ANÁLISIS ESTRUCTURAL Y SEMÁNTICO DEL TESAURO DE CIENCIAS DE LA DOCUMENTACIÓN "DOCUTES"

Blanca Rodríguez Bravo, Mª Luisa Alvite Díez, Ángela Diez Diez, Josefa Gallego Lorenzo, Amparo López García, Mª Antonia Morán Suárez, Carmen Rodríguez López y Lourdes Santos de Paz

Área de Biblioteconomía y Documentación, Universidad de León
Facultad de Filosofía y Letras; Campus de Vegazana s/n 24071 León (España)
Tfno.: +34. 987291469; E-mail: dphbrb@unileon.es

ABSTRACT

Se evalúa el Tesauro de Ciencias de la Documentación "Docutes", elaborado por el Área de Biblioteconomía y Documentación de la Universidad de León, atendiendo al análisis de los siguientes parámetros: composición, tamaño y cobertura, tasa de equivalencia, tasa de relación, tasa de enriquecimiento, tasa de precoordinación, notas de aplicación, valores de profundidad, morfología y aspectos estéticos y tipográficos.

Keywords: Tesauro de Ciencias de la Documentación, Docutes, Evaluación, Análisis.

1. INTRODUCCIÓN

El objetivo de este trabajo se centra en el análisis del Tesauro de Ciencias de la Documentación "Docutes" realizado por el Área de Biblioteconomía y Documentación de la Universidad de León gracias a las ayudas concedidas por la Junta de Castilla y León en el seno de las convocatorias de elaboración de *Recursos de Apoyo y Experiencias innovadoras de las Universidades de Castilla y León* de los años 2000 y 2002.

La construcción de un Tesauro en Ciencias de la Documentación pretende englobar la terminología necesaria para el tratamiento de la información de los distintos campos abarcados por el concepto científico de "Documentación" en un sentido amplio. El estudio y reflexión sobre los conceptos que integran las Ciencias de la Documentación, la elección de la terminología que mejor representa a esos conceptos y su estructuración semántica resulta igualmente útil, máxime por tratarse de un conjunto de ciencias jóvenes que tienen todavía mucho por definir en este terreno terminológico en el que muchos conceptos reciben variedad de denominaciones, resultando ciertamente recomendable un mayor grado de normalización y univocidad.

Para la construcción del Tesauro nos hemos servido del programa MultiTes, software específico de creación, mantenimiento y publicación de tesauros propiedad de la empresa ame-

ricana Multites http://www.multites.com. El programa nos ha facilitado sobremanera las tareas mecánicas de construcción del tesauro, ha resultado imprescindible para la validación de relaciones y la consistencia en la asignación de términos, asimismo ha generado automáticamente los correspondientes índices. En este momento nos hallamos en la segunda fase del proyecto dirigida, de un lado, a la extensión semántica y validación de la primera fase y, de otro, a la difusión del tesauro, que proporcionará realmente sentido al proyecto.

Hemos de reseñar la introducción en el tesauro de terminología de campos colaterales como la Bibliología, Historia del libro, Conservación e Informática que creemos complementan adecuadamente la base conceptual de "Docutes".

Durante los cursos académicos 2001/2002 y 2002/2003 se ha empleado el tesauro en el marco de la docencia de las asignaturas de Lenguajes documentales y Tesauros. Se ha evaluado de este modo su utilidad en la indización de literatura especializada, se ha detectado la necesidad de profundizar en algunos campos, abreviar otros, intercambiar sinónimos sustituyendo descriptores por no descriptores, etc. Hay que hacer notar que un trabajo de base terminológica como el que nos ocupa, y que refleja contenidos científicos, tiene que evolucionar paralelamente con las ciencias cuyos conceptos representa y con sus denominaciones, en nuestro caso las ciencias documentales. Un tesauro exige siempre, por tanto, un trabajo de mantenimiento.

Por último, el proyecto contempla la creación de una sede web sobre herramientas terminológicas dentro de la sede oficial del Área de Biblioteconomía y Documentación http://www3.unileon.es/dp/abd, hospedada en el servidor de la Universidad de León que nos permita introducir el tesauro en línea e insertar contenidos relacionados con la organización del conocimiento que proporcionen valor añadido. El tesauro será publicado en HTML para su consulta navegacional, desarrollando, asimismo, un CGI o tecnología ASP que permita la búsqueda de términos mediante el correspondiente formulario.

2. OBJETIVOS Y METODOLOGÍA

El objetivo de este trabajo se centra, como hemos señalado, en el análisis del Tesauro de Ciencias de la Documentación resultado de la primera fase del proyecto 2000/2001. Pretendemos que el estudio nos proporcione pautas para su perfeccionamiento que nos permitan cerrar la segunda fase en diciembre de 2003 y realizar la difusión correspondiente. Los resultados derivados de los indicadores propuestos nos darán idea de la calidad de nuestro trabajo y pondrán de manifiesto las posibles fortalezas y debilidades del mismo.

Partiendo de la revisión de la literatura sobre valoración de tesauros, Gil Urdiciain y Lancaster tomando como referencia las medidas cuantitativas propuestas por el Bureau Marcel Van Dijk y el trabajo de Kochen y Tagliacozzo, proponemos aplicar los siguientes criterios evalutivos:

Composición: se analiza la estructura del tesauro, los campos semánticos que lo componen, las formas de presentación e índices utilizados.

Tamaño y cobertura: volumen total de entradas del vocabulario, tanto descriptores como no descriptores así como volumen de descriptores correspondiente a cada una de las familias semánticas propuestas. Permite valorar el equilibrio o desequilibrio de las facetas y la cobertura o exhaustividad terminológica de las mismas.

Tasa de equivalencia: ratio entre el número total de no descriptores o términos equivalentes y el número total de descriptores del tesauro. Los valores recomendados se sitúan entre 0,5 y 2.

Tasa de relación -o *Razón de conexión* según el Bureau Marcel Van Dijk-: ratio entre el número total de relaciones asociativas recíprocas y el número total de descriptores. Los valores aconsejados se sitúan entre el 0,5 y 2.

Tasa de enriquecimiento: mide la proporción entre el número de relaciones de equivalencia, jerárquicas y asociativas y el número total de entradas. Kochen y Tagliacozzo denominan a esta medida *Razón de accesibilidad* señalando que indica la amplitud de las uniones, es decir, las referencias cruzadas entre los términos del vocabulario, siendo probable que altas cifras se correspondan con un tesauro más útil por su capacidad para inducir la caracterización de documentos y preguntas, siempre que no superen el valor de 5 que indicaría un exceso de reenvíos por descriptor que podría entorpecer el uso del tesauro. El valor recomendado oscila entre 2 y 5. Para hallar la tasa de enriquecimiento se ha empleado una muestra de 407 términos (descriptores y no descriptores) correspondientes a las entradas del índice alfabético por a, b y c.

Tasa de precoordinación: mide la media de términos significativos por descriptor contabilizando el número de descriptores (a) unitérminos, (b) bitérminos, (c) tritérminos, etc. y aplicando la fórmula:

$$t = \frac{a + 2b + 3c + 4d}{a + b + c + d}$$

Los valores recomendados se sitúan entre 1,5 y 2. La ratio se halla empleando una muestra de 343 descriptores correspondientes a las entradas del índice alfabético por a, b y c.

Notas de aplicación: se incluyen en este apartado las notas de alcance, históricas y de definición. Se considera que si en el lenguaje incluyen notas menos del 10% de los descriptores se puede producir cierta ambigüedad. Se halla dividiendo el número total de notas de aplicación empleadas en el tesauro entre el total de descriptores.

Valores de profundidad: tienen en consideración el nivel de sistematización del tesauro permitiendo conocer el número de niveles jerárquicos de los distintos campos semánticos y la especificidad del vocabulario empleado.

Morfología: se analiza la adaptación a las directrices establecidas en la norma UNE 50-106-90 sobre las características morfológicas y sintácticas de los términos que conforman el tesauro: orden de las palabras, técnicas de fraccionamiento, uso de singular y plural, uso de acrónimos y siglas, barbarismos y extranjerismos, etc.

Aspectos estéticos y tipográficos: análisis formal del tesauro atendiendo a los recursos utilizados para visualizar las jerarquías, diferenciar descriptores de no descriptores, etc.

3. RESULTADOS Y DISCUSIÓN

Composición

El tesauro de Ciencias de la Documentación se halla estructurado en siete campos semánticos que trascienden la división tradicional por disciplinas, Biblioteconomía, Archivística y Documentación, reflejando las facetas que las componen: materia sobre la que se trabaja, agentes, herramientas, etc.

Las familias son las siguientes:

01 Ciencias de la Documentación: Historia. Teoría. Sistemas.

02 Investigación y Metodología documental.

03 Información. Documentos. Fuentes de Información.

04 Tratamiento y recuperación de la información.

05 Sistemas de información.

06 Tecnologías de la información.

07 Profesionales y usuarios.

"Docutes" cuenta con tres índices. El índice jerárquico que permite visualizar globalmente de forma arbórea la estructura completa del tesauro, se inicia con el término cabecera correspondiente, para, a continuación, siguiendo un orden alfabético, mostrar los sucesivos niveles jerárquicos de los descriptores que conforman el tesauro.

El índice alfabético relaciona todos los descriptores y no descriptores indicando bajo cada uno de ellos las relaciones preferenciales que remiten de descriptores a no descriptores por medio del operador UF (Used For = Usado Por) y de los no descriptores a los descriptores con el operador USE. Aparecen, asimismo, las relaciones jerárquicas establecidas en el índice arbóreo: BT (Broader Term = Término Genérico) y NT (Narrower Term = Término específico), y las relaciones asociativas indicadas por el operador RT (Related Term = Término Relacionado).

El índice permutado KWOC permite localizar los descriptores sintagmáticos o precoordinados por el segundo o tercer término significativo.

Aun resta pendiente la elaboración de una introducción que describa el objetivo, el campo temático, el significado de los signos convencionales, las características del tesauro, el método de construcción y su desarrollo, etc.

"Docutes" se adapta, por tanto, a las directrices que recomiendan el uso de índices y la mínima doble estructura alfabética y jerárquica.

Tamaño y cobertura

En la fase analizada el tesauro ha quedado configurado en 7 familias que incluyen un total de 1520 términos de los cuales 1375 son descriptores y 145 se corresponden con no descriptores. Dado que se circunscribe a un único ámbito temático correspondiente a una ciencia todavía joven, consideramos suficiente el nivel global de cobertura que aporta.

Se aprecia uniformidad en la mayoría de las familias, si bien se observa un claro desequilibrio de los campos 06 y 07. Con respecto al campo de las Tecnologías de la información 06 se constata la vinculación estrecha entre la Documentación y la Informática y las Telecomunicaciones y, tal vez, se haya tendido al exceso buscando exhaustividad, no obstante, es evidente la notable influencia de las tecnologías de la información en el tratamiento documental así como la rápida obsolescencia de sus conceptos en consonancia con el vertiginoso desarrollo de sus herramientas.

Con respecto a la faceta 07 Profesionales y usuarios, el análisis nos hace poner en cuestión la posibilidad de enriquecer la terminología para lograr un equilibrio aceptable con el resto de las familias semánticas o quizá la perentoria necesidad de eliminar este campo temático e integrar su terminología en el 05 Sistemas de información.

Tabla 1. Cobertura de los campos semánticos

	Descriptores
01 Ciencias de la Documentación	213
02 Investigación y metodología	221
03 Información. Documentos. Fuentes	149
04 Tratamiento y recuperación	201
05 Sistemas de información	178
06 Tecnologías de la información	390
07 Profesionales y usuarios	23
Total	1375

Tasa de equivalencia

Aparecen 145 relaciones de equivalencia lo que supone una tasa del 0,105 que se aleja de los valores recomendados. Parece evidenciarse la necesidad de incrementar el número de términos no preferentes en el índice alfabético.

Tasa de relación

La tasa de relaciones asociativas es del 0,127 lo que manifiesta la necesidad de enriquecer las conexiones entre descriptores de diferentes campos semánticos.

Tasa de enriquecimiento

Sobre la muestra utilizada la tasa es de 1,889, cercana, por tanto, al mínimo recomendado. Como se deduce de las medidas anteriores, "Docutes" presenta una ratio aceptable de relaciones jerárquicas y la evaluación manifiesta que el incremento de las relaciones asociativas y de equivalencia permitirán alcanzar los resultados recomendados.

Tasa de precoordinación

La tasa resultante es del 2,002. Los descriptores que cuentan con dos términos significativos son mayoritarios, suponiendo el 64,1%, los unitérminos representan el 18,6% y los descriptores tritérminos un 15,4%. En la muestra analizada solamente un 1,7% de los descriptores tienen cuatro términos significativos.

El alto nivel de precoordinación obedece, de un lado, a las propias características del vocabulario tratado, caracterizado por su generalidad y polisemia y, de otro, a la opción que nos ha llevado a preferir este mecanismo para evitar la ambigüedad sobre el uso de aclaradores o notas de aplicación.

Tabla 2. Tasas

	Tasa de equivalencia	**Tasa de relación**	**Tasa de enriquecimiento**	**Tasa de precoordinación**
Docutes	0,105	0,127	1,889	2,002
Valores aconsejados	0,5-2	0,5-2	2-5	1,5-2

Notas de alcance

Aparecen 32 notas de aplicación lo que nos permite hablar de una tasa del 0,023. Se trata de un resultado pobre que creemos obedece a la alta tasa de coordinación y al buen funcionamiento de las relaciones jerárquicas, ya que como señala la norma UNE 50-106-90, el tesauro presenta, a través de su estructura, una serie de relaciones que establecen en general

el contexto de "significado" de un término dado, con especial referencia a términos de connotación más amplia o más restringida. Normalmente, esto basta para indicar la interpretación que se hace de un término. La presentación de relaciones jerárquicas ya informa, con frecuencia, del significado de un término sin necesidad de añadir una definición o una nota aclaratoria.

Profundidad

En general es elevado el grado de especificidad alcanzado, descendiendo cuatro de los campos semánticos hasta el séptimo nivel de jerarquía, si bien se observa un predominio de 3 y 4 estratos de profundidad. No obstante, tal vez resultara más consistente reducir a 5 o 6 los niveles. En tal caso, se habría de dotar de profundidad al campo 03 Información. Documentos. Fuentes de información y al 07 Profesionales y usuarios. Se manifestaría igualmente necesaria la generalización en las familias semánticas 01 Ciencias de la Documentación, 02 Investigación y Metodología documental y 06 Tecnologías de la información.

Por último, llama la atención el vértice que muestra la faceta 04 Tratamiento y recuperación de la información con 100 descriptores de tres niveles de jerarquía. Quizá fuera conveniente enriquecer el segundo nivel jerárquico para equilibrar el árbol semántico de este campo.

Tabla 3. Nivel de profundidad de los campos semánticos

	1	2	3	4	5	6	7
01	2	9	37	62	33	6	4
02	6	44	56	48	46	20	1
03	3	41	81	24	—	—	—
04	3	15	100	47	33	3	—
05	8	27	64	65	24	—	—
06	4	32	91	77	56	42	8
07	3	15	4	1	—	—	—
Total	29	183	433	324	192	71	13

Morfología

El tesauro se atiene a las directrices establecidas en la norma UNE 50-106-90 en lo que respecta al uso de singulares y plurales, no existe inversión, siendo el orden de entrada directo en todos los casos. Se ha procurado elegir como término preferente el término más aceptado. Se han establecido relaciones de equivalencia entre los acrónimos y sus desarrollos. Se ha intentado no introducir terminología anglosajona, si bien en el campo 06 ha sido inevitable.

Aspectos estéticos y tipográficos

Nuestros esfuerzos hasta el momento se han orientado hacia la construcción del tesauro dirigiendo la atención al contenido y a su usabilidad como producto en línea, de ahí el interés por facilitar la consulta navegacional, ya resuelta, y la interrogación directa, en fase de desarrollo.

La tipografía y el aspecto estético vienen determinados por el software empleado y de cara a una potencial edición impresa deseamos personalizar determinados aspectos. Aunque la visualización de la profundidad jerárquica se garantiza, gracias a la sangría correspondiente, sería deseable la introducción de notaciones que clarificasen la misma. Con respecto al índice alfabético preferiríamos utilizar los operadores en castellano y una distinción tipográfica entre descriptores y no descriptores. Se debería introducir, igualmente, un mismo código que permita la conexión desde el índice alfabético al jerárquico. En el índice KWOC se mejoraría la legibilidad destacando los términos de entrada.

4. CONCLUSIONES

Consideramos que el tesauro cumple su finalidad pues como apunta Lancaster (2001): "*En la práctica (...) un tesauro sólo puede ser juzgado por el uso que se hace de él –es decir, su validez para fines de recuperación- (...). Los términos deben ser lo suficientemente específicos como para permitir la recuperación de los documentos deseados pero sin recuperar, al mismo tiempo, una gran cantidad de documentos no deseados. De igual modo, la estructura de referencias cruzadas del tesauro debe ofrecer una ayuda positiva para que el usuario seleccione los términos más apropiados para su necesidad concreta de información; y, en el caso de una búsqueda exhaustiva, debe conducir al usuario a todos los términos que podrían ser relevantes*".

En este sentido hemos constatado la eficacia del tesauro en las prácticas docentes de indización. La actual evaluación complementa la valoración previa obtenida de su uso.

Los resultados alcanzados nos permiten realizar la revisión del tesauro incidiendo en la corrección de los aspectos menos favorables como son los relacionados con la escasez de relaciones de equivalencia y asociativas así como el desequilibrio del campo temático 07.

Como se ha señalado, el mantenimiento de un tesauro debe ser un proceso continuo que permita detectar carencias e, igualmente, términos sobreutilizados o de raro uso, ambos ineficaces en la recuperación de información. Como establece la norma UNE 50-106-90, los términos muy usados pueden subdividirse en términos más específicos y los no usados, por el contrario, eliminarse dejando únicamente su genérico. Deseamos que la difusión de este tesauro amplíe su número de usuarios y ello nos permita obtener información de la

utilización de su terminología en contextos reales de indización y recuperación, es decir, en unidades de información.

Mientras tanto habremos de limitarnos, para su actualización, a los resultados obtenidos de su aplicación en las prácticas de indización de literatura especializada actual en la Diplomatura de Biblioteconomía y Documentación, y a nuevos trabajos de análisis como el presente, que ha resultado de inestimable ayuda.

5. BIBLIOGRAFÍA

Bureau Marcel Van Dijk (1976). *Definition of Thesauri essential characteristics*. Brussels.

Gil Urdiciain, B. (1998a). Evaluación del rendimiento de tesauros españoles en sistemas de recuperación de información. *Revista Española de Documentación Científica,* vol. 21, n. 3, pp. 286-302.

——— . (1998b). Evaluación semántica y estructural de tesauros. *Revista Española de Documentación Científica,* vol. 21, n. 2, pp. 173-192.

Kochen, M. y Tagliacozzo, R. (1968). A study of cross-referencing. *Journal of Documentation,* vol 24, n. 3, pp. 173-191.

Lancaster, F. W. (1995). El *control del vocabulario en la recuperación de información.* Valencia: Universitat de València, pp. 171-171.

——— . (2001). Elaboración y mantenimiento de tesauros. En Lancaster, F. W. y Pinto Molina, M. (coord..) *Procesamiento de la información científica*. Madrid: Arco Libros, pp. 182-193.

Mochón Bezares, G. y Sorlí Rojo, A. (2002).*Tesauro de Biblioteconomía y Documentación*. Madrid: Consejo Superior de Investigaciones Científicas.

Rodríguez Bravo, B. y Alvite Díez, M. L. (2002). Construcción de un tesauro en Ciencias de la Documentación aplicado a la docencia de técnicas documentales. En *1ª Jornadas de tratamiento y recuperación de información (JOTRI).* Valencia: Universidad Politécnica, pp. 151-156.

UNE 50-106-90. (1999). *Directrices para el establecimiento y desarrollo de los tesauros monolingües.* Madrid: AENOR.