

EL THESAURUS

COMUNICACION presentada por M.^a ANGELES AMORÓS DEVESA.

M.^a DOLORES MERÁS MENÉNDEZ.

JOSÉ AGUILO REGLA

Facultad de Derecho. Universidad de Alicante

1. IMPORTANCIA DE LA INFORMACION DE CARA AL SIGLO XXI

Se puede decir ya, que las gentes y las naciones han tomado conciencia de la importancia que tiene la información en nuestras sociedades contemporáneas.

Aunque se venían acumulando datos estadísticos relativos al crecimiento de la ciencia y se habían efectuado interpretaciones descriptivas del mismo, fue el historiador y físico Derek J. de Soya Price, quien llegó a formular de forma precisa la ley del crecimiento exponencial de todos los aspectos medibles de la ciencia; y una de las características, quizás la más importante, de este crecimiento es su ritmo —mucho más rápido que el de la mayoría de los demás fenómenos sociales—, de tal forma, que el investigador tiene que ocupar mucha parte de su tiempo en la puesta al día de sus conocimientos y técnicas, mediante la lectura de los documentos primarios.

Así pues, es totalmente prioritario tratar de organizar y hacer asequible al investigador todo ese caudal de información que le será preciso para un trabajo más productivo y de mayor calidad.

De la misma manera que la consecución del bienestar social e individual debe apoyarse en los logros de la actividad científica y tecnológica, esto no podría organizarse ni desarrollarse sin la aportación adecuada y continua de aquellos servicios que, como la información científica, producen, tratan y difunden, para su mejor percepción, asimilación y transferencia, los últimos avances del conocimiento científico y tecnológico, independiente de la forma y disciplina en que se produzcan.

Cada día es más evidente la significación económica y cultural de la información científica y técnica. La incidencia de estos servicios en los procesos productivos y empresariales, donde las tecnologías y los mercados alcanzan elevadísimos niveles de dificultad e inversión económica, es

fundamental. Por otra parte, la extensión y profundización cultural como factor coadyuvante del bienestar, necesita el auxilio instrumental de las técnicas, recursos y sistemas de información científica y técnica, que permiten al ciudadano y a la colectividad acceder más fácilmente a las fuentes del conocimiento.

2. LAS BIBLIOTECAS COMO CANAL DE INFORMACIÓN

Entre los recursos materiales de información están las instituciones implicadas en el acopio, proceso y transmisión de la información científica y técnica en su sentido más amplio, es decir, la información no sólo referida a las ciencias puras y aplicadas, sino también a las ciencias humanas y sociales. Entre estas instituciones figuran las Bibliotecas, los Archivos, los Museos, los Centros de Análisis de la Información, los Centros de Referencia, las bases de Datos y las Redes Públicas de transporte de Datos.

Así pues, las Bibliotecas son uno de los medios de transmisión de conocimiento; a ellas les están confiadas las funciones de depósito, preservación, conservación, organización y difusión del proceso de comunicación. Para lograr estos objetivos, las Bibliotecas deben estar bien organizadas y puestas al día lo que no ocurrirá si no son consideradas como instrumentos esenciales en cada uno de los niveles de la educación y de la investigación, y si sus servicios no están orientados a los grupos de usuarios o lectores implicados en cada caso.

Debido a la gran explosión de información que se ha producido en este siglo —y todo indica que la producción intelectual no sólo no va a disminuir sino más bien a crecer a un ritmo cada vez mayor— el carácter adecuado de los anteriores servicios solo podrá garantizarse por la vía de su mecanización. Es completamente necesario agilizar el acceso del lector a cualquier información que precise y poder ofrecerle con prontitud y de la manera más completa posible la bibliografía que requiere su objeto de estudio.

3. UNA BIBLIOTECA ESPECIALIZADA DE LA UNIVERSIDAD: LA DE LA FACULTAD DE DERECHO

Desde la Baja Edad Media, cuando la Institución Universitaria aparece como tal, los Centros de Enseñanza del Derecho han sido las Instituciones que en las sociedades occidentales han producido y consumido una mayor cantidad de información.

Su función principal hoy, como institución de estudios superiores, debería ser la de crear una atmósfera propicia para la creatividad intelectual, bajo la cual el individuo puede aprovecharse de recursos tales como bibliotecas bien surtidas y atendidas. En los últimos tiempos, el papel de la Universidad en la producción de información, se ha intensificado de forma apreciable mediante las publicaciones que sus investigadores rea-

lizan cada vez en mayor número; pero es evidente que, para producir información, estos investigadores necesitan, a su vez, consumir información.

Esta información es también indispensable para la realización de su trabajo docente que consiste esencialmente en transmitir a los estudiantes los conocimientos previamente adquiridos por ellos.

Por otro lado, durante las últimas décadas, las necesidades informativas de la comunidad universitaria han cambiado de manera significativa al necesitar los profesores e investigadores recurrir a los materiales de estudio que pertenecen a diversas disciplinas y rompiendo así las fronteras tradicionales. Estas necesidades interdisciplinarias han creado problemas al consumidor de información. Artículos relevantes para su trabajo pueden aparecer en publicaciones relacionadas sólo periféricamente con su disciplina nominal. Como la elaboración convencional de índices tiende generalmente a cubrir sólo una única materia, se hace necesario recurrir a instrumentos de referencia para poder llevar a cabo la investigación.

Así pues, las exigencias de los estudios interdisciplinarios, la proliferación de información y la necesidad creciente de disponer de conocimientos actualizados, han afectado también a las bibliotecas universitarias que se han convertido (o deben convertirse) en centros vivos donde se proporcionan servicios basados en una información selectiva y actual.

Es evidente que la finalidad de la mecanización de cualquier centro de documentación (como por ejemplo, una biblioteca) no es otra que la de recuperar automáticamente la información que en él se contiene. Es de sobra conocido que, a grandes rasgos, existen dos métodos distintos para realizar dicha recuperación de información: el método del texto íntegro y el método de las palabras-clave⁽¹⁾. El primero consiste en asumir como clave de recuperación y búsqueda todas las palabras contenidas en el documento memorizado (salvo las palabras vacías, que una vez seleccionadas y catalogadas constituirán el Thesaurus negativo). El segundo método consiste en anteponer a un cierto texto un conjunto de palabras que identifiquen su contenido; cuando esta selección se hace aplicando las técnicas de los índices KWOC (*Key-word out of context*) se crea un léxico documental (artificial-normalizado), en otras palabras se elabora un Thesaurus positivo.

Por falta de espacio no podemos adentrarnos en la especificación de ambos métodos ni en la explicación de las razones que hacen pertinente la opción por el método de las palabras-clave para la recuperación de información en una biblioteca. Respecto a esto último, sólo indicar que esta opción viene impuesta por la naturaleza del documento a memorizar. En efecto, en una biblioteca se trata fundamentalmente de recuperar referencias bibliográficas (o, en el mejor de los casos, *abstracts*) y como es obvio este tipo de documento, que requiere elaboración y manipulación por parte del documentalista, es incompatible con el método del texto íntegro. Por otro lado, la gran ventaja que presenta el método de las palabras-

(1) Véase, p.a., Pagano, Rodolfo. «Nota per un sistema di informatica giuridica». — En: *Informatica e Diritto*. — Mayo-diciembre (1980).

clave frente al otro, en el ámbito de una biblioteca es que permite la comparación conceptual y no solo formal (control de identidad gráfica) entre el lenguaje de interrogación y el lenguaje contenido en los documentos.

Pues bien, visto lo anterior dedicaremos esta parte de la comunicación a explicar en qué consiste un Thesaurus (de ahora en adelante cuando nos refiramos al Thesaurus lo haremos en el sentido del Thesaurus positivo) y a explicar algunos de los problemas con que se enfrenta el documentalista para su elaboración. También incluiremos una extensa referencia explicativa, a modo de ejemplo, del THES-BID.

3.1. EL THESAURUS

Si aceptamos que por Thesaurus cabe entender un vocabulario específico o documental no encontraremos grandes dificultades para comprender que la mayor parte de los problemas que surgen en su construcción son de orden lingüístico⁽²⁾. Por esta razón, nos interesa distinguir aquí entre un lenguaje natural, que abarca todo el patrimonio lexicológico; un lenguaje jurídico, entendido como aquella parte del lenguaje natural que comprende el léxico utilizado en los documentos jurídicos (leyes, sentencias y doctrina); y un lenguaje documental que consiste en un léxico artificial fundamentado sobre el lenguaje natural y que está formado por palabras-clave, es decir, por palabras cuyo significado está normalizado.

Construir un thesaurus jurídico consiste, pues, en elaborar un lenguaje documental en el campo del Derecho, esto es, en unificar el léxico del lenguaje jurídico atribuyendo a cada palabra del lenguaje documental un significado normalizado. Lo que quiere decir que cada palabra del lenguaje documental debe tener siempre el mismo significado con independencias de los contextos. Así, cuando el thesaurus utiliza dos palabras diferentes se puede afirmar con toda rotundidad que pretende recuperar conceptos diversos.

Siguiendo a Viktor Knapp⁽³⁾ podemos realizar en torno al léxico del lenguaje natural la siguiente diferenciación: a) de una parte, un grupo de palabras que llamaremos «grupo fundamental», que está formado por conjuntos de términos que expresan conceptos que no requieren para su utilización y comprensión ninguna formación profesional o cualificación especial; b) de otra parte, grupos específicos que incluyen términos relativos a ciertos campos del pensamiento y de la actividad humana. No cabe duda de que dentro de estos grupos específicos podemos incluir al Derecho; ahora bien, sería un error identificar el grupo específico de vocablos que expresan conceptos jurídicos con el lenguaje jurídico, ya que éste, comprende: 1) términos que expresan exclusivamente conceptos jurídicos; 2) términos que designan un concepto jurídico, pero que simultáneamente tienen un significado común (exotérico); y 3) términos que representan ti-

(2) Knapp, Viktor. «Alcuni problemi relativi alla costruzione di un thesaurus giuridico». — *En: Informatica e Diritto*. — Enero-marzo (1979); p. 175.

(3) *Ibidem*, p. 176 y ss.

pos neutros, puesto que a pesar de que se utilicen en documentos jurídicos no expresan ningún concepto jurídico.

El grueso del lenguaje jurídico pertenece al segundo grupo y en líneas generales, podemos afirmar que caben dos posibilidades en cuanto a la relación entre su significado jurídico y su significado común: la concordancia o la divergencia. La primera supone que estos términos, tanto en contextos jurídicos como en contextos generales (en su sentido común), son aproximadamente unívocos, esto es que los significados son iguales o muy similares (por ejemplo: padre, contrato, matrimonio, etc.). La segunda alternativa, la divergencia, quiere decir que los significados de dichos términos no son iguales y ni siquiera aproximados, lo que plantea grave problemas de cara a la construcción de un thesaurus jurídico (por ejemplo: carga, acción, causa, etc.). El origen de esta divergencia semántica entre estos términos se debe fundamentalmente a la polisemia y a la homonimia.

Por polisemia cabe entender una pluralidad de significados de una misma palabra. Ciertamente es que el lenguaje jurídico, se esfuerza por excluir o limitar el fenómeno de la polisemia, sin embargo, también en ocasiones él mismo lo genera, puesto que utiliza determinados términos como polisemias cuando éstos no lo son en el grupo fundamental (el ejemplo típico es el de «acción»). Se dice, por otro lado, que son homónimas las palabras que teniendo la misma forma gráfica poseen significados diferentes. La diferencia entre la polisemia y la homonimia radica en que, en el primer caso se trata de una palabra con varios significados y en otro de varias palabras con la misma forma gráfica. Desde la perspectiva de la informática jurídica ambos fenómenos reciben igual tratamiento.

Otro de los fenómenos lingüísticos al que debemos prestar especial atención es el de la sinonimia. En sentido estricto, la sinonimia hace referencia a aquellos términos que tienen significados idénticos, lo que los hace intercambiables. Ahora bien, de cara a la construcción de un thesaurus jurídico resulta más conveniente utilizar un concepto de sinonimia menos rígido y así incluir, además de los términos entre los que se da una relación de identidad absoluta, aquellos entre los que se establece una relación de vecindad o proximidad.

Si se acepta todo lo anterior, se verá como el problema fundamental en la construcción de un thesaurus radica en la transformación del lenguaje jurídico en un lenguaje documental (artificial-normalizado). En cualquier caso, esta transformación supondrá siempre una reducción de términos, es decir, las dimensiones del lenguaje documental. A continuación nos referiremos al tratamiento de los problemas derivados de la polisemia, la homonimia y la sinonimia.

El lenguaje documental pretende eliminarlas, pero sólo consigue reducir su número. La técnica más habitual consiste en unir a la palabra-clave o descriptor una breve definición que permita identificar el sentido que debe darse a este término. Así, por ejemplo, podrían adjuntarse al término ACCION, las siguientes definiciones:

- ACCION (derecho que se tiene a una cosa).
- ACCION (facultad de acudir a los tribunales y recabar una decisión).

- ACCION (acto humano voluntario constitutivo de delito).
- ACCION (porción en que se divide el capital de una sociedad) (4).

Del tratamiento de la sinonimia surge el concepto de equivalente. El equivalente se apoya en la idea de que el thesaurus jurídico en cuanto léxico artificial carece de descriptores sinónimos y que los sinónimos de un descriptor se agrupan bajo éste como equivalentes, pero sin ser descriptores. La idea de fondo, no es otra que la de que cada descriptor del thesaurus recupere una información diferente. De esta manera, el equivalente puede definirse como aquella palabra-clave, no-descriptor, que se considera sinónimo de un descriptor determinado. Ahora bien, no debemos olvidar que los equivalentes son sinónimos artificiales de los descriptores, por lo que el significado de un determinado descriptor en cuanto lenguaje de máquina vendrá delimitado por el significado del descriptor en el lenguaje jurídico y por el significado de cada uno de los términos que se establezcan como equivalentes. De este modo, al ser artificial la relación de sinonimia entre los descriptores y sus equivalentes, ésta puede calificarse de verdadera o falsa según se dé o no sea relación de sinonimia en el lenguaje jurídico. Cabe, entonces, preguntarse qué objeto puede tener la creación de falsos sinónimos. Para responder a esta cuestión es conveniente que recurramos a los ejemplos (5):

a) Puede ser útil establecer una relación de falsa sinonimia para aquellos términos del lenguaje jurídico que raramente se encuentran en los documentos jurídicos o que son de menor importancia y que desde el punto de vista de la informática jurídica no es conveniente que adquieran la cualidad de descriptores, pero que, sin embargo, no pueden ser ignorados al thesaurus no como descriptores, sino como equivalentes bajo un descriptor cuyo significado sea similar o aproximado (por ejemplo caducidad y prescripción).

b) Existen, también, en el lenguaje jurídico parejas de términos que podríamos calificar de antinónimos o complementarios, y que no necesariamente tienen por qué incluirse ambos descriptores en el thesaurus, sino que por el contrario puede resultar más oportuno no incluirlos. En estos casos, es bastante frecuente considerar a uno de los términos como descriptor y al otro como equivalente. Este tipo de sinonimia no sólo es falsa, sino que además nos resulta intuitivamente paradójica (ejemplos: derecho subjetivo-deber jurídico, cosa hallada-cosa pérdida, etc.).

Por último, de todo lo anterior podemos concluir que dos notas caracterizan al Thesaurus como lenguaje documental: a) están copuestos por palabras-clave, que representan la unidad lexicológica del lenguaje documental; y b) expresan ciertas relaciones semánticas entre las palabras-clave, como por ejemplo relaciones de jerarquía (de particular a general o viceversa), de equivalencia, de clasificación, etc. Pero esto último se verá con más claridad y mayor detenimiento en el siguiente apartado.

(4) Las definiciones incluidas en el ejemplo, las hemos extraído de la voz «Acción». En: Gómez de Liaño, F. *Diccionario Jurídico*. — Salamanca : Cervantes, 1979.

(5) Los ejemplos los extraemos de Knapp, Viktor *Alcuni problemi...*, ob. cit., p. 188 y ss.

3.2. EL THES-BID: UN EJEMPLO DE THESAURUS

En este punto, y después de haber tratado el problema de la universalidad del thesaurus, no nos resistimos a la tentación de referirnos, aunque sea brevemente, al magnífico trabajo realizado por tres investigadores del «Istituto per la documentazione giuridica» del CNR (Florencia): nos referimos al THES-BID elaborado por C. Ciampi, E. Fameli y G. Trivisono (6). THES-BID: A Computer-based Thesaurus of Terminology in Computers and the Law» es un thesaurus en lengua inglesa para utilizar en el ámbito de la «Bibliografía internazionale d'informatica e diritto» (BID), que periódicamente publica el propio «Istituto per la documentazione giuridica». Este thesaurus constituye el primer intento orgánico a nivel internacional de normalización del lenguaje o, si se prefiere, de sistematización general de los conceptos que aparecen tanto en el campo de la informática jurídica como en el del derecho de la informática. Sirve por tanto, para recuperar el material bibliográfico relativo a la informática, a sus aplicaciones en el mundo del Derecho y a los problemas jurídicos que ella plantea. A continuación, nos parece oportuno explicar en qué consiste y cómo se organiza el THES-BID, ello se debe a que, a estas alturas del trabajo, consideramos conveniente incluir la exposición de un thesaurus concreto con el fin de que el lector alcance una perfecta comprensión de lo que supone este instrumento de trabajo. Para ello, pasamos a explicar cada una de las listas de que está dotado el THES-BID (7).

1) *La lista alfabética estructurada («The Structured Alfabetic List»)*

La lista alfabética estructurada es el índice del thesaurus, que incluye todos los términos (palabras-clave y expresiones compuestas) ordenados alfabéticamente. En este índice se indican las relaciones existentes entre los diferentes términos. Las relaciones consideradas en este índice son las siguiente:

— US = USE; A US B: el término A no está incluido en el thesaurus y en su lugar se debe utilizar el término B. Esta relación es asimétrica, irreflexiva e intransitiva.

— UF = USED FOR; B UF A: el término B ha sido utilizado en sustitución del término A. La relación se opera automáticamente desde la relación US y así (\times US y = y UF x).

— TT = TOP TERM; A TT: el término A se define como top term cuando constituye el vértice de una pirámide conceptual o, en otras palabras, la raíz de una estructura en forma de árbol. Por tanto sólo los descripto-

(6) Ciampi, C.; Fameli, E. y Trivisono, G. «THES-BID : A Computer-based Thesaurus of Terminology in Computer and the Law». En: *Informatica e Diritto*. — Enero-abril (1982).

(7) La información relativa a las listas incluidas en este Thesaurus la hemos extraído fundamentalmente de la presentación que los autores hacen de la obra en *Ibidem* y de Ciampi, C. «THES-BID : un Thesaurus per informatica giuridica e il diritto dell'informatica, redatto con l'ausilio dell'elaboratore elettronico». — En: *Informatica e Diritto*. — Mayo-agosto (1983); p. 155 y ss.

res situados en dicho vértice o raíz pueden definirse mediante la relación TT y ninguno de ellos puede quedar falto de tal definición. Esta relación es monádica.

HT = HEADING TERM; B HT: el término se definirá como Nodal Heading cuando represente un nudo conceptual importante al que se une un grupo de términos no inferior a cinco.

Esta relación también es monádica.

SC = SCOPE NOTE; A SC n: al término A le sigue una nota que aclara su significado y su uso, nota que se recupera mediante la utilización del número n situado a su derecha.

BT = BROADER TERM; A BT B: el término A tiene como superior jerárquico al término B. La relación BT es irreflexiva, asimétrica e intransitiva. Es la relación opuesta a NT.

NT = NARROWER TERM; B NT A: el término B tiene como inferior jerárquico al término A (\times BT y $=$ y NTx).

RT = RELATED TERM; A RT B: el término B está asociado al término A porque entre ellos se establece una relación de sinonimia, de casi-sinonimia o de asociación genérica. La relación es simétrica.

2) *KWOC índice de términos permitidos (KWOC Index of Permitted Terms)*

El índice KWOC guía al usuario desde los términos que componen las expresiones compuestas recogidas en el thesaurus a dichas expresiones de las que forman parte. Es un índice alfabético construido de acuerdo con las técnicas de los índices KWOC: en la columna de la izquierda se ordenan las palabras singulares y en la columna de la derecha las expresiones compuestas del thesaurus en que dichas palabras aparecen.

3) *El índice de jerarquía («Index to the Hierarchy»)*

Los descriptores del thesaurus, palabras-clave y expresiones compuestas, vienen ordenados alfabéticamente al tiempo que a continuación de cada uno de ellos se expresan los vértices (Top Terms) de las pirámides conceptuales a las que pertenecen.

4) *La lista de vértices («List of Top Terms -TT-»)*

Contiene los descriptores considerados como vértices de las 52 pirámides conceptuales construidas para estructurar el thesaurus.

5) *La lista de términos de encabezamientos («List of Heading Terms -HT-»)*

Es una lista compuesta por 156 descriptores que, dentro de la estructura piramidal en que se organizan los conceptos, recoge aquellos nudos importantes por el número de relaciones que dependen de ellos (nunca menos de cinco). La lista consta de dos partes, la primera incluye los descriptores en orden alfabético y la segunda, los divide sobre la base del número de relaciones conectadas con ellos.

6. *Lista jerárquica («Hierarchy List»)*

La lista jerárquica contiene la descripción general de la estructura jerárquica del thesaurus, es decir contiene todo el léxico agrupado en torno a 52 árboles conceptuales (o pirámides) en cuyas raíces (o vértices) hay un número igual de descriptores, que se consideran de una importancia semántica general. El nivel jerárquico se presenta gráficamente mediante un determinado número de puntos-indicadores que preceden a la palabra o a la expresión compuesta. Así pues, para buscar el superior o inferior jerárquico de un determinado concepto, debemos seguir la lista hacia arriba o hacia abajo (según el caso) hasta encontrar un descriptor precedido por un determinado número de puntos-indicadores, uno más o menos que el concepto del que partíamos. Los conceptos con igual número de punto se consideran del mismo nivel jerárquico.

7) *El esquema de clasificación («Classification Table»)*

Es un esquema de clasificación decimal que representa una construcción sistemática por materias o disciplinas relevantes en el campo de la informática jurídica y del derecho de la informática. Se establecen nueve clases generales (numeradas del 0 al 9) que a su vez subdividen en subclases. Lo que se propone este esquema es una colocación sistemática de los descriptores dentro de un cuadro general de la materia objeto del thesaurus.

8) *La lista por clases («Class List»)*

En esta lista se encuentran desarrollados los encabezamientos que aparecen en el esquema de clasificación.

9) *Lista geográfica («Geographic List»)*

Es una lista en orden alfabético de nombres de países políticamente relevantes que se usa para examinar la documentación identificada geográficamente en áreas donde se investiga o se aplican los ordenadores.

10) *Lista de siglas («Acronim List»)*

Recoge las abreviaturas que aparecen en el thesaurus y que son frecuentes en el campo de la informática jurídica y en el derecho de la informática.

Hemos intentado dar la filosofía de un thesaurus, tema prioritario a la hora de querer informatizar los fondos bibliográficos. Todos tenemos elaborado *catálogos de materias* más o menos extensos, pero esto no es suficiente.

El contenido debe estar relacionado de modo que siempre haya una salida que remita a un nuevo concepto, más específico o más amplio, según la necesidad del investigador.

En estos momentos conocer los buenos thesaurus existentes dentro y fuera de nuestro país es la tarea de muchos. Cualquier información que nos faciliteis en este sentido nos sería de gran utilidad.

