Tesauro ISOC Economía	1995	ES, EN, FR	6.792	5.464	1.328	SÍ	SÍ	SÍ	SÍ	SÍ
Tesauro ISOC Psicología	1995	ES	4.400	2.788	1.612	SÍ	SÍ	SÍ	SÍ	SÍ
Tesauro ICYT Biología Animal	1995	ES	3.781	2.782	999	SÍ	SÍ	SÍ	SÍ	SÍ
Tesauro de Derecho	1997	ES	19.494	14.179	5.315	SÍ	SÍ	SÍ	SÍ	SÍ
Tesauro Historia Contemporánea España	1999	ES	4.139	3.776	350	SÍ	SÍ	SÍ	SÍ	SÍ
Tesauro de Derechos de Autor	1999	ES				SÍ	SÍ	SÍ	SÍ	SÍ
Tesauro de Geología	2001	ES	2.108	2.065	43	SÍ	SÍ	SÍ	SÍ	SÍ
Tesauro de Biblioteconomía y Documentación	2003	ES, EN, FR	1.153	914	239	SÍ	SÍ	SÍ	SÍ	SÍ
Tesauro ISOC de Topónimos	2004	ES	14.879	12.287	2.592	SÍ	SÍ	SÍ	SÍ	SÍ
Glosarios	Última edición	Idiomas	Términos			Índices			Edición	
			Total			Alfab.	Jerárq.	Permut.	Papel	Internet
Diccionario de Alimentos	1994	ES, EN, FR, DE	3.814			SÍ	NO	SÍ	SÍ	SÍ
Glosario de Términos de Acuicultura	1996	ES, EN, FR, DE, IT	2.200			SÍ	NO	SÍ	SÍ	SÍ
Terminología Multilingue Drogas	1997	ES, EN, FR, DE	2.000			SÍ	NO	SÍ	SÍ	SÍ
Terminología Multilingue Máquinas Herramienta	1999	ES, EN, FR, DE	1.800			SÍ	NO	SÍ	SÍ	SÍ
Glosario de Gestión de Patrimonio Cultural	2004	ES, EN	1.000			SÍ	SÍ	SÍ	SÍ	SÍ
Glosario de Prospectiva Tecnológica	2004	ES, EN	1.000			SÍ	NO	SÍ	SÍ	SÍ