Compilación de un macrotesauro conceptual para los centros españoles de información juvenil

Miguel-Ángel López Alonso
Facultad de Comunicación y Documentación
Departamento de Informática, Universidad de Extremadura

1.1. Resumen

Se presenta el proceso de construcción de un tesauro para los centros de información juvenil de las diecisiete autonomías españolas. Se partió de la macroestructura de la anterior Clasificación Temática de la Red Española de Servicios de Información Juvenil, analizando y modificando profundamente su primer nivel estructural. Los demás términos fueron incorporados tras su evaluación como individuales dentro de la nueva estructura jerárquica. Como resultado, el tesauro quedo configurado en diez macrofamilias que incluyen un total de 3.384 términos, de los cuales 2.543 son descriptores y 841 no descriptores, 183 notas de alcance, 1.638 relaciones asociativas y 1.684 reenvíos.

Palabras clave: Información juvenil. World Wide Web. Tesauros. España.

1.1. Abstract

The process of building a thesaurus for the Spanish Youth Information Centres of the seventeen autonomous governments is presented. The previous Thematic Classification of the Spanish Youth Information Net of Centers was used as the starting point. The firs structural model was analysed and deeply, and, after their assessment, the other terms were treated as descriptors inside the new hierarchical structure. As a result, the thesaurus has ten macrohierarchies, a total of 3,384 terms, 2,543 descriptors, 841 linguistic equivalents, 183 use notes explicative, 1,683 associative relations and 1,684 cross-references.

Keywords: Youth Information. Thesauri. World Wide Web. Spain.

1. Introducción

En estos momentos se vuelve a pensar seriamente en los tesauros como herramienta de precisión para la recuperación de la información relacionada semánticamente, que describe el mismo concepto en fuentes diferentes, mediante la identificación de sus diferencias semánticas —por ejemplo, la terminología,

la estructura, el contexto, etc.— en las numerosas bases documentales de la Web. Estos tesauros difieren de los compilados en los años 60 y 70 en que se compilan a partir del sublenguaje científico tomado del lenguaje natural en contextos concretos, y se utilizan principalmente en la recuperación de documentos no indizados previamente con ninguna otra herramienta lingüística documental. Se integran con ventaja en los sistemas de gestión de la información para mejorar la pertinencia de las búsquedas, debido a sus numerosas relaciones asociativas contextuales. Estos nuevos tesauros son una superserie de sublenguaje controlado en un dominio científico específico, que se usan durante el proceso de indización —como ayuda para la identificación de los conceptos— y en el proceso de recuperación —como fuente de nuevos términos que identifiquen conceptos y aumenten la precisión y exhaustividad de las búsquedas booleanas.

Cada día más, la dicotomía entre lenguajes controlados y lenguaje natural está dando paso a la integración del lenguaje del usuario en los lenguajes controlados. Estos lenguajes controlados parten de las clasificaciones facetadas —Ranganathan, Bliss, y ahora Yahoo o Google en Internet— para autogenerar los hipertesauros conceptuales de Internet, mediante la extracción de la terminología de los documentos en texto completo y su integración en las bases de conocimientos descentralizadas de la Web (Muddamalle, 1998). La gran utilidad de los tesauros durante la búsqueda y recuperación de la información estriba en que sirven de puente de "comunicación conceptual" entre los metadatos aportados por los indizadores y los conceptos demandados por los usuarios. Su proceso de reutilización integrada pasa por añadir "conocimiento" a los tesauros jerárquicos existentes en cualquier subárea del conocimiento y convertirlos en tesauros "conceptuales", enriqueciéndolos con multitud de relaciones asociativas entre los descriptores y numerosas relaciones cruzadas entre los descriptores y los no descriptores. Tras su posterior conversión en hipertextuales, se les relaciona en redes semánticas en Internet de manera que constituyan una poderosa ontología terminológica en el hiperespacio (Shiri y Revie, 2000).

2. Antecedentes y conceptos previos

En general, un tesauro interrelaciona los términos de un vocabulario especializado para construir un lenguaje documental que es utilizado por los indizadores para la representación abreviada de los documentos de las bases de datos, y como guía terminológica para la normalización de las entradas de los encabezamientos de materias en la clasificación documental. Además, posibilita las tareas de recuperación documental: ecuaciones de búsqueda, navegación, asociación terminológica, comprensión del entorno de los descriptores del tesauro, etc.

Los primeros desarrollos prácticos enfocados a la "recuperación contextual" datan de los años ochenta en EE. UU. Se desarrollaron dentro de los sistemas

integrados de gestión de la información de grandes instituciones como la National Library of Medicine (1972), el American Petroleum Institute (1987), o la NASA (1994). Aunque las investigaciones se iniciaron para mejorar la indización, después derivaron hacia la recuperación automatizada en grandes bases de datos documentales, y con el desarrollo de las herramientas de la Inteligencia Artificial se fueron aplicando con distinto éxito al desarrollo de agentes inteligentes de campos específicos del conocimiento científico. Autores como Schmitz-Esser (1991), Bates (1998) o Milstead (2000) nos vienen introduciendo en las características de este tipo de tesauros conceptuales que palian en parte la indeterminación de las búsquedas en lenguaje natural de los usuarios no expertos con el sublenguaje científico de los profesionales de un área específica del conocimiento. Estos nuevos tesauros conceptuales, a diferencia de los tesauros documentales usados para la indización, incluyen entre sus requisitos los siguientes: listar todos los términos no "vacíos", usados en cualquier momento en el catálogo o en la base de datos; distinguir cuidadosamente los términos realmente usados de los no usados; añadir notas de alcance que aclaren las dudas a los posibles usuarios, incluso aportando algunas definiciones; contener equivalencias auto explicativas de los términos y/o sus relaciones cruzadas y asociativas; aportar un extenso vocabulario superior al número de términos controlados; e incluir los términos coloquiales, variaciones de los términos reconocidos e incluso truncamientos.

En la materia específica de la juventud, se disponía de dos tesauros generales bien conocidos a nivel mundial, pero que no tenían aplicación directa para la compilación de un tesauro utilizable en los sistemas de información de los Centros de Información Juvenil españoles: uno en lengua inglesa, elaborado en 1981 por Aitchison y su equipo para el National Youth Bureau de Inglaterra, denominado *Thesaurus on Youth*; y otro en lengua francesa, elaborado en 1988 por la profesora Szpakowska y otros profesores de la Escuela de Biblioteconomía y Ciencias de la Información de Quebec (Canadá) para la Universidad de Montreal, denominado *Thésaurus Jeunes Gens*.

Para la compilación del *Tesauro de Juventud para Centros de Información Juvenil* se formó un grupo de trabajo con los documentalistas responsables de los 17 Centros Coordinadores Autonómicos y/o Regionales de Información Juvenil y el Servicio de Información y Difusión del Instituto de la Juventud Español (INJUVE, Ministerio de Trabajo y Asuntos Sociales), que el 7 de octubre de 1999 encomendó la ejecución del proyecto al Prof. Miguel-Ángel López Alonso. El equipo fue tomando las sucesivas decisiones sobre las diferentes familias jerárquicas y los términos que se integrarían en cada una de ellas. Desde el 21 de septiembre de 1999 se celebraron reuniones periódicas de trabajo de dicho grupo con el profesor López Alonso, doctor del área de Lenguajes Documentales por la Universidad Carlos III de Madrid, y actual profesor titular de universidad de la

asignatura "Diseño de Sistemas de Indización" en la Facultad de Biblioteconomía y Documentación de la Universidad de Extremadura, que, en los períodos de trabajo entre dichas reuniones, se entregó a la tarea de crear e integrar en los índices del tesauro tanto sus propias aportaciones terminológicas como las indicadas por los responsables de los Centros Coordinadores.

De las dos soluciones teóricas más viables como punto de partida —compilar un tesauro totalmente nuevo o aprovechar la *Clasificación Temática* existente en los Centros de Información Juvenil españoles como fuente de segmentaciones temáticas y de terminología— se optó por la segunda para "evitar que ello supusiera una revolución para los servicios de información a los jóvenes que llevan trabajando con ella más de diez años y la han adaptado a sus Bases de Datos". Se decidió, pues, tomar su macroestructura, analizando y modificando profundamente su primer nivel estructural, y utilizando los demás términos como meros descriptores individuales dentro de la nueva estructura jerárquica, siempre que éstos fueran considerados útiles para la representación de los conceptos sobre juventud. Para el desarrollo del tesauro, el citado profesor ha utilizado una metodología que trata de integrar las directrices de las Normas ISO y UNE de confección de tesauros con su propia experiencia, siguiendo una metodología propia asimilada de la del CINDOC durante la elaboración de su tesis doctoral, *La construcción de un Tesauro en Derecho Comercial* realizada con una beca de investigación entre 1994 y 1997, y los avances más recientes en la construcción de tesauros. Una vez finalizados los trabajos de ordenación jerárquica se ha procedido al reenvío de los términos sinónimos no descriptores a sus correspondientes términos descriptores normalizados, al acompañamiento a aquéllos que lo precisen de sus correspondientes notas de alcance, y a la búsqueda intensiva de las referencias cruzadas y las relaciones asociativas entre ellos.

Para evitar la debilidad de las asociaciones terminológicas inherentes a todos los sistemas de información documental, derivadas de la menor riqueza conceptual de los lenguajes controlados en relación con el lenguaje natural, se intentó mejorar la subjetividad tradicional de las relaciones asociativas mediante la emulación de un modelo de "composición inferencial facetado ad hoc", que toma el concepto informático de las reglas establecidas para los esquemas de representación del conocimiento del tipo Description Logic (DL) (Svenonius, 1986). Este desarrollo conceptual aplica las técnicas de representación del conocimiento a la gestión y al razonamiento con los conceptos terminológicos que, durante el proceso de construcción de la clasificación, ayudan a producir jerarquías coherentes y a asegurar que las asociaciones representadas en el tesauro sean sensibles al contexto. Este caso de debilidad de la estructura de los lenguajes controlados —que se ven como un subconjunto restringido de su correspondiente lenguaje natural del que toman prestado su terminología básica y su estructura conceptual, y, por

tanto, el significado dado a los términos puede no corresponder exactamente con el que tienen en su emplazamiento "natural", o no utilizan todos los significados de los lenguajes naturales para expresar sus asociaciones estructurales—, provoca que algunos sistemas documentales no logren representar sus asociaciones de manera adecuada y deban recurrir a métodos inferenciales para la identificación de las relaciones, sus participantes y sus papeles (López Alonso, 2000).

Para paliar esta debilidad se han venido utilizando diversas estrategias estadísticas de ponderación —por ejemplo, la búsqueda por proximidad, relación histórica entre términos afines, uso de indicadores de significado, etc.— que pueden evitarse con la estructura integrada de la aproximación semántica propuesta por Soergel (2001). En el caso más simple, la asociación se forma por todos los objetos que pueden alcanzarse desde un objeto inicial, mediante vínculos de un tipo determinado con sus vecinos. Por ejemplo, en un sistema documental la asociación puede estar formada por todos los documentos que critican un determinado documento, el cual puede alcanzar a los demás documentos siguiendo vínculos binarios del tipo "criticado por". En este contexto se da una relación directa entre asociación y pregunta. Una pregunta conduce a una asociación, y cada asociación corresponde a una pregunta. Halasz (1988) habla de los "nodos compuestos virtuales" que resultan de la formulación de una pregunta y aclara que "ésta puede permitir que el lenguaje usado para descripciones estructurales virtuales —o Description Logic (DL) de base semántica/conceptual— sea el mismo que el lenguaje de la pregunta usado para las búsquedas y para los filtros del interfase".

3. Objetivos

Se trataba de reorganizar la estructura del índice de contenidos de la Red Española de Servicios de Información Juvenil, no como otra lista de encabezamientos de materias o su mera adaptación a cualquiera de los sistemas de clasificación universal, sino como un tesauro conceptual que permitiese la recuperación de documentos en las bases de datos de información de dicha red tanto a partir del lenguaje natural de los usuarios como del lenguaje controlado que se venía utilizando, tomara las clases del citado índice de contenidos (Busch, 1999).

Los objetivos más inmediatos que se esperaban obtener fueron, por un lado, evitar la repetición innecesaria de los mismos descriptores en las distintas familias del índice de contenidos; y, por el otro, permitir la utilización de operadores booleanos en las búsquedas de documentos, de manera que aumente la precisión.

Los objetivos a medio plazo, tras una etapa previa de equiparación informática de las clasificaciones numéricas de la lista de contenidos actual con las del nuevo tesauro conceptual fueron, en primer lugar, poder cargar en cualquier PC un programa simple de "gestión de tesauros" que incluyera el *Tesauro de*

Información y Documentación Juvenil, y permita su utilización integrada en todas las indizaciones y recuperaciones futuras de documentos de la red del INJUVE; y, en segundo lugar, una mayor simplicidad, rapidez y, por tanto, deseo de utilización de las bases de datos de información de los citados Servicios de Información Juvenil. De manera más específica se pretendía:

1º) La reutilización de los epígrafes existentes en el índice de contenidos actual como términos controlados o descriptores del nuevo tesauro conceptual, con la finalidad de que los documentos indizados hasta la actualidad no se pierdan al recuperar con los términos del nuevo tesauro.

2º) La actualización de los términos caídos en desuso, así como su ampliación con todo tipo de términos controlados existentes en otros "tesauros en materia de juventud" —los Servicios de Información Universitaria, las propuestas de los distintos centros de documentación integrados en el INJUVE, etc.—, tras su previa normalización.

3º) La relación, primero jerárquica y después asociativa, tanto de los descriptores (existentes y añadidos) como de los términos pertenecientes al lenguaje vulgar de los usuarios. Estos últimos que, a pesar de no permitir una normalización como términos controlados, permitan el reenvío a los descriptores normalizados mediante relaciones del tipo USE o "Véase".

La voluntad del citado grupo de trabajo ha sido la confección de una herramienta conceptual de indización y recuperación eminentemente práctica para los sistemas de información juvenil, aunque no vinculada a ninguno de ellos en concreto, con el objetivo de presentar niveles medios de especificidad gracias al amplio desarrollo de cada una de las áreas conceptuales.

Es preciso destacar la ardua tarea de homologación de las colaboraciones terminológicas de los responsables de los diecisiete Centros Autonómicos de Información Juvenil, debida mayoritariamente a la falta de integración previa de sus sistemas de información, y a las dificultades en la adopción en estos sistemas de una herramienta normalizada del tipo del macrotesauro conceptual que presentamos, integrado por amplias áreas del conocimiento de las que cualquiera de ellas podrían dar lugar a un tesauro específico bien diferenciado.

4. Metodología de trabajo y fases de realización

La fase de compilación ha llevado casi tres años de trabajo conceptual continuado, y ha constado de cuatro subfases distintas, cada una de seis meses de trabajo; las dos primeras de carácter documental y las dos últimas de carácter cognitivo.

La primera subfase se dedicó a recopilar todos los términos posibles, sugeridos o encontrados, y a integrarlos en las diferentes segmentaciones temáticas correspondientes a la cabecera del primer nivel jerárquico en el que se divida

teóricamente el tesauro, asignándoles a cada uno de ellos una clasificación de dos dígitos, de acuerdo con el siguiente ejemplo:

A: Información juvenil
A1: cultural
A2: deportiva
A3: informativa
A4: orientativa
[...]
B: Temas afines sobre la juventud

La segunda subfase consistió en integrar los términos agrupados en las diferentes segmentaciones temáticas, a tenor del siguiente ejemplo:

A1: cultural
A1: Cultura
A1: Educación
[...]
A2: deportiva
A2: Deportes
A2: Tiempo libre

Al finalizar esta fase se redactaron dos listas alfabéticas. En primer lugar, una lista sistemática con los dos niveles de profundidad y los descriptores en orden alfabético dentro del segundo nivel. Ambas se repartieron entre los miembros de la Comisión Coordinadora de Centros de Información Juvenil para su verificación, depuración y mejora sucesivas, como paso previo a las fases cognitivas tercera y cuarta del proyecto de *Tesauro de Información y Documentación Juvenil*.

La tercera subfase, de carácter cognitivo, consistió en: a) la reasignación de los descriptores reenviados de unas cabeceras a otras más idóneas, para su inclusión en las mismas; b) la reelaboración de la situación completa de las diferentes subcabeceras para armonizar la concreción y profundidad de las mismas; c) la elaboración del tercer nivel jerárquico en cada una de las subsegmentaciones que así lo precisaran, integrando sus descriptores correspondientes en las nuevas divisiones establecidas en las mismas; y d) la repetición de esta forma de proceder en cada subcabecera del tesauro, cuantas veces sea procedente, hasta completar la jerarquización completa de todos sus descriptores.

Por último la cuarta subfase, de trabajo cognitivo-informático, se dedicó, en primer lugar, a acompañar a aquellos descriptores normalizados que lo precisaran de sus correspondientes notas de alcance, y a buscar las referencias cruzadas y las relaciones asociativas entre ellos, así como a asignar a los términos no descriptores los reenvíos a sus correspondientes términos descriptores; y, en segundo lugar, a generar mediante un programa informático de "gestión de tesauros", el índice permutado Kwic o Kwoc, y el índice conceptual o completo.

5. Resultados

Como resultado final, el tesauro ha quedado configurado en diez macrofamilias que incluyen un total de 3.384 términos, de los cuales 2.543 son descriptores relacionados jerárquicamente del tipo TG/TE y otros 841 términos corresponden a no descriptores. Las notas de alcance de los términos más específicos ascienden a 183, las relaciones asociativas entre descriptores del tipo TR suman 1.638 y los reenvíos entre los no descriptores y sus correspondientes descriptores del tipo USE/UP se elevan a 1.684. Finalmente, los anexos de acrónimos y auxiliares de lugar constan de 1.100 términos.

6. Apuestas de futuro

El proyecto que aquí presentamos pretende inscribirse en el trabajo de sistematización conceptual y terminológica que se está llevando a cabo en numerosas áreas del conocimiento, con el objeto de que la recuperación de la información documental descentralizada —haya sido indizada previamete o no— de la Web pueda realizarse con unos niveles de eficacia cada vez mayores (Hoy, 1998). Se han tenido presentes los estudios previos sobre la introducción de las nuevas tecnologías informáticas en la difusión de los corpora terminológicos (Valle et al., 2000) que se utilizan tanto para la indización de textos científicos, como para la preparación de perfiles de búsqueda en sistemas automatizados interactivos. En dichos trabajos se exponen los desarrollos para un tratamiento informático que resuelva la puesta en formato HTML de los citados vocabularios controlados (ISO-5964), permitiendo obtener una estructuración de la información terminológica que facilite su consulta mediante un navegador de Internet en las labores de indización y recuperación documental. Se trata de diseñar y estructurar un árbol de directorios hipertextuales que permita la creación de los ficheros relativos a los vocabularios para resolver consultas en tiempo real o difundir los contenidos mediante CD-ROM.

Las aproximaciones teóricas conceptuales sugieren una estructura simple en lenguaje natural que incluya en las bases documentales un fichero de punteros a los textos completos de los documentos, como modelo de arquitectura para el almacenamiento de la información. La estructura ontológica de representación del conocimiento sugerida es una red asociativa de nodos de conceptos o proposiciones cuya estructura semántica no haya sido rígidamente preestablecida, sino que se enriquezca en el contexto de la tarea para la que se utilice y que sea lo suficientemente potente para inducir relaciones inferenciales y enlaces a dichos punteros.

Además, para la navegación dinámica e interactiva por el hiperespacio vectorial se deberán definir mecanismos hipertextuales interactivos capaces de relacionar las ontologías con los diferentes espacios documentales. Dicha estructura

supondrá un medio virtual de relación entre las diferentes bases de datos de los sistemas de información documental descentralizados de Internet, que permitirá a los usuarios adaptarse con suficiente rapidez a la evolución permanente del entorno documental (Fugmann, 1997).

7. Agradecimientos

Este Proyecto ha sido financiado por el Instituto de la Juventud del Ministerio de Trabajo y Asuntos Sociales (INJUVE) durante el periodo del 21/9/1999 al 30/11/2001.

8. Referencias

Bates, M.J. (1998). Indexing and access for digital libraries and the Internet: human database and domain factors. // JASIS. 49:13 (1998) 1186-1205.

Busch, J.A. (1999). From Authority Files to Ontologies: Knowledge Management in a Networked Environment. // Proceedings of the Subject Analysis and Retrieval Working Group Conference: Controlled Vocabulary and the Internet, Sep. 29. Bethesda, MD, 1999.

Fugmann, R. (1997). Bridging the gap between database indexing and book indexing. // Knowledge Organization. 24:4 (1997) 205-212.

Halasz, F. (1988). Reflections on notecards: Seven issues for the next generation of hypermedia systems. // Communications of the ACM. 31:7 (1988) 836-852.

Hoy, M. (1998). Understanding official government terminology: natural language searching and government thesauri. Canberra: National Archives of Australia, 1998.

López Alonso, M.A. (2000). Las estructuras conceptuales de representación del conocimiento en Internet. // Scire. 6:1 (2000) 107-123.

Martínez, C.; Lucey, J.; Linder, E. (1987). An Expert System for Machine-Aided Indexing. // J. Chem. Inf. Comput. Sci 27:4 (1987) 158-162.

Milstead, J.L. (1995). Invisible Thesauri: the year 2000. // Online & CDROM Review.19: 2 (1995) 93-94.

Muddamalle, M.R. (1998). Natural language versus controlled vocabulary in information retrieval: a case study in soil mechanics. // JASIS, 49:10 (1998) 881-887.

National Library of Medicine (1972). Medical subject headings. Washington DC: Government Printing Office, 1972.

Schmitz-Esser, W. (1991). New Approaches in Thesaurus Application. // International Classification. 18:3 (1991) 143-147.

Shiri, A.A.; Revie, C. (2000). Thesauri on the Web: current developments and trends. // Online Information Review. 24:4 (2000) 273-280.

Soergel, D. (2001). Report from the ASIST SIG/CR 11th Classification Research Workshop 2000: Classification for User Support and Learning. // Bulletin of ASIST.27:4 (April-May 2001).

Silvester, J.P.; Genaurdi, M.T.; Klingbiel, P.H. (1994). Machine-aided indexing at NASA. // Information Processing Management. 30 (1994) 631-645.

Svenonius, E. (1986). Unanswered questions in the design of controlled vocabularies. // JASIS. 37:5 (1986) 331-340.

Valle et al. (2000). Tesauros en HTML: Un modelo de diseño y estructura para su consulta en la malla mundial (WWW) // Rev. Esp. Doc. Cientif. 23:2 (2000) 159-178.