SELECCIÓN DE LECTURAS DE INDIZACIÓN Y RESUMEN

Compiladora: MSc. Ania R. Hernández Quintana

Facultad de Comunicación

Departamento de Bibliotecología y Ciencia de la Información

2004

PREFACIO

La indización y el resumen forman parte de un subproceso fundamental dentro del ciclo de vida de la información en las instituciones de información, aquel que identifica, describe y dispone los contenidos esenciales de los recursos de información a través de un lenguaje que sirva de base para la búsqueda y la recuperación de la información, es decir, para la comunicación entre los sistemas de información y las demandas informacionales.

Las posibilidades tecnológicas que permiten la búsqueda en texto libre, han hecho pensar que el proceso de indización tendería a desaparecer; sin embargo, los servicios de información están demostrando todo lo contrario. El proceso de infoxicación digital obliga cada día a que los buscadores incorporen técnicas del análisis documental en sus plataformas de intercambio y que sus interfaces simulen con mayor transparencia las fórmulas mentales, cognitivas y operacionales con que los usuarios acceden a la información. Los estudios sobre la web semántica, entre otras cuestiones, confirman que las tradicionales técnicas de indización son el principio para una búsqueda lógica y cooperada en la red.

Habitualmente nuestra especialidad se ha encargado de preparar profesionales capaces de disponer los contenidos denotados, explícitos, en las fuentes documentales y de preparar productos con una fuerte influencia clasificatoria. La indización se equipara con la verticalidad de las clasificaciones, pero se ha superado a sí misma al intentar ocuparse también de las connotaciones, cuando adopta configuraciones reticulares, incorpora las ventajas espaciales y de la graficación, se empeña en vincular cada vez más términos y más realidades conceptuales y cuando, de hecho, se horizontaliza.

Desde este punto de vista, la indización ya no trata solamente del "contenido" en un sentido teórico, al decir de Lancaster, y de productos analítico-sintéticos idénticos para todos, sino los rasgos o las características de los documentos que lo harían interesante para grupos particulares de usuarios, lo que redimensiona su pragmatismo. Si hasta hace poco tiempo se consideraba que la indización abarcaba un conjunto de operaciones para etiquetar puntos de acceso relativamente fáciles e inequívocos, la orientación al usuario y a demandas específicas suponen entrenamiento y estudio de áreas

complejas y diversas. Las herramientas de la lingüística documental, por ejemplo, se han sofisticado en virtud de esas tendencias, y son el eje de una indización eficiente.

Por otro lado, la indización tradicional y los distintos resúmenes han estado vinculados sobre todo al texto escrito, bidimensional y particularmente "científico". Ahora, el reto de la indización y el resumen es el hipertexto, como portador múltiple y dinámico de información de cualquier naturaleza, en un espacio abierto, intangible y manipulable que permite otras interacciones entre el/los emisor/es y el/los receptor/es. En este espacio de convivencia del texto con imágenes, sonidos y movimientos como un todo, en que los tiempos de creación y de difusión se equiparan y en el que el lapso de obsolescencia de lo publicado es incontrolado, las técnicas propias de la indización, y también los resúmenes, se adecuan a la virtualidad y perfeccionan su modo de operación con una multiplicación de las probables necesidades y asociaciones, cada vez más cercanas a las formas de actuación de la mente y de la gestión del conocimiento.

Esta Selección de Lecturas es apenas el primer intento por comenzar a publicar para nuestra especialidad en Cuba, trabajos que reflejen esta cambiante realidad. Por supuesto, el estudiante encontrará trabajos básicos, como el de la Dra. Rosa Giráldez, absolutamente agotados en librerías y almacenes, pero con el gran mérito de sistematizar los principios, entiéndase, los elementos esenciales que sobre la indización deben ser de total dominio del futuro profesional. De hecho, durante muchos años esta ha sido la obra fundamental que ha contribuido en la preparación de nuestros egresados. Sirva esta Selección de Lecturas, al presentar al menos un extracto de aquella, un homenaje a la labor de esta profesora.

Pero también están aquí artículos muy actuales, que son resultado de las investigaciones de profesionales de varias universidades extranjeras que serán referentes de algunas de las ideas antes expuestas. Uno se debe resaltar por su resonancia en el ámbito teórico e inédito hasta ahora en Cuba. La profundidad de las ideas de Birger Hjörland, su amplio aparato conceptual y el ser uno de los más reconocidos teóricos de la Ciencia de la Información, convierte su lectura en un ejercicio de profundidad y, a la vez, en un ejemplo del componente multidisciplinar que envuelve a la indización.

Otros autores son ejemplos de la estrechísima relación entre información, documentación y comunicación; por ello los artículos de Antonio Luis García Gutiérrez nos parecieron imprescindibles. Sobre la temática del resumen la pluma más certera en lengua española quizás sea la de María Pinto Molina. En definitiva, cada uno de los compilados está en la cima de las reflexiones sobre estos tópicos, pero quedan muchos, por razones obvias, que no aparecen en esta Selección, y queda, por nuestra parte, ofrecérselas.

Una sugerencia a los estudiantes: Relean, desde la altura del año que cursan y de los conocimientos adquiridos, los artículos que la Dra. Dolores Vizcaya agrupó en la Selección de Lecturas de Organización de la Información sobre análisis, lenguaje documental e indización. Allí aparecen autores y reflexiones que complementarán la visión que necesitan para enfrentar con éxito esta etapa de sus estudios.

MSc. Ania R. Hernández Quintana
Marzo, 2004

ÍNDICE

SOBRE EL ANÁLISIS Y REPRESENTACIÓN DE DOCUMENTOS

Ramiro Lafuente López
Universidad Nacional Autónoma (México)

DEL ANÁLISIS DE DOCUMENTOS

El término análisis documental alude al conjunto de conocimientos relativos a los principios, métodos y técnicas que permiten examinar, distinguir y separar cada una de las partes de un documento, para determinar la categoría a que pertenece, su estructura formal, propiedades y significado de sus contenidos temáticos.

Se trata, pues, de un método de conocimiento que facilita el estudio de los documentos ya sea en grupo o aisladamente.

Los resultados del análisis de grupos de documentos se expresan en forma de *categorías de documentos*, es decir, en conceptos abstractos que definen sus propiedades comunes y sus relaciones más generales. Estas categorías son el resultado de una abstracción que generaliza los aspectos particulares o singulares de los documentos que produce y utiliza una comunidad.

Las categorías son un elemento para el estudio y clasificación, y pueden referirse a los modos de organización de la producción y uso de documentos que generalmente se expresan en forma de tipologías. Asimismo pueden crearse categorías que eluciden las propiedades comunes y las relaciones más generales de la estructura formal de los documentos. Normalmente esto se logra a través de sistemas de l término análisis documental alude al conjunto de conocimientos relativos a los principios, métodos y técnicas que permiten examinar, distinguir y separar reglas para describirlos, de manuales de estilo para producirlos y publicarlos, o de sistemas de normas para regular su apariencia.

Las formas para la creación de documentos ha sido motivo de estudio de diversas disciplinas y ha generado una amplia literatura relativa a los conocimientos necesarios para construir y validar un determinado tipo de documentos.

En la sociedad contemporánea la difusión, publicación y uso de documentos es de tal magnitud que ha dado lugar a la creación de documentos denominados *secundarios*, los cuales describen sistemáticamente a otros documentos denominados, obviamente, *primarios*.

La descripción de los documentos se hace mediante la expresión abreviada de sus características formales y del significado de sus contenidos para facilitar la formación de acervos y el acceso a los mismos. Las descripciones de forma y contenido de un documento intentan representarlo en forma unívoca y singular, tanto para distinguirlo individualmente como para expresar sus relaciones generales dentro de la producción documental, y con ello crear medios para su localización y adquisición. De hecho se da por sentado que ante la ingente producción de diversos tipos de documentos, es prácticamente imposible que una sola persona pueda tener un acceso directo y simple a todos ellos. El sujeto difícilmente puede tener noticia de todo lo que circula y se enfrenta al problema de ubicar un documento en particular.(1)

La creación y construcción de *documentos secundarios,* como índices, catálogos y bibliografías, se realiza básicamente con fines de orientación científica e informativa para representar, sintéticamente, tanto el continente como el contenido de los documentos originales.

Se presupone que los documentos secundarios, al contener información concentrada y sistematizada, ofrecen cierta facilidad para obtener referencias de los materiales originales o primarios. Los documentos secundarios, pues, pueden contener la producción general o parcial de documentos, como es el caso de los catálogos editoriales y bibliográficos, o bien, pueden referir la existencia de colecciones documentales administradas por instituciones, como es el caso de los catálogos de bibliotecas.

El diseño, construcción y operación de bibliotecas está sustentado en la selección, formación y administración de colecciones de documentos, pero también en el diseño y construcción de documentos secundarios, como índices, catálogos y bibliografías, con el objetivo de facilitar la ubicación y localización de materiales dentro de los acervos de las instituciones –dedicadas a formarlos para utilizarlos en forma colectiva– o en el mercado editorial.

Los conocimientos sobre estos procesos están referidos básicamente a los impresos, no obstante, con la automatización y digitalización de publicaciones ha cambiado la estructura formal de los documentos secundarios, toda vez que se tiende de manera creciente a construirlos como documentos en línea para su uso vía intranets o Internet.

La difusión de la producción editorial se sustenta en la idea de autores propietarios de impresos, que si bien son independientes entre sí, constituyen un monopolio; sin embargo no están aislados unos de otros, puesto que el significado de los contenidos temáticos de los impresos que generan mantienen un vínculo en razón de una serie de creencias –sustento del desarrollo de la actividad científica, académica, de divulgación, difusión o entretenimiento– que facilitan el que las publicaciones producidas adquieran sentido y significado para distintas comunidades. Así por ejemplo, la actividad científica considera que la publicación debe servir al propósito de recrear y acumular el conocimiento, de manera que un autor retoma partes de otro autor y los incorpora a su propia publicación y así sucesivamente.

La actividad académica establece que el aprendizaje debe apoyarse en el conocimiento de los significados de esas publicaciones y crea bibliotecas como un instrumento para construir acervos clasificados de uso colectivo que faciliten su acceso.

El análisis documental se desarrolla con la finalidad de crear métodos y técnicas para analizar documentos, clasificar acervos y con ello facilitar su uso colectivo. Ello se fundamenta en la idea de que los significados de los contenidos temáticos de alguna manera se ligan con los significados de otros, lo cual permite crear esquemas de clasificación capaces de abarcar los significados comunes a varios tipos de documentos. Esta idea de representación de documentos, independientes unos de otros, posibilita particularizar su contenido temático y establecer los elementos necesarios para vincularlos con los contenidos de otros documentos.

La creciente presencia de la automatización obligó a revisar las formas y estructuras de los documentos secundarios convencionales, de los que incluso se generaron versiones en línea. Sin embargo, la innovación tecnológica que representa la automatización hace inevitable reconsiderar los conceptos acerca del análisis y representación de documentos, sobre todo porque hasta ahora únicamente se ha estudiado la estructura de los impresos y no existen acuerdos o convenciones generalizadas acerca de los documentos digitales, aunque es previsible que la tendencia en la estructuración de éstos, sustentada en el uso de

lenguajes de marcado como el SGML y sus derivados, sea la tendencia que prevalezca.
Los sistemas para almacenar información cuentan con elementos que permiten:

1) almacenar los documentos que se adquieren;
2) organizar y controlar los documentos que se adquieren en función de las diferentes demandas de los usuarios (catalogación, indización, clasificación) de modo que puedan ser identificados fácilmente;
3) describir físicamente el documento para analizar su contenido conceptual y traducirlo a un vocabulario determinado –que constituya una representación del mismo– para finalmente almacenarlo en una base de datos que sirva como instrumento de salida del sistema. (2)

ÁMBITO DEL ANÁLISIS DOCUMENTAL

El ámbito del análisis documental comprende el estudio de los principios, conceptos, técnicas y métodos que permiten formular enunciados cuya función es expresar una idea acerca de un documento por medio de palabras, signos y códigos convencionales, con la intención de que éstos constituyan una representación que haga las veces del documento a fin de poder identificarlo, clasificarlo y localizarlo.
La representación de un documento a través de símbolos está condicionada por el sujeto que la realiza, es decir por los objetivos que persigue, así como por las formas generales de la representación que pertenecen al sujeto y no a los elementos del documento. (3)

REPRESENTAR DOCUMENTOS

Al representar un documento se construye una imagen o un símbolo del mismo y convencionalmente se acepta que éste permite saber acerca del material original. Lo cual no implica que los enunciados que se edifican lleven a conocer el documento, simplemente son formulaciones que facilitan el saber acerca del mismo.

Conocer un documento implica un nivel de mayor profundidad, a saber: leerlo, comprenderlo, ubicar sus contenidos temáticos como parte de otros contenidos temáticos; supone, además, integrar todos los elementos acerca de un documento en una sola unidad. La condición para conocer un documento es captarlo en sus diferentes matices, bajo diferentes perspectivas y, eventualmente, en situaciones distintas. Supone también tener experiencias variadas sobre el mismo y poder hacer, de algún modo, una serie de inferencias a partir de ellas, referidas al documento en cuestión.
Cuando nuestro conocimiento es circunstancial y hablamos de “conocer” conforme a su significado semántico común, sólo podemos referirnos a aspectos superficiales y aun ocasionales del documento. Sin embargo, conocer un documento implica poder contestar múltiples cuestiones de diversa índole. Pensemos en el siguiente enunciado: “Conoce las obras de Villoro”. En esta frase se supone que quien conoce puede ser una fuente de información variada sobre los libros, artículos, ponencias, etc., escritos por Villoro, y por tanto ser capaz de

resolver problemas que le sean presentados respecto y orientar a otros. Porque conocer la obra de Villoro, en este sentido, no es sólo saber su descripción externa, sino captar su forma y manera, su estilo, el modo como sus partes están relacionadas en un todo, y ser capaz de relacionar entre sí las partes de su doctrina. Conocer un documento en su sentido más rico es poder integrar en una unidad cualquier experiencia y cualquier saber parcial sobre un documento, por variados que estos sean.

El que sabe muchas cosas acerca de un libro como *El Quijote* tiene con él una relación cognoscitiva diferente a quien realmente lo conoce. Quien sabe del *Quijote* podrá citar frases del texto, dar noticias de sus diferentes ediciones y características de las mismas, incluso distinguir tipografía, papel y cambios en el texto, organización de los capítulos e ilustraciones que aparecen en distintas ediciones, y saber de las opiniones de distintas personas sobre el mismo. No obstante, quien conoce *El Quijote* tal vez no sepa nada de lo anterior, en cambio puede comprender su mensaje central, captar su espíritu, interpretar diversos asuntos a partir del mismo, saber responder preguntas acerca del alcance y aplicación de lo expuesto. Conocer *El Quijote* no es saber muchas cosas acerca de él, sino poder distinguir lo esencial de sus contenidos temáticos y literarios, "núcleo" del que puede desprenderse cualquier formulación parcial.

Conocer un documento implica integrar lo que se sabe del mismo a modo de comprender lo central de sus contenidos y captar su articulación interna. Es distinto "**saber** sobre el libro *El Juego de la lógica* escrito por Lewis Carroll" a "**conocer** el libro *El Juego de la lógica* escrito por Lewis Carroll". Lo primero es poder describir lo que se sabe, o bien, exponerlo parte por parte; lo segundo es haberlo comprendido en su estructura y poder, en consecuencia, distinguir en él lo relevante y exponerlo como un conjunto coherente.

Conocer no es una suma de saberes, sino una fuente de ellos, implica tener un modo de relacionar cualquier saber específico con otros saberes. El saber, en cambio, es necesariamente parcial, mientras que el conocer aspira a captar una totalidad.

Distinguir entre saber y conocer (4) permite establecer diferentes niveles o modos de representación de un documento, e implica distintas modalidades de aproximación al mismo. El saber proporciona una interpretación descriptiva del documento; el conocer, en cambio, conduce a su comprensión.

La representación de un documento se expresa de diversas formas, por ejemplo una ficha catalográfica, un resumen, una reseña, una ficha bibliográfica, una referencia, palabras clave, entre otras. Los enunciados y formulaciones que se realizan para nombrar un documento como un todo, constituyen una representación del mismo.

Una colección, agrupamiento, agregado o lista de representaciones de documentos conforman un universo cuya definición depende del contexto en que se utiliza o del problema que se trata de resolver. Existen diversos tipos de ellos, como pueden ser: catálogos, índices, bibliografías, bases de datos bibliográficos, entre otros.

Para facilitar la representación de un documento se definen conceptos y procedimientos en forma de **sistemas de reglas** que establecen conceptos, principios generales, procesos y técnicas para representar a un documento con la intención de incluirlo como parte de un universo siempre en proceso de

construcción, y que podrá estar constituido por un número indeterminado de representaciones diversas.
Las reglas se pueden organizar:
- **conforme al tipo de representación**, en cuyo caso las reglas se orientan a la construcción de algún tipo de representación, como puede ser la simple descripción física de un documento, o bien, tratarse de formas más complejas como es el caso de los resúmenes o las reseñas.
- **conforme a las características de una clase de documentos**, en cuyo caso se agrupan y jerarquizan los elementos que se consideran peculiares para representarlos dentro de una clase específica.
La representación de un documento no debe contemplarse como un hecho aislado, al contrario siempre debe concebirse como parte de otras representaciones, con objeto de facilitar su agrupación de distintas maneras, para posibilitar la creación de diversas clases de colecciones de representaciones, como puede ser un catálogo de biblioteca, una bibliografía, una base de datos, entre otras. Cada una de las cuales cumple diversas finalidades y define las características de la colección en cuestión.
La representación de los documentos y sus contenidos debe estar orientada a patentizar frente al público tanto las características físicas y/o contenidos temáticos como las relaciones que guardan entre sí los diferentes documentos que forman una colección. La representación de documentos puede expresarse de dos formas:

a) por medio de símbolos que representan los contenidos documentales y sirven para acomodar los documentos en un orden predeterminado, y/o
b) a través de un registro que contiene la descripción de las características físicas y/o temáticas.

Tanto los símbolos creados a partir de un sistema de clasificación (Dewey, LC, CDU, u otros) como los registros realizados en forma de fichas o registros electrónicos utilizados para representar un documento (libro, material hemerográfico, revista, vídeo, etc.), muestran las relaciones entre diferentes fenómenos vinculados a la producción y uso de documentos (pueden referirse a sus características físicas y/o a sus contenidos temáticos).
Al representar algún tipo de documento por medio de un registro, construido específicamente para figurar en lugar de éste, se pretende que tenga las cualidades necesarias para relacionar cualquiera de los elementos incluidos como parte de un registro con los elementos de otro. Estas relaciones tienen la finalidad de crear un orden que explicite los vínculos.
Por ejemplo, cuando decimos “el libro escrito por García Márquez se titula *Cien años de Soledad*”, estamos expresando una relación entre un sujeto: “un libro escrito por García Márquez”, y un objeto en particular “el libro titulado *Cien años de soledad*”.
Además, esta relación tiene un orden específico: es el libro escrito por García Márquez el que tiene el título *Cien años de soledad*, y no es el título el que tiene un libro.
Algunas relaciones que aparecen en los registros no siempre mencionan todos los elementos que entran en juego. Así, cuando decimos “el libro titulado *El Hombre Gramatical* es un libro de *divulgación científica*”, queremos decir que existe una categoría, denominada *divulgación científica*, que se aplica al libro *El Hombre Gramatical*.

Sin embargo, no se señala quién considera que esta obra es de divulgación científica ni por qué. Así pues, detallar todos los fenómenos implicados en una relación depende de qué es lo que se quiere decir, y de la finalidad que se persigue al construir una representación.

Existen diversos modos, métodos y técnicas para construir representaciones, pues la invención humana no sólo modifica la selección y agrupación de los elementos que las constituyen, sino que la conformación misma puede organizarse sobre la base de la homologación de distintas formas, en atención al cumplimiento de propósitos diversos relacionados con la organización de colecciones documentales y con la creación y operación de redes de información.

En este sentido es factible retomar conceptos y principios de la lógica, así como las técnicas establecidas en diversos sistemas de reglas y normas –siempre y cuando se eviten posibles contradicciones– y emplearlos para construir políticas, lineamientos o sistemas normativos específicos. Sería el caso, por ejemplo, de los estilos bibliográficos que utilizan diferentes revistas científicas, donde cada uno de ellos obedece a una lógica y al cumplimiento de finalidades específicas.

Para evitar que un sistema normativo quedase aislado es necesario homologarlo con otros, es decir hacer sus reglas y resultados equiparables a las de otros sistemas semejantes. El conocimiento compartido sobre los modos de representación documental permite cifrar y descifrar mensajes entre distintos tipos de interlocutores.

Estos modos son producto de conocimientos acerca de la naturaleza de distintas clases de documentos que se expresan en forma de sistemas de reglas cuyos enunciados permiten producir e interpretar sus representaciones.

Los principios generales y las reglas a que nos referimos sirven al propósito de generar sistemas de significación que permitan organizar las representaciones en clases o tipos. Sin embargo no es factible la existencia de un solo tipo, forma o modo para representar documentos, toda vez que, en términos generales, cada comunidad tiene un conjunto de enunciados con los cuales, los practicantes de una disciplina, la definen, trazan sus bordes externos e internos y sus rutas interiores por medio del lenguaje y sus significados. Estos enunciados delimitan la función del sujeto (poeta, bibliotecario, ingeniero, médico, artista) y la definición de los principios generales de la actividad de la disciplina, estableciendo lo que le es pertinente, así como los conceptos y estructuras de discursos pertenecientes a la disciplina; en su caso, también determinan la metodología y los principios epistemológicos. Son enunciados en los que reconocemos una actividad disciplinaria y con ellos se genera una tipología útil para la representación documental.

Representar, pues, adquiere sentido únicamente en el contexto de necesidades y conductas de los miembros de una comunidad. El proceso intelectual de interpretación de la forma y/o contenidos temáticos de los documentos –para representarlos por medio de códigos o palabras– únicamente adquiere sentido cuando está orientado al cumplimiento de intencionalidades o finalidades que se expresan a través de algún tipo de producto o servicio (formación de acervos, préstamo, servicio de información, o el clásico catálogo). Las finalidades dotan a la representación de significados relevantes para los miembros de una comunidad, al hacer patente el medio que el público tiene para beneficiarse de la organización de documentos que se logra por medio de su análisis y representación.

Ahora bien, si nos atenemos al hecho de que a los servicios es necesario expresarlos por medio de sistemas de control administrativo, de instrumentos de acceso y de obtención de documentos, nos encontramos entonces frente a un asunto álgido. Si el proceso organizador de los servicios se orienta exclusivamente hacia la búsqueda de la eficiencia y margina la disponibilidad de medios para hacer explícitas las intenciones, significados y finalidades de la representación de documentos, indudablemente estaremos frente a la presencia de una maquinaria administrativa que puede ser impresionantemente eficaz para resolver los problemas estándar de almacenamiento y recuperación de información, pero que poco tendrá para ofrecer al individuo que busca información o conocimientos. Por ello es fundamental que exista coherencia, así como vínculos y relaciones entre los servicios, el análisis y la representación de documentos, situación que se rompió al dividir el trabajo y separar a uno de otro.

El análisis de documentos implica preguntarse acerca de la naturaleza del diálogo que se presenta o debe presentarse con la comunidad para la cual se realiza, asunto que evidentemente está relacionado con aspectos inherentes a la lógica de la construcción de sistemas de clasificación y a la determinación de significados y relevancia que se le otorguen a los contenidos de los documentos. Pero sobre todo, implica establecer las cualidades que deben reunir los servicios por medio de los cuales se pretende responder a las necesidades de un mundo que induce al uso de la información en forma rápida y precisa; y, al mismo tiempo, requiere de elementos y espacios para que el individuo pueda generar e integrar conocimientos que le permitan no sólo explicarse lo que sucede en el mundo, sino que contribuyan a enriquecer su concepción individual del mismo; a fin de cuentas la lectura de compresión en primer término le sirve para estructurar o reestructurar sus propias concepciones.

En las últimas décadas ha prevalecido la idea de sustentar el análisis y la representación de documentos en la concepción del almacenamiento y recuperación de la información, cuya intencionalidad esencial radica en la construcción de técnicas que sean funcionales para hacer eficiente el proceso de búsqueda, por medio de servicios orientados a proveer de información al usuario. La organización de colecciones de documentos así como los instrumentos de acceso se dirigen a la localización de información y datos específicos, relegando a un segundo plano el establecer y mostrar las relaciones que existen entre diferentes documentos. Al sustentar la búsqueda y recuperación de información en las ideas acerca de la relevancia y la pertinencia de ésta, se producen dos fenómenos: la sobreinformación, porque la eficacia de los sistemas de almacenamiento y recuperación produce más información de la que un individuo puede asimilar; o bien la desinformación, dado que al sujeto no le es fácil recuperar información pertinente. Esta situación se presenta porque no es fácil determinar la relevancia y la pertinencia de un documento, ya que éste adquiere esos valores en función de las necesidades específicas de información de un sujeto, y no por sí mismo.

Estos conceptos se introdujeron como parte de las técnicas de búsqueda de información. Fueron derivados de las concepciones de la escuela lingüística de Praga, y se han definido de la siguiente forma:

- **Relevancia**: califica los rasgos significativos y operaciones lógicas (exclusión, suma, pertenencia, etc.) que tienen un valor diferencial en la selección de términos para una búsqueda de información.

- **Pertinencia**: califica los rasgos significativos que tienen un valor diferencial en la selección de los datos referentes a la descripción de un documento, los cuales se obtienen por medio de una búsqueda de información y a la vez determinan el propósito de esta última.

Empero, consideramos que el aspecto fundamental que debe orientar el análisis para obtener representaciones de documentos no debe radicar exclusivamente en la idea de almacenar para recuperar, sino en una constante interrogación del texto que se trata de representar. La cuestión radica en encontrar los métodos y principios que permitan preguntarse acerca de las posibles relaciones entre los contenidos temáticos. Se trata de realizar procesos de inferencia que conduzcan a describir en forma de una secuencia los contenidos conceptuales del documento y sus relaciones con el contenido de otros, a fin de que las representaciones sirvan para que un sujeto pueda formarse un esquema de referencia acerca de los documentos analizados.

SOBRE LOS MODOS PARA REPRESENTAR DOCUMENTOS

El análisis y representación de un documento se apoya en un conjunto de principios, normas, técnicas y tecnologías que establecen lo que debe incluirse o no como parte de la descripción de un documento, o bien, indican los contenidos que debe tener para ser acreditado como perteneciente a un tema o considerarlo como parte de un tipo de documentos, o aun prescriben aquello que es necesario considerar para resumirlo o reseñarlo.

La construcción de categorías sistemáticas para establecer las condiciones o cualidades que deben considerarse para incluir un documento dentro de una clase perteneciente a una tipología, juega un papel destacado en la determinación de elementos que se utilizan para representarlo.

La tarea de la tipología de documentos, en el terreno de la representación de los mismos, no consiste en proponer definiciones ni en establecer clasificaciones ajenas al uso y características que diversas comunidades le asignan, sino en comprender los criterios empleados por ellas. A este respecto, y con referencia a la clasificación típica del libro en manuales, monografías, libros de texto, obras de consulta, etc., lo fundamental sería comprender porque se acepta y valida esta clasificación, además de que nos proporcionaría conocimientos para normar los criterios acerca de la representación.

Los criterios bajo los cuales deben clasificarse los documentos, para agruparlos en clases representativas –de los modos que tienen diferentes comunidades para registrar información y/o conocimientos–, al margen de la naturaleza y calidad de su contenido, tienen su origen en:

- **Conceptos inherentes al documento**, por ejemplo las reflexiones de diferentes autores sobre la idea del valor del libro; las definiciones de la naturaleza del artículo científico (esto ha ocupado a diversos autores); o las opiniones sobre el documento digital.
- **Reglas regulativas para su producción y uso**, como es el caso de los manuales de estilo que establecen enunciados que describen las características formales que debe reunir determinado tipo de documento; los criterios de

validación académica a tomar en cuenta para publicarlo; o las directrices editoriales de una revista.
- **Normas de carácter institucional o disciplinario**, es el caso por ejemplo de los manuales técnicos como el Marc, las reglas de catalogación, el ISBD, o los manuales de políticas de una biblioteca, entre otros. Cuya finalidad es establecer reglas de uso común que faciliten el control de acervos.
La tipología de documentos, como un modo de clasificación, adquiere distintos significados y formas conforme tiende a establecer sistemas de organización útiles para la representación de documentos cuyos contenidos se refieren a diferentes ramas del conocimiento, expresan distintos tipos de saberes, o el uso de la representación obedece a finalidades inherentes a una actividad específica, como puede ser la de identificar los documentos que se citan en un artículo científico.
Aquí, los sistemas de reglas, normas, técnicas y tecnologías abordan universos limitados, por ejemplo la descripción de monografías, o la forma de crear asientos para un catálogo. Aún cuando sistemas como las *Reglas para Catalogar*, o las *AACR2* pareciera que abordan un conjunto de principios muy amplio, su universo es tan limitado como puede ser la construcción de catálogos para bibliotecas. Estas reglas se organizan para solucionar problemas específicos. Fundamentalmente se responde a la pregunta ¿cómo puedo? que se expresa en forma de enunciados muy precisos: ¿cómo puedo registrar el asiento de un autor corporativo? ¿cuáles son los elementos para describir un video?
Para facilitar la enseñanza en esta materia algunos autores han establecido denominaciones para las diversas formas de agrupar conocimientos que sirvan a la representación de algunos tipos de documentos con fines específicos; sería el caso por ejemplo el caso de la *catalogación descriptiva,* que alude a los conocimientos para describir documentos e incluir su representación como fichas de un catálogo.
Como parte del análisis de documentos podemos identificar tres niveles o clases de principios para:
- describir
- clasificar
- evaluar

La aplicación de estos principios produce como resultado distintos tipos de representaciones que pueden emplearse en forma combinada. Así por ejemplo, una reseña necesariamente incluye tanto la descripción del documento, como los elementos para clasificarlo e incluirlo como parte de un tema o corriente de pensamiento, e incluso los juicios acerca del valor del mismo.

REPRESENTACIÓN DOCUMENTAL Y COMUNICACIÓN

El estudio de las finalidades de la representación documental involucra problemas de significados en cuanto al uso que se le pretenda asignar a la producción de esas representaciones.

Un acervo de documentos es un acervo de conocimientos teóricamente disponibles para todos los hombres comunes, expertos o los bien informados, pues es la acumulación de la experiencia práctica, o la ciencia y la tecnología como concepciones fundamentales. Pero este acervo no está integrado, consiste

en una mera yuxtaposición de sistemas de conocimiento que en sí no son coherentes, ni siquiera compatibles unos con otros y que para poder describirlos e integrarlos como una colección, es necesario formular representaciones que interpreten el significado de los diversos contenidos temáticos para poderlos relacionar o yuxtaponer como parte de un esquema de clasificación documental. Podríamos pensar en terrenos de significación en los niveles de comunicación entre el emisor y el receptor de información, lo cual plantearía una visión clara y evitaría caer en un laberinto del que no sería fácil salir porque nos permitiría saber y no precisamente creer lo que se informa.

En el primer terreno, el sujeto que informa y el que se informa tiene una significatividad primaria con una estructura clara y nítida, pues representa el mundo que está a nuestro alcance y que podemos observar de modo inmediato, y también, al menos en parte, dominar, o sea cambiar y reordenar mediante nuestras acciones. Es la zona dentro de la cual nuestros proyectos pueden ser materializados y concretados. El sujeto bien informado tiene claramente estructurado en su pensamiento el conocimiento de lo que puede transmitir.

En el segundo terreno de significación se plantea a nivel de pensamiento la existencia de campos no abiertos a nuestro dominio, pero vinculados de modo mediato al terreno de significatividad primaria, porque brindan las herramientas ya creadas que deben emplearse para alcanzar el fin proyectado, o porque establecen condiciones de las cuales depende la planificación misma o su ejecución. Basta con estar familiarizados con el objeto de conocer las posibilidades, probabilidades y riesgos que pueden contener con referencia a nuestro interés principal. De esta manera, el sujeto bien informado y el hombre que se informa implican las relaciones de significado de forma autónoma e inmediata para conducirlo a una sola significación.

El siguiente terreno de significación podemos decir que por el momento no tiene vinculación con el significado de la información final; es relativamente no significativo, porque podemos seguir presuponiéndolo mientras no tenga lugar dentro de la información que pretendemos saber, ésta puede influir en los terrenos de significación que se manifiestan por medio de nuevas e inesperadas probabilidades o riesgos. Dadas las múltiples relaciones que pueden originarse en el pensamiento del sujeto que informa o del que se informa, estructuramos la jerarquía de interés a través de lo que se quiere saber de la información, eliminando aquello que por el momento no es útil para significar.

Existe otra forma de significación que podemos llamar asociativa, en donde ningún cambio posible puede influir en el significado de la información que queremos saber. Para todos los fines prácticos, basta creer ciegamente en el por qué y en el cómo del significado que pretendemos perseguir.

Las diversas zonas de significatividad se superponen en busca de una precisión y actúan de forma autónoma, de ahí que posibilitan diversos grados de interpretación.

La constante elección y combinación de elementos crea zonas de significación transitoria e inestables, dado que es un proceso de aproximación constante a la significación primaria que se supone conduce a encontrar los conceptos que satisfagan los planteamientos que motivan la búsqueda de la información.

En el proceso de búsqueda se presenta un momento inicial dirigido a elegir los elementos necesarios para crear un grado de significación a partir del cual profundizar las indagaciones que servirán de guía. En este primer momento las

elecciones establecen el problema o fijan los objetivos respecto de los conceptos o contenidos temáticos que se buscan.
En un segundo momento todos los elementos que definen el problema de búsqueda se distribuyen por medio de zonas de significatividad, dando lugar a un proceso que termina en el momento en que se encuentran los conceptos que satisfacen los significados que se plantearon inicialmente.
En la significación de búsqueda, el "interés a mano" –es decir las intenciones primarias que motivan a un sujeto a realizar un análisis– juega un papel determinante, puesto que como no existe aislado sino que es producto de contextos o situaciones sociales, tiene la característica de ser inconstante, de moverse junto con la dinámica del proceso social que lo origina; además tiene un peso diferente para cada sujeto en cada momento, es decir que un individuo puede tener "intereses a mano" dispares, incluso heterogéneos, en cuyo caso el intento por explicar estas características motiva la búsqueda de información.

DE LOS PRINCIPIOS PARA ANALIZAR DOCUMENTOS

Los modos para representar documentos aluden a cada una de las distintas maneras de darle significado a la representación. El modo más general consistiría en describir un documento con objeto de identificarlo y diferenciarlo de los demás, pues no está de más recordar que la representación de un documento siempre se realiza en función de otras representaciones.
Mencionamos ya que el conocimiento compartido sobre los modos para representar documentos, permite cifrar y descifrar mensajes entre distintos tipos de interlocutores. Estos modos son producto del conocimiento acerca de la naturaleza de distintas clases de documentos, y pueden expresarse en forma de principios generales para describirlas. A partir de ello es factible construir sistemas de reglas o normas cuyos enunciados faciliten la producción e interpretación de representaciones documentales, dado que crean una base común de conocimiento utilizable como punto de referencia para normalizar criterios útiles que sirvan a la descripción de un documento en forma consistente.
Como en la descripción es importante la capacidad de evocación de la palabra convertida en código alfabético para representar algún elemento del documento, resulta entonces que la eficiencia y la economía de códigos se convierten en un factor determinante en la selección de los mismos. De esta manera, el apellido de un autor conformado por unas cuantas letras es eficiente y económico porque puede evocar no sólo a una persona, sino incluso una gama completa de temas –dependiendo de la experiencia del receptor–, aún cuando puede presentarse la ambigüedad en el momento en que un apellido hace referencia a varios sujetos que se desempeñan en distintos campos de la actividad humana.
La descripción de un documento no debe pretender únicamente enumerar datos, sino buscar convertir sus características en elementos productores de información al servicio de la interrogación del público que utilice la representación. Por tanto, los elementos a incluir en la descripción no se determinan siguiendo la lógica interna de las posibles relaciones entre los datos que la conforman, sino en razón de elementos externos como pueden ser los fines de recuperación establecidos como valiosos, relevantes o pertinentes para que la descripción de un documento cumpla su objetivo.
Por ello, para el estudio de lo que significa describir un documento, no tiene interés el simple abordaje de las formas de aplicación de reglas contenidas en

algún sistema –para determinar los elementos y el orden que debe tener una cadena de datos que represente un documento–, porque todo se reduciría a la búsqueda de la solución predeterminada por los enunciados de las reglas. Este tipo de abordaje ha conducido a una concepción técnica de la descripción de documentos que acaba transformándose en un fin que se agota en sí mismo. El estudio de la descripción de documentos debe empezar por cuestionar la naturaleza de los sistemas de reglas, de no ser así ¿cómo indagar el por qué de los mismos?

Cualquier investigación al respecto debe empezar por un por qué y un para qué, cuestionamientos necesariamente referidos a la naturaleza de las finalidades y función que deberá desempeñar la descripción de un documento, así como a las características del mismo. Este tipo de descripción tiene una lógica propia relacionada con el uso y finalidades de la representación documental.

Uno de los objetivos primarios de la descripción radica en poder diferenciar claramente a uno de otros documentos, para lo cual es necesario obtener, como producto del análisis, aquellos elementos que les son propios. Una primera cuestión salta a la vista: ¿cómo iniciar el análisis del documento para determinar sus rasgos particulares?

El significado de la descripción formal es consecuencia de los conocimientos que se tienen acerca de los elementos que se considera contribuyen para lograr que ésta sea relevante. Cuando tales conocimientos se transforman en sistemas de reglas o normas, éstas enuncian lo que se estima significativo para representar un documento. Por ejemplo, considerar que el autor y el editor son elementos significativos para la descripción formal daría como consecuencia la creación de varias descripciones. Los nombres de autores tienen en común su calidad de autores, lo cual crea una relación entre todos ellos, pero además existe un vínculo entre éstos y los editores, así como con la naturaleza del mercado a que se dirigen.

Al delimitar los fenómenos que crean relaciones significativas para un documento, se establece una consistencia que hace factible la relación entre los contenidos de distintos registros. Un registro aislado únicamente permite saber acerca de un documento, de manera que son las relaciones entre los contenidos de varios registros las que dan la posibilidad de establecer un orden documental que hace factible conocer acerca de una colección de documentos.

Anteriormente indicamos que el análisis para representar un documento a través de símbolos está condicionado por el sujeto que realiza la representación, es decir por las finalidades que persigue, así como por las formas generales de la representación que pertenecen al sujeto que la elabora y no a los elementos del documento.

Esto quiere decir que el análisis debe obedecer en una primera instancia a esas formas generales de la representación; por tanto, si para representar documentos es necesario establecer clases para agruparlos, entonces la primera condición del análisis es determinar si existe una clase dentro de la cual sea factible incluir el material que se pretende analizar y en caso de no existir debe empezarse por determinarla.

LA INFERENCIA EN EL ANÁLISIS DOCUMENTAL

Cuando nos enfrentamos a un tipo de documentos sobre el cual no existen elementos que permitan identificarlo y describirlo es necesario establecer

principios generales para su análisis y descripción. A este respecto Ranganathan (5) menciona que el desarrollo del método para la construcción de principios y de sistemas de reglas que sustenten el análisis y descripción de documentos está condicionado por la presencia de un método científico constitutivo de un ciclo infinito que podríamos caracterizar de la siguiente manera:

1. Aprovechar las experiencias particulares en cuanto a la descripción de documentos como una suerte de conocimiento previo que sirva al propósito de iniciar la construcción de principios generales por medio de la inducción y la aplicación de principios de normatividad documental.
2. Reducir los conocimientos generados a principios normativos, por medio de la intuición y la imaginación.
3. Crear cánones derivados de los principios normativos, por medio de la inferencia y la semántica.
4. Confirmar por medio de la aplicación particular que los principios normativos creados son validos.
5. Volver a iniciar el ciclo.
6. Continuar hasta generar nuevos principios.

Asimismo, Ranganathan señala que la aparición de nuevos tipos de materiales de lectura que aparentemente podrían trascender el ámbito de los principios de análisis documental en uso, deberían abordarse a través de una interpretación apropiada de los principios existentes. En caso de que dicha interpretación resulte inadecuada, sería necesario la construcción de nuevos sistemas de reglas, necesariamente a partir de los principios normativos existentes, a fin de lograr una acumulación de conocimientos.

Sería importante revisar los principios del análisis documental en el momento en que se presentara la necesidad de generar servicios –distintos a los existentes en cuanto a organización, finalidades y técnicas–, que no fuese posible construir a partir de los principios existentes. Esta situación haría necesario crear nuevos principios para reemplazar a los anteriores. A este respecto, consideramos que en el caso de la creciente circulación y uso de documentos digitales, nos encontramos frente a una situación que obliga a recrear los principios del análisis documental, pues si bien es cierto que los documentos siguen teniendo esencialmente las mismas finalidades, también lo es que cambian sus medios de producción, consulta y tipos de servicios a prestar.

Un documento digital tiene una naturaleza diferente a los impresos, y su sola utilización por medio de tecnologías distintas genera nuevos conceptos como las relaciones hipertextuales.

Los primeros documentos digitales derivaron de la automatización de los impresos para prestar servicios en línea, y hasta ahora la versión digital ha seguido los mismos principios normativos establecidos para su predecesor, y en algunos casos ni siquiera ésos.

TIPOLOGÍA DE DOCUMENTOS Y DESCRIPCIÓN FORMAL

Las diferentes propuestas para agrupar documentos por clase provienen de los estudios realizados por distintos autores, y se encuentran sistematizadas y vinculadas a la descripción de documentos en los sistemas de reglas y normas; asimismo, los manuales de estilo para edición y publicación establecen su

propia tipología que sirve de modelo para la construcción de documentos académicos. Aún cuando existen diversos estudios dedicados a la clasificación documental, es un campo que requiere de profundización, sobre todo respecto a la tipología de documentos aplicada al análisis y representación de los mismos –en áreas temáticas específicas o en cuanto a los nuevos soportes digitales–, pues la transformación que se plantean en cuanto a los criterios de publicación y uso de documentos hacen necesario reflexionar sobre las finalidades del análisis en el contexto de una sociedad informatizada.

La determinación de la clase de un documento implica el conocimiento del significado semántico de las palabras y los términos que se utilizan para aludir a éste; por ejemplo, si el documento en cuestión es un programa de cómputo, el conocer el significado de este término (*programa de cómputo*) ayuda a comprender su naturaleza, ya que para establecer el significado se alude a las características del objeto o concepto, así como a sus relaciones de sinonimia. Esta comprensión previa del documento permite incluso identificar las variaciones que pudieran existir, es decir las probables especies dentro de una clase de documentos.

Los manuales de estilo para editar y publicar constituyen un paradigma acerca de los documentos académicos que resulta de gran utilidad para comprender la naturaleza de diversos de tipos de materiales, dado que establecen definiciones que se consideran como modelos a seguir para la construcción de documentos académicos.

A este respecto tenemos por ejemplo, el caso del *Manual de Estilo* de la *Universidad de Chicago* que tiene un carácter general, sin embargo existen manuales de estilo destinados al uso de comunidades académicas específicas.

La tipología de documentos, tanto la expresada por los autores como la contenida en normas y sistemas de reglas, y manuales de estilo, maneja una terminología que pretende estar libre de ambigüedades y facilitar con ello la comprensión. Pero esta terminología no necesariamente corresponde a las formas en que socialmente se utilizan palabras y términos para denominar a distintos tipos de documentos, incluso los nombres que se emplean no son del dominio del lenguaje común. Se trata de un lenguaje especializado útil para comprender y explicar la naturaleza de los documentos, razón por la cual es necesario distinguir entre la terminología propia y la tipología documental; es decir, diferenciar la terminología que es útil para estudiar, comprender y analizar los documentos, y la que se debe usar como parte de la descripción. Así por ejemplo, López Yepez emplea la denominación *documentos sonoros* para referirse a los que comúnmente están identificados con el audio (audiocasete). En este caso utilizaremos para describir el documento en particular aquel término que tenga una significación unívoca para el público al cual se dirige la descripción.

Las categorías utilizadas por la tipología documental son de utilidad porque permiten, además, organizar el análisis de un documento. Si se considera, por ejemplo, la clásica división en documentos primarios y secundarios, es fácil percatarse de que a partir de ella se desprenden distintas formas de análisis. En otros términos, aún cuando en esencia las formas de análisis serían semejantes,

el uso de cada uno de estos tipos de documentos imprime formas peculiares a la descripción de cada uno de ellos. Mientras que en el caso de los documentos primarios es imprescindible singularizar sus contenidos temáticos para relacionarlos con otros de su misma clase o tema, en el caso de los secundarios resulta deseable describir el alcance de los datos que contiene, así como la confiabilidad de los mismos como instrumento de búsqueda.

Así pues, al analizar un documento para representarlo es necesario, en primera instancia, ubicar algunas categorías y conceptos que no están explícitas en ninguna de sus partes (por lo regular en los libros o programas de cómputo no se dice qué es un libro o un programa de cómputo). No obstante, la naturaleza de un documento se encuentra implícita en él mismo, debido a que su creación, reproducción y uso forma parte de la cultura de una comunidad, cultura que es necesario conocer para poder determinar la clase a la que pertenece. De esta manera, nos encontramos con que las categorías a partir de las cuales se determinan los elementos para representar un documento pertenecen a los modos que tiene una comunidad para crear, reproducir y apropiarse de sus contenidos temáticos.

SISTEMAS DE REGLAS Y NORMAS

El propósito de un sistema de reglas o una norma es crear una base común, utilizable como punto de referencia para la normalización de criterios que sirvan a la descripción de un documento en forma consistente y fundamentada en principios generalizados.

La representación documental se sustenta en sistemas de reglas y normas que funcionan como puntos de referencia para su producción e interpretación con la intención no de uniformar las representaciones, sino de imponerles consistencia, es decir hacerlas coherentes estableciendo relaciones entre los diversos elementos que las componen, a fin de que puedan cumplir con las funciones, actividades o fines para las cuales se realizan.

Gran parte de los conocimientos y principios desarrollados en esta materia se han convertido en reglas normativas que tienden a ser codificaciones de las mejores prácticas conocidas. Empero, la influencia de la ciencia, las tecnologías de todo tipo, y el cumplimiento de ideales sociales o modas, han influido en la creación de normas que no necesariamente expresan esas prácticas.

La pretensión de los sistemas de reglas no se reduce a enumerar los requerimientos de la representación, su intencionalidad está dirigida a propiciar la producción de representaciones significativas, productoras de información al servicio de la interrogación del lector. Establecen enunciados lingüísticos que expresan la aplicación de un concepto, o bien proporcionan parámetros para normar la ejecución de un procedimiento. No son ideales a cumplir porque no expresan todas las posibilidades de calidad deseables en la puesta en marcha de una tarea, pues la calidad depende de la competencia y cualidades intelectuales de quien la efectúa.

Los objetivos que se pretende lograr con los sistemas de reglas y normas se centran en la expresión formal de métodos para la representación documental que

es necesario que sea, al menos en principio, socialmente deseable.
Un sistema de reglas expresa por medio de formulaciones lingüísticas el proceso a seguir para cumplir los objetivos que se persiguen con una representación documental. Cada una de estas formulaciones tiene una estructura compuesta de los siguientes elementos:
- **Un antecedente** en el cual se hace mención del objetivo deseado.
- **Un consecuente** en donde se señala algo que tiene que o no tiene que hacerse (hay que, debe de).
- **Un sistema de relaciones** entre las reglas que componen el sistema.
La verificación de las formulaciones contenidas en un sistema de reglas está referida a su carácter empírico funcional, mientras que en el caso de una norma su estructura contempla únicamente un antecedente en el que se expresa un deber ser –para logro de los objetivos deseados– y un consecuente en donde se menciona algo que tiene que o no tiene que hacerse.
Así también las reglas para describir relaciones entre los elementos utilizados para representar un documento. Por ejemplo, la regla que nos dice que “la descripción física de un libro debe contener por lo menos el autor, el título y el año de publicación”, nos dice “algo” acerca de lo que significa un libro, a saber: que lo escribe una persona denominada autor, que tiene un nombre denominado título, y una fecha de publicación. También nos dice cómo podemos averiguar acerca de su existencia: simplemente hay que preguntar por su autor o el título. Es fundamental destacar que las reglas casi siempre están demasiado simplificadas, pero que son aceptables como “definiciones nominales”. Después de todo, no puede esperarse que una definición nos diga todo acerca de algo. Por ejemplo, la mayoría de las personas estaría de acuerdo en que un libro implica mucho más de lo expresado por la regla antes citada. No obstante, cuando enfrentamos problemas de representación documental necesitamos apoyarnos en aquellas reglas que nos ayuden a resolverlos.
Siempre debemos tener en consideración el uso de las reglas, a pesar de su simplificación, porque éstas no sólo ayudan al propósito de resolver los problemas de representación documental; su utilización implica, además, lograr una consistencia tal que nos permita relacionar o comparar los elementos vertidos en la representación de documentos. Por ejemplo, si empleamos la regla que dice que para describir un libro debemos referirnos al autor, título y fecha, todas las descripciones que se sujeten a ella estarán relacionadas entre sí, en razón de que en todos los casos intervendrán los mismos elementos y relaciones. De otra forma, si utilizamos una regla para describir el libro 1 y otra para el libro 2, obtendremos dos tipos de descripción que pueden o no relacionarse dependiendo de los elementos y relaciones que intervengan en cada caso.
Es importante señalar que si las reglas aplicadas no relacionan los fenómenos de las descripciones efectuadas, se perdería el sentido de la representación documental: establecer relaciones entre distintos documentos.
La construcción de representaciones documentales tiene un carácter complejo, ya que para expresar las relaciones entre los diversos elementos de la descripción es indispensable recurrir a distintos sistemas de reglas. Así por ejemplo:
- Para la descripción física de los documentos están las reglas para descripción establecidas por la International Standard Book Description (ISBD’s), que es un sistema de uso generalizado.
- Para determinar los asientos para la estructuración de catálogos existen diversos

sistemas de reglas, algunos producidos por organismos internacionales de normalización como la ISO y otros de carácter regional como las reglas angloamericanas de catalogación.
- Para describir los contenidos podemos optar entre varios sistemas de reglas orientados a la clasificación de los contenidos: encabezamientos de materia, tesauros, descriptores, resúmenes, y otros.
Cada una de las normas o sistemas de reglas especifica y delimita su campo de aplicación, con lo cual se establecen criterios que permiten homologar y combinar las formulaciones de distintas normas o sistemas de reglas. Las reglas para representar documentos tienen una característica en común: su estructura lógica permite combinar diferentes sistemas de ellas con el propósito de solucionar problemas específicos.
Los sistemas de reglas para representar documentos están constituidos por expresiones lingüísticas a las que se les denomina reglas, cada una de las cuales enuncia los fenómenos referidos a determinado tipo de documentos y los representa mediante *términos* que son la expresión lingüística de todo lo que pueda ser objeto de pensamiento o de aquello que pueda darse en cualquier proposición verdadera o falsa.
La representación de la forma y/o contenidos de un documento nos refiere a sistemas de reglas que determinan los fenómenos y las relaciones que se estiman significativos para la descripción. Sin embargo, la representación únicamente nos permite conocer acerca de un documento, pero las relaciones que puedan establecerse entre la forma y/o contenidos de varios materiales depende de la organización catalográfica y de la clasificación.
La clasificación delimita un universo documental que permite, en primera instancia, determinar si un documento en particular puede o no incluirse como parte de ese universo, pero además crea un espacio en donde se establecen relaciones lógicas entre los fenómenos representados en los registros que describen a dicho documento; relaciones que dotan de significado a cada uno de esos fenómenos y significados que sirven de base para la organización física de los documentos en sí y sus representaciones. En otras palabras, podríamos decir que la clasificación genera un núcleo de conocimientos, porque al crear relaciones lógicas entre los contenidos de diversos documentos, se obtienen también las relaciones entre los conceptos vertidos en esos contenidos.
El análisis para describir la forma y/o contenidos temáticos de un documento presenta varias facetas que se realizan simultáneamente, fundamentalmente radica en la habilidad intelectual para reconocer los contextos y significados de un documento, para, en un segundo momento, determinar la aplicabilidad de una regla específica con la intención de obtener una representación simbólica de los contextos y significados del documento en cuestión.
La clasificación, al delimitar un universo documental y crear un espacio donde se establecen relaciones lógicas con la finalidad de crear un orden que haga explícitos los vínculos entre diversos documentos, facilita la organización de éstos y sus contenidos por medio de la formación de colecciones en un lugar específico como puede ser una biblioteca, un centro de información, un archivo, entre otros.

LOS PRINCIPIOS GENERALES DE LA NORMALIZACIÓN

Al estudiar los tópicos relacionados con la normalización en el ámbito de servicios

bibliotecarios, bibliográficos, de información y redes de información, no ha dejado de llamar mi atención el constante esfuerzo de la mayoría de los estudiosos del tema por reducir el fenómeno y aislarlo, en un vano intento por tratar de convencernos de la naturaleza obsesivamente técnica del tema. No obstante, considero que la explicación de los fenómenos involucrados en la realización y uso de normas trasciende los estrechos límites de la apreciación técnica.

La normalización cobra importancia conforme la actividad económica y social se guía más por el cumplimiento de fines y metas que por la referencia a valores; la normalización se convierte así en una necesidad imperiosa destinada a servir de instrumento para establecer un sistema de valoración jerárquica respecto de los fines y metas a cumplir en el ejercicio de una actividad profesional. En la medida en que se abandona el estudio y comprensión de la naturaleza de la representación documental, se hace necesario suplir estos conocimientos con normas y procedimientos técnicos.

El imperativo de la normalización como instrumento para orientar el desarrollo de las actividades profesionales ha conducido a algunos autores a considerarla como objeto de estudio de una disciplina independiente que abarcase los problemas de normalización en todos los ámbitos de la actividad humana, propuestas que han corrido con poca fortuna, puesto que no existen realmente argumentos para sostenerlas. Sin embargo, sí es de llamar la atención el aumento de la actividad normativa que obliga a diversas disciplinas a desarrollar un campo de estudio dedicado a las normas y a la actividad de normalización.

En este campo de estudio se emplean palabras como norma, reglas, sistema de reglas, normalizar, homologar, normativo, normalización, competencia, actividad, procedimiento, técnica, tecnología, uniformidad, consistencia, transferir, protocolos, normar, formatos, estándar, standard, standardizar, standardización, para referirse a objetos, conceptos y métodos inherentes a este campo. Por ello, al utilizar estas palabras desde la perspectiva de la normalización les atribuiremos un significado unívoco con el fin de establecer una relación de significado entre un objeto o un concepto y la palabra en cuestión, independientemente del que puedan tener en términos semánticos o en cualquier otra disciplina o campo de estudio.

¿Qué es normalizar?

En un sentido general normalizar, según los diccionarios de la lengua, significa regularizar o poner en buen orden aquello que no lo está. Si normalizar implica poner en buen orden, entonces, cuando nos referimos a normalizar en el ámbito de estudio de la representación documental, estamos implicando la idea de ordenar lo desordenado. Cabría preguntarse si este significado semántico de normalización coincide con los fenómenos propios de normalizar en el campo de la representación documental. En términos generales podríamos responder afirmativamente, aún cuando habría que hacer algunas precisiones.

Normalizar alude fundamentalmente a la pretensión de lograr productos y servicios equiparables, es decir que sea factible distinguir semejanzas y diferencias con otros productos y servicios. Para que éstos puedan compararse fácilmente se

recurre a productos o servicios tipificados y ordenados por clase en torno a elementos comunes claramente identificados. En este sentido, normalizar implica la acción de organizar conforme a determinada tipología que ha sido aceptada como norma para simplificar y obtener una mayor eficiencia en el rendimiento del producto o en los resultados del servicio.

Este concepto implica varios elementos, tales como la aceptación de normas para simplificar el diseño, la construcción y uso de productos y servicios, así como la presencia de instituciones o círculos sociales encargados de producir y validar normas –conforme a metas y fines que se establecen como comunes y social o comercialmente deseables–.

En otros términos, normalizar incluye un conjunto de acciones orientadas a organizar productos y servicios conforme a metas y fines comunes, establecidos en una norma avalada por una institución o un círculo social, con objeto de lograr una mayor eficiencia en el uso de productos y servicios.

Normalizar tiene que ver con el establecimiento de enunciados cuyo propósito es crear referentes que sirvan como guía para el desarrollo de actividades profesionales específicas, con la finalidad de hacer funcional un producto o un servicio; implica una conciencia de que la actividad profesional que se realiza no sólo incumbe a quien la efectúa, y por tanto debe estar referida a fines y metas de fácil comprensión y acceso a todos los involucrados en la actividad profesional en cuestión.

La belleza de los estándares radica en su volumen, es decir en la posibilidad de contar con una amplia gama para escoger. El poder seleccionar estándares parecería un contrasentido, sin embargo no lo es ya que implica enfrentar el problema de la normalización como un proceso dinámico y dialéctico, dado que se necesita de estabilidad para poder desarrollar procesos tecnológicos. Ello no obstante, la estabilidad tendería a propiciar el estancamiento del conocimiento y los procesos tecnológicos, de ahí que la inestabilidad sea necesaria para el desarrollo. Por ello los procesos de normalización implican procesos inestables, que a lo sumo pueden aspirar a establecer criterios para facilitar la homologación de procesos tecnológicos.

El campo de la normalización comprende el estudio de las acciones necesarias para aplicar normas y para generar políticas y procedimientos que lleven a la obtención de productos o servicios normalizados a través de métodos apropiados.

Evidentemente, también es materia de estudio los métodos y procedimientos para construir y validar normas, así como el abordaje de las normas en sí mismas.

Como la idea de normalización surgió y se desarrolló en el mundo anglosajón, los términos para designarla se tomaron del inglés, se castellanizaron y dieron lugar a las siguientes denominaciones:

-*standard*, que se le atribuye el significado de tipo o modelo, es sinónimo de *norma,*

-*standardizar*, que significa producir conforme a normas, es sinónimo de *normalizar*, y

-*estandardización* cuyo significado es estandardizar, es sinónimo de *normalización*.

Preferentemente utilizaremos los términos norma, normalizar, y normalización, porque su fonética es más bella, aunque existen autores que se inclinan por utilizar standard para diferenciar el campo de estudio y referirlo exclusivamente a

lo tecnológico, dado que el uso del vocablo norma y normalizar tiene otros contenidos semánticos en Filosofía y en Derecho. Situación sobre la que sí vale la pena abundar.

Principios de normalización

Los principios generales de normalización deben considerarse al margen de cualquier finalidad tecnológica y atender sobre todo al cumplimiento de finalidades socialmente deseables en el sentido de lograr productos útiles para el bienestar de la comunidad.

Entendemos por productos útiles aquellos que sean fáciles de comprender, aprender, usar y aplicar a la solución de problemas o a la satisfacción de necesidades.

Para ello es necesario construir sistemas documentales conforme a los siguientes principios:

Consistencia

- Que en los sistemas documentales una misma cosa se haga de la misma manera.

- En el uso de signos y símbolos es necesario atender a los significados culturales.
- No importa la manera en que un sistema documental se encuentre técnicamente organizado, sino que esté construido de tal manera que su contenido y operación sea fácilmente predecible para quien intente utilizarlo.

Retroalimentacion

- Debe contar con elementos de guía y ayuda para el usuario (todo tipo de manuales, ayudas, tutores, preguntas frecuentes, etc.).

Estructura

- La estructura del sistema debe de ser consistente en el sentido de la descripción y asignación temática, aun en los errores.

- Debe tener una estructura lógica fácilmente comprensible en su totalidad.

- Debe hacer explícitas las finalidades que trata de cumplir y las funciones por medio de la cuales es factible hacer operativo su cumplimiento,
- Debe ser eficiente para los propósitos que fue hecho.

A continuación se describen algunas técnicas que ayudarían a configurar un buen sistema normativo para el análisis documental.

Los sistemas deben construirse teniendo en mente la homologación a otros sistemas de manera que ésta se convierta en una fácil y rápida transformación de elementos, incluso es necesario señalar aquéllos que no tienen equivalencia en otros sistemas, también si el sistema, por sus características, no es homologable o sólo lo es parcialmente. En cualquiera de los casos se requiere señalar las formas para homologar los datos y procesos.

No podemos ignorar desde luego la importancia que reviste para algunos el que los sistemas sean diferentes. En ello se encuentran envueltas estrategias de mercado orientadas al logro de mayores beneficios, o bien intereses institucionales que apuntan a la intención de mantener una determinada posición considerada estratégica.

Sin embargo, inevitablemente, los días del oscurantismo en cuanto al proceso de datos bibliográficos está llegando a su fin y a los usuarios realmente les tiene sin cuidado el porque los sistemas tienen que ser diferentes; no les interesan ni las estrategias de mercado, ni las posiciones de las instituciones, y han empezado a generar sus propios sistemas, situación que debería ser motivo de reflexión tanto para los bibliotecólogos, en su carácter de profesionales de estos asuntos, como para las bibliotecas –instituciones en pleno acomodo de su funcionalidad social–.
Existen sistemas de reglas con una larga y venerable tradición que se utilizan como medio para formalizar y expresar las condiciones indispensables para construir un procedimiento que satisfaga el cumplimiento de un fin deseado (clasificar, describir un documento, elaborar un catálogo, etc.). Difieren esencialmente de las normas en tanto no son patrones o modelos a seguir, ni establecen metas o fines.
Los sistemas de reglas son enunciados procedimentales, no describen, ni explican, ni predicen ningún hecho, sencillamente tienen el propósito de orientar la ejecución de determinado tipo de procedimiento y especifican las condiciones para que ésta pueda tener lugar.
La palabra regla tiene varias acepciones: instrumento para trazar líneas, precepto o prescripción, instrumento para medir, etc.; también significa regular, poner en orden una cosa. Se habla de reglas para referirse a los principios que rigen la enseñanza de un arte o ciencia, acepciones estrechamente vinculadas al sentido que se le da en la práctica bibliotecaria, no obstante, aunque el sentido semántico de la palabra regla nos acerca al concepto de instrumento de regulación para obtener un determinado orden documental, consideramos que sólo una aproximación a la naturaleza lógica de los sistemas de reglas puede servirnos para comprender y explicar su razón de ser.
Si consideramos las reglas independientemente de factores como la forma que adoptan, el sujeto creador o el destinatario, encontramos que una regla sólo existe como tal desde el momento en que adquiere un carácter verbal. Esto quiere decir que la regla lo es en tanto es susceptible de expresarse. La regla es el significado de una expresión lingüística. Es una **proposición**. No toda proposición es una regla.
Para tener el carácter de regla es indispensable que se *inserte en un sistema proposicional* expresivo de un ámbito en el cual (tiene que) tener lugar determinado tipo de procedimientos.
Una regla considerada en forma aislada tiene una significación ambigua o carece de significado, por ejemplo la regla que asigna la notación 200 para el tema religión, únicamente es comprensible y explicable en razón del sistema de clasificación Decimal de Melvil Dewey. Las reglas que componen un sistema no pueden desligarse de él, ya que sólo tienen significado en cuanto se entrelazan para formar el sistema. Viceversa: el sistema no puede ser pensado sin las reglas, puesto que el sistema no es sino un conjunto de expresiones lingüísticas dirigidas, directa o indirectamente, a orientar la creación de procedimientos específicos. Todas las reglas de un sistema son proposiciones constitutivas del mismo, y por tanto son proposiciones **necesarias**, o reglas que establecen un **tener que**, cuyo contenido se orienta a la instauración de procedimientos que es necesario realizar si se desea cumplir determinado fin. Para poder describir un procedimiento tipificado por las reglas de un sistema, habremos de valernos forzosamente de las mismas reglas que lo constituyen, sistematizando e interpretando el contenido de

éstas. De la misma manera, la definición de un sistema en particular, o bien una clase de sistemas, la podemos efectuar mediante el ordenamiento e interpretación exhaustiva del contenido de sus reglas.
Las reglas para representar contenidos temáticos y descripciones de documentos no describen ni explican. El contenido de las reglas, aquello que debe o puede o tiene que hacerse o no hacerse, indica el campo de aplicación y las condiciones que tienen que darse, de tal modo que la acción de representar documentos pueda ser calificada como tal. Estos contenidos tienen un carácter *necesario* ya que la finalidad deseada –representar un documento– sólo es posible si se cumplen las reglas del sistema escogido para ese propósito.
Un *procedimiento* implica una forma de acción, incluye una serie de elecciones y presupone un campo de aplicación (la biblioteca, una red de información) el cual puede estar sumamente formalizado mediante el establecimiento de límites estrictos o de caracteres fijos e invariables (como en Las Reglas Angloamericanas); o puede ser, por el contrario, vago en su delimitación, como en el caso de los tesaurus.
Los procedimientos que pueden desprenderse de un sistema de reglas dan por sentado que los sujetos que habrán de constituirlo y ejecutarlo existen y que son competentes para hacerlo; tienen además un carácter dinámico que hace posible su identificación con una acción y sus resultados.
El que el procedimiento pueda elucidarse con la acción y sus resultados da lugar, en ocasiones, a conclusiones erróneas acerca de la naturaleza de los sistemas de reglas. Normalmente la ejecución de un procedimiento produce un resultado, es por ello que cuando expresamos: “alguien efectuó un procedimiento para clasificar documentos, conforme a las reglas de un sistema”, suele traducirse por: “alguien ha conseguido clasificar estos documentos”. Si pensamos en términos ideales en la acción de clasificar –como realización de un acto o conjunto de actos considerados unitariamente–, lo que **resulte** de la ejecución de estos actos (la organización documental) no es lo mismo que su **consecuencia** (la posibilidad de buscar determinados documentos, o su posible contenido, y conocer su ubicación física). El **resultado** es un componente intrínseco del acto o actos que conforman el procedimiento para clasificar, mientras que la **consecuencia** es el efecto del resultado y, por ello, no constituye un elemento del acto o actos realizados sino que se sitúan fuera de ellos y en relación directa con las finalidades perseguidas.
Una definición parcial de *norma* podría ser que es aquella que tiene como resultado que algo deba o pueda o tenga que ser o no ser hecho. Las formulaciones de normas, en un sentido lingüístico, son un grupo muy variado; utilizan varios tipos gramaticales de sentencias, sin agotar ni ser exclusivamente agotados por ninguno en particular. Debemos, por tanto, estar prevenidos frente a la idea de basar el análisis conceptual de las normas en un estudio lógico de determinadas formas lingüísticas de discurso.
Que una sentencia sea o no la formulación de una norma jamás podría decidirse sobre fundamentos “mórficos”, es decir, sólo con base en el signo. Esto sería así, sólo si existiera una clase precisa y delimitada gramaticalmente (morfológica o sintácticamente) con expresiones lingüísticas cuya función “normal” o “propia” fuera la de enunciar normas. Pues aun en este caso sería el *uso* de la expresión y no su *aspecto* lo que determinaría si es la formulación de una norma u otra cosa. Cuando decimos que es el uso y no el aspecto de la expresión lo que muestra si

es la formulación de una norma, estaríamos de hecho diciendo que la noción de norma es previa a la noción de la formulación.
Es un requisito lógico de las normas el que sea posible que las personas puedan cumplir las exigencias que éstas le imponen. Aunque de mayor interés para la discusión de las normas, en el presente trabajo, es la aplicación, a las normas técnicas, del principio: *debe*, *entraña*, *puede*.
Las normas técnicas en términos aproximados guardan relación con los medios a emplear para alcanzar un determinado fin. Las reglas contenidas en ellas presuponen que las personas que las siguen aspiran a un fin o resultado. Veamos.
La formulación tipo de las normas técnicas encierra oraciones condicionales, en cuyo antecedente se señala alguna cosa que se desea, y en las que en su consecuente se menciona algo que tiene que (hay que, debe de) o no tiene que hacerse.
Supongamos que la norma sea:
si quiero conseguir un determinado fin **E** tengo que hacer un determinado acto **A**.
Como existe la posibilidad de que quiera conseguir este fin independientemente de que pueda o no hacer cualquier acto necesario para su logro, sucede entonces que **E** es algo que deseo, pero que puede significar varias cosas.
Puede significar por ejemplo:
Que **E** es algo que recibiría "con agrado" si me sucediera, ya sea como favor del destino o gracias a la acción de algún otro agente. En este sentido **E** puede ser una cosa que yo deseo, aún cuando no pueda hacer lo necesario para su obtención.
Que **E** es algo que deseo, puede significar que ansío que **E** me suceda. Esto puedo tenerlo sin ser capaz de poner los medios para obtener **E**. Pero desear algo puede también significar perseguirlo como **fin de la acción.** Una inferencia práctica se da cuando una persona extrae de una norma técnica una prescripción para su propia conducta.
A las directrices las llamamos también normas técnicas, dado que presuponen fines de la acción humana y relaciones necesarias de los actos con estos fines.
Una norma representa requerimientos mínimos aceptables en la realización técnica de un trabajo, no representa un ideal a cumplir porque no expresa todas las posibilidades de calidad o deseables en la ejecución de una tarea, muchas de las cuales dependen de la calificación profesional de quien realiza el trabajo y de su habilidad para ligar los procedimientos técnicos con los requerimientos.
La idea de una norma es la de imponer consistencia y no uniformidad, entendiendo por consistencia que existe cohesión, relación entre varias cosas. Las normas se construyen y justifican en razón del cumplimiento de una función determinada.

¿Funciona o no? es constantemente la pregunta. Una de las consideraciones a tomar en cuenta es que el deber ser de una norma técnica, en cuanto a los objetivos que se pretende lograr con su aplicación, tiene que ver con ser necesaria (al menos en principio socialmente deseable) para que pueda cumplir sus funciones sociales previstas. De otra manera caería en el vacío social.

Normalización

La normatividad tendría como finalidad crear normas para ayudar a regular la aplicación, transformación, el desarrollo y la adaptación de la tecnología, y no para establecer un control del desarrollo tecnológico, ya que un acto así tendría por consecuencia eliminar la creatividad en la aplicación de la tecnología y dificultaría su transformación.

Las normas abordan universos limitados y tienen reglas explícitas para relacionar los procedimientos tecnológicos empleados en la solución de un problema o en la realización de una actividad. Tienen como finalidad solucionar los posibles conflictos que surgen al aplicar procedimientos distintos a los habituales.

La mayor parte de los problemas de normalización no son en sí mismos conflictos técnicos, sino que derivan de la forma en como las personas aplican los procedimientos técnicos para resolver determinadas necesidades en torno a la creación de acervos y medios para manejarlos.

La normatividad estaría orientada a la formación de conceptos a partir de los cuales puedan derivarse actividades formales de carácter unificador de técnicas (normas) y conceptos (finalidades), así como al fomento de las actividades de normalización informal (docencia, investigación, terminología, acuerdos institucionales).

Las normas son instrumentos esenciales para la normalización, pero como representan respuestas concretas a problemas específicos, es menester establecer generalizaciones que definan los principios con base en los cuales se habrían de articular diversas técnicas derivadas de distintas normas, y poder así formular un todo coherente. Las normas deben cumplir la misma función de puntos de referencia que cumplen los diccionarios de una lengua.

Los componentes de la normatividad deben sustentarse en la construcción de principios generales cuyo objetivo primordial sería hacer hincapié en la acción unificadora de conceptos por encima de cualquier noción tecnológica específica. No sería ajena tampoco al enunciado de juicios cualitativos que sirvan para evaluar la forma y objetivos de un sistema de información documental, y abarcar los diferentes aspectos de la amplia gama de fines e intereses a cubrir por una organización, porque la normalización únicamente adquiere sentido cuando se le vincula al uso y destino de los documentos.

Aunque estamos acostumbrados a evaluar sobre la base de elementos cuantitativos, es indispensable encontrar los medios cualitativos adecuados para que en conjunto nos ayuden a determinar si un sistema de información es el indicado para disponer de la información documental.

La configuración de conceptos que sirvan como medios para decidir acerca del uso de determinadas normas, no se reduce a un problema de enumeración de criterios de selección en torno a la aplicabilidad monolítica de una norma en detrimento de otra, pues abarca una gama más amplia de conflictos que tienen que ver con la definición de criterios y lineamientos para el uso de técnicas y procedimientos de almacenamiento y recuperación.

La creación de conceptos normativos dirigidos a obtener la eficacia, así como la creación de normas técnicas, apoyan fundamentalmente el desarrollo de la industria de la información. Es el caso por ejemplo del International Standardized Book Number (ISBN) que crea un número normalizado para facilitar la identificación de los títulos publicados; o el International Standardized Serials Number (ISSN) que otorga un número normalizado a cada una de las revistas que

se publican; o el caso del CODEN, un número normalizado que facilita la codificación de publicaciones seriadas.

Todas las normas técnicas relacionadas con los números normalizados están dirigidas a facilitar la codificación de datos para usos automatizados; sus objetivos son estrechos y quedan sujetos a las posibilidades del desarrollo tecnológico en este terreno. Un caso típico a este respecto es la norma ISO–2709 que establece un procedimiento para realizar formatos de captura de datos bibliográficos; a partir de ella se han realizado los formatos MARC y CCF, pero ambos han sido superados por las posibilidades técnicas actuales, y como son procedimientos ligados a determinada tecnología, su modificación resulta complicada, por lo que hoy crean más problemas que soluciones.

El incremento en las cantidades de información disponible trajo consigo la idea de que el volumen, en mayor o menor medida, es un factor importante en la transmisión de información. Como consecuencia, en el campo del control bibliográfico o documental se llegó a estimar que necesariamente la información reduce la incertidumbre. Esta manera de pensar se ve reforzada sobre todo porque los sistemas de control hacen demasiado énfasis en la organización de la información y su provisión al usuario, pero excluyen cualquier consideración respecto a los fenómenos de comunicación presentes en el proceso de análisis y representación de documentos.

Prestar una mayor atención a los hechos que gobiernan la producción, provisión y recepción de información documental, considerándolos como parte de un proceso de comunicación orientado a satisfacer la necesidad de conocer determinados documentos, nos permite situar el análisis y representación de documentos en el centro de un ambiente informativo determinado; y, con ello, aumentar las posibilidades de que sea útil en la familiarización con distintos campos del conocimiento –campos relacionados con el cumplimiento de objetivos institucionales o de los posibles usuarios de un acervo documental–.

En este sentido, la normatividad de clasificación y descripción documental debe proveer instrumentos para determinar y evaluar la eficiencia técnica de un sistema de control bibliográfico o documental en razón de las respuestas que pueda proporcionar al público, y dirigir la evaluación hacia efectos favorables (rentabilidad, productividad, etc.), pero más en razón de las pérdidas que evitan, que de las ganancias o provecho obtenido. Por ejemplo, si un pasajero tiene una tabla de horarios y rutas de trenes y aviones, eso no hará que el sistema de transporte vaya más rápido, pero puede evitar esperas innecesarias. La información sobre rutas y tiempos puede ayudar a facilitar el viaje.

Los servicios de información documental, medidos por el grado de prevención de pérdidas, ayudan a configurar un esquema normativo del control bibliográfico o documental cuya intención principal sea prevenir pérdidas innecesarias.

Medir los efectos de un sistema de información documental por medio de las pérdidas que evita, cuando la comunidad a la que sirve está bien informada, y no sólo con base en la cantidad de información que es capaz de captar y organizar, significa crear esquemas normativos de evaluación cualitativa.

Generalmente los sistemas de información documental responden a criterios normativos cuantitativos con los que se evalúan sus actividades en función del número de servicios prestados. Ahora bien, si la normatividad pretende incluir medios cuantitativos y cualitativos, sería necesario definir los procesos en torno a los cuales se habrá de fundamentar la percepción de la información. Tenerlo claro

es indispensable para codificar y representar –tanto la descripción del documento como sus contenidos temáticos–, así como para derivar aquellos elementos cualitativos que habrán de incluirse en el deber ser de la normatividad durante el proceso de análisis y representación de documentos.

Normalización y desarrollo tecnológico

A pesar de los indiscutibles avances que constituyen las normas como instrumentos para representar documentos, únicamente ahondan en un problema ya conocido, sin proporcionar posibles soluciones a cuestiones derivadas del momento en el cual el público tiene necesidad de definir los elementos necesarios para orientar el rastreo de información.

La normatividad en el ámbito de la representación documental implica la definición de un *deber ser* ineludiblemente vinculado a la determinación de objetivos a perseguir. Su intención debe encaminarse a convertir la representación en un medio para facilitar la distribución social del conocimiento y lograr que el público esté bien informado.

En última instancia, la normatividad para representar documentos forma parte del *deber ser* acerca de las finalidades y funciones de la preservación y organización del conocimiento registrado por medio de todo tipo de documentos, es decir, fundamenta su actuación en fines y valores socialmente aceptados; de estos últimos se derivan los elementos constitutivos de principios encaminados a establecer una normatividad cuya intención sea servir de guía para instrumentar una organización documental funcionalmente dirigida a cumplimentar los fines y valores de la preservación y difusión del conocimiento.

Por esta razón, la definición de elementos teóricos en torno a la representación documental debe concebirse como una función evolutiva que permita englobar en un todo coherente los elementos relevantes para el cumplimiento de los fines asignados a la representación de un documento, sin perder de vista su destinatario final, porque es el público el que en ciertas condiciones y en algún momento le habrá de dar significado a la funcionalidad, formas y modos de la representación documental.

Cuando se piensa en la representación documental únicamente como un medio para describir, y clasificar y formar colecciones de documentos como parte de un simple formalismo, aislado o al margen de cualquier consideración sobre la estructura que se está creando, sin ocuparse de las condiciones sociales que facilitan o impiden su utilización, se está contribuyendo a formalizar estructuras documentales carentes de sentido, porque las colecciones, por bien estructuradas que estén, por sí mismas, aisladas de su contexto social, no tienen significación de ningún tipo.

La creación de nuevas opciones para representar documentos es un imperativo derivado de la dificultad de satisfacer las demandas de información generadas por el público, porque en los hechos la representación se ha desplazado hacia la captación y representación del contenido del documento, con base en su representación por medio de códigos descriptivos.

Como indica Schutz, “la principal característica de la vida de un hombre en el mundo moderno es su convicción de que, en conjunto, su mundo vital no es totalmente comprensible para él ni para ninguno de sus semejantes. Existe un acervo de conocimiento teóricamente disponible para todos, acumulado por la experiencia práctica, la ciencia y la tecnología como concepciones fundamentales.

Pero este acervo de conocimiento no está integrado; consiste en una mera yuxtaposición de sistemas de conocimiento más o menos coherentes, que por su parte no son coherentes, ni siquiera compatibles unos con otros."

REFERENCIAS

Material bajado de Internet. Publicado en: Investigación Bibliotecológica, v.15, No. 30 enero/junio de 2001.

(1) *Cfr*. María Rosa Garrido Arilla. *Teoría e historia de la catalogación de documentos*. Madrid : Editorial Síntesis, 1996, pp. 17-24.

(2) *Cfr*. Frederick W. Lancaster. *El control del vocabulario en la recuperación de información*. 2a. ed., Valencia, España: Universitat de València, 1995, 286 p.

(3) Partimos de la idea del documento como un todo integrado. El distinguir forma y contenido únicamente tiene sentido para fines metodológicos de análisis. Algunos autores conciben el documento como poseedor de una doble naturaleza: soporte más contenido, es decir, el documento estaría integrado por dos elementos dicotómicos, la información en él contenida y su soporte documental. Un resumen sobre estas ideas puede consultarse en: Adelina Clauso García. "Análisis documental: el análisis formal". *Revista General de Información y Documentación*, vol. 3(1), 11-19, 1993.

(4) *Cfr*. Villoro. *Saber y conocer*. México, Siglo XXI, p. 197 y ss.

(5) Ranganathan. *Classified Catalogue Code*. India : Sarada Ranganathan Endowment, 1989, pp. 20-21.

ANÁLISIS DOCUMENTAL

Eugenio Tardón
Universidad Complutense de Madrid (España)

ANALISIS DOCUMENTAL (AD)

DEFINICIÓN Y OPERACIONES DE AD

Consiste en extraer de un documento los términos que sirvan para una representación condensada del mismo. Su objetivo es identificar el documento mediante puntos de acceso e indicar su contenido para permitir su recuperación posterior por parte del usuario.

El resultado es la producción de un nuevo documento diferente al original, un documento secundario: la referencia bibliográfica.

Las operaciones que implica el AD refieren un conjunto de técnicas bibliotecarias tradicionales: catalogación, indización, clasificación y resumen.

PROBLEMAS LINGÜÍSTICOS Y DOCUMENTALES DEL AD

El AD afronta dos tipos de problemas:

1. Lingüísticos, puesto que hay que traducir un texto en lenguaje natural a otro normalizado.

La traducción se realiza a través de los llamados **lenguajes documentales**, que mediante vocabularios controlados limitan la ambigüedad conceptual (tarea más difícil en el área de humanidades).

2. Documentales. Son los que afectan a:
   - el nivel de profundidad del análisis, que afecta al grado de silencio (por exceso de superficialidad) y el ruido (por una profundidad excesiva),
   - la dificultad de normalizar operaciones de un alto contenido subjetivo: dos analistas pueden elegir para el mismo documento pocos descriptores idénticos.

LA REFERENCIA BIBLIOGRÁFICA (RB)

El producto final del AD es la RB, que contiene la descripción bibliográfica del documento original, su clasificación, indización y, eventualmente, un resumen y otras informaciones.

Los servicios secundarios de información (bibliotecas, servicios de resumen, productores de BDs...) tienen como misión elaborar RBs.

Estas RB se realizan de acuerdo a una normativa de carácter internacional, fundamentalmente las ISBD, AACR2 y la ISO 690/1975 (Referencias bibliográficas. Elementos esenciales).

DESCRIPCIÓN BIBLIOGRÁFICA (DB) DE LOS DOCUMENTOS

DEFINICIÓN DE DB

Es el conjunto de información destinada a dar una referencia única que identifique y localice un documento. La tipología de estos es variada: libros, informes, tesis, patentes...

NIVELES DE DB

La información sobre el documento puede tener distintos niveles de descripción en función de la profundidad del análisis. Los niveles más generales son:
- Analítico. El documento se analiza como formando parte de un conjunto más amplio: capítulo de un libro, un artículo de revista...
- Monográfico. El documento se analiza como una unidad entera: fascículo de revista, informe...
- Colectivo. El documento se analiza como un conjunto de entidades físicas: libro en varios volúmenes...

FORMATOS DE INTERCAMBIO DE LA DB

Los principales formatos son:

1) MARC y UNIMARC para bibliotecas;
2) Manual UNISIST de referencia de la UNESCO para centros de documentación;
3) Formato Común de Intercambio.

INDIZACION DE DOCUMENTOS

Definición de indización

Indizar consiste en extraer uno o más conceptos que representan el contenido temático del documento con el objetivo de recuperarlo posteriormente (por ejemplo: distribución de las cuotas de pesca de 1996).
Implica dos tareas:
- Asignar uno o más códigos, numéricos o alfanuméricos, que representan el tema del documento.
- Asignar significantes que corresponden al tema y que suelen extraerse de listas *ad hoc*.

Términos de indización

Son las palabras o números que indican el contenido de los documentos. El número de términos de una referencia varía según las BDs. Cuantos más, mayor exhaustividad en la búsqueda, pero menor precisión por exceso de ruido (documentos no pertinentes); y cuantos menos, mayor precisión y silencio (documentos que serían pertinentes pero que no son recuperados). La solución es buscar el equilibrio entre el ruido y el silencio.

FASES DE LA INDIZACIÓN

Hay tres fases u operaciones a realizar durante la indización:

Examen del documento

Permite establecer su contenido. Hay que prestar atención las partes más informativas (título, resumen, introducción, conclusiones y títulos de los capítulos) y preguntarse qué, cómo, cuándo y dónde.

Extraer conceptos para identificarlo

Se trata de extraer los conceptos que mejor concreten el tema del documento. Se recomienda el uso de listados controlados.

Selección de los términos de indización

Si se utiliza un lenguaje documental, hay que traducir los conceptos extraídos a los términos del lenguaje. Si se trata de texto libre, conviene que los términos sean aceptados en fuentes de referencia: diccionarios, manuales...

SISTEMAS DE INDIZACIÓN

Los sistemas de indización son diversos y responden a exigencias concretas. Podemos distinguir los siguientes.

Indización por materias

Encabezamientos de materias

Su representación típica son los encabezamientos de materias empleados en casi todas las bibliotecas públicas. Los más importantes son: las *Subject Headings* de la LC (1909), la *Sears List of Subject Headings* (1923) y en España, la *Lista de Encabezamientos de materias para bibliotecas públicas* del Ministerio de Cultura y los de algunas universidades (Sevilla, UCM).

Productos. De esta indización se obtienen productos como los catálogos alfabéticos de materias y los índices y bibliografías impresas por materias.

Inconvenientes: Los principales incovenientes son: 1) falta de flexibilidad; 2) inadecuados para las BDs informatizadas; y 3) escasa exhaustividad o profundidad.

Indización por unitérminos (Mortimer Taube, 1955)

Sistema ideado por Mortimer Taube (1955). Consiste en utilizar un sólo término o palabra, el unitérmino, para representar los contenidos de un documento. Pese a sus inconvenientes supone un avance importante respecto de la indización por materias.

Inconvenientes. Los principales son: 1) exceso de falsas combinaciones; y 2) abundancia de palabras polisémicas, homonímicas, sinónimas, ambiguas y vacías.

Indización por palabras-clave y descriptores (C. Mooers, 1941)

Es una indización relacionada con los primeros tesauros. Hay una ligera diferencia entre palabra-clave, que es una indización en lenguaje libre extraída del texto del documento y descriptor, que es un término sacado de un lenguaje documental y que puede ser unitérmino, sintagmático (varias palabras), identificador (geográfico, personal, acrónimo). La indización basada en descriptores la inició el norteamericano Calvin N. Mooers en 1941.

Indización automática

Consiste en contrastar los vocablos de un documento con un diccionario invertido del programa, que puede ser un tesauro. Tras ello se asignan los términos seleccionados.

Indización vectorial

Es un tipo de indización automática con importantísimas consecuencias en el terrero documental. Parte de las insuficiencias de la lógica booleana empleada en los motores de búsqueda y que se sustenta en el uso de técnicas binarias, donde

los términos de búsqueda están o no, y no existe ponderación de términos en los documentos o registros, sino solamente operaciones booleanas (y, o, no). Este método es insuficiente y ha sido criticado desde los 80 por su lógica no intuitiva, la ausencia de lenguaje natural y la necesidad de FU, a lo que se añade el hecho de que los operadores son muy restrictivos (y) o muy inclusivos (o) y rígidos, pues consideran todos los documentos igualmente pertinentes. Todo ello lleva a un alto índice de búsquedas sin respuestas (seach failure), casi el 50% generan silencio (por desconocer los puntos de acceso y otras dificultades) o respuestas excesivas del sistema (information overload).
La indización vectorial resuelve este problema, se basa en la ponderación de entradas, sobre todo en los trabajos de Shalton, que se apoyan en las formulaciones de Zipf y S. Jones, y que establecen, básicamente, la relación entre la frecuencia de un término y su importancia para la representación del documento. Shalton elaboró un modelo vectorial que comparaba la similaridad de la petición del usuario con la de los documentos de la base de datos. Cada documento de la BD tiene un coeficiente que resulta del peso de cada uno de sus término, y cada pregunta del usuario es otro vector con un coeficiente análogo. El resultado son dos coeficientes vectoriales: D (del documento) y Q (de la pregunta). De esta forma, recuperar información es determinar el coeficiente de similaridad de los vectores D y Q.
Las ventajas son tremendas:

1) Los documentos se ordenan según su pertinencia, que se deriva de su puntuación (ahorrando tiempo al lector);
2) El tamaño de los conjunto recuperados es predefinible, lo que supone el fin del overloap; y 3) No es necesario conocer el lenguaje de información, pues la consulta se puede efectuar en lenguaje natural. El éxito de este sistema ha llevado a su implantación en muchas BDs documentales: Lotus Notes, Personal Librarian, Wais, el host Dialog con su orden *target*, e incluso los motores de búsqueda de Internet Lycos y Altavista.

LENGUAJES DOCUMENTALES (LD)

DEFINICIÓN DE LENGUAJE DOCUMENTAL (LD)

Un LD es un conjunto de términos o frases nominales convencionales empleados para representar el contenido de un documento con el fin de facilitar su recuperación. Es un lenguaje artificial, controlado, para diferenciarlo del lenguaje natural. Permite la comunicación entre usuario e información al emplear la misma representación formalizada. Se denominan también sistemas de clasificación, lenguajes de indización y léxicos terminológicos.

TIPOLOGÍA DE LOS LD: PRECOORDINADOS Y POSTCOORDINADOS

1. Los precoordinados se elaboran antes de su aplicación a los documentos, para normalizar los conceptos que forman una materia. Ejemplo: CDU, encabezamientos de materia.
2. Los postcoordinados son los que yuxtaponen los conceptos y los coordinan después del almacenamiento. Se desarrollan al unísono con el conocimiento científico. La incorporación de los nuevos conceptos es casi inmediata.

ESTRUCTURA DE LOS LD

Por su estructura, los LD pueden dividirse en lenguajes jerárquicos y asociativos.

Lenguajes jerárquicos o clasificatorios

Agrupan los conceptos desde lo general a lo específico, lo que lleva a clasificaciones sistemáticas lineales. Cada concepto se representa por un código (numérico, alfabético o alfanumérico). Las categorías son inamovibles y carecen de flexibilidad.
Los tipos de lenguajes jerárquicos más conocidos son:

Sistemas o clasificaciones enciclopédicas

Permiten organizar documentos de cualquier materia, pues son universales y multidisciplinares. Dividen el conocimiento en clases y susbclases. El prototipo es la CDU, que es la Clasificación Decimal de M. Dewey (1876). Pese a sus inconvenientes, ha mostrado su utilidad, pues nació para normalizar la clasificación bibliotecaria y evitar la babel terminológica. Cubre bastante bien las necesidades de la mayoría de las bibliotecas, facilitando, sobre todo, el libre acceso y el estudio del uso de la colección por categorías. Es un sistema ineficaz para documentos especializados y para los nuevos tipos documentales: video, cartografía, patentes, archivos de ordenador. Otros sistemas son: Library of Congress Classfication (1904), originada en A. Cutter; la clasificación de Henry Bliss - BC -, y la de la antigua URRS, BBC.

Clasificaciones especializadas

Abarcan diferentes disciplinas: medicina, derecho, economía. Ejemplo de ellas: Excerpta Médica.

Clasificaciones facetadas

Son de origen enciclopédico, pero su organización permite construir áreas concretas del conocimiento, siendo un nexo de unión entre los sistemas jerárquicos y los asociativos.
Funcionan asignando índices parciales, que se yuxtaponen, de cada una de las facetas que caracterizan al documento. El modelo es la Colon Classification (CC), ideada por Ranganathan en 1933, que divide el conocimiento en cinco grandes familias: personalidad, materia, energía, espacio y tiempo.

Lenguajes de estructura asociativa

Organizan, por lo general, las nociones por orden alfabético mediante términos que describen los conceptos. Estos descriptores se combinan entre sí libremente. Se organizan desligados unos de otros, salvo en las operaciones de indización (preguntas y análisis).
Pertenecen a esta categoría: encabezamientos de materia, unitérminos, descriptores y tesauros, constituyendo estos últimos el modelo más completo y hacia el que tienden los vocabularios, taxonomías, etc.

Tesauros

La ISO 2788/1974 lo define desde el punto de vista funcional y estructural. Funcionalmente es un instrumento para controlar la terminología al trasladar a un lenguaje más estricto la lengua natural de los documentos. Estructuralmente es una lista de autoridades compuesta por descriptores relacionados entre sí semánticamente (jerarquía, asociación, equivalencia).

ESTRUCTURA DEL TESAURO

Descriptores
Son los términos de un tesauro que representan un concepto sin ambigüedad. Se diferencian de los unitérminos y las palabras-clave en que éstos son parte del lenguaje natural.

Identificadores
Son descriptores referidos a nombres geográficos, de personas, entidades, acrónimos.
Relación entre descriptores:
1. Relaciones de sinonimia, polisemia y homografía. Se establece mediante la notación USE (que remite del descriptor no admitido al admitido), y UF (que informa de términos sinónimos, polisémicos..., no admitidos).
2. Relación jerárquica, que define los descriptores más genéricos y más específicos, dando lugar a relaciones recíprocas. Emplea las notaciones BT y NT.
3. Relación asociativa o de afinidad. Indica las relaciones de cierta equivalencia en dirección horizontal de los términos. Se representa por la notación RT, en español VT (véase también).
4. Notas de definición o Scope Note. Explican brevemente la utilización que debe asignarse al descriptor para: limitar su empleo, desarrollar acrónimos, excluir dobles sentidos. Se emplea con la notación SN, en español NA (nota de alcance).

Partes de un tesauro
Pueden distinguirse cuatro apartados:
1) Grandes familias o microtesauros, con la relación de todas las facetas de cada uno;
2) Descriptores ordenados por facetas;
3) Alfabético de descriptores;
4) Índices permutados KWIC y KWOC.

EL RESUMEN O ABSTRACT
Es la representación abreviada y precisa del contenido de un documento, sin interpretación crítica y sin distinción del autor del análisis. La norma ISO 214 proporciona reglas para preparar y presentar los resúmenes.

- TIPOS DE RESÚMENES

Informativo o analítico
Es un resumen completo, con información cuantitativa y cualitativa, de unas 250 palabras.
Descriptivo o indicativo
Es más breve, entre 50-100 palabras. Describe el tipo de documento y tema tratados de forma breve.
De autor o documentalista
Cada vez más, las normas exigen a los autores un resumen del texto en el momento de su aparición.

- OBJETIVOS DEL RESUMEN

Tres objetivos principales: determinar el interés del documento de una forma rápida, ayudar a la selección de la información, y difundir la información.

- CÓMO HACER EL RESUMEN

Elementos del resumen

Al elaborar el resumen se deben mencionar los siguientes aspectos:

1. Finalidad. El resumen debe recoger los objetivos principales o el tema del estudio, salvo que aparezca en el título
2. Metodología del estudio. Los métodos de investigación no deben describirse salvo que ayuden a explicar el texto o sean técnicas nuevas.
3. Resultados y conclusiones. Deben estar representados claramente en el resumen.

Redacción del resumen: disposición y estilo

En cuanto a cuestiones estilísticas, debe redactarse como mínimo en el idioma del documento original. Si forma parte de una revista, se dispondrá al principio del artículo; si es de un libro o tesis, en el reverso de la página del título o en la siguiente; y si es una referencia bibliográfica, tras la descripción bibliográfica de ésta.

Debe iniciarse con una frase que contenga en lo posible la idea esencial, con pocas abreviaturas, con verbos en forma activa y palabras significativas que sean útiles al interrogar al sistema.

Como pautas a seguir:

a) leer las partes principales del texto, tomar nota de las ideas más significativas y apuntar palabras clave;

b) redactar un borrador a partir de las notas tomadas evitando copiar del documento original, sino con el estilo del redactor;

c) pulir el estilo, sintaxis, puntuación y gramática.

REFERENCIA

Material bajado de Internet.

BIBLIOGRAFÍA

Amat Noguera, N. *Documentación científica y nuevas tecnologías de la información*. Madrid: Pirámide, 1987.

Gimeno Perelló, J. “Sistemas de indización aplicados en bibliotecas: clasificaciones, tesauros y encabezamientos de materias”. En: *Tratado básico de Biblioteconomía*. Madrid: Síntesis, 1996.

Guinchat, C.; Menou, M.; Blanquet, M-F. *Introducción general a las ciencias y técnicas de la información y documentación*. Madrid: CINDOC, UNESCO, 1990.

CRITERIOS E INDICADORES PARA EVALUAR LA CALIDAD DEL ANÁLISIS DOCUMENTAL DE CONTENIDO

José Antonio Moreiro González
Universidad Carlos III de Madrid (España)

JUSTIFICACIÓN Y PROPÓSITOS

La calidad es un asunto de importancia creciente en los servicios de información. De su seguimiento preocupa ante todo cuanto se refiere a la gestión normalizada de la calidad en los sistemas de información, junto a la propia medición de la calidad que presentan los productos informativos (1). La norma ISO 8402 define la calidad como "la totalidad de rasgos distintivos de un producto o servicio que tienen que ver con su capacidad para satisfacer necesidades manifiestas o implícitas" (2) . Esa satisfacción respecto al producto debe verse desde una doble perspectiva: la que se refiere a administrar externamente las expectativas de los usuarios respecto a un producto, y la que desde dentro intenta reducir las consecuencias de los fallos humanos y empequeñecer los defectos.

Quienes procesan la información se han preocupado siempre de suministrar buenos productos y servicios a sus usuarios. Los resultados del tratamiento documental deben garantizar la calidad y eficacia de los productos que ofrecen si se quiere justificar y asegurar la propia existencia de los repertorios documentales. Entendemos como tales a los documentos referenciales o secundarios que agrupan los registros analíticos de los documentos originales o primarios. Y que lo hacen tanto en el modo más clásico de difusión, sirviéndose de las ediciones impresas (bibliografías analíticas, boletines de resúmenes, sumarios analíticos de publicaciones periódicas), como mediante formatos digitalizados (esos mismos productos fijados en bases de datos referenciales que pueden distribuirse en línea, mediante CD-ROM, e incluso difundirse por Internet). Los repertorios han sido los productos más comunes elaborados en los sistemas de información, y continúan ocupando un lugar fundamental en la difusión de la información científica, técnica y especializada.

Para ello debemos decidir la presencia de unos criterios explícitos en las diferentes fases de confección de los repertorios, que a la hora de hacer la evaluación se plasmarán en unos indicadores determinados. La complejidad del proceso es mucha, afectando como elementos a evaluar a:

- La atención, interés, número y pericia de los analistas.
- Las características de las fuentes a incluir en el repertorio.
- Los costes, métodos, procedimientos, y tiempo en que se efectúa el análisis.
- El producto obtenido y su adecuación a los objetivos documentales.
- El esfuerzo que deba hacer el usuario.
- E incluso la forma de presentación.

La calidad de las tareas correspondientes al análisis de contenido documental (especificadas en la indización y el resumen) resulta, pues, fundamental para permitir una satisfactoria recuperación de información y una adecuada explicación de los contenidos a los usuarios. Los rasgos que determinan ambos procesos son paralelos a las exigencias cualitativas que vamos a describir a continuación.

En este sentido, es notable el esfuerzo que actualmente realiza Europa. Sirva de ejemplo la actitud de la European Association of Information Services (EUSIDIC) que se ha convertido en responsable de la excelencia de sus productos y servicios (3), igual que de la inglesa National Federation of Abstracting and Information Services (NFAIS), así como de la Association des documentalistes et bibliothècaires specialisés francesa por asegurar la calidad de los productos generados por sus industrias de la información, en especial por los que conforman las bases de datos.

Nuestra propuesta se centra en la evaluación del proceso de análisis documental, pues sabemos que, dada la complejidad de los procesos de indización y resumen, nunca alcanzan una exhaustividad y precisión plenas. No consideramos aquí otros factores que intervienen en el buen funcionamiento del sistema de información, como los relacionados con los usuarios. Atendemos tan solo a los factores derivados del análisis documental de contenido.

Precisamente por tratarse de textos, la medición estadística y numérica de los factores cualitativos no siempre puede hacerse. Máxime si consideramos que hablamos de calidad, concepto que conlleva la consideración de rasgos que solo se pueden apreciar mediante la observación y juicio personales. Por ello, la aplicación de estos criterios no tiene por qué ser cuantitativa en exclusiva. A causa de la gran cantidad de información que se acumula en las colecciones documentales, y a la intervención de múltiples factores que hacen de la indización un proceso muy complejo, los procedimientos de recuperación no pueden ser nunca íntegramente exhaustivos y precisos. Así pues, a la hora de determinar el nivel de calidad alcanzado por los productos documentales, debemos considerar unos factores cualitativos que denominaremos criterios, junto a unas unidades de medida o indicadores (4). Consideramos indicadores de calidad los que miden la coherencia, la pertinencia o precisión, la exhaustividad o respuesta, la consistencia, la densidad informativa, la profundidad, la extensión o tamaño, así como los indicadores temporal, de costes (recursos invertidos en un servicio), del esfuerzo del usuario y de errores.

El indicador temporal es la medición del tiempo de respuesta que está determinado por la organización y el tipo de archivo en que se custodien los datos, la ubicación del sistema de información, la saturación a que el servicio se halle expuesto y el tamaño de los ficheros manejados.

Mientras que el esfuerzo del usuario vendrá determinado por la ayuda que le puedan prestar los profesionales de la información, la cantidad de información que se le sirva tras efectuar una búsqueda, el formato en que la información se presente, así como la facilidad de manejo e interacción con el sistema, y las propias habilidades para buscar en general o para hacerlo en un sistema en concreto.

El indicador de errores quiere valorar la introducción de errores que modifiquen los rasgos originales de la información analizada. Suceden en cada eslabón de la cadena documental: en la identificación de las fuentes, en los procesos de extracción, clasificación y codificación, e incluso en la grabación de los datos.

PROPUESTA METODOLÓGICA

A. Tareas de **identificación** de los repertorios existentes en el área o las áreas estudiadas (incluyéndose las Bibliografías y las Bases de Datos - ya sean estas on line, en CD-ROM, o distribuidas por Internet-). Confección del listado de los repertorios identificados.

La identificación pretende obtener conocimiento sobre estos aspectos del repertorio:

A.1. Antecedentes del mismo. Su evolución.

A.2. Naturaleza de la Bibliografía o Base de Datos:

– ¿Qué campo cubre?
– ¿Quiénes son los editores y quiénes los usuarios potenciales?
– ¿Qué tipo de producto concreto nos ofrece?
– Contenidos.
– Formato.
– Acceso.
– Periodicidad.
– ¿Cómo podemos obtener la información de esos productos?

A.3. Medición del porcentaje de exhaustividad de la **cobertura del repertorio** (CR): es el grado de cobertura temática (Coverage) o proporción de información existente sobre una materia, publicada dentro de un período de tiempo concreto, que está incluida en la Base de Datos. Se realiza mediante la construcción de una lista de publicaciones periódicas del área estudiada, comparándola con las que recoge el repertorio. De acuerdo con la fórmula:

***CR**= C/i x 100,*

donde **C** es el número de publicaciones periódicas del área recogidas en el repertorio e **i** el número total de publicaciones periódicas del área. Esta unidad de medida tiene una apreciación cualitativa, al tener que considerarse si las publicaciones periódicas recogidas son después analizadas en su integridad o lo son de manera selectiva.

Cuando se hagan estudios comparativos de cobertura entre varios repertorios estaremos intentando delimitar la cobertura que efectúa cada uno de ellos y, en consecuencia, también las repeticiones y lagunas contenidas en los listados de referencias respecto de los originales de las áreas que pretenden cubrir. En este caso hablaremos de cobertura relativa, en la que dados dos repertorios A y B, la cobertura de A será (5):

$$\frac{\text{Número de referencias en } \boldsymbol{A}}{\text{Número de referencias en } \boldsymbol{A} \cup \boldsymbol{B}}$$

Asimismo, podremos utilizar la medida complementaria del **solapamiento**, por la que se perciben aquellas referencias comunes a dos o más sistemas de información.
Dados dos repertorios A y B, el solapamiento global saldrá del cociente:

Número de referencias comunes en A y B
_____
*Total de referencias en **A** U **B***

Mientras que el solapamiento relativo será el de un repertorio (A) respecto a su propia cobertura (6):

Número de referencias comunes en A y B
_____
Total de referencias en A

Contraria al solapamiento global es la medida del aporte específico de un repertorio (A), resultante de oponer

Número de referencias no solapadas comunes en A
_____
*Total de referencias en **A** U **B***

Una consecuencia del estudio de las revistas de un área mediante la cobertura y el solapamiento en los repertorios de esa misma área es saber cuáles son las revistas más analizadas y, por tanto, las que forman el núcleo dentro de una disciplina.
B. Preparación profesional de los técnicos que elaboran el repertorio. Debe entenderse que el primer criterio de calidad se deriva de la actitud y preparación específicas con que los analistas abordan su trabajo:
• Si el repertorio está realizado por profesionales o no.
• Si son graduados en Biblioteconomía y Documentación.
• Si lo son en el área de la Base de Datos o Repertorio.
• Si no tienen titulación universitaria, pero han sido formados en la editorial.
• Cuál es su nivel de experiencia, conocimiento de la terminología y de las reglas de uso del lenguaje utilizado.
• Cuál es la relación entre el número de analistas y el trabajo a realizar.
• Grado de atención e interés.
Obtener información de este tipo es una tarea difícil que muchas veces conllevará la interrogación mediante encuesta a los propios servicios de información, ya que nada suele explicitarse de este y otros asuntos en las explicaciones técnicas que ofrecen los repertorios.
C. Existencia de normativas o directrices: Observar si se guarda respeto a las normas generales y a las directrices existentes en el centro productor; cuál es la estabilidad de los criterios fijados en ellas; qué seguimiento o supervisión de su cumplimiento se realiza; si existe o no algún control periódico de la calidad (7). Todos ellos son factores que definen la homogeneidad en los procesos de indización y resumen, así como en la publicación de los resultados.

D. Tipo de lenguaje utilizado para efectuar la indización:
– Mediante lenguaje libre:
• Extraído de las palabras del título.
• A partir de la información contenida en el resumen.
• Considerando el texto completo.
Se pretende saber de qué manera se indizó en lenguaje libre. En este caso, se deben contrastar las palabras-clave con el título, el resumen o el propio texto para determinar si se deben a un proceso de extracción, a una determinación de los documentalistas, o bien si son las mismas palabras-clave utilizadas por el autor del original.
– O usando algún lenguaje controlado, cuya identificación y características deben determinarse, y que puede ser:
• Tesauro.
• Lista de palabras clave.
• Glosario de descriptores.
• Lista de encabezamientos de materia.
• Otros (Mapas conceptuales, CDU…).
La **accesibilidad** al lenguaje documental vendrá determinada por los recursos utilizables por el usuario para la selección de los términos de recuperación más adecuados.

Fig. 1 Comparación de las características fundamentales entre los tipos básicos de resumen.

	Estructuras	**Expresión**
Informativo	Mg + (Mp1+ Mp2...+ Mpn) =	
Explicativa	Mg +5	
Selectivo	Mg + (Mp1+ Mp2...+ Mpn) =	
	Mg + Sp	Indicativa
Indicativo	Mg (+ otra información)	Indicativa

E) Tareas relativas a la **evaluación del análisis de contenido** (indización y resumen, cuando haya). Dado que se trata de una apreciación cualitativa, se requiere la selección de un número suficientemente representativo de registros de cada Base de Datos o Bibliografía atendida, y su posterior comparación con los artículos originales a los que esos registros se refieren. La evaluación se hace mediante inspección directa de la muestra de los registros seleccionados, que es considerada representativa de cada repertorio. La aplicación de este método muestra mucha lentitud, pero nos asegura que el proceso sea crítico y deductivo.

CRITERIOS A EVALUAR EN LA ELABORACIÓN DE LOS RESÚMENES

1) Debe considerarse el grado de **reutilización** de los resúmenes hechos en origen (de autor). Hay que apreciar si se aprovechan tal cuál estaban en el original, o si por el contrario se han mejorado técnicamente, buscando las deseables coherencia y normalización (8). El seguimiento de la reutilización de los resúmenes debe discernir entre aquellos copiados literalmente del autor, de aquellos que basándose en el resumen analítico han sido corregidos o modificados. En el caso de que el original no contenga resumen de autor, sería muy positivo comparar el resumen de la referencia con los primeros párrafos del

texto o alguna frase significativa en la que se represente la información sustancial. Ya que muchas veces es esta la fuente de la que se extraen los datos que luego conforman un resumen, casi siempre indicativo.

2) Es fundamental identificar cómo se ha efectuado el **traslado de la superestructura** del original (de manera informativa, selectiva o indicativa) (9):

– Los resúmenes informativos deben trasladar la Macroestructura global del texto (Mg), además de la superestructura (Sp), o lo que quiere decir además de las macroestructuras parciales (Mp) ordenadas, y tener una expresión explicativa (con oraciones suficientemente informativas). Es en realidad una versión abreviada de la idea central de un documento (macroestructura global), así como de las ideas que vinculan cada una de las partes que lo componen (macroestructuras parciales), y éstas en orden (superestructura). Así pues, el resumen informativo representa: Macroestructura global + (Macroestructura parcial 1 + Macroestructura parcial 2 + Macroestructura parcial n), o lo que es lo mismo Macroestructura global + Superestructura, junto a una descripción concisa.

Esquemáticamente, y en el caso de un artículo de ciencia experimental, las oraciones del resumen informativo indican:

1. Macroestructura global / objetivos
2. Metodología: Fase de descripción / Fase de análisis.
3. Resultados – Discusión.
4. Recomendaciones – Conclusiones.
5. Bibliografía – Anexos.

– Los resúmenes selectivos deben trasladar la Macroestructura global del texto (Mg), además de la superestructura (Sp), o lo que quiere decir además de las macroestructuras parciales (Mp) ordenadas), pero hacerlo en expresión indicativa. Resumen selectivo = Mg + (Mp1+ Mp2.+ Mpn), o lo que es lo mismo Mg + Sp, con explicaciones para esta última tan sólo indicativas.

– Los resúmenes indicativos se limitan a trasladar la Macroestructura global del texto (Mg) y, a veces alguna de las macroestructuras parciales (Mp), siempre mediante una expresión indicativa (Mg + otra información, explicada ésta solo de forma indicativa).

3) Calidad técnica de los resúmenes: Viene dada por aquellas presencias inútiles y ausencias evitables, antes explicadas (10): Presencia de redundancias (frases inútiles, repetición del título, presencia de expresiones como: El autor analiza... , Este artículo..., El documento que...); presencia de faltas ortográficas o sintácticas. Debe hacerse explícito en la frase introductoria del resumen la **naturaleza y enfoque** del documento original (11).

Así, debe comenzarse el resumen indicando la naturaleza del documento original: si se trata de un artículo, crítica histórica, crónica, entrevista, editorial, ensayo, estudio estético, examen de un caso, nota, presentación de resultados, reportaje, tesis,... Luego, debe constatarse el enfoque específico que caracteriza al trabajo analizado, cómo el autor ha tratado el asunto: desarrollo complementario, estadístico, en exhaustividad, expositivo, de réplica, revisión, técnico-experimental, teórico...

4) Tamaño de los resúmenes: cantidad de palabras que los conforman, pareciendo excelente que su extensión se sitúe entre las 100 y las 250 palabras (12). Considerando el número total de palabras que conforman un resumen, la **densidad** es el indicador que mide las que son nocionales, es decir, aquellos

términos que representan un concepto relacionado con el tema o materia de la que se trata. La fórmula de la densidad es:

$$\frac{\text{Número de palabras nocionales}}{\text{Número total de palabras}} \times 100$$

5) Indicadores de legibilidad de los resúmenes. La legibilidad mide la claridad de expresión, y en el caso de los resúmenes es un requisito de su calidad expresiva, por lo que asimismo puede afectar a la indización.
Consideramos que actúan como factores indicativos de legibilidad (13):

Dificultándola	*Facilitándola*
– La presencia de la voz pasiva	*– La presencia de enlaces sintácticos*
– La subordinación excesiva de oraciones dentro de una frase.	*– La presencia de oraciones cortas*
– La presencia de abreviaturas y acrónimos sin normalizar.	

La legibilidad debe alcanzar a la medición del número de sílabas por palabra, siguiéndose por el número de palabras por frase, y concluyendo por el número de frases por párrafo. Si bien creo que debe centrarse exclusivamente en el recuento del número de palabras por frase, de acuerdo con estas consideraciones:
Tendremos por oraciones cortas las inferiores a 15 palabras. Por oraciones medias, las que tengan entre 15 y 20 palabras. Y por oraciones largas, las que tengan más de 30 palabras (El porcentaje se obtendrá por repertorio)
6) La **cohesión** de los resúmenes consiste en comprobar que las oraciones estén bien unidas gramaticalmente Es un aspecto puramente de coherencia superficial que permite la lectura seguida, de manera discursiva, sin trabas y que se considera, por tanto, dentro de los valores expresivos de la legibilidad.

CRITERIOS E INDICADORES CON LOS QUE SE HIZO LA INDIZACIÓN

El proceso de indizar consiste en describir y caracterizar un documento con la ayuda de representaciones de los conceptos contenidos en dicho documento, con la finalidad de permitir una búsqueda eficaz de las informaciones contenidas en una colección documental.
Debe destacarse aquí la relevancia que alcanza una indización correcta a la hora de evaluar un repertorio. Ya que la indización se establece como instrumento referencial del conocimiento de los contenidos específicos de cada repertorio. Razón por la que cumple el papel protagonista en la recuperación de los contenidos.
El indizador debe buscar y utilizar un descripción que traduzca lo más de cerca posible el contenido del documento (**especificidad**), rechazando los descriptores demasiado generales o demasiado particulares con relación a las nociones que expresa el documento. Su logro supone la **relevancia**, concepto que puede atribuirse a la recuperación, cuando un documento es útil para los propósitos que causaron una búsqueda por parte del usuario (14). Lo ideal sería encontrar todos los documentos relevantes y evitar los no relevantes (obtener a la ver exhaustividad y precisión). Pero que en nuestro caso lo aplicamos a la carga de significación del descriptor.

1) La **procedencia de los términos** de indización es un elemento que marca diferencias para valorar la calidad de un repertorio. Lo es por la enorme importancia de una recuperación exacta, sin la cual las demás tareas en la construcción del repertorio no tienen sentido. Respecto a este punto puede darse varias situaciones:
– Que los indizadores reutilicen la indización hecha en origen. En cuyo caso habrá que ver si fue corregida o no. Y dependerá también de cómo se efectuó esta indización en origen: si lo fue con descriptores o con palabras clave. En este último caso sería adecuado revisar desde el original la pertinencia y exactitud de las palabras escogidas para representar los conceptos. Una indización en palabras clave que traslade literalmente la propuesta por el original puede dejar pasar incoherencias e imprecisiones.
– Si percibimos que la indización fue totalmente elaborada en el centro de análisis, habrá que plantearse la procedencia de los descriptores, por una parte, respecto al lenguaje documental utilizado (comentado antes, y que puede darse una presencia simultánea de descriptores y palabras clave en un mismo registro) y, por otra, intentar determinar si en cualquiera de los dos casos fueron obtenidos a partir de la consulta al texto íntegro, desde los epígrafes del original, consultando tan solo la introducción o el resumen, o si exclusivamente se tomó el título como fuente de consulta. Estas últimas consideraciones guardan una relación muy estrecha con el punto siguiente.

2) Respecto a la **profundidad** en la representación del contenido textual (hasta qué nivel se representa la superestructura), dependerá de si los términos de indización se refieren a todo el texto reflejando las macroestructuras parciales, si fueron extraídos a partir solamente del resumen analítico y, en dependencia del tipo de este, que alcancen a representar solo algunas de las macroestructuras parciales, o si son tan genéricos que reflejen tan solo los conceptos del título.
El indicador de profundidad resulta del cociente:

$$\frac{\textit{Número de palabras del original}}{\textbf{Número de palabras x 10}}$$

Esta cualidad se asocia directamente con la pertinencia o relación entre los contenidos del original y los términos de la indización (15).

3) Índice de consistencia, que solo sería posible en aquellas áreas con varios repertorios, la consistencia del análisis documental se refiere a que un concepto o tema aparece siempre expresado de la misma forma (16) :

$$\textbf{\textit{IC (\%):}} \frac{100A}{A + M + N}$$

Donde A son los términos comunes a los repertorios M y N.

En el caso de que la descripción fuese comparativa entre varios indizadores o usuarios estaríamos hablando de la **uniformidad** u homogeneidad con la que todos ellos deben describir el mismo documento, o documentos sobre el mismo tema, de la misma manera.

4) Otros indicadores y factores de evaluación

a) Indicador de pertinencia o precisión
Tras obtener la respuesta a una búsqueda, es un cociente (precisión ratio) que resulta de dividir el:

Número de documentos relevantes recuperados
———————————————————
Número total de documentos recuperados

Desde la perspectiva opuesta a la pertinencia, puede hablarse también de una tasa de silencio: los documentos no recuperados, pero que son relevantes. El cociente de precisión nos indica que cuanto más se acerque a 1 mejor será la recuperación.

85 referencias pertinentes
ejemplo: ——————— = 0,6 - 0,7 (resultado bueno)
100 referencias totales

Es un concepto relativo que se mide con respecto a un referente (serie de datos) que sirve como punto de partida.
Se puede hablar de pertinencia de un descriptor, de la indización, del lenguaje...

b) Indicador de exhaustividad o de respuesta.
El indicador de exhaustividad (recall ratio) busca que todos los temas, objetos y conceptos que encierra el documento estén bien determinados en la indización, por lo que habrá una respuesta ajustada a una búsqueda dada, que se mide en porcentaje a través de la relación entre:

Número de documentos relevantes recuperados
———————————————————
Número total de documentos relevantes existentes en la Base de Datos

Es más difícil de corregir que en el caso de la pertinencia, pues para comprobarlo habrá que examinar toda la colección documental. También un porcentaje de respuesta entre el 0,6% y el 0,7% se considera un buen resultado. Respecto a este indicador, y desde la perspectiva opuesta a la exhaustividad, hablaremos de una tasa de ruido que valora la proporción de documentos recuperados que no son relevantes para la búsqueda propuesta:

Número de documentos no-relevantes recuperados
———————————————————
Número de documentos recuperados

Se ha demostrado que la eficacia es mayor utilizando lenguajes de términos simples que cuando se manejan lenguajes de términos sintagmáticos (17). Así como se ha observado la existencia de una relación inversa entre la precisión y la exhaustividad.

c) Finalmente, podemos hablar del esfuerzo requerido en la recuperación (Retrievability): Cantidad de información contenida en la Base de Datos, sobre un tema concreto, que puede recuperarse usando una estrategia de búsqueda razonable.

Así como de la facilidad para juzgar la pertinencia (Predictability): Recursos utilizables por el usuario para juzgar con rapidez si los registros son adecuados a sus necesidades.

Y de la actualización (Timeliness): Presencia y proporción de la información más moderna dentro del conjunto de registros consultables en la Base de Datos.

PRESENTACIÓN COMPARATIVA DE LOS RESULTADOS OBTENIDOS

Una parte de las conclusiones ha de expresarse en textos exponiendo valorativamente los resultados tras hacer observaciones en los repertorios siguiendo los criterios antes expresados.

Deben distinguirse los resultados de la aplicación de los criterios a los juicios de calidad del repertorio entendido globalmente de aquellos propios del resumen o de los que afectan a la indización. Es conveniente separar la aplicación de los criterios en la observación de un hecho o descripción técnica de las mediciones en que se aplicaron índices o indicadores.

Es aconsejable utilizar tablas para recoger los resultados de la evaluación de cada una de los repertorios o incluso de cada uno de los registros analizados. Como ejemplo:

La utilización de gráficas comparativas y de tortas por porcentajes otorga mucha claridad a las conclusiones y facilita la comprensión de los resultados por parte de los lectores.

NOTAS Y REFERENCIAS

Material bajado de Internet. Publicado en: Ciencia de la Informaçao, Brasilia, v. 31, no. 1, pp. 53-60, jan/abr. 2002.

(1) SWINDELLS, N.- *Managing the quality of Information products*, en *Managing Information*, (1995), 4:

(2) UNE-EN *ISO 8402: 1995. Gestión de la calidad y aseguramiento de la calidad: Vocabulario*. Madrid: AENOR, 1995.

(3) DENIS, S. et al.- *Liability in the provision of information services. EUSIDIC Research Project 1989*. Brussels: EUSIDIC, 1990.

(4) Un buen indicador debe ser pertinente hacia su objeto, operativo y cuantificable.

(5) ABAD GARCÍA, F.- *Investigación evaluativa en Documentación*. Valencia: Universitat de Valencia, 1997: 130-132.

(6) LABOIRE, T.; HALPEIN, M. y WHITE, H.- *Library and InformationScience Abstracting and Indexing services: Coverage, Overlap and Context*, en *Library and Information Science Abstracts*, (1985), 7: 183-195.

/7) LANCASTER, F. W.- *El control de vocabulario en la recuperación de la información*. Valencia: Universitat de Valencia, 1996.

(8) MOREIRO, J.; MELO, D.; GARCÍA, J.; DUARTE, E.; ALBUQUERQUE, E.; MELO, L. e NEVES, D.- *Avaliação dos repertórios brasileiros em Agricultura, Ciência da informação e Direito: A qualidade da análise de conteúdo*, en *Ciência da Informação*, (1998), 4: 44.
(9) MOREIRO GONZÁLEZ, J. A.- *Aplicación de las Ciencias del textoal resumen documental*. Madrid: Universidad Carlos III – Boletín Oficial del Estado.
(10) MOREIRO GONZÁLEZ, J. A.-*La técnica del resumen científico*, en LÓPEZ YEPES, J.- *Manual de Información y Documentación*. Madrid: Pirámide, 1996: 383.
(11) MOREIRO GONZÁLEZ, J. A.- *Aplicación de las Ciencias del texto al resumen documental*. Madrid: Universidad Carlos III – Boletín Oficial del Estado.
(12) TENOPIR, C. y JACSÓ, P.- *Quality of abstracts*, en *OnlineReview*,*(1993),* 5: 46.
(13) Ibid. Id.: 50.
(14) PÉREZ ÁLVAREZ.OSORIO, J.- *Introducción a la información y Documentación científica*. Madrid: Alhambra, 1988.
(15) LANCASTER, F. W. – *Indización y resúmenes: teoría y práctica*. Buenos Aires, EB publicaciones, 1996: 54.
(16) ABAD GARCÍA, F.- *Investigación evaluativa en Documentación.* Valencia: Universitat de Valencia, 1997: 130.
(17) ELLIS, D.- *New horizons in information retrieval*. London: Library Association, 1991.

BIBLIOGRAFÍA

ABAD GARCÍA, F. Investigación evaluativa en documentación. Valencia : Universitat de Valencia, 1997. p. 130-132.
AENOR (Madrid). UNE 50-121-91: métodos para el análisis de documentos, determinación de sua contenido y selección de los términos de indización. Madrid, 1991.
_____. ISO 8402:1995: gestión de la calidad y aseguramiento de la calidad del vocabulario. Madrid, 1995.
AGUADO BLANCO, Maria B. Técnicas avanzadas de indización y clasificación de información. Madrid : Aguado, 1997.
BAKER, S. The measurement and evaluation of library services. Arlington: Information Resources, 1993.
BLAIR, D. Language and representation in information retrieval. Amsterdam: Elsevier, 1990.
CHAUMIER, J. Análisis y lenguajes documentales. Barcelona : Mitre, 1986.
DENIS, S. et al. Liability in the provision of information services. Brussels: EUSIDIC, 1990. (EUSIDIC Research Project, 1989).
ELLIS, D. The effectiveness of information retrieval systems: the need for improved explanatory frameworks. Social Science Information Studies, n. 4, p. 261-272, 1984.
_____. New horizons in information retrieval. London : Library Association, 1991.
EISENBERG, M. Measuring relevance judgments. Information Processing and Management, n. 24, p. 373-389, 1988.
FOLSTER, M. B. A study of the use of information sources by social science researchers, Journal of the Academic Librarianship, n. 1, p. 7-11, 1989.

GRIFFITHS, J.; KING, D. A manual on the evaluation of information centers and services. Neuilly-sur-Seine : North Atlantic Treaty
Organization, 1991.
INFORMATION MARKET OBSERVATORY. The quality of electronic Information products and services. Luxembourg, 1995. (Working Paper, 95/4).
LABOIRE, T.; HALPEIN, M.; WHITE, H. Library and information science abstracting and indexing services: coverage, overlap and context. Library and Information Science Abstracts, n. 7, p. 183-195, 1985.
LANCASTER, F. W. El control de vocabulario en la recuperación de la información. Valencia : Universitat de Valencia, 1996.
_____. Indización y resúmenes: teoría y práctica. Buenos Aires: EB Publicaciones, 1996.
MICHEL, J.; SUTTER, E. Valeur et compétitivité de l'information documentaire. Paris : ADBS, 1991.
MOLINA, M. Indicadores de calidad descriptiva en la gestión de los procesos analítico-documentales. In: ACTAS DE LAS JORNADAS ESPAÑOLAS DE DOCUMENTACIÓN, 1994, Gijón. Oviedo: Universidad, 1994. p. 184-204.
MOREIRO GONZÁLEZ, J. A. Aplicación de las ciencias del texto al resumen documental. Madrid : Universidad Carlos III, [1996?]. Boletín Oficial del Estado.
_____. La técnica del resumen científico. In: LÓPEZ YEPES, J. Manual de información y documentación. Madrid : Pirámide, 1996. p. 373-390.
_____. et al. Avaliação dos repertórios brasileiros em agricultura, ciência da informação e direito: a qualidade da análise de conteúdo. Ciência da Informação, Brasília, v. 27, n. 3, p. 284-292, set./dez. 1998.
O'CONNOR, B. Explorations in indexing and abstracting: pointing, virtue and power. Englewood : Libraries, 1996.
PÉREZ ÁLVAREZ, J. Osorio. Introducción a la informaciíon y documentaciíon científica. Madrid : Alhambra, 1988.
ROLLING, L. Indexing consistency, quality and efficiency. Information Processing and Management, v. 17, p. 69-76, 1981.
SUTTER, E. Maîtriser l'information pour garantir la qualité. Paris: AFNOR, 1992.
SWINDELLS, N. Managing the quality of information products. Managing Information, n. 4, p. 36-48, 1995.
TENOPIR, C.; JACSÓ, P. Quality of abstracts. Online Review, n. 5, p. 44-55, 1993.
UNESCO (Paris). Principes directeurs pour l'évaluation des systèmes et services d'information. Paris : UNESCO, 1978.

ELEMENTOS DE LINGÜÍSTICA EN SISTEMAS DE INFORMACIÓN Y DOCUMENTACIÓN

Antonio Luis García Gutiérrez
Universidad de Sevilla (España)

INTRODUCCIÓN

Las bibliotecas, los archivos, los servicios de documentación y las redes de información son instituciones que tienen ya una larga trayectoria práctica, de varios siglos a varios decenios. La tecnología ha impuesto cambios y hábitos que, potenciados por la globalización, han permitido la transformación de pequeños ficheros manuales en potentes memorias de datos consultables remotamente. Sin embargo, la verdadera revolución de los sistemas de información no proviene esencialmente del factor tecnológico, a pesar de la importancia del mismo. En ellos, la materia prima, o el producto transportado y almacenado, es la información misma y su manipulació no ha acusado una transformación semejante a la operada en los soportes. A tenor de las observaciones en algunas "redes de información", a mi entender apenas redes telemáticas, podemos concluir que el cambio debe provenir de la aproximación documentológica, más concretamente, de la lingüística documental (LD). (1)

A cualquier usuario de Internet, antes beneficiario privilegiado de bases de datos "on line" suministradas por multinacionales situadas en los países occidentales más industrializados, no se le escapa que, si bien se abre ante sus ojos un prometedor escaparate de posibilidades de obtención de información, ora la promesa es efectivamente sólo un escaparate sin trastienda, ora surgen conflictos lógico-semánticos y sintácticos en la localización de datos pertinentes, ora el nivel de ruido se dispara en relación a la demanda planteada. Muchos (los creyentes de la panacea tecnológica) atribuyen los problemas a la corta edad de la tele-informática (para otros "cortedad") aunque los errores en la era del ordenador multimedia en lo que se refiere a recuperación de información y satisfacción del usuario son idénticos a los conocidos hace treinta años. Nuevamente estamos ante una ausencia de aprovechamiento de los recursos de la lingüística aplicados a la documentación.

Nos hallamos, en consecuencia, ante el reto histórico de acompañar el necesario desarrollo tecnológico arbitrado por ingenieros y tecnólogos con modelos, aparato conceptual y metodologías aportadas por humanistas y científicos sociales. En este contexto, debemos resaltar la importancia de la consolidación y expansión de la lingüística documental como disciplina que enraíza sus fundamentos en los postulados de las ciencias del lenguaje, semánticos y gramaticales, esencialmente, en los cruces habidos con campos afines como análisis del discurso, análisis de contenido y, en general, en las denominadas, y en construcción, ciencias cognitivas.

En efecto, el objetivo de la ciencia de la información/documentación es la sistematización de principios de operación sobre el conocimiento con la finalidad pragmática de organizarlo, representarlo y ponerlo al alcance de la mayor cantidad posible de usuarios (entroncando ahí con la tecnología, si bien me refiero al interfaz amigable del lenguaje de representación y búsqueda y no al entorno informático). Pues bien, parece que los vientos apuntan en otra dirección: la aparente familiaridad de los software, el trabajoso desmenuzamiento del sistema

intuitivo llevado ya a la drástica reducción icónica (todo ello muy beneficioso en el estricto campo de la relación con la computadora) crean la falacia de un, igualmente, acceso familiar e intuitivo a la información contenida en los ordenadores. Se confunde, por tanto, una vez más, el soporte con el contenido y se suscita, entre los usuarios menos avezados, la falsa realidad de que recuperar información es trivial porque, en todo caso, se recupera mucha información (incluso más de la solicitada).

El problema de la recuperación de información en Internet ha sido parcialmente evidenciado por las mismas máquinas que han pretendido su solución. Los llamados "robots" o motores de búsqueda como Yahoo, Altavista, Olé, Webcrawler, etc. que han pasado del rastreo sistemático a convertirse en una puerta a la que las ofertas web deben llamar, dirigen la más simple pregunta hacia masas incoherentes de información incluso con hiperbotones que cacarean una "advanced research" y que no es más que el viejo operador booleano usado desde los albores de la computación. Sobre este tema, particularmente, es urgente realizar una investigación que formule con precisión el problema y apunte soluciones. En ese caso, también, el marco teórico sería fundamentalmente lingüístico.

MODELOS Y PRINCIPIOS

Los investigadores que, a principios de los ochenta, decidimos romper con el paradigma tecnicista imperante en la bibliografía norteamericana y europea, vinculado a la estadística, la empresa y la evaluación de sistemas, forma de entender la información derivada del modelo conservacionista originario y evolucionando hacia lo que algunos denominan hoy el paradigma digital, en definitiva el triunfo del soporte sobre el contenido, nos agrupamos (sin conciencia grupal desde luego) en torno a los análisis semánticos y sus aplicaciones a la documentación. Esta perspectiva fue aprovechada tanto por quienes veníamos de los problemas de la información como por lógicos y lingüistas interesados en los mismos, buscando una salida aplicada a sus conocimientos. Inicialmente, como es obvio, se produjeron desencuentros: excesivo celo en los postulados, falta de visión global de los procesos, extrañamiento respecto a los fines pragmáticos de la documentología… Como puede deducirse, posiciones propias de la desconexión de investigadores entre ellos y respecto al objeto, de un lado, actitud rupturista con el statu quo oficial de las investigaciones y crisis normal en el nacimiento de un nuevo enfoque.

El arqueólogo francés Jean Claude Gardin llevaba muchos años dando pistas sobre el camino a seguir en obras publicadas en los sesenta, por lo que se le puede considerar verdadero precursor de la nueva aproximación, especialmente, en su trabajo Les analyses de discours (2) de 1974 parcialmente publicado como

artículo en inglés: Document analysis and Linguistic Theory (3). Esta línea le lleva a materializar sus postulados en aplicaciones en el área de conocimiento en la se halla especializado su grupo de investigación y, como consecuencia, publica conjuntamente con sus colaboradores varios libros en los ochenta que consolidan pragmática y magistralmente la aproximación que vengo exponiendo: Systemes experts et Sciences humaines (4), La Logique du Plausible (5) y, ya en los noventa, Le calcul et la raison (6), entre otros. Esta obra viene a confirmar la concepción de una epistemología práctica como sinónimo de documentología y su imbricación, desde la teoría lingüística, en el nuevo paradigma cultural/cognitivo que impregna a los que nos reconocemos en esta corriente de pensamiento científico.

De la apropiación del modelo estructuralista del signo lingüístico por la documentación, podemos decir que surgen las nuevas tendencias que observan los procesos documentales, formulan los problemas y proponen procedimientos de manera distinta a la oficialista si bien ensamblando, en esta nueva forma de pensar la información, los métodos, las normas y los autores clásicos de nuestra disciplina, bien con consideraciones muy críticas, bien reconduciendo y aprovechando ciertos bagajes. Como ilustración de lo expuesto baste mencionar las aportaciones esenciales de autores como Ranganathan y Vickery en cuanto a la clasificación del conocimiento o Salton y Ellis en recuperación, y en el lado opuesto el anquilosamiento de la norma ISO 2788 sobre elaboración de thesaurus o la ausencia alarmante de técnicas basadas en reglas para la objetivación del análisis del contenido de los documentos.

A pesar de la limitada bibliografía sobre la concepción que propugno, escasos pero fundamentales libros, tesis doctorales o contribuciones en revistas han hecho consistente la idea de que los problemas derivados de la obtención de información en los nuevos sistemas de información son problemas de lenguaje y, por tanto, la solución a los mismos proviene de las disciplinas que se ocupan tradicionalmente de estos, por emplear una expresión, las semánticas y gramáticas aplicadas a la gestión de la información y, por poner una etiqueta, la lingüística documental.

De esta forma, la lingüística documental (LD)se ocupa de ordenar los procedimientos de captación de los mensajes (lectura), de las transformaciones resultantes de la actividad anterior y de la organización y estructuración de dispositivos de representación a fin de que la obtención de conocimiento se dé eficaz y satisfactoriamente. Para ello, la disciplina introduce elementos de actuación (reglas procedimentales) y mecanismos de explicitación de los raciocinios, condición indispensable para que los procedimientos adquieran fiabilidad, sean verificables y, en consecuencia, científicos. La LD se distancia de las normativas, que persiguen el mismo fin normalizador, al conferir credibilidad científica a sus propuestas enmarcándolas en la lógica del proyecto investigador.

Así, tanto la elaboración de una técnica de análisis documental deberá ser montada sobre corpus rigurosamente verificados y validados experimentalmente, como la construcción de un thesaurus debe realizarse lanzando hipótesis metodológicas y epistemológicas, describiendo las variables consideradas en el vocabulario, en la estructura o predeterminables en el uso. Con ello, las

herramientas de organización y representación citadas constituyen artefactos científicamente consistentes y sus aplicaciones o actualizaciones no quedan a merced de la arbitrariedad ni dependen de la casualidad, la memoria, la intuición, la experiencia o la lógica personales.
Veamos los principios básicos de trabajo que debemos considerar al desarrollar un procedimiento o lenguaje en un sistema de información:

Aplicación

La documentación implica especialización o aplicación temática, tanto a nivel de investigación como de práctica. El documentólogo debe manejar un referente temático como marco real en el que se confirman o rechazan sus conjeturas y, del mismo modo, el documentalista trabaja sobre contenidos acotables. Excepcionalmente, el documentalista de prensa es un generalista, en el sentido de abarcar un área enciclopédica pero el discurso periodístico, siendo peculiar, responde a estructuras de producción sistemáticas y cualquier producto periodístico es reconocible como tal por lo que, en su caso, la aplicación viene determinada por la compensación de la intensión o profundidad en favor de la extensión temática.
En consecuencia, todo documentalista, incluido el de prensa, debe ser formado en el discurso sobre el que pretende trabajar así como instrumentado con los métodos y teorías que le ayudan a entender las claves y elementos propios del mismo. La documentación general debe formarse como constructo teórico a partir de las teorías parciales obtenidas y aplicadas sobre discursos especializados.

Experimentación

El método prioritario de observación y descripción en investigación documentológica es el experimental.
En cuanto al método empírico, se justifica por la necesidad de encontrar soluciones a universos de datos que precisan una canalización bien sea a través de mecanismos de selección, bien de análisis o de representación. Puesto que el problema habitual del documentalista es metodológico con relación a datos manipulables, puede construir "observables" artificiales mediante el muestreo y la simulación. En todo caso, las extrapolaciones generales son inviables y el grado de aprovechamiento de un discurso a otro estará sometido a un escrupuloso aumento de las muestras, de tal forma que el nuevo universo asuma los procedimientos extrapolados con todas las garantías.
Puesto que la experimentación vincula excesivamente un método a un corpus, las modelizaciones de más alto nivel resultan impracticables o insuficientes. Así, es de poca utilidad modelizar los principios de selección o los usuarios del discurso químico y establecer extrapolaciones hacia el discurso sociológico y, de éste, al discurso periodístico. Dentro de este último, los elementos de un método empírico de lectura, por ejemplo, deben ser modificados según la variable de género o extensión. En ese sentido, el reconocimiento artificial de estructuras sintácticas del discurso doxológico de la prensa se asemeja más al aplicado sobre el discurso de la argumentación y del saber científico, a pesar de la brevedad de un editorial por ejemplo, que a otros géneros de su propio ámbito: noticias, entrevistas o reportajes.

No solamente la necesidad de trabajar en corpus reales o simulados nos impone el método empírico. El obligado marco tecnológico sin el cual los procesos documentales modernos no son posibles (transmisión de millones de datos desde / hacia millones de usuarios potenciales) marca, también, el método a seguir. De hecho, la mayoría de los procedimientos de índole metodológica o reglada como lectura, síntesis o representación convergen necesariamente en una tecnologías que los hace viables o inviables.

Pragmatismo

El objetivo de la documentación es de orden pragmático, es decir, todos los esfuerzos se dirigen a la obtención de un producto. Este hecho, que afecta a las vías de construcción teórica y a la misma epistemología documentológica, obedece al carácter históricamente práctico y manual de las actividades documentales: organización, ordenación, dosificación, representación, difusión, recopilación, son palabras claves del universo del documentalista y, por tanto, también de la perspectiva de su investigación.

La documentación es disciplina instrumental o auxiliar de otras ciencias o discursos, lo mismo que la terminología o la normalización. Así, la documentación hace suyo el objetivo de organizar y divulgar los conocimientos en otros campos y, en consecuencia, la instrumentalidad determina nuevamente su carácter pragmático.

En nombre del pragmatismo, pues, se investiga la documentación, se buscan y recortan contribuciones de otras disciplinas, se edifica una superestructura epistemológica a prueba del sismo constante que provoca la praxis, hasta tal punto, que la misma naturaleza de esa estructura central se basa en una renovada transformación.

Validación

El método experimental sobre muestras hace necesaria una metodología de validación de resultados. La validación debe producirse sobre corpus en los que cualquier elemento del universo discursivo en cuestión haya tenido la misma oportunidad de participar. La evaluación de métodos documentales ha conocido un gran desarrollo en su vertiente tecnológica, merced al interés de las multinacionales por el rendimiento de los módulos de consultas de las bases de datos. Sin embargo, estos procedimientos verifican el sistema y sus prestaciones y no las relaciones que mantiene el discurso matriz con su representación documental y el rol que desempeñan productores, mediadores y usuarios en el proceso.

En consecuencia, es necesaria la investigación de métodos de validación documentológica y la elaboración de los mismos para los discursos específicos en el trabajo científico a la vez que, en el docente, el estudiante de documentación aplicada debe conocer metodologías evaluadoras para rectificar procedimientos profesionales y estar en condiciones de modificar las conductas en la adquisición de materiales y en la actualización de fondos.

DOCUMENTACIÓN Y COMUNICACIÓN SOCIAL

Además de sus vitales relaciones con la lingüística, también la documentación mantiene conexiones con otras disciplinas del mismo ámbito, que enumeraré más adelante, además de su entronque con las ciencias de la comunicación de las cuales se declara partícipe. Nuestra disciplina es, para la mayoría de tratadistas y en la mayoría de las lenguas científicamente relevantes, la ciencia de la información (Information Science). En consecuencia, las referencias que sugiere el epígrafe imponen una matización al pluralizar: las denominadas en España ciencias de la información (más conocidas como ciencias de la comunicación, al menos, en Europa y América), o conjunto de disciplinas que tienen por objeto la descripción y la extracción de los postulados y leyes que rigen los procesos comunicativos promovidos por los mass media, su evolución, causas y efectos. La ciencia de la información, o documentación, tiene por objeto el establecimiento de metodologías y la explicación de los procesos de comunicación en los que interviene la información documental (información registrada reutilizable) y es un área de conocimiento de las ciencias de la comunicación. Hoy, las redes telemáticas dan a la documentación el carácter de mass media que poco tiempo atrás se reservaba a la prensa o la TV.

Un sistema documental comprende unos modos y unos medios de tratamiento y circulación de la información contenida en documentos. El objetivo esencial del sistema es informar sobre contenidos localizables en documentos de cualquier tipología. Tal vez, las diferencias más notables entre los especialistas de cualquier materia y sus documentalistas, serían resumibles en dos puntos:

1) la condición de permanencia del soporte como elemento indispensable para la selección de información (en consecuencia, no son documentación: hechos, observaciones, reflexiones, deducciones, discursos, gritar, interpelar, dialogar, etc., fuente y discurso, salvando las distancias, propios de periodistas, historiadores, científicos, juristas, etc.).

2) el aprovechamiento derivado del proceso que realiza el documentalista. A diferencia de los especialistas que leen y observan para su propio conocimiento y producción, el documentalista es un delegado informativo que lee para otros, en sentido análogo al periodista: la captación de la información cobra sentido si hay inmediata difusión y recepción.

La documentación se ocupa del proceso de un discurso fragmentado en unidades físicas (soportes) y no tiene, por tanto, constancia directa de los hechos ni de la realidad. El soporte permite la manipulación de datos para su proceso y es, en consecuencia, un anclaje del conocimiento pero el documentalista abandona su suerte a la credibilidad y fiabilidad de la fuente.

Puesto que el proceso documental no tiene sentido sino es para culminar un ciclo comunicativo (al igual que el periodístico) dotando al usuario de información sobre fuentes que han sido intervenidas en diversos momentos por distintos agentes (políticos, agencias, redactores, analistas, clasificadores) se genera una mediación y el canal transmisor impone una codificación y decodificación de señal, tanto en el sentido semántico como técnico o telemático. Véanse las analogías del proceso documental y el periodístico en una sociedad moderna.

Es más, el proceso documental es un tipo de proceso comunicativo en el que los documentalistas son los emisores (persuasores en palabras de Lozano) (7), el mensaje es el producto que genera (resúmenes, datos factuales, índices...) o discurso documental, el código es el lenguaje de representación (semántico) y la

señal del módem (telecomunicación), el medio es la infraestructura telemática y los receptores son los usuarios (interpretadores), habitualmente especializados.
Así, la documentación se inscribe en los modelos generales de la comunicación a la vez que produce sus propias teorías parciales. Documentación es un modo informativo que materializa sus productos a través de medios convencionales: libros, revistas, ordenadores, discos ópticos, páginas web, auxiliado por un necesario marketing que dé a conocer la oferta de información que proporciona un centro documental.
Shannon y Weaver publican en 1949 su modelo matemático (8) fijando el concepto de entropía sobre la suma de información requerida en una situación dada para eliminar la incertidumbre. Se aplica inicialmente sobre los procesos de transmisión electrónica, aspecto que incumbe esencialmente a ingenieros y tecnólogos más interesados en la capacidad de transmisión del canal que en la información transmitida.
Para De Bonville, "a partir de las teorías de Shannon y Weaver cristalizan los modelos comunicativos aportando un cuadro conceptual en el que sería reducido el conjunto de fenómenos de la comunicación humana" (9). El investigador canadiense examina el paradigma haciendo extrapolaciones hacia la documentación y describe sus cinco componentes: fuente que produce el mensaje, transmisor que adapta el mensaje de la fuente haciéndolo compatible con el canal, canal que transporta la señal, receptor que interpreta el mensaje mediante la captación de formas transformadas en datos para ofrecerlo al usuario, y usuario a quien se destina el mensaje. Concordamos, con De Bonville, en que este modelo fundamentado en los procesos de telecomunicación es perfectamente ajustable a las necesidades documentológicas por lo que nuestra disciplina se halla inserta en la epistemología comunicativa.
A pesar de la extrapolación elemental, el proceso documental genera sus propios instrumentos y métodos creando una idiosincrasia que, para los detractores de este modelo, proporciona en la documentación otras vinculaciones (con la ciencia de la ciencia y la epistemología). Zunde, por ejemplo, no critica la extrapolación, pero señala que la documentación debe tener un mejor conocimiento de sus propias leyes y de los fenómenos medibles antes de aceptar el modelo: "el objeto de estudio de la ciencia de la información son fenómenos empíricos asociados con procesos de información tales como la generación, transmisión, transformación, condensación, almacenamiento y recuperación. El objetivo último consiste en alcanzar mejor comprensión sobre la naturaleza de la información. Comenzando como hacen todas las disciplinas empíricas -con una descripción de los fenómenos en el dominio de su interés- "la ciencia de la información pretende establecer principios generales a través de los cuales puedan explicarse fenómenos observados" (10).
Con esta afirmación, Zunde reduce en gran medida el carácter auxiliar y aplicado de la documentación poniendo como objeto de estudio los procesos generados por la misma. Esto es razonable, como ocurre con el caso de las ciencias publicitarias, ya que el producto que genera el proceso documental, el discurso documental, ha sido poco observado y analizado por los teóricos de la documentación. Este discurso se compone de elementos o constructos elaborados para la comunicación y de él, nos interesan las condiciones de producción (actitudes, posición y limitaciones de los agentes productores), las estrategias y estructuras transmitidas, las interacciones y los efectos del propio discurso

documental y, relacionado con el discurso "natural" que representa, las distorsiones, reducciones y simplificaciones, que lleva a cabo y los mecanismos (modos y medios) que emplea para su realización. Esto es un campo de investigación específicamente documentológico.
Ahora bien, no podemos olvidar el fin prioritario que tiene asignada la documentación en el conjunto de las ciencias instrumentales: generar procesos de organización y circulación de todo tipo de conocimientos. El mensaje debe ser codificado y decodificado en su recorrido, de tal forma, que puede producirse ruido en la recepción. El concepto de ruido es uno de los esenciales incorporados por la documentación en el control de su proceso (11).
Junto a su carácter de disciplina instrumental para el desarrollo científico pensamos, con De Bonville, que la función social del documentalista no se limita a crear y organizar memorias sino que, fundamentalmente, tiende a dar a conocer, a poner en circulación esos fondos sobre los que se pueden establecer consideraciones de índole cognitiva (en cuanto que son motores de nuevo conocimiento, no sólo soportes de conocimiento) y social (en cuanto el proceso alcanza cotas de difusión pública, restringida a públicos especializados).
La dependencia documentológica respecto a los medios de comunicación ya fue enunciada por Otlet en 1934: documento es, para quien consolidó la documentación como disciplina académica, un sistema de signos sobre un soporte que se elabora con vistas a su transmisión (12). El documento nace en sociedad y a ella va destinado lo que indica el carácter social de la documentación, y en una dimensión inferior, Otlet reflexiona sobre la dependencia funcional de tres factores: lectores, libros y autores, lo que introduce un aspecto psicológico y psicosociológico (psicología bibliológica) que debe estudiar esta dependencia entre "perceptores, agentes y medios (tiempo y espacio)" (13).
La relatividad de los procesos documentales la establece Otlet sobre la pragmática receptiva, del mismo modo que Wittgenstein desde una concepción funcionalista sobre el uso del significado (14), aproximación en la que coinciden Eco, Foucault, Sartre y muchos otros pensadores y lingüistas (15). Para Otlet, "el libro no existe más que en función del lector, es decir, lo que no percibe el lector no existe para él, por tanto, su contenido desde la perspectiva de la recepción no es más que la expresión de las facultades del lector" (16). Algo que ya dijera Platón en *El banquete* muchos siglos antes.
Este carácter individualista del documento, como entidad social, es uno de los elementos esenciales del proceso documental, puesto que su objetivación, en teoría, desvincula el contenido de los usuarios. Sin embargo, documentación implica "socialización" y, en consecuencia, máxima objetivación de los procedimientos a fin de atender a mayores audiencias. En este sentido, las metodologías documentales se rigen por el principio del pragmatismo, se vinculan con los fines sociales del proyecto y se inscriben en los modelos generales que se ocupan del proceso de la comunicación en sociedad.

DOCUMENTACIÓN Y MASS MEDIA

La presencia de medios o instrumentos usados para la difusión masiva en el ámbito documental es lo que dota a esta disciplina de su dimensión massmediática. Es la transmisión y uso social (Mijailov) lo que convierte definitivamente a la documentación en disciplina social y la ajusta a los postulados de las teorías comunicativas. El individuo ya no necesita buscar información

documental porque ésta le sale al paso en carteles, periódicos, teletexto, vídeo o microordenadores ligados a Internet. Esta mutación de la posición del usuario de la documentación, de activo a pasivo, le convierte en foco de consumo de datos y, en consecuencia, en el objeto de gran parte de los estudios de aproximación psicosociológica y comunicológica imperantes hace veinte años en otras disciplinas. En este sentido, y salvo algunas peculiaridades propias, la documentología no tiene que inventar nuevos métodos de observación sino, por el momento, extrapolar y adaptar los experimentados por los científicos sociales, en general, y los de la comunicación, en particular, sobre sus distintos intereses. No cabe duda, que el hecho que culmina la transformación de la documentación tradicional (práctica antiquísima) en disciplina moderna del ámbito de la comunicación de masas, es la nueva tecnología de conservación y transmisión de conocimiento y el nuevo campo de posibilidades, y también de nuevos problemas, que comporta.

La documentación adopta dos modalidades de expresión social:

1) a través de los medios considerados masivos como prensa, radio, TV en los que se halla mezclada con datos informativos. Incluyamos también en esta categoría el libro y las revistas especializadas y científicas de cierta circulación. La información documental obtenida por un usuario, básicamente pasivo, presenta altas cotas de elaboración (y en consecuencia de mediación).

2) a través de medios de difusión individualizada, ante los que el usuario adopta un rol aparentemente activo, como videotex, redes telemáticas, CD-ROM, es decir, medios que permiten la interacción y que presentan los datos en un estado falaz de materia prima pero, no por ello, menos mediados. Los perfiles, la difusión selectiva de información -DSI- y la revista electrónica son ejemplos de servicios documentales a la carta, si bien Internet esta revolucionando en los últimos años las distintas concepciones y actuaciones documentológicas.

Ambos sistema de difusión documental tienen espacios públicos reservados y, lejos de hacerse competencia, se complementan y refuerzan incluso entre los más homogéneos: revistas y prensa, televisión y videotex, creando distintos espacios de consumo de información para los nuevos media. En la heterogeneidad de medios documentales, observamos como en el caso publicitario, el refuerzo que generan mutuamente: las redes referenciales apoyan al sector librero o de revistas, la documentación de congresos y reuniones realiza un marketing de determinadas publicaciones...

Así, la información documental es utilizada por muchos medios como un sistema publicitario más, además de servir como producto con valor comercial propio. No es de extrañar la existencia de pleitos millonarios sobre la propiedad y los derechos de autor en documentación (ej. Le Monde contra una sociedad canadiense que resumió y vendió los resúmenes de textos del diario en los ochenta) o el proteccionismo y advertencias legales de algunos medios sobre las transformaciones y venta de sus contenidos transformados, algo que no se observaba veinte años atrás.

DOCUMENTACIÓN Y ANÁLISIS DEL DISCURSO

El documentalista es un lector de textos, realiza una lectura dirigida (o que debe estar dirigida) por unas reglas específicas para la obtención de un resultado: la esencia del discurso o macroproposición global del productor. Por lo tanto, el objeto de su lectura no es el nivel de palabra o frase, sino el de texto o discurso.

Este cambio de orientación disciplinar para el análisis documental ha supuesto una revolución en las investigaciones y en las prácticas y ha estado motivado por la incapacidad de la máquina de entender sentidos a partir de palabras fuera de contexto.
A pesar de la dificultad de obtener significados globales de forma mecánica es posible la convivencia de mecanismos reductores aplicados por el ser humano y el reconocimiento automático de los productos obtenidos hasta conseguir una liberalización de la lectura simulada, toda vez que exista una extraordinaria memoria empírica que dicte al motor de inferencia los comportamientos a seguir en función de elementos y construcciones memorizadas miles de veces.
Este problema a resolver en los próximos años no es, sin embargo, prioritario puesto que existe una necesidad previa que paso a describir: el documentalista no es capaz de extraer las mismas conclusiones de un texto que otro colega que se aplica a la lectura del mismo texto. Incluso advertimos disparidad en los resultados obtenidos por el mismo agente lector a partir de un mismo texto en épocas distintas. Esta afirmación es constatable en cualquier centro de documentación.
Si bien el sentido común y la experiencia contribuyen a la construcción de reglas virtuales e intuitivas que el documentalista y sus colegas aplican mecánicamente, los resultados siguen presentando importantes divergencias además de otros problemas: las reglas empíricas se adaptan a textos específicos con mecanismos difíciles de explicitar lo que imposibilita la adaptación de un nuevo lector o equipo al esquema de trabajo y dificulta el acercamiento del usuario al sistema, desconocedor de los modos de segmentación textual.
El análisis del discurso proporciona un importante instrumental a la documentación para la resolución de algunos de estos problemas. Puesto que la misma disciplina está impregnada de un rico cruce interdisciplinar en el que intervienen teorías que explican los procesos mentales de la interpretación de la realidad (desde el cognitivismo), las estrategias de producción de textos y los contextos comunicativos y socioculturales en los que se desarrolla el discurso (teoría de la comunicación y pragmática discursiva), conecta con el mismo corpus epistemológico que constituye la documentación.
Al explicarnos, en consecuencia, cómo se produce y usa el texto junto a las condiciones y contextos involucrados a la vez que nos facilita herramientas para la detección de las estrategias discursivas que ocultan o refuerzan determinados elementos a la vez que afloran las proposiciones del autor de entre cientos de lexias y estructuras gramaticales de superficie, el análisis del discurso presta un auxilio de máxima importancia a la documentación, lo que ya ha sido demostrado en varias investigaciones teórico-prácticas (17).

DOCUMENTACIÓN Y ANÁLISIS DE CONTENIDO

El análisis de contenido (AC) aporta a la documentación una larga experiencia en descripciones pragmáticas (18), muchas de ellas de vinculación social, de universos cargados de significados, los cuales, debidamente depurados y sometidos a referentes construidos (tablas de indicadores) permiten hacer inferencias y extrapolaciones sustentadas en sólidos métodos de validación procedentes de la socioestadística.
El análisis documental es una metodología de lectura o captación de elementos a partir de textos (descripción y universo pragmáticos) que pretende la representación de los mismos en lenguajes controlados (tablas de descriptores)

para posibilitar la recuperación ulterior. Vemos que la mayor divergencia acontece en los objetivos: inferir (AC) y recuperar (AD).
En cualquier caso, ambas disciplinas coinciden durante un largo trayecto común y poco explotado, de ricos y posibles intercambios: si el AC nos enseña cómo fabricar muestras, unidades operativas y métodos de observación y verificación, el AD le ofrece métodos de construcción, organización y ordenación de bases de datos, registros y campos, normalización semántica del vocabulario y amplias conexiones con otras disciplinas recortadas por el análisis documental y aprovechables para el AC.
Pero, a pesar de las aportaciones referidas, tal vez la más importante para la documentación, por la ausencia en sus investigaciones, sea la práctica constante de la validación en los trabajos sometidos al análisis de contenido (en sentido documental). En efecto, tanto en la construcción de métodos de lectura como de representación documentales, se trabaja sobre muestras (texto, vocabulario) compuestas por unidades menores. Según observamos en la bibliografía documentológica, las metodologías de creación de muestras y de distribución aleatoria que garanticen los resultados, así como los procedimientos de validación final que hagan fiables las conclusiones y pronostiquen posibles extrapolaciones, brillan por su ausencia. Por lo tanto, y dada la cercanía de ambas áreas, es necesario recortar la experiencia validadora del AC en aras de la consolidación que comportaría para la investigación documental.

LÓGICA Y DOCUMENTACIÓN

La documentación tiene una necesidad imperiosa, en su engarce con la tecnología, de formalización de elementos y enunciados tanto en el nivel de entrada como en el de salida y proceso de datos en un sistema. En determinadas operaciones de laboratorio, el investigador no está interesado en el significado real de los términos sino en su verosimilitud dentro del corpus que utiliza para la simulación. En este caso, la lógica proposicional o enunciativa es útil en cuanto que ha alcanzado grandes cotas de formalización en la representación de sentencias declarativas.
La declaración supone una reducción de la estructura sintáctica natural pero, como dice Allwood, "hay categorías morfosintácticas que no tienen contrapartida lógica" (19). La lógica predicativa, por ejemplo, no tiene en cuenta los enunciados imperativos o las interrogaciones a pesar de que teorías como la "hipótesis performativa" defiende que bajo estas formas subyace una afirmación en sus estructuras profundas y, por tanto, son objeto de análisis lógico. Por el momento, en las aplicaciones lógicas adoptadas por tecnólogos y, probablemente, a la espera de la confirmación de nuevos logros (especialmente de la Fuzzy Logic o lógica difusa), la enunciación tópica de los sistemas expertos es declarativa (al menos en aquellos sistemas que ofrecen garantías).
Los investigadores de la llamada "semántica lógica" (Lewis, por ejemplo) trabajan para aplicar el análisis lógico a la lengua natural. Ese es el mayor punto de confluencia de la terna documentación - lógica - tecnología. La formalización pasa por la reducción, a inventarios controlados, de todas las equivalencias de cualquier categoría léxica posible. Claro está que la lengua natural, en un campo especializado, ofrece una morfosintaxis, distinta a la de la lengua coloquial, facilitando su simbolización.

La lógica construye lenguajes formales para evitar la vaguedad, la ambigüedad y la dependencia del contexto haciéndolos exactos y unívocos. Cualquier constructor de lenguaje documental sabe que esos mismos son los objetivos que deben cumplir los vocabularios, en consecuencia, la lógica formal contribuye específicamente a la elaboración de lenguajes desambiguados.
En cuanto a la creación de prototipos inteligentes para la gestión documental hemos de recordar que la lógica estudia las reglas de inducción y deducción de elementos no necesariamente reales pero, a pesar del desinterés del lógico por la realidad psicosemántica, extraemos un importante aparato de inferencias posibles y extrapolables a enunciados reales, a fin de constituir en la máquina una base de reglas, es decir, un conjunto de procedimientos inferenciales humanos simulados.
El análisis que efectúa el documentalista sobre los textos se rige por dos lógicas: la lógica general, en cuanto organiza los procesos de adquisición del conocimiento científico, la construcción de hipótesis, de las leyes y teorías y la lógica formal, en cuanto nos informa de cómo están montados los razonamientos desde el punto de vista formal. Nótese que me refiero a discursos científicos cuya estructura responde, desde la primera concepción y por sus objetivos, a un alto grado de formalización. En ese sentido, la epistemología científica debe ser parte de la formación de los documentalistas puesto que les ayuda a comprender el discurso "logicista" (en palabras de Gardin) de la ciencia y a "mapear" las construcciones específicas del conocimiento.
Finalmente, la lógica matemática y, concretamente, la teoría de conjuntos y las aplicaciones del álgebra de De Boole ha sido de gran utilidad en los sistemas de recuperación de las bases de datos convencionales sobre conocimiento científico básico y experimental aunque de poca eficacia sobre los discursos humanos y sociales expresados en lengua natural y con sintaxis de cierta complejidad. La reducción de los operadores lógico-matemáticos en la combinatoria de búsqueda de datos es uno de los mayores problemas que debe resolver la documentación a partir de la superación de los mismos por símbolos formalizados que expresen todos los sentidos de los enunciados naturales. Este problema se ha potenciado al masificarse los datos y las demandas, los sistemas y los analistas en redes telemáticas.

El caso del thesaurus de Patrimonio Histórico andaluz: una aplicación de la teoría lingüística a los sistemas de información

Las ventajas de la concepción científica de los procedimientos documentales defendida aquí no radican exclusivamente en el más que beneficioso fin de la objetivación que nos lleva a la posibilidad de programación y a la indispensable fiabilidad. Además, pueden darse infinidad de circunstancias positivas derivadas de la conversión de la raigambre científica de unas técnicas consideradas, en su aplicación además de en sus fundamentos, meramente profesionales.
Por ejemplo, lo expuesto puede ser ilustrado por el proyecto de construcción del thesaurus andaluz de Patrimonio Histórico, encargado a quien suscribe por el Instituto Andaluz de Patrimonio Histórico de la Consejería de Cultura, organismo de la comunidad autónoma de Andalucía (20).
En ese proyecto, no solamente han debido superarse las reducciones recogidas por la normativa internacional sobre construcción de thesaurus dado el calado y la extensión del objeto, la complejidad y multiplicidad de actores involucrados, sino

que, además, la consistencia de las metodologías ya arbitradas por la LD ha hecho posible salir de un atasco corporativista a historiadores del arte, arqueólogos, arquitectos, antropólogos, conservadores y restauradores merced a la redistribución del mapa conceptual que nos propone la gramática de casos (o teoría de los casos universales) aplicada a la elaboración de lenguajes documentales. En efecto, el mismo thesaurus se construye a partir de unas conjeturas previas que afectan a su contenido y a su estructuración y que devienen hipótesis en un momento dado, junto a un inventario de variables que se erige como sistema de anclaje del constructo en elaboración a una realidad determinada, concebida como observable (aún en el nivel de confección teórica de la herramienta) para más adelante ser desarrollado sujeto a un método, la gramática de casos conceptuales enunciada por Fillmore y Pottier y recortada por Cunha, a su vez reconducido y modificado en constantes contrastaciones con los corpus iniciales.

Tras varios meses de cotejo de los casos de la gramática (nivel macroestructural) con la base léxica (nivel microestructural), se procede a realizar una división del trabajo por encima de las especialidades de los diez componentes del grupo de trabajo (especialistas en las distintas disciplinas que confluyen en el patrimonio histórico, tanto en su vertiente de investigación -universidad- como de conserva y explotación -museos-). La estrategia metodológica urdida por la gramática de casos nos ha permitido no solamente proceder dentro de un marco científico y por lo tanto fiable, si bien sujeto al condicionamiento de la variables, y en ese sentido hemos obtenido resultados convencionales pero no arbitrarios, sino además dejar de lado los desencuentros de las disciplinas mencionadas incapaces de articularse por sí mismas en un todo global pragmático (agruparse en un mismo foro) superado por la realidad: ya existían bases de datos interdisciplinares que urgían la sistematización de un vocabulario común y la normalización de las formas de análisis y acceso a la información.

En la concepción auténticamente patrimonialista de los bienes históricos y de interés cultural, la LD a través de uno de sus dispositivos, la metodología de casos conceptuales ha sido de crucial importancia para dar al traste con disputas sectaristas que afectan la recuperación global de información y el interés de los usuarios de las bases de conocimiento sobre patrimonio histórico y hacer posible un lenguaje integrador de todas las disciplinas concernidas.

REFLEXIÓN FINAL

Hemos visto, a lo largo de esta exposición, algunas frases y términos, que fuera de contexto nos harían parecer que no hablamos de documentación: macroestructura, microestructura, semántica, gramática, anclaje, base léxica, representación y todos los que de ellos dependen no citados en una comunicación con las pretensiones de ésta: sema, enunciado, archisemema, lexia, infraconceptos, estructura lógico- semántica, eje paradigmático y sintomático, acepción, lexicografía, terminología, proposiciones lógicas, etc. Todos ellos, combinados con el vocabulario más tradicionalmente documental: descripción bibliográfica, análisis y lenguajes documentales, búsqueda y recuperación, usuario, demanda, etc. constituyen el mapa conceptual de la lingüística documental, o genéricamente análisis documental para mis colegas del Departamento de Biblioteconomía de la Universidad de Sao Paulo. Ahora bien,

muchas teorías lingüísticas abandonadas o en desuso o plenamente vigentes pueden ser recortadas y aplicadas a nuestros fines siempre que haya indicios de utilidad: desde la documentalmente sobre- explotada semántica hasta las inexploradas sintaxis, lexicología y, de interés más reciente para los documentalistas por las máquinas captadoras y emisoras de fonemas que se nos avecinan, fonología.

Desde los postulados clásicos hasta los generativistas, y las derivaciones como la lógica semántica, el análisis del discurso, la semántica estructural o la semiótica textual por citar algunos campos en los que se han realizado incursiones o los intuimos prometedores, se constituyen los límites de esta vasta disciplina, en simbiosis con las que se ocupan de cómo se construye el raciocinio, su representación (ciencias cognitivas) y sus procesos de transferencia masiva mediante artefactos mecánicos (informática y telecomunicaciones). Sin la presencia sintética y simultánea de todo ese marco multidisciplinar en la mente del investigador de la documentación, generador de procedimientos e instrumentos útiles y pragmáticos, la organización y el acceso ordenado en los depósitos de conocimiento actuales nunca alcanzará mayores niveles de fiabilidad y satisfacción que en épocas pasadas.

NOTAS Y REFERENCIAS

Texto de la comunicación presentada en el V Simposio internacional sobre Comunicación social celebrado en Santiago de Cuba, del 21 a 25 de enero de 1997. Actualizado en junio de 1998.

Publicado en: Revista LATINA de Comunicación Social La Laguna (Tenerife) - julio de 1998 - número 7 D.L.: TF - 135 - 98 / ISSN: 1138 – 5820

http://www.lazarillo.com/latina [Junio de 1998]

(1) Argumentos sostenidos en mis trabajos: Lingüística documental. Aplicación a la comunicación social. – Barcelona: Mitre, 1984. – 279 p.; Estructura lingüística de la documentación: teoría y método. – Murcia: Universidad de Murcia, 1990. – 166 p.; Análisis documental del discurso periodístico. – Madrid: CTD, 1992. – 160 p. y Procedimientos de análisis documental automático: estudio de caso. – Sevilla: Instituto Andaluz de Patrimonio Histórico, 1996. – 88 p.

(2) Obra que no ha perdido vigencia, vid. Gardin, J.C.: Les analyses de discours. – Neuchâtel: Delachaux et Niestlé, 1974. – 178 p.

(3) Véase ídem: Document Analysis and Linguistic Theory. – In: Journal of Documentation, v.29, 2, 1973. – p. 137-168 y Document Analysis and Information Retrieval. --In: Bol. Unesco bibliotecas, v.16, 1, 1960. – p.2-5

(4) Idem: Systèmes experts et Sciences humaines. – Paris: Eyrolles, 1987. – 269 p.

(5) Idem: La Logique du plausible. Essais d'Epistémologie pratique. – 2éme ed. – Paris: Maison des Sciences de l'Homme, 1987. – 330 p.

(6) Le calcul et la raison:essais sur la formalisation du discours savant. – París: Ecole des Hautes Etudes en Sciences Sociales, 1991. – 293 p.

(7) Lozano; Jorge: El discurso histórico. – Madrid: Alianza Editorial, 1987. – 223 p.

(8) Vid. la clásica obra de Shannon, C y Weaver, W.: Teoría matemática de la comunicación. – Madrid: Forja, 1981. – 159 p.

(9) El investigador quebequés de la Universidad Laval Jean de Bonville sienta, en su artículo, las bases de la adopción del modelo en documentación: Application du

Paradigme du Shannon à la Bibliothéconomie et à la Documentation. – In: Revue canadienne des Sciences de l'Information". – v.3, mai 1978. – p.13-27
(10) Zunde, P.: Information Theory and Information Science. – In: Information Processing and Management, 17, 6, 1981. – p.341
(11) García Gutiérrez, A.: Lingüística documental… op. cit.
(12) Consúltese la obra imprescindible de Otlet, Paul: Traité de Documentation. – Bruxelles: Mundaneum, 1934. -- p. 426a
(13) Ibid., p. 34b
(14) Wittgenstein apud Geckeler, Horst: Semántica estructural y teoría del campo léxico. – Madrid: Gredos, 1984. – 389 p.
(15) Todos ellos se refieren, en sus respectivos ámbitos, a la participación del receptor en la construcción del significado, principio elemental de la documentación.
(16) También Otlet destaca la figura del usuario como pieza clave en op. cit. p.33b
(17) Por ejemplo, en la investigación realizada por mi colega de la Universidad de Sao Paulo Regina Obata: Contribuçao da Análise do Discurso para à análise documentária: o caso da documentaçao jornalistica. – Sao Paulo: Escola de Comunicaçoes e Artes de la USP, 1991. – 87 p. y anexos.
(18) Véanse, al respecto, los trabajos de Bardin, L.: Análisis de contenido. – Madrid: Akal, 1986. –183 p. o Krippendorf, Klaus: Metodología de análisis de contenido: teoría y práctica. – Barcelona: Paidós, 1990. – 279 p.
(19) Véase el excelente recorte teórico conceptual que realizan en su obra: Allwood, Jens; Lars, Gunnar y Dahl, Osten: Lógica para lingüistas. – Madrid: Paraninfo, 1981. – p. 183
(20) Lenguaje construido desde distintas disciplinas para servir como herramienta de análisis de objetos (ánforas, indumentaria, armas, medallas, monedas o cualquier artefacto móvil), inmuebles (plantas de edificios, fachadas, cubiertas, cercas, motivos ornamentales, túmulos, etc.) imágenes (audiovisuales y fotografías de objetos o realidades de interés patrimonial), textos (bibliografía sobre patrimonio) con el fin de ofrecer un instrumento central de referencia para los bienes históricos de Andalucía. Vid. Thesaurus de Patrimonio histórico andaluz. –Sevilla: IAPH, Consejería de Cultura, 1998 y su evaluación, junto a nuevas propuestas metodológicas en García Gutiérrez, A.: Principios de lenguaje epistemográfico: la representación del conocimiento sobre patrimonio histórico andaluz. – Sevilla: IAPH, 1998.

LENGUAJES DOCUMENTALES E INFORMACION DE ACTUALIDAD

Antonio Luis García Gutiérrez
Universidad de Sevilla (España)

CONCEPTO Y FUNCIONES DEL LENGUAJE DOCUMENTAL (1)

Entendemos por lenguaje documental -ld- un constructo artificial de elementos léxicos y reglas realizado con el fin de normalizar y facilitar la entrada y salida de datos en un sistema de información. Tales elementos suelen ser palabras de la lengua natural o códigos ligados y formalizados mediante reglas morfológicas y combinatorias. Gracias a esas convenciones, los analistas y usuarios se comunican en tiempo diferido con mayor precisión y eficacia. Desde otro punto de vista, un ld es un instrumento de representación del conocimiento con el objetivo de su organización, conservación y recuperación. Representar es una de las funciones generales de estos lenguajes, es decir, soportar y transportar, mediante un conjunto finito y controlado de signos, ideas, imágenes o sonidos (a su vez representaciones cognitivas) que pertenecen a planos más abstractos, amplios y libres como las acciones o la creatividad.

Controlar o normalizar sería una segunda función básica de los ld. En un depósito de conocimiento sin control léxico la búsqueda de datos queda abandonada al azar y a la adivinanza puesto que los criterios empleados por documentalistas y usuarios, acaso introducidos en diferente lugar, tiempo, cultura e idioma, deben coincidir mediante etiquetas formales o formas de representación documental. Además, los lenguajes de acceso a la información deben disponer de recursos que inspiren y sugieran, a los usuarios, itinerarios de búsqueda y localización dentro del sistema. Sin coincidencia no hay posibilidad de comunicar, tercera y definitiva función general de los ld; y si el sistema no es capaz de establecer una comunicación satisfactoria con el usuario ¿Para qué sirve?

Todo ld cuenta con un conjunto de signos que representan los conceptos cubiertos y con una estructura, o mapa conceptual e intersígnico, responsable de la distribución de esos elementos de acuerdo a un método y a unos objetivos predeterminados. En las primeras clasificaciones, cuya consideración se remonta a la época protodocumental y, ya más sistemáticamente, a partir de finales del siglo XIX, los elementos léxicos eran códigos numéricos, alfanuméricos o alfabéticos si bien la tendencia, de los últimos cincuenta años, se derivó hacia el uso de la lengua natural. Por otro lado, las estructuras pasaron, de la arborescencia o férrea jerarquización de los conceptos, a esquemas más flexibles que aparecían parejos al proceso de naturalización del vocabulario, facilitando su manejo. Se observa, por tanto, una "familiarización" del interfaz de estos lenguajes en consonancia, sin duda, con los nuevos tiempos y modos introducidos por la automatización y la telecomunicación.

Las posibilidades del acceso remoto, desde el conocido "online" de las pasadas décadas al web actual, han contribuido, también, a pensar, desde posiciones

documentológicas, en ld más naturalizados y amigables, tanto en vocabulario como en gramáticas, y más intuitivos de manera que el usuario pueda ejercer de propio documentalista. Ahora bien, en tanto no existan robots de búsqueda especializada en la red, que incorporen la filosofía de los ld y las formas y estructuras de los discursos sobre los que informan, los niveles de ruido (información no deseada) y silencio (información no recuperada) se verán incrementados, provocando la deserción de los usuarios.

Los ld tradicionales, por otro lado, cualquiera que sea su modalidad de elaboración, son estructuras lógico-semánticas, artificios simbólicos y teóricos llenos de convenciones en los que el usuario reconoce, vagamente, los enunciados del discurso que le interesa. Están presentes términos que comprende, se le informa de las reglas de combinación pero, pese a todo, persiste un extrañamiento de sus estructuras cognitivas y discursivas respecto a las teóricas y convencionales del lenguaje. En otras palabras, es complicado, por no decir imposible, recoger satisfactoriamente en un lenguaje documental que utiliza términos descontextualizados, y ubicados en contextos adoptivos, todos los comportamientos y roles de los conceptos en textos reales. Por tanto, la naturalización es sólo un recurso de carácter interno del lenguaje ya que en absoluto supone reflejar la combinatoria real de los conceptos de un texto en algún lugar del mismo lenguaje.

La verdadera transformación de los ld, en herramientas tanto teóricas como discursivas, provendría de la conjunción empírica de los miles de puntos de encuentro que hay entre los textos adscritos a un sector del conocimiento: una estructuración hipertextual controlada y sometida a la normalización de los términos y de las gramáticas de análisis y búsqueda. No quiere esto decir, sin embargo, que los lenguajes documentales sean inoperantes en su configuración actual sino tan sólo que estamos ante el desafío y en los albores de una nueva concepción y aplicación de los ld ligados, indisolublemente, a las prácticas discursivas y a los avances tecnológicos.

LOS LENGUAJES DOCUMENTALES EN EL DISCURSO PERIODÍSTICO: TIPOLOGÍA

Sin duda, una de las áreas del conocimiento más complicadas y polémicas en cuanto al control lingüístico-documental es la información de actualidad. Sus características, como el enciclopedismo, la superficialidad, la dispersión o el acelerado ritmo de envejecimiento de los datos, provocan que unos consideren inviable o, en el mejor de los casos, nada rentable crear un ld propio en tanto que los ajenos nunca satisfacen los intereses institucionales. Sin embargo, pocos sectores del conocimiento necesitan, como la información de actualidad, de tan alto grado de normalización y estructuración para poder convertirla en un producto manejable y reutilizable.

Desde los inicios de la documentación periodística en las "morgues" de los diarios norteamericanos (2) se hizo necesario ordenar los fondos del archivo y, privados de ordenadores, los documentalistas de prensa organizaban pacientemente las noticias recortando o fotocopiando los textos tantas veces como asuntos relevantes contenían para, a continuación, ubicar las, copias en sus respectivas carpetas y estanterías. El crecimiento del número y grosor de las carpetas, de los metros de estantes y del espacio adjudicado al archivo pronto obligó a los responsables del medio, bien a ampliar el cubicaje del servicio de documentación

y del archivo, bien a condenarlo al estancamiento y a adquirir obras de referencia y recursos ajenos.
La racionalidad que introducía la clasificación de materias en la ordenación de los recortes hizo que los medios de mayor prestigio conservaran sus fondos mínimamente analizados lo que comenzó a tener cierto tipo de rentabilidad: un periódico documentado era un medio más riguroso y fiable para los lectores. A esa filosofía se sumaron, no hace tantos años, el New York Times (también modelo de base de datos y vocabulario controlado en los setenta), Le Monde en Francia, cuya documentación curiosamente se vio potenciada por la sociedad de redactores que se hizo con el control del diario (y actualmente en la edición web de Le Monde Diplomatique encontramos nuevas formas documentales) y El País, en España, desde sus inicios (1976).
Los avances informáticos y de telecomunicaciones de los noventa irrumpen, no obstante, en las redacciones con un halo de panacea que ha perjudicado seriamente el pensamiento documentológico y, seguramente por ello, el desarrollo sostenible de los fondos documentales. En efecto, puesto que la tecnología aparentemente soluciona el problema de la búsqueda de información, ofreciendo al usuario una gran cantidad de la misma, sin entrar en su calidad o pertinencia, la documentación ha sufrido un bloqueo debido al desvío o recortes de presupuestos para investigación y desarrollo en beneficio de los proyectos tecnológicos. Además, los propios usuarios perciben que la nueva tecnología no ofrece lo que esperaban y que los problemas de la recuperación persisten cuando no se han incrementado.
Por otro lado, los lenguajes documentales convencionales -clasificaciones y tesauros- no han evolucionado desde hace medio siglo, hecho que alimenta el rechazo de este tipo de herramienta. En consecuencia, es patente que la tecnología no soluciona todos los problemas de la documentación, particularmente los de contenido y lenguaje, por lo que es necesario crear procedimientos y productos para el control del acervo que acompañen los avances de las máquinas y sus softwares.
La extensión enciclopédica de la actualidad crea dificultades, a la hora de compilar y estructurar el vocabulario, que se compensan gracias a la superficialidad del tratamiento. Si es cierto que la información y el vocabulario periodísticos presentan índices de obsolescencia acelerada por lo que cualquier lenguaje documental necesitaría un equipo humano que realizara actualizaciones constantes, lo que significa una asignación presupuestaria (3) que no todos se pueden permitir. Ahora bien, incluso los costes de gestión del lenguaje, de cuya existencia depende en gran parte la eficacia del sistema y, por tanto, la satisfacción de sus usuarios, son incomparablemente menores que los gastos ocasionados por el software, el equipo informática, su mantenimiento y permanente renovación. Y si cualquier medio económicamente solvente considera, hoy día, como gastos ordinarios ineludibles la actualización del parque de ordenadores ¿Cuál es el motivo de excluir la herramienta léxica de la que depende el acceso a la información?
Existe una variada tipología de lenguajes documentales relacionados con el discurso periodístico a partir de distintas perspectivas como la estructura, la composición o el uso, entre las más notables, que podemos resumir en los siguientes bloques:

- según el nivel de tratamiento:

- lenguajes libres: no se aplican convenciones ni reglas específicas sobre la terminología utilizada en la base de datos.
- lenguajes controlados: el lenguaje se organiza y utiliza mediante condiciones y pautas convencionales que pueden afectar a los significados, a los significantes y a la combinatoria.

- según el tipo de grafía de los elementos léxicos:

- lenguajes naturales: los elementos del vocabulario adoptan formas de expresión de la lengua natural si bien puede aparecer controlado el alcance de los significados: capital (dinero).
- lenguajes codificados: los elementos del vocabulario son códigos compuestos por signos numéricos o alfabéticos que, convencionalmente, representan conceptos o materias: 300 Ciencias sociales.

- según la extensión temática:

- lenguajes enciclopédicos: abarcan todas las áreas del conocimiento o son altamente intertemáticos.
- lenguajes especializados: centrados en un sector del conocimiento o en cruces interdisciplinares y acotados.

- según el momento de la composición:

- lenguajes precoordinados: se basan en composiciones terminológicas realizadas a priori a fin de prever y cubrir cualquier enunciado. Estos sistemas tenían gran aplicación antes de la irrupción del ordenador en la escena documental.
- lenguajes poscoordinados: se sustentan en la economía sígnica de los elementos del vocabulario y en el manejo, a posteriori, de gramáticas de búsqueda incrementándose, por tanto, el riesgo de distorsión.

- según el tipo de estructura:

- lenguajes jerárquicos: el lenguaje (generalmente precoordinado) adopta forma de árbol a pirámide, haciendo de la dependencia su principal filosofía. Las relaciones se manifiestan mediante criterios hiper e hiponímicos (clase/especie) y partitivos (todo/parte) como primordiales rasgos de la estructura. Los elementos léxicos, artificiales o naturales, son forzados a ocupar un solo espacio en el árbol determinando su localización, su significado y sus combinaciones a un modo de entender la realidad por lo que el código ideológico impregna el lenguaje.
- lenguajes asociativos: el lenguaje (generalmente poscoordinado) se estructura horizontalmente de forma que desaparece, total o parcialmente, el criterio de dependencia o adscripción terminológica. Relaciones del tipo materia prima/producto final, técnica/instrumento o agente/técnica asumen la lógica de la estructura aproximándola a formas discursivas más reales.

Las tipologías precedentes no son puras ni excluyentes. Así, en los lenguajes precoordinados hay posibilidad de una mínima poscoordinación y en los poscoordinados abundan los descriptores con apariencia de enunciado; cualquier lenguaje jerárquico permite asociaciones y los asociativos, en la práctica, se basan en una jerarquía previa, camuflada en mayor o menor grado; la codificación también es independiente del tipo de estructura o composición aun estando, tradicionalmente, más ligada a la jerarquización y al enciclopedismo. En la era de la Documentación moderna, desde 1895, se han elaborado centenares de lenguajes documentales que mezclan los tipos citados.

La Clasificación Decimal Universal -CDU- es el principal exponente occidental de lenguaje jerárquico (4). Fue compilada por los belgas Paul Otlet y Henri La Fontaine a finales del XIX, a partir de la Clasificación del norteamericano Melvin

Dewey, sobre una estructura arborescente (desglosando el conocimiento en diez jerarquías) con vocación enciclopédica (todas las áreas científicas), de carácter codificado (números y divisiones decimales) y precoordinado aunque ofreciendo ciertos recursos para realizar combinaciones. Si la CDU tuvo y sigue teniendo una gran aceptación, en el mundo bibliotecológico, para el control bibliográfico superficial del ámbito científico, la extrapolación de su filosofía a la organización documental del discurso periodístico sería un error ya que el enciclopedismo aparece como único rasgo común y tan sólo en el nivel extensional. De hecho el enciclopedismo que interesa al mass media queda marcado por intereses e ideología institucionales de los que la CDU carece a pesar de ser un producto del pensamiento positivista.

Las restantes características de la actualidad eliminan la posibilidad de adoptar esquemas encorsetados, codificados, y de imposible puesta al día. Los tesauros se sitúan en el otro extremo de la tipología. Surgieron en la segunda mitad del siglo XX como reacción contra la escasa flexibilidad de las clasificaciones, sustituyendo lo universal por la especialización, lo codificado por la naturalización y lo jerárquico por asociaciones, si bien todo ello tan solo en teoría. En la realidad, el vocabulario compuesto por significantes naturales da una falsa y peligrosa sensación de acercamiento a las estructuras mentales de los usuarios porque, efectivamente, la mayoría de los tesauros publicados se construyen a partir de un árbol, la relación significante/significado es biunívoca, no admitiéndose el contexto como anunciador de sentidos y abundan las precoordinaciones al viejo estilo clasificatorio y, como las clasificaciones, más parecen instrumentos para la síntesis que para el análisis.

A partir de los años sesenta proliferan centenares de tesauros en el mundo, en todas las lenguas, sobre los sectores más imprevisibles por iniciativa de organismos públicos y privados, locales o internacionales y aparecen tesauros a la carta, en la instituciones, y de vocación intercultural (UNESCO, OCDE, OTT) (5).

En 1974 se edita la norma internacional IS 2788 que regula la construcción de Tesauros monolingües, influida por e influyendo en las normas nacionales (BSI, AFNOR, AENOR). Con ello, la Organización Internacional de Normalización –ISO- sentenciaba una manera de concebir los tesauros, ratificada en la 2ª edición de la norma en 1986 (6), que aun perdura a pesar de las perspectivas abiertas por las tecnologías y los nuevos productos que solicitan los usuarios de la información, elementos suficientes para justificar una urgente actualización.

De todo lo anterior podemos extraer dos conclusiones:

1. la necesidad de controlar y estructurar los enunciados y significados del discurso periodístico como única vía de optimización de las tecnologías y de sus productos.
2. la necesidad de investigación permanente en el área de los lenguajes documentales, y concretamente en sus aplicaciones al discurso periodístico, parcela que ha sufrido la incidencia de imposibles extrapolaciones, a fin de incorporar y explotar los recursos de análisis y recuperación ofrecidos por la tecnología en constante renovación.

Y entre los puntos de reflexión sobre los lenguajes documentales de los mass media, destacamos:

- El problema de la representación en los documentos fotográficos y audiovisuales. La incorporación de categorías léxicas no sustantivas (como adjetivos, que

expresan cualidad, gerundios, que indican movimiento y participios, que refieren estado o situación).
- La interacción entre la maeroestructura lógico-semántica y teórica del lenguaje con la estructura discursiva y real de textos y usuarios.
- Conexión y traslación de la estructura de los lenguajes documentales a las hojas de trabajo de las bases de datos.
- Las conexiones entre lenguajes documentales y lenguajes y tecnología hipermedia: el lenguaje documental como hipertexto.

CONCEPTOS INSTRUMENTALES

Para elaborar Id, es necesario un aparato conceptual y terminológico procedente de disciplinas próximas a los análisis semánticos que enriquece el propiamente documentológico (vid tipologías de descriptores y de Id): Análisis del Discurso, Análisis de Contenido, Lógica proposicional, Teoría de la Comunicación y las diversas ramas de la Lingüística, entre otras, que configuran el nuevo metalenguaje de la Documentación. Veamos las expresiones más relevantes en lo que se refiere al área de los lenguajes documentales (7):
- macroestructura: esquema temático global de un Id que recoge, organizadas, las grandes etiquetas o denominaciones de los campos conceptuales.
- microestructura: estructura mínima establecida en un lenguaje mediante la interacción de dos conceptos.
- campo conceptual: conjunto jerarquizado de conceptos que atiende a un rasgo distintivo común (archisemema) definido convencionalmente en extensión y profundidad. Lo conceptual desborda lo semántica puesto que, en el Id, las relaciones exceden lo lingüístico constantemente (pueden ser empleados criterios ideológicos, culturales, sociológicos, antropológicos o, simplemente, consuetudinarios además de los léxicos). Los Id tradicionales se estructuran a partir de la jerarquía de campos conceptuales.
- clase conceptual: conjunto horizontal formado entre los conceptos que mantienen relaciones asociativas (no jerárquicas). Un lenguaje documental puede ser construido a partir de una estructura de clases conceptuales asociadas.
- jerarquización: estructura de la dependencia inmediata entre los conceptos. Puede ser ascendente, de lo particular a lo general, o descendente, de lo general a lo particular.
- coordinación: estructura horizontal inmediata establecida entre elementos del mismo campo conceptual (o con la misma notación).
- asociación: estructura horizontal inmediata establecida entre elementos de distinto campo conceptual (o con distinta notación). Las asociaciones, del mismo nivel de profundidad forman clases conceptuales.
- profundidad de campo: grado numérico de profundidad en que se encuentra un campo respecto al nivel de superficie de la macroestructura.
- profundidad de término: grado numérico de profundidad en que se encuentra un término respecto al nivel de superficie del campo conceptual.
- notación: código de identificación atribuido a un campo y a todos los términos incluidos en el mismo.
- macrocategoría: macroetiqueta organizativa abstracta que da nombre a los campos conceptuales de primer nivel.
- categoría: etiqueta organizativa abstracta que da nombre a los campos conceptuales de cualquier nivel de profundidad.

- macrodescriptor: etiqueta organizativa concreta que da nombre a los campos conceptuales de cualquier nivel y es utilizable en el análisis y la búsqueda.
- concepto: asunto o ente nítidamente recortado y representado por un descriptor.
- descriptor: significante autorizado por un lenguaje que representa un sólo concepto en el sistema semántico.

Puesto que, entre los objetivos de este libro, figura la presentación de técnicas e instrumentos utilizados y familiares en la gestión de la documentación periodística, nos limitaremos en el epígrafe siguiente, a las fases de desarrollo del tesauro, herramienta más extendida actualmente en las bases de datos, aunque apuntando las modificaciones que acentúan su rendimiento y empleando el necesario metalenguaje antes descrito.

FASES DE CONSTRUCCIÓN DE UN TESAURO PERIODÍSTICO

A continuación se exponen las fases y actuaciones que han de llevarse a cabo para la elaboración de un tesauro aplicado a la Documentación periodística (8):

1. Determinación de la cobertura: acotación de la profundidad y de la extensión del área de conocimiento o, en su caso, del cruce interdisciplinar. Cálculo aproximado del número de términos previstos para realizar el cronograma y planificar el método y el desarrollo, crear el equipo de trabajo y decidir la infraestructura informática necesaria. (9)
2. Elaboración del método:
    a) de compilación:
        - procedimiento analítico o "a posteriori": basado en fuentes secundarias (ya procesadas documentalmente) y aproximación inductiva. De los términos reales, hallados en los índices de las bases de datos y en las bibliografías, emerge una macroestructura empírica muy próxima a la realidad pero exenta de previsión y capacidad heurística.
        - procedimiento global o "a priori": cimentado en una base especulativa o hipotético-deductiva, sobre la que se construye un esquema teórico que ha de ser desarrollado con terminología de cualquier procedencia, existan o no experiencias reales de trabajo documental. Las etiquetas surgen como hipótesis que deben ser ratificadas por el vocabulario y, más tarde, consolidadas o modificadas por la práctica.
        - procedimiento mixto: consiste en partir de una mínima macroestructura teórica cotejada con términos procedentes de la actividad documental y modificar aquélla en función del comportamiento y las necesidades observados en los elementos del vocabulario.

    b) de estructuración de los términos:
        - precoordinación: los términos se combinan "a priori" (en el lenguaje documental) a fin de cubrir cualquier concepto o demanda. Este procedimiento de uniones morfológicas facilita la organización terminológica piramidal pero provoca un aumento excesivo de vocabulario y deja escaso margen de expresión al usuario.

- poscoordinación: cada concepto tiende a ser representado por un unitérmino excepto en los casos de ambigüedad en que se emplearán sintagmas nominales. Este procedimiento reduce considerablemente el número de descriptores compuestos (ya que las composiciones las realiza el usuario en la búsqueda) si bien crece el número de falsas combinaciones y de distorsiones debido a la polisemia. El vocabulario suele resistirse a las formas canónicas de organización por lo que es preciso construir esquemas más flexibles. En cualquier caso, la poscoordinación (restando la precoordinación como un recurso para evitar la polisemia) es el único procedimiento razonable y acorde con los sistemas automatizados.

c) de organización estructural:

- esquematización temática: se parte de una organización de primer nivel mediante etiquetas temáticas concretas o macrodescriptores (1. Derecho, 1.1 Derecho Civil, 1.2 Derecho Penal).

- esquematización categorial: se parte de macrocategorías o etiquetas de mayor abstracción (1. Modos, 1.1 Técnicas, 1.2 Procesos). Aunque la estructuración temática es más sencilla para documentalistas y usuarios ya que suele reproducir los organigramas institucionales o las nomenclaturas (más familiares pero hechos con otros fines), la terminología se deja organizar difícilmente por este procedimiento, especialmente en las áreas interdisciplinares, por lo que la categorización se ofrece como un método de mayor flexibilidad para absorber el vocabulario salvando, mediante instrucciones y notas, la extrañeza inicial que experimentará el usuario ante una clasificación desconocida y abstracta.

3. Recopilación de fuentes terminológicas, en función de las decisiones adoptadas en el punto 2: a) primarias: manuales, enciclopedias, revistas, tesis, monografías, actas de congresos, etc., procediéndose a la extracción y depuración documental de términos; b) secundarias: índices, bibliografías, resúmenes, perfiles, etc., que contienen términos filtrados por el análisis documental y e) terciarios: glosarios, tesauros, terminologías, clasificaciones, diccionarios y nomenclaturas, es decir, documentación que ofrece un vocabulario previamente estructurado con fines documentales u otros.
4. Compilación del corpus terminológico: se recogen los términos en función de los procedimientos y objetivos determinados distribuyendo la compilación, mediante criterios explícitos, entre los miembros del grupo de trabajo. Debe crearse un registro terminológico en el que se transcribirá cada uno de los términos junto a las modificaciones necesarias. El conjunto de registros, una vez depurado, conformará la base léxica o vocabulario del lenguaje.
5. Verificación terminológica. Sobre cada uno de los términos, han de comprobarse los siguientes extremos:

a) relación de biunivocidad entre significante y significado. En caso de polisemia o sinonimia se aplicarán los criterios descritos en 6.

b) pertinencia, o grado de adecuación temática del término, en relación con el contexto del lenguaje. En caso de no existir pertinencia, se provocará artificialmente, mediante sustitución o sintagma o se eliminará el término.

c) relevancia, o grado de autosignificación, y suficiencia expresiva del término en relación con el contexto. Se procede según los recursos del punto 6.1 sobre desambiguación.

d) alcance: estimación de la adecuación del nivel de profundidad del término (siendo la profundidad una variedad de pertinencia) respecto a los niveles de profundidad de campo previstos.

e) todos los términos deben cumplir las normativas de construcción generales del lenguaje documental elegido y las reglas de Normalización, establecidas por el grupo de trabajo, en cuanto a composición, género, número y cualesquiera otras convenciones oportunas.

6. Depuración y consolidación de la base léxica: como resultado de los controles anteriores, obtenemos un vocabulario más reducido que debe cubrir un alto porcentaje de la base léxica final a fin de poder trabajar sobre un corpus estable. Al mismo se accede mediante dos actuaciones sobre los accidentes habituales del vocabulario en lengua natural: polisemias y sinonimias.
   a) eliminación de polisemias o de los diversos sentidos de un término mediante:
      - calificador: /manzanas/ (bloques)
      - nota de alcance (NA): /manzanas/ NA fruta
      - sintagmatización: auto + fe = /auto de fe/
      - pluralización: /joyería/ como actividad y /joyerías/ como lugar.
      - uso de sinónimo: /auto/ USE /coche/

   b) control de sinonimia o de las diferentes entradas (significantes) posibles de un mismo concepto:

   - sinonimia lingüística o natural causada por:
      • variante ortográfica: cesio, cesium, caesium, caesio
      • transliteración: khomeini, Jomeini, Mao Tse Tung, Mao Ze Dong
      • préstamo: interviú, interview, entrevista
      • variante histórica o política: Hispanoamérica, Iberoamérica, América Latina
   - cuasisinonimia o sinonimia documental provocada por:
      • variante lexemática: documento, documentalista, documentación
      • variante clasemática o hiponímica: navaja, cuchillo, puñal, arma blanca
      • variante antonímica: estabilidad, inestabilidad

   Los operadores de sinonimia, en los tesauros, son:

   - USE: reenvía a la entrada autorizada o descriptor: Suráfrica USE Unión surafri- cana

- UP (usado por): da cuenta de los términos no autorizados o no descriptores: Unión surafricana UP África del Sur, República surafricana, Sudáfrica, Suráfrica.

El operador NA (nota de alcance o aclaratorio) de los tesauros tiene varias funciones:

- definir, pragmáticamente, el descriptor y atribuirle el valor o sentido que va a tener en el lenguaje: /documento/ NA unidad de información compuesta por un soporte material y un mensaje.
- focalizar o ampliar el sentido de un descriptor: /cine/ NA sólo español.
- realizar indicaciones de combinación o modificaciones efectuadas: /cine/ NA sólo español. Restricción incluida a partir de febrero de 1997. Puede combinarse con los identificadores de país, región o provincia.

7. Construcción y consolidación de la macroestructura empírica generada por el procedimiento analítico o, en su caso, ajuste de la macroestructura teórica, obtenida por el método global, ambos descritos en 2. a). Determinación de las rnacrocategorías y macrodescriptores de, al menos, hasta tercer nivel y codificación, mediante notaciones numéricas, alfanuméricas o alfabéticas, de todas las denominaciones de campo.
8. Formación del campo conceptual : adscripción de los términos depurados de la base léxica, obtenidos tras el procedimiento 6, a las etiquetas correspondientes de la macroestructura, mediante el principio de pertinencia, o inclusión inmediata en la categoría más próxima. Verificación de la correspondencia y adecuación recíprocas: Macroestructura Término. Ejecución de las expansiones, reducciones o modificaciones necesarias ya sean estructurales o terminológicas. Ajuste de los elementos organizativos y léxicos y creación del campo conceptual como conjunto autónomo de significación. Una vez formados los campos, el grupo de trabajo pasa, de operar en el nivel macro, al nivel microestructural.
9. Jerarquización: se procede, dentro de un mismo campo conceptual, a establecer en dos direcciones, el árbol de los términos que lo componen partiendo de la formación de la microestructura jerárquica (relación inversa e inmediata de dependencia entre dos elementos del campo conceptual):

   a) relación jerárquica ascendente mediante operadores TG (reenvían al término más amplio):

      - de generalidad TGE (término genérico de especie): /trenes/ TGE /vehículos de transporte/
      - de totalidad TGP (término genérico partitivo): /vagones/ TGP /trenes/

   b) relación jerárquica descendente mediante operadores TE (reenvían al término más concreto):

   - de especificidad o clase TEE (término específico de clase): /bicicletas/ TEE /bicicletas de carreras/, /bicicletas de montaña/
   - de parte o elemento constituyente TEP (término específico partitivo): /bicicletas/ TEP /manillar/, /sillín/.

10. Coordinaciones: mediante el operador TR (término relacionado) de los tesauros se incluyen las asociaciones, expuestas en el punto siguiente, y las coordinaciones o relaciones horizontales habidas entre los términos

pertenecientes al mismo campo conceptual siempre que en éste haya un número elevado de descriptores que impida una visualización rápida y sugerente. Ejemplo: /medios de comunicación social/

- TE diarios
  radio
  revistas
  televisión

Coordinación: diarios TR revistas (por ejemplo, porque ambos son medios impresos dentro del listado de específicos), y

11. Asociaciones: cuando la relación horizontal se establece entre distintos campos conceptuales, la microestructura creada se denomina asociación. Si la macroestructura se compone de categorías en el primer nivel, es posible establecer combinaciones teóricas entre las mismas, denominadas vectores (10), que ayudan a establecer la microestructura asociativa en todos los niveles inferiores.
    Macrocategorías:
    Agentes-Técnica
    Materia prima-Producto final
    albañil-enfoscado petróleo-gasolina
12. Finalmente, se redacta la introducción al lenguaje documental, manual de uso en el que se explican los recursos y fuentes manejadas, las claves, siglas y convenciones, las acotaciones y conexiones del universo temático cubierto, las formas de representación y el método y software utilizado.

Una vez construido, y antes de ser usado y editado, el lenguaje debe superar un test de coherencia estructural, de consistencia y suficiencia de la base léxica y ser probado sobre una muestra de documentos con el fin de adquirir un cierto rodaje sobre textos reales. En cualquier caso, a los pocos meses de manejo, debe sufrir un ajuste para, a partir de ese momento, ser actualizado en los plazos dictados por las necesidades de cada área de conocimiento.

PARTES DEL TESAURO

Un tesauro convencional puede constar de las siguientes partes o formas de presentación:

- clasificación general o macroestructura: es la tabla de materias del tesauro que ofrece, por tanto, las grandes etiquetas temáticas y categoriales con notación desarrolladas en las secciones siguientes alfabética o sistemáticamente. Esta parte permite una aproximación al vocabulario por focalización (de lo general a lo particular) indispensable cuando no se conocen los campos que abarca el tesauro o los términos que necesitamos. (Anexo 1)
- listado jerárquico: los términos aparecen mostrando su dependencia y sus dependientes en toda la cadena mediante sangrado u otras marcas físicas que indican el nivel de la jerarquía y la posición del descriptor respecto a los demás de su campo conceptual. (Anexo 2)
- listado alfasistemático: repertorio alfabético general o por campos de todos los términos con indicación de sus relaciones inmediatas de sinonimia, jerarquía, coordinación y asociación. Esta es la sección neurálgica del tesauro y a la que

debe llegarse para realizar la estrategia de análisis o búsqueda de los descriptores. (Anexo 3)
- índice permutado: repertorio alfabético general de los descriptores (y a ser posible también de los no descriptores con el operador de reenvío USE) sin relaciones semánticas y con la particularidad de que los descriptores compuestos se repiten tantas veces como términos llenos contienen. El índice permutado tiene la función de verificar o autorizar una forma concreta buscada por el usuario que no necesita sugerencias de la sección alfasistemática. (Anexo 4)
- índices auxiliares: repertorios de términos, habitualmente identificadores onomásticos o locativos, no incluidos en las clasificaciones temáticas regidas desde la microestructura. Pueden presentarse alfabetizados o estructurados. (Anexo 5)
- otras representaciones: esquemas flechados y terminogramas son formas de representación gráfica de los tesauros que ofrecen una especie de mapa o fotograma de los campos conceptuales. Estos productos pueden obtenerse de los software de construcción que disponen de tales formatos de salida. (Anexo 6)

NOTAS Y REFERENCIAS

(1) Tomado de: Introducción a la Documentación Informativa y periodística. Sevilla: Editorial Mad, S.L., 1999, capítulo XV, del cual el autor es editor.
(2) Véase, bien compendiada y estructurada, la historia y evolución de los servicios de documentación de prensa en el mundo en Galdón, Gabriel: Perfil histórico de la documentación en la prensa de información general (1845-1984). - Pamplona: Eunsa, 1986.- 167 p.
(3) Sobre equipo humano, calendario y costes en la construcción de tesauros vid Slype, George Van: Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales. - Madrid: Fundación Germán Sánchez Ruipérez, 1991. - p.105-108
(4) Sobre clasificaciones, consúltense los trabajos de San Segundo, Rosa: Sistemas de organización del conocimiento: la organización del conocimiento en las bibliotecas españolas. - Madrid: Universidad Carlos III de Madrid; BOE, 1996 . - 317 p. y de Gil Urdiciain, Blanca: Manual de lenguajes documentales. - Madrid: Noesis, 1996. - 269 p.
(5) Una tipología de tesauros en García Gutiérrez: Lingüística documental: aplicación en la comunicación social. - Barcelona: Mitre, 1984. - p. 177ss.
(6) ISO 2788: Directrices para el establecimiento y desarrollo de tesauros monolingües. - Ginebra: ISO, 1986. Publicada, en dos partes, en la Revista Española de Documentación Científica: parte 1, 12, 4 (1989), 463-483 y parte II, 13,1 (1990) p. 601-629.
(7) Aparato conceptual explicado y utilizado, con profusión, en García Gutiérrez, A.: Principios de lenguaje epistemográfico, la representación del conocimiento sobre Patrimonio histórico andaluz. – Granada: Junta de Andalucía; Comares, 1998.- 91 p.
(8) Sobre planificación de tesauros existe muy poco y disperso material. Véase el ya citado, trabajo de George Van Slype: Los lenguajes de indización ... op.cit. y, de manera más pragmática, la Introducción del Thesaurus del Patrimonio histórico andaluz. - Granada: Junta de Andalucía, Instituto andaluz del Patrimonio histórico; Comares, 1998.- p. 9-35.

(9) Ídem.
(10) Un cuadro de modos estructurales, representados como vectores, en: García Gutiérrez, A.: Estructura lingüística de la Documentación: teoría y método.- Murcia: Servicio de Publicaciones de la Universidad de Murcia, 1990. Capítulo 5.

BIBLIOGRAFÍA

- Aitchison, J. and Gilchrist, A.: Thesaurus Construction: a Practical Manual. - 21,d ed. - London: Aslib, 1987. - 173 p.

- Amaro, Regina K. Obata: Contribuçao da análise do discurso para á análise documentária: o caso da documentaçao jornalistica. - Sao Paulo: Ecal Usp, 1991. - 87 p., anexos (ined. tesis de maestría).
- Cintra, A. M. et al. Para entender as linguagens documentárias. - Sao Paulo: Polis- APB, 1994. - (Colegao Palavra Chave, 4). - 72 p.
- Cunha, 1. M. R. F.: Do mito a Análise documentária. - Sao Paulo: Edusp, 1990. - 163 p.
- Currás, E.: Thesauros: lenguajes terminológicos. - Madrid: Paraninfo, 1991. - 284 p.
- García Gutiérrez, A.: Lingüística documental: aplicación a la Comunicación social. - Barcelona: Mitre, 1984. - 279 p.

- Ídem: Estructura lingüística de la Documentación: teoría y método. - Murcia: Servicio de Publicaciones de la Universidad de Murcia, 1990. - 166 p.

- Ídem: Análisis documental del discurso periodístico. - Madrid: CTD, 1992. - 160 p.

- Ídem: Procedimientos de Análisis documental automático: estudio de caso. - Sevilla: Instituto andaluz de Patrimonio histórico, Consejería de Cultura, 1996. - 88 p.

- Ídem: Principios de lenguaje epistemográfico: la representación del conocimiento sobre Patrimonio histórico andaluz. - Granada: Junta de Andalucía; Comares, 1998.- 91 p.- (Cuadernos técnicos, 3).
- Gardin, J. C. et al: Systémes experts et Scíences humaines. - Paris: Eyrolles, 1987. - 269 p.
- Gil Urdiciain, B.: Manual de lenguajes documentales. - Madrid: Noesis, 1996.- 269 p.
- Greimas, A. J.: Semántica estructural. Investigación metodológica. - Madrid: Gredos, 1976.- 398 p.
- IS2788: Principes directeurs pour l'établissement et le developpement de thesaurus mo- nolingues. - Généve: Iso, 1974. - 111 + 14 p. (2a ed. 1986).
- ISO 704: Principles and methods of Terminology. - Généve: ISO, 1987. - ISIOS7: Principes de Terminologie. - Généve. ISO, 1990.

- ISO: Directrices para el establecimiento y desarrollo de tesauros monolingües. Norma internacional 2788-1986. Publicada en dos partes:

- p.1 En: Revista española de Documentación científica. 12, 4 (1989). - p.463-483.
- p.2 En: Revista española de Documentación científica. 13, 1 (1990). - p.601-629.
- Kobashi, N. Y.: Análise documentária: consideraçoes sobre um modelo lógico-

semántico. - In.-. Análise documentária: consideraçoes teóricas e experimentaçao Cunha 1. (org.). - Sao Paulo: Febab, p. 31-44.
- Maniez, Jaeques: Los lenguajes documentales y de clasificación: concepción, construc- ción y utilización en los sistemas documentales. - Madrid: Fundación Germán Sán- chez Ruipérez, 1992. - 231 p.
- Pécheux, M.: Hacia el análisis automático del discurso. - Madrid: Gredos, 1978. - 269 p.
- Pottier, B.. Lingüística general: teoría y descripción. - Madrid: Gredos, 1976. - 426 p.
-Ranganathan, S.R.: Prolegomena to Library Classificcation. –3rd ed. .- Bombay: Asian Publishing House, 1967.
- San Segundo, Rosa: Sistemas de organización del conocimiento: la organización del conocimiento en las bibliotecas españolas. - Madrid: Universidad Carlos III de Madrid; BOE, 1996. - 317 p.
-Smit, J: (org.) Análise documentária: a análise da síntese. –Brasília: Ibict, 1987. – 133 p.
- Tálamo, F. G: A definiçao semantica para a elaboraçao de glossários. – En: Análise documentária: a análise da síntese. – Smit, J. Brasília: Ibict, 1987. – 87-98 p.
- Tesauro de Patrimonio histórico andaluz/ A. García Gutiérrez (comp.). – Sevilla: Instituto andaluz de Patrimonio histórico, 1998. 1035 p.
- Van Dijk, T.: Texto y contexto: Semántica y pragmática del discurso. – Madrid: Cátedra, 1980. – 357 p.
- Van Dijk, T.: La noticia como discurso: comprensión, estructura y producción de la información. – Barcelona: Paidós, 1990. – 284 p.
- Slype, George Van: Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales. - Madrid: Fundación Germán Sánchez Ruipérez, 1991. – 198 p.

Anexos

MACROESTRUCTURA

1000000 . Acontecimientos. Actividades. Procesos. Técnicas *

1100000 .. Acontecimiento
1110000 ... Acontecimiento natural
1120000 ... Acontecimiento sobrenatural
1130000 ... Acontecimiento social

1200000 .. Actividad
1210000 ... Actividad doméstica
1220000 ... Actividad en organizaciones sociopolíticas *
1230000 ... Actividad festivo-ceremonial *
1231000 .... Rito de paso
1240000 ... Actividad lúdica
1241000 .... Actividad deportiva
1241100 ..... Deporte
1241110 ..... Clases de deportes *
1242000 .... Baile. Cine. Música. Teatro *
1242100 ..... Baile
1242200 ..... Música
1242210 ..... Flamenco
1242211 ..... Cante flamenco
1243000 .... Espectáculos *
1244000 .... Juego
1250000 ... Actividad mágico-religiosa
1251000 .... Prácticas devocionales *
1251100 ..... Ceremonias cristianas *
1260000 ... Actividad socioeconómica
1261000 .... Actividad constructiva *
1262000 .... Actividad de servicios *
1262100 ..... Actividad de gobierno *
1262200 ..... Actividad en seguridad-defensa *
1262210 ..... Actividad militar
1262220 ..... Actividad penitenciaria
1262300 ..... Actividad financiera
1262400 ..... Actividad jurídica
1262500 ..... Actividad legislativo-normativa *
1262600 ..... Comercio
1262700 ..... Comunicación
1262800 ..... Enseñanza
1262900 ..... Gestión administrativa
1262910 ..... Expediente administrativo
1262920 ..... Procedimiento sancionador *
1262A00 ..... Sanidad
1262B00 ..... Transporte
1263000 .... Actividad de transformación *
1263100 ..... Transformación de materia animal *
1263200 ..... Transformación de materia mineral *
1263300 ..... Transformación de materia vegetal *
1263310 ..... Transformación de fibras hiladas *
1263320 ..... Transformación de fibras sin hilar *
1264000 .... Actividad primaria
1264100 ..... Actividad forestal
1264200 ..... Agricultura
1264300 ..... Caza
1264400 ..... Ganadería (Actividad)
1264500 ..... Minería
1264600 ..... Pesca
1270000 ... Delincuencia

TG Figuras geométricas * A600000
TE Curva plana
Polígono geométrico
TR Diseño 1400000
Perspectiva 14E1000
Técnica de dibujo 14E3000

Figuras tridimensionales * A630000
UP Figuras espaciales
TG Figuras geométricas * A600000
TE Cilindro
Cono
Elipsoide
Esfera
Hiperboloide
Paraboloide
Pirámide
Prisma
TR Diseño 1400000
Perspectiva 14E1000
Técnica de dibujo 14E3000

Figurinistas 22A0000
TG Agentes en representación gráfica *
TR Agentes en representaciones dramáticas * 22B0000
Cinematografía 1242000
Decoración 1262000
Diseño 1400000
Teatro 1242000

Fijación (Protección) 1431300
TG Protección preliminar

Fijación de estratos 1431300
UP Asentado (Color)
Pacing
TG Protección preliminar

Fijado 14H0000
TG Técnica de unión *
TR Adherido 14H0000
Inmersión 1400000
Microfisuración 1340000
Revelado en positivo 14E2100
Separación en películas 1340000
Técnica de collage 14E0000
Técnica de restauración 1432000
Utensilios de perforación * 72ID000

Fijasellos 72B2200
TG Componentes de los sellos *

Fijo 3700000
TG Atributos de movimiento *
TR Llaves mecánicas 72I6000

Filacteria A820000
TG Motivos de cinta *

Filatelia A400000
TG Disciplinas
TR Sellos postales 72B2120

Filete (Molduras) 7113612
TG Molduras planas

Fileteado 1450000
TG Técnica de corte *

Filiación 5350000
TG Relaciones de parentesco *

Filiación agnática USE: Patrilinealidad 5350000

Filiación uterina USE: Matrilinealidad 5350000

Filibusterismo USE: Piratería 1270000

Filibusteros USE: Piratas 2244000

Filiforme 3510000
TG Atributos de forma *
TR Fibroso 3520000

Filita 6212200
TG Roca metamórfica

Filmación USE: Rodaje 14E2000

Filmaciones (Cine) 7231100
UP Película proyectable
TG Documentos audiovisuales
TR Cámaras de televisión 7286000
Cámaras de video 7286000
Celuloide 7252100
Cinematografía 1242000
Cines 71124B2
Cintas (Documentos) 7234000
Discos 7234000
Filmotecas 71124B1

Filmografías 7231400 7231500
TG Documentos secundarios
TG Documentos sonoros
TR Documentos visuales 7231600

Filmotecas 71124B1
TG Depósitos documentales (Edificios)
TR Celuloide 7252100
Cinematografía 1242000
Cines 71124B2
Filmaciones (Cine) 7231100

Filología A400000
TG Disciplinas
TE Etimología
Lingüística
TR Estructuralismo 5120000
Funcionalismo 5120000
Idiomas 5000000
Literatura A400000

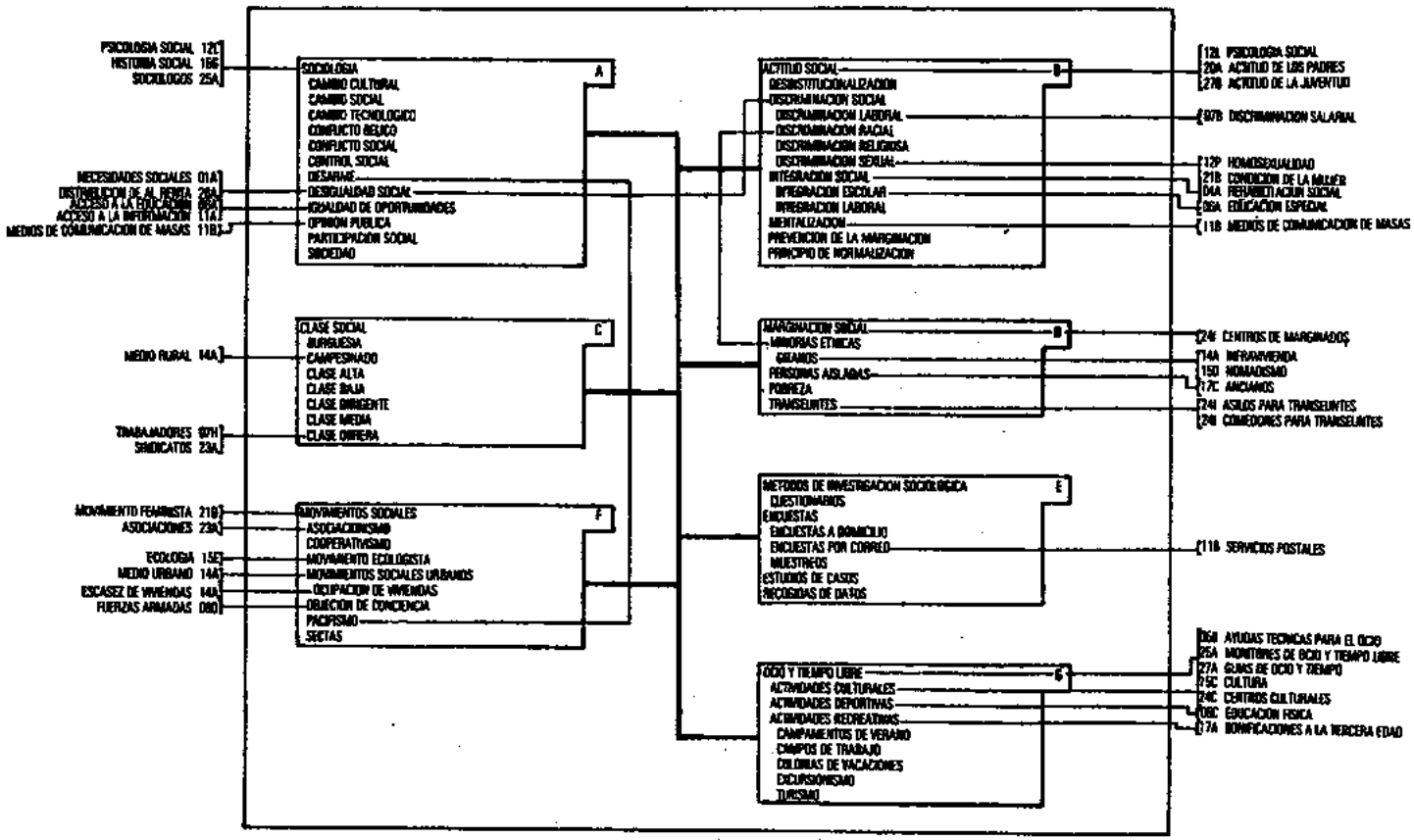
Filosilicato 6211820
TG Silicato 6211800
TE Arcilla
Clorita
Mica
Serpentina (Mineral)

Filosofía A400000

ÍNDICE PERMUTADO

ÍNDICE AUXILIAR

B730000 Temas mitológicos*
Brujas
B731000 Animales mitológicos*
Animal Grifo
Animal Quimera
Aretos
Argios
Arpía
Ave fénix
Bianor
Cancerbero
Centauro
Drialos
Esfinge
Eurinomos
Eurito
Folos
Licos
Minotauro
Neso
Oureios
Pegaso
Perro de Argos
Petraios
Quirón
Sirena
B732000 Lugares mitológicos grecorromanos*
El aqueronte
El estigio
El leteo
El olimpo
El parnaso
B733000 Personajes mitológicos*
Bacante
Duende
Gorgora
Hada
Huri
Nereida
Anfitrite
Tetis
Parca
Putti
Sátiro
Tres gracias
Victoria
B733100 Dioses paganos*
Dioses germánicos*
Dioses peninsulares*
B733110 Dioses grecorromanos*
Afrodita
Apolo
Ares
Artemisa
Aslepio
Atenea
Boreas
Demeter
Dionisos
Dios Hermes
Eros
Cibeles
B733121 Dioses egipcios*
Amón
Anubis
Atón
Atun
Bast
Bes
Enanos patecos
Hapi
Hathor
Horus
Isis
Jonau Haractes
Khepri
Khonsu
Knum
Min
Montu
Mut
Neferten
Neftis
Neit
Nekhebet
Num
Osiris
Pakhet
Path
Ra
Sebek
Sekhmet
Set
Shu
Sokar
Tefnut
Thot
Wadjet
B733122 Dioses fenicios*
Astarté
Melkart
Tanit
Musa
Caliope
Clio
Erato
Euterpe
Melpómene
Polonia
Talía
Terpsícore
Urania
B733300 Ninfa
Alseida
Aretusa
Cabiride
Náyade
B733400 Personajes de la mitología grecorromana*
B733410 Héroes mitológicos*
B733411 Hércules
Trabajos Hércules



TERMINOGRAMA

13. SOCIOLOGIA

EL CONCEPTO DE "MATERIA" EN LA CIENCIA DE LA INFORMACIÓN

Birger Hjörland
Real Escuela de Bibliotecología (Dinamarca)

1. EL CONCEPTO INGENUO DE MATERIA

Según el punto de vista ingenuo el concepto de "materia" o de "asunto" no representa ningún problema: está muy claro lo que son las materias. La materia del libro *General Psychology* es naturalmente la "sicología", y la del *Cambridge history of England* es la "historia", pudiéndola posteriormente subdividir si uno desea en "historia mundial" y en "historia de Inglaterra".

Un punto de vista ligeramente menos simple reconocería que necesariamente no tiene que haber una correspondencia entre, por ejemplo, el título del libro y su "materia" real. No todos los libros (por ejemplo el *Handbook of Psychology)* utilizan este término en sus títulos, ni todos los títulos necesariamente se corresponden con el criterio que tiene el usuario del contenido del libro. Los autores que tienen una base de conocimiento en una disciplina particular(por ejemplo la sicología, la psiquiatría o la sociología) pueden tener la tendencia a darle a sus trabajos títulos que nombren a sus propias disciplinas, aún cuando el contenido de los trabajos puedan fácilmente justificar la mención de alguna otra disciplina. *A history of dynamic psychiatry* pudiera también propiamente titularse *A hisytory of dynamic psychology,* y cuál es su materia real. El punto de vista ingenuo entró en dificultades.

El punto de vista ingenuo se corresponde, en parte, con la falta que tiene un niño de diferenciar entre las formas lingüísticas y los significados. Es aparentemente típico de una percepción del lenguaje que una palabra y su construcción fonética se vean como atributo de la cosa en sí, que no puede separarse de sus otras características. (Vygotsky/1, pp. 358-359/). La persona ingenua ve típicamente la materia como parte de, por ejemplo, los atributos de un libro, una concentración como si fuera lo que aparece en su título y que no puede separarse de los otros atributos del libro. Esta actitud, de cierta forma, está relacionada con el concepto filosófico de *realismo ingenuo*(de acuerdo a cual la experiencia de los sentidos proporcionan el acceso directo a la realidad: el realista ingenuo, por ejemplo, ve que las estrellas son más pequeñas que la luna, y por tanto asume que son más pequeñas).

Una caracterización más detallada, un escrutinio o investigación de la concepción ingenua del concepto de materia requiere que nosotros mismos hayamos alcanzado una sólida concepción de materia, este es precisamente el objetivo de este trabajo.

2. IDEALISMO SUBJETIVO

El idealismo es un concepto fundamental en la filosofía, cuya principal característica es que el proceso mental o conciencia se ve como elemento primario, o determinante, en relación con la realidad o el mundo material. En oposición al idealismo están las diferentes variedades de filosofía realista o materialista, en las que lo mental se concibe como algo secundario, o derivado de la

realidad o del mundo real. Algunos investigadores y filósofos se proclaman idealistas, pero es mucho más común que los investigadores no se consideren idealistas, ni tampoco asumen un punto de partida conscientemente idealista (y, por ejemplo, ven el choque entre el idealismo y el materialismo como un tema irrelevante), pero en sus pensamientos inadvertidamente caen en los modos idealistas de pensamiento. Esto precisamente se cumple en el campo de la bibliotecología y la ciencia de la información, por ejemplo, en lo referido al concepto de "asunto". Frohmann (2) recientemente ha publicado una amplia crítica a las tendencias mentalistas (y por ende idealistas) en la teoría de la "recuperación de la información".

Mis propios intentos por esclarecer la ciencia de la información son formas definitivas idénticas al punto de partida de Frohmann.

El concepto idealista de asunto significa que la "materia" es una "idea", ya sea en un sentido objetivo (es decir Platónico), o en un sentido más subjetivo. En esta sección observaremos más de cerca los conceptos subjetivos-idealistas de "materia", en la próxima sección se tomaran en consideración los conceptos objetivos-idealistas.

El idealismo subjetivo asume que los conceptos y las materias son la expresión de las percepciones o criterios de uno o más individuos (sujetos). Los conceptos y las materias son aquello que es subjetivamente comprendido o entendido por los individuos. La clave del concepto de materia por tanto descansa en el estudio de los pensamientos de algunas personas, por ejemplo, los autores o los usuarios de los documentos. Desde el punto de vista de la epistemología, el idealismo subjetivo se caracteriza por hacer que la percepción y el pensamiento sean independientes de manera subjetiva. El positivismo es el representante más común del idealismo subjetivo.

Si se trata del asunto o temática de un libro, existen muchas posibilidades: la versión del autor (según se expresa frecuentemente en el título o en el texto, ya sea de forma implícita como explícitamente), la versión del lector (aquí se hace posible una gran variación), la versión del redactor, como se indica con frecuencia en una serie de títulos (por ejemplo, "European Monographs in Social Psychology"), y la versión del bibliotecario, que puede muy bien expresarse en términos de la clasificación de la biblioteca.

Bentle Ahler Moller (3) ha publicado un breve trabajo en el cual compara la clasificación que para los mismos libros utiliza la State University Library en Aarhus, Dinamarca, con el sistema de Clasificación Decimal Dewey. Esto demuestra que pueden existir diferencias sorprendentes entre las percepciones subjetivas acerca de lo que son las materias de los libros. Pero esta subjetividad puede estar extremadamente bien fundamentada: *la subjetividad no es un ruido ni un error, es una tendencia analítica pensada, apoyada y consistente.* No estamos hablando simplemente de las diferentes estructuras que dan a las materias los diferentes sistemas de clasificación (es decir más o menos subdivisiones), sino diferencias inequívocas en la concepción de la materia de un libro, en donde un criterio coloca al libro bajo la temática de "libros", y otro criterio coloca al mismo libro bajo la temática de "comercio".

En conexión con el idealismo subjetivo se le da una consideración especial a las intenciones del autor, su criterio acerca de su temática, y cuáles son las cosas nuevas que debe relacionar. Esto ha dado lugar al surgimiento del concepto de "cercaneidad" en la literatura de bibliotecología y ciencia de la información, interés que, según mi opinión, representa un aliado ciego, un intento de escapar de las dificultades en el concepto de materia (Nota 1). Los seguidores del concepto de "cercaneidad" le asignan a este una claridad y una significación especial en el análisis de las materias, pero evidentemente desconocen su posición epistemológica como elemento subjetivo-idealista.

En relación con la teoría subjetiva-idealista del "asunto" demostraré que ni el punto de vista ni la comprensión subjetiva del autor, ni del lector, ni del bibliotecario/especialista de información, ni de ninguna otra persona (por ejemplo el del redactor) puede tener cierto conocimiento objetivo acerca de la temática de un documento, ni tampoco puede definir el concepto de "materia". Cada uno de estos puntos de vista puede contribuir en algo a determinar la materia, pero la concepción idealista subjetiva de materia hace un énfasis excesivo en ciertos aspectos del documento, ya sea desde el punto de vista del autor, del lector o de un intérprete.

1. Un libro puede, pero no necesariamente tiene que contener una aseveración acerca de lo que es su materia. El autor puede explícitamente discutir su obra, por ejemplo en la introducción y puede resaltar su relación con otras materias. Si un libro se titula "sicología general" puede contener un análisis de "¿qué es la sicología general?" Como las bases de la sicología constituyen un problema teórico complejo, el criterio del autor naturalmente no tiene que ser por necesidad cierto, simplemente la expresión de sus más o menos bien fundamentadas (subjetivas) ideas. Lo que es sicología para algunos puede, después de consideraciones teóricas, ser más bien sociología o fisiología. El libro puede no tratar en lo absoluto acerca de lo que el autor piensa que trata, ni sobre lo que indica el título.

Con mucha frecuencia, sin embargo, una obra no contiene ninguna discusión explícita de su materia. "*The History of Dynamic Psychiatry"* asume implícitamente que el psicoanálisis forma parte de la ciencia médica (psiquiatría) y no de la sicología. Se puede hablar mucho acerca de esto, pero la clasificación dada a un libro determinado puede no ser correcta. Un libro no tiene necesariamente que tratar sobre la materia de psiquiatría porque dice que lo hace.

El análisis científico de las materias de los documentos para las bases de datos tendría que asumir ciertas definiciones consistentes que a veces, pero no siempre, estarían de acuerdo con la versión de la materia dada en el propio libro.

2. En relación con el *usuario*, un documento puede solicitarse teniendo en mente las estructuras conceptuales y las percepciones temáticas del usuario. El usuario puede incluso tener una visión subjetiva de lo que es la materia del libro.

Algunos teóricos de la recuperación de la información parecen trabajar a partir de la premisa de que un sistema de recuperación de información debe solicitar las materias de acuerdo al significado subjetivo que da cada lector. Estos se inclinan a crear investigaciones psicológicas de las percepciones que tienen los usuarios de la materia, sus "estructuras de conocimiento". Existen también ejemplos de investigaciones realizadas sobre esas bases (Mark 4,5 es claramente un ejemplo de esto). Un modo de consideración relacionada es, por Pejtersen /ejemplo, el modelo ASK de Belkin (6-8). A pesar de ello J.E Farradane (9,10) asume un enfoque psicológico explícito dentro de la literatura de bibliotecología y ciencia de la información, una interpretación más detallada de su obra parece llevar implícito un modelo más objetivo que subjetivo-idealista.

Nosotros afirmamos que existen tipos de sistemas de información que evidentemente tratan de hacer una descripción de las materias para las percepciones subjetivas del

usuario. Ejemplos de esto son los sistemas de bibliotecas para niños o sistemas pedagógicos en los cuales se pueden describir el punto de partida y el objetivo tanto para el proceso docente como para el asesoramiento de los estudiantes. Ambos tipos de sistemas expresan un cierto *paternalismo*, digamos que alguien asume la responsabilidad de la dirección de las búsquedas de información de otros. Esto se hace presumiendo crear conexiones entre documentos determinados y el universo temático del usuario, es decir, información de los documentos a partir de una evaluación asumiendo la interpretación de las materias o del contenido de la psicológica o pedagógica de las necesidades y objetivos.

¿Fuera de esos enfoques paternalistas, deben entonces las descripciones temáticas tomar en cuenta la sicología del usuario? Si, de cierta forma esto es conveniente. Los sistemas de recuperación de la información deben resultar favorables para el usuario, y esto puede lograrse teniendo un conocimiento del lenguaje y de las percepciones subjetivas del usuario y utilizar este conocimiento por ejemplo, para *buscar* referencias de los términos preferidos. Pero esto no significa que se interprete el contenido temático de los documentos sobre la base del conocimiento de las percepciones subjetivas de los usuarios, sino que estas percepciones se emplean para crear las referencias e instrucciones necesarias, es decir, hacer que el sistema sea favorable al usuario. En mi opinión, esta condición favorable del sistema para el usuario no es *el* tema teórico central en la recuperación de la información. El tema central lo constituye la representación del conocimiento, cómo representar el conocimiento en los documentos. La cuestión del carácter amigable del sistema al usuario es una cuestión cognoscitiva-ergonómica que debe ejecutarse en el sistema, pero es de interés secundario si la comparamos con la representación adecuada del conocimiento en las bases de datos.

En mi opinión los sistemas de información científica deben presuponer que el usuario tenga las categorías, la terminología y las clasificaciones de la ciencia, el conocimiento y los sistemas de información, y no lo contrario. La adopción de las categorías y la terminología del usuario por parte de la ciencia y de sus sistemas de información constituye una función para la popularización, y no de forma primaria para la ciencia de la información. La referencia frecuentemente se hace para utilizar los principios de la sicología y de la lingüística para el diseño del sistema, pero esos principios a veces representan problemas o contradicciones que se oponen a las consideraciones puramente disciplinarias. Aquí llegamos a la conclusión de que *el que busca la clave del concepto* “sicologismo materia” en la mente del usuario comete un error..

3. Existe una tercera concepción subjetiva que puede ser expresada por el bibliotecario o el especialista de información en una descripción temática de los documentos en la base de datos. En el mejor de los casos se utiliza un sistema (de clasificación, un tesauro o alguna otra cosa) que hace posible un alto grado de bases explícitas y consistentes para el posterior análisis. Como ha sido demostrado, por ejemplo en Moller (3) los sistemas diferentes emplean diferentes principios (subjetivos) de análisis y por tanto diferentes determinaciones de materias. Esta situación no se documentará posteriormente en este trabajo, ya que forma parte significativa del argumento de esta sección sobre la teoría materialista del asunto. Aquí simplemente estableceré que tanto el trabajador individual de la información como los diferentes sistemas de IR despliegan considerables variaciones en sus descripciones de las materias de documentos determinados. Estoy hablando acerca de la concepción subjetiva-idealista, en tanto esta subjetividad se convierte en una cualidad del propio sujeto.

De ahí que sea típico de la concepción subjetiva-idealista de materia hacer un énfasis excesivo en ciertos aspectos del documento, ya sea desde el punto de vista del autor, del lector o del intérprete. Cuando este ejemplo subjetivo en su papel relativo al documento no puede garantizar el análisis correcto del asunto**,** ese análisis es subjetivo, y esto puede llevar *a una concepción agnóstica de*

"materia": resulta imposible decir de qué se trata la materia, y cómo puede determinarse. Ese criterio ha sido expresado por Patrick Wilson (11).

Patrick Wilson investiga, especialmente por medio de experimentos, la disponibilidad de diferentes métodos para determinar la materia de un documento. Entre estos métodos tenemos 1. identificar el objetivo del autor al escribir el documento, 2. valorar el relativo dominio y subordinación de diferentes elementos en la visión dada por la lectura del documento, 3. agrupar o contar el uso en el documento de conceptos y referencias y 4. inventar una serie de reglas de selección para los que constituyen elementos esenciales (en contraposición a los no esenciales) del documento en su totalidad. Patrick Wilson demuestra convincentemente que cada uno de estos métodos por sí solo es insuficiente para determinar la materia de un documento, y concluye: "la noción de materia de un escrito es indeterminada..."(p. 89); o (sobre lo que un usuario puede esperar encontrar bajo una posición particular en un sistema de clasificación de una biblioteca): "Porque nada definitivo *puede* esperarse de las cosas que encontramos en cualquier posición determinada"(p. 92). Vinculado con este último señalamiento Wilson incluye una nota interesante al pie de página, en la cual llama la atención hacia el frecuente uso impreciso que se hace de los conceptos por parte de los autores de los documentos (la"hostilidad" se menciona como un ejemplo). Aún cuando el bibliotecario personalmente pudiera lograr una comprensión muy precisa del concepto, no podría utilizarla en su clasificación ya que ninguno de los documentos utiliza el concepto de la misma forma precisa. Por tal motivo Wilson concluye que: "si la gente escribe acerca de lo que son para ellos fenómenos mal definidos, la correcta definición de sus materias debe reflejar su mala precisión".

El rechazo de una determinación de uno de los conceptos básicos de la bibliotecología y la ciencia de la información es un asunto aún cuestionable. No pensamos que el agnosticismo que expresa Patrick Wilson en sus citas previas sea una solución aceptable.

Como podremos ver posteriormente, *es* posible definir las materias. Pero no es posible determinar las materias examinando los pensamientos de los autores, los usuarios o cualquier otro grupo específico de personas. Hacer esto sería cierta forma de "mentalismo".

Los intentos por ir más allá de esto hace que surja la interrogante: ¿Cuáles son los criterios objetivos de la materia de un documento? Si las materias no son percepciones o "ideas" en las mentes de algunas personas, qué otra cosa pueden ser. ¿Qué debe entenderse por el planteamiento de que "el documento A pertenece a la categoría temática X?.

3. IDEALISMO OBJETIVO

La teoría subjetiva-idealista de materia ve a las materias como categorías subjetivas, para las que la persona X y la persona Y cada una tiene su captación subjetiva de la materia de un documento dado. (Estas categorías pueden ser más o menos idénticas- este es otro tema; el principio es que son individuales, dependientes de una concepción subjetiva.)

El idealismo objetivo no considera una materia como subjetiva en esta forma: las personas X y Y llegarán, si realizan un análisis correcto, a la misma materia de un documento dado, la materia que puede entonces denominarse objetiva (al menos en un significado particular del término). Mientras que el idealismo subjetivo en general se caracteriza por un énfasis excesivo en las percepciones de los sentidos, el idealismo objetivo tiende a hacer un énfasis excesivo en ciertos aspectos del análisis teórico y hacer que estos aspectos sean absolutos.

La concepción idealista indica que la materia es una designación de una *idea. En el sistema de Ranganathan esto se hace explícito, como citara uno de* sus estudiantes, Gopinath:

"Materia, un cuerpo organizado de ideas cuya extensión e intención son propensas a caer coherentemente dentro del campo de interés de la especialización inevitable de una persona normal"; "Una materia es un cuerpo de ideas organizado y sistematizado. Puede consistir en una idea o en una combinación de varias ideas..." (12). Esto se acerca mucho a la propia concepción de Ranganathan, aún cuando este a veces evade el problema, como en *Documentation and its facets* (13, p. 27), en donde declara que la materia es un "término asumido".

Para elucidar de forma más estrecha el criterio que asume del concepto de materia el idealismo objetivo, partiremos de su criterio de los conceptos en general. El idealismo objetivo (según se representa, por ejemplo, por Platón o el realismo escolástico) considera que un concepto es una entidad mental o psíquica abstracta (una idea), que existe en si y por sí, y la relación de esto con las cosas concretas es tal que estas cosas forman parte de las entidades mentales que las representan a través de los conceptos. El realismo (según el significado anterior) considera, en otras palabras, que los conceptos generales representan algo universal, que existe fuera e independientemente de la conciencia humana, y que al mismo tiempo existe previo a las cosas separadas (originalmente con referencia a Dios, hoy más bien una forma de cognición *a priori* en el sentido Kantiano).

Traducido a los términos del problema de la "materia", esto quiere decir que los documentos concretos toman parte de las "ideas" expresadas en una materia dada. Estas ideas existen fuera de la conciencia humana (o dentro de esta como percepciones *a priori*) y son también previas a los conceptos individuales expresados en los documentos individuales. Estas ideas o materias tienen propiedades universales o fijas; ellas pueden analizarse de una sola vez en un sistema universal, o separadas en partes individuales.

Este punto de partida teórico sigue teniendo mucha influencia en las teorías actuales acerca de las materias que pueden delinearse a partir de los criterios de Ranganathan (12), Tranekjaer Rasmussen (14, p. 26) siguiendo al filósofo danés Harald Hoffding, Thomas Johansen (15-19) y otros, sobre la materia como idea que puede analizarse en sus partes individuales.

La "Colon Classification" de Ranganathan se analiza en un artículo escrito por Gopinath, en el que afirma (12, p. 60):

> 2.7 Sintaxis absoluta de las ideas
>
> Una materia es en gran medida el producto del pensamiento humano. Se presenta un patrón organizado de *ideas* creadas por los especialistas en cualquier campo de investigación. El trabajo a nivel casi elemental y la proposición de las secuencias útiles entre las facetas y los elementos aislados *han llevado a la conjetura de que puede haber una "sintaxis absoluta" entre las partes constituyentes de las materias dentro de una materia básica, quizás de forma paralela a la secuencia del propio proceso del pensamiento, independientemente del lenguaje en el cual puedan expresarse las ideas, independientemente de la base cultural u otras diferencias en los entornos en los cuales pueden estar ubicados tanto los especialistas, como los creadores, como los usuarios de la materia...* (un énfasis añadido).

Este criterio de que el pensamiento humano, el lenguaje humano, la conciencia humana, el universo material humano tienen una "sintaxis absoluta", es decir, que

es fundamentalmente independiente del contexto funcional de los procesos mentales, es un patrón de concepción idealista, un contraste directo del criterio de que los procesos mentales son herramientas formadas por, y ajustadas a, tareas y condiciones en las que funcionan. Como no existen interrogantes acerca de si las personas X y Y tienen una sintaxis diferente, este es un idealismo objetivo, no subjetivo.

El idealismo objetivo se expresa en su proceso de clasificación con el criterio de que la clasificación de los documentos puede hacerse independientemente del contexto en el cual se utilice la clasificación. La "sintaxis", en el sistema de Ranganathan es la fórmula PMEST (Personality, Matter, Energy, Space and Time)/personalidad, materia, energía, espacio y tiempo/. Gopinath/12p.60/ da un ejemplo del análisis del documento. La temática "el ejercicio de la encartación por los ciudadanos indios en los años 1960" se analiza de la siguiente forma en el sistema Colon:

Historia(temática básica)
Comunidad India (ronda 1, personalidad, nivel 1)
Ciudadano (Ronda 1, personalidad, nivel 2)
Encartación (Ronda 2, materia, nivel 2)
Ejercicio (Ronda 1, energía)
años 1960s (nivel 1, tiempo)

Yo planteo que este tipo de análisis, que determina las prioridades de los criterios a tomarse en consideración en un documento, no es óptimo en todas las situaciones. Podemos imaginarnos a los investigadores trabajando sobre aspectos técnicos del proceso de elección que quieren establecer una comparación entre estos aspectos en varios países. Para esa persona la elección sería la materia o temática central y no sería conveniente que esta fuera un subtópico de La historia y la India. (La búsqueda automatizada en gran medida ha establecido secuencias fijas entre las facetas superfluas; el problema sigue presentándose solamente en los catálogos impresos y en otros sistemas de ordenamiento unidireccional, pero esto es otro asunto).

En realidad afirmamos que el concepto objetivo idealista de asunto tiende a encaminarse a las descripciones que sólo tienen una relación abstracta con las necesidades de descripción de la materia y los contextos en los que se utilizan, porque esas descripciones se basan en las propiedades de las ideas dadas *a priori*. Uno puede también expresar esto como si las materias o temáticas se vieran como "propiedades innatas" de las cosas o los documentos. Esta es una consecuencia del concepto de ideas objetivas de la teoría, separado de los elementos individuales de la realidad. En otras palabras, esta es también una expresión de la concepción especial del idealismo objetivo de las relaciones que existen entre lo general y lo particular: que lo general existe fuera e independientemente de lo particular. Esto se opone al concepto de que una materia sólo existe en documentos específicos, y que toda descripción de materia contiene un análisis con sus puntos de partida dentro de los propios contextos que utiliza, y que se analizará más detalladamente debajo. *El concepto idealista de "materia" tiene además la consecuencia de que ni los criterios del mundo, ni la disciplina académica y las prioridades políticas expresadas en los sistemas de información son reconocidos,* criterio que ha sido criticado entre otros por Steiger (20).

Para resumir: el punto de vista objetivo idealista no se ajusta, como lo hacía el punto de vista subjetivo idealista, al concepto de materia presente en las mentes de algunas personas. En su lugar presume que pudiera utilizarse cierto tipo de análisis abstracto o procedimiento establecido para penetrar en la superficie de los documentos, revelando así sus materias o temáticas reales. Como podremos ver posteriormente, ninguno de estos procedimientos establecidos puede garantizar un análisis correcto de la materia. Entre otras cosas, este enfoque no toma en consideración los aspectos pragmáticos de las materias: el uso potencial de los documentos.

4. CONCEPTO PRAGMATICO DE ASUNTO

Un usuario tiene una necesidad (específica) de información, un problema que resolver para lo cual requiere información. Esta información se busca en las bibliotecas o en las bases de datos en las cuales los documentos(portadores de la información) están registrados por *materia.*

El registro de las materias por parte de los bibliotecarios o especialistas de la información debe, para que el proceso tenga significación, anticiparse a las necesidades del usuario: debe hacer posible que el usuario encuentre lo que busca. Los datos de las materias en las bibliotecas y en los sistemas de información tienen una función instrumental o pragmática. Como escribieran Bookstein y Swanson (21): "los documentos se indizan con el fin de su recuperación, y podemos alcanzar un procedimiento teóricamente bien establecido de indización siendo fieles a este fin".

Dagobert Soegel (22) ha establecido una distinción entre "la indización orientada al contenido" y la "indización orientada a la solicitud" que ha probado ser muy estimulante en mi filosofía sobre el concepto de materia. Aquí no estamos investigando si Soergel ideó realmente la "indización orientada a la solicitud" ni siquiera el nombre simplemente. El señala que en la literatura sobre bibliotecología y ciencia de la información sólo se describe la primera y que lo segunda difícilmente se conoce en teoría, aunque en la práctica si existen ejemplos (por ejemplo la base de datos Ringdok, que describe la literatura química de forma diferente al *Chemical Abstracts*, porque Ringdok presta una atención especial a las necesidades de la industria farmacéutica).

La indización orientada al contenido es una descripción de las materias que debe concebirse puramente como una función de los atributos del documento: como en la observación de que "este documento contiene la fórmula química del ácido sulfúrico" (y su consecuente categorización como "química inorgánica").

La indización orientada al usuario o a la necesidad es una descripción de una materia que debe percibirse como la relación entre las propiedades de un documento y la necesidad real o anticipada del usuario. "Este documento trata sobre el ácido sulfúrico. El ácido sulfúrico corroe. Los señalizadores necesitan agentes corrosivos", siguiendo de esta forma una categorización, por ejemplo, "Literatura sobre químicos a utilizar en la señalización". *La indización orientada a la necesidad es una relación instrumental (medios y fines) entre un documento y la necesidad de un usuario.*

Dentro de la ciencia de la información los medios como el *Science Citation Index, el Social Science Citation Index y el Atlas of Science* (todos publicados por el Institute of Scientific Information en Filadelfia) ofrecen los nexos existentes entre las materias, o la categorización de los documentos sobre las bases de una relación previa puramente

instrumental o de medios y fines: se asume que los documentos citados por el mismo documento están relacionados por su materia, ya que todos han contribuido a los resultados del documento en cuestión. En otras palabras, estos atlas (o el concepto de nexo bibliométrico y cita) son expresiones implícitas de un concepto de "materia" en el cual hay una relación instrumental previa y factual (como se refleja en la práctica de las citas) que proporciona las bases de la definición.

La anexión bibliométrica es un método de búsqueda de literatura que ha ocupado un lugar en el sistema, y que tiene sus ventajas y sus desventajas. Ocupa un lugar especial: *no es el hecho de simplemente mapear esos nexos instrumentales previos y producir así una medicina patente para la búsqueda literaria, ni se trata de reducir el concepto de materia a estas relaciones empíricas.*

En esto hay varias razones que juegan su papel. Primero, de una relación instrumental previa no se puede extraer una relación potencial instrumental. En la ciencia de la información, la literatura sobre "telecomunicaciones" puede estar vinculada (co-citada) con la literatura sobre "recuperación de la información", porque las telecomunicaciones en cierta etapa de desarrollo constituyeron un problema crucial para la recuperación de la información. Pero posteriormente, los problemas de las telecomunicaciones pueden considerarse triviales, y estos nexos bibliográficos pueden ser una mala expresión de una "relatividad temática". Segundo, hay ciertas condiciones, culturales o sociológicas dentro del ambiente investigativo, que cambian o tergiversan la imagen, en tanto que hay documentos epistemológicamente fértiles que no se citan tanto como los documentos que fácilmente nos llevan a investigaciones concretas(esto es como decir, hay un énfasis marcado en el empiricismo). Hay una tercera y última razón, un documento particular la mayor parte de las veces contiene tipos de información esencialmente diferentes, algo que resulta útil para categorizar de formas diferentes a la que nos conduciría a la práctica puramente orientada al uso. Por ejemplo, muchas investigaciones psicológicas citan literatura estadística y metodológica tanto como literatura de carácter psicológico. Sería oportuno operar con ellas como materias diferentes, aún cuando aparezcan unidas (mediante nexos bibliométricos) dentro de la literatura psicológica de un período dado.

La teoría pragmática de la materia cae en otras dificultades: Si asumimos que un documento dado tiene que incluirse en relación con todos sus posibles usos, entonces esto daría lugar a toda una gran cantidad de repeticiones o de clasificaciones múltiples. En el ejemplo anterior con el ácido sulfúrico sería imposible para una biblioteca universal clasificar el ácido sulfúrico según todos sus usos potenciales. De ahí que el concepto de Soergel de indización orientada a la solicitud sea de hecho significativo, y para los servicios de información especializados resulte importante clasificar de acuerdo a la necesidad del grupo clave.

Por supuesto, el problema con el concepto pragmático de materia descansa, en el sentido más básico, en la condición que este comparte con la filosofía pragmática: aunque el objetivo es desarrollar la práctica humana, la orientación práctica limitada es muy poco previsora y muy superficial en sus criterios de verdad. El pragmatismo no tiene criterios profundos significativos que puedan orientarnos hacia la prioridad de las propiedades de un documento.

Una vaca puede describirse zociológicamente como un mamífero y pragmáticamente como un animal doméstico o ganado. Dalhberg (23, p.194) designa la última relación como la relación que existe entre el hombre y el objeto, pero a la primera le asigna otra clasificación, o sea la "ontológica". No estamos de acuerdo con esta distinción absoluta: toda cognición es fundamentalmente instrumental para el hombre. El concepto de "animal doméstico" tiene una conexión más *inmediata* con la práctica humana, mientras que el concepto "mamífero" es una abstracción con una relación *menos inmediata* con la práctica humana. La clasificación de un libro sobre vacas en la categoría temática de "mamíferos" o de "animales domésticos" no depende de la propiedad más significativa del libro(el objeto central es la vaca en ambos casos). Depende básicamente de la evaluación de si el libro es más útil para las personas que buscan literatura de zoología o para las que buscan literatura de agricultura, es decir, si el libro es más útil para un biólogo o para un campesino. Este es un juicio que se basa en las propiedades del libro en relación con la percepción de los intereses en todo su sentido epistemológico. Este juicio se toma quizás principalmente sobre las bases del contenido del libro, pero cuando la descripción de la temática o materia está dirigida a otro grupo clave, se deben tomar otras decisiones (cf.el ejemplo del *Chemical Abstracts* y Ringdok).

El conocimiento abstracto y general de la biología y de otras ciencias ha demostrado su significación para el hombre, aún cuando su designación de las funciones útiles es menos inmediata que la de "animal doméstico" La sistematización y la terminología científicas facilitan una organización limitada de conocimiento que en un nivel superior asegura la comunicación más eficaz en el desarrollo del conocimiento humano. Esa organización del conocimiento es difícil de justificar a partir de una filosofía pragmática en la comprensión usual de este concepto en la filosofía.

Aunque la teoría pragmática de la materia tiene sus limitaciones, hace una importante contribución a la percepción de las propiedades centrales del concepto de materia, al señalar su carácter de *medios y fines*(y repudiar así el criterio que asume que las materias son "cualidades inherentes"; las materias no son cualidades más inherentes que el *valor* de la cosa).

Esto está apoyado por la etimología de "materia"(especialmente en los idiomas escandinavos, pero también en el inglés y el alemán, ver Nota 2). "Materia" (el "emne" escandinavo) significa "materia prima", entre otras cosas. El hierro es materia para el herrero. Una vaca es materia para el zoólogo y el campesino. La epistemología es materia para el filósofo y para el investigador de la información. La materia es, por tanto, siempre materia para alguien o para algo.

5. TEORIA REALISTA Y MATERIALISTA DE LA MATERIA

Según el criterio realista y materialista las cosas existen objetivamente y tienen propiedades objetivas. Este es un punto crucial de partida que debe tomarse por hecho en este artículo. (Nota 3) En este trabajo no se hará ningún esfuerzo por demostrar las diferencias que existen entre el "realismo científico" y el "materialismo".

Los documentos (en este contexto) son un problema teórico. Por una parte, naturalmente, los documentos reflejan el criterio subjetivo del autor de la materia que se trata. Por otra parte, el documento tiene propiedades objetivas. Si un documento afirma que la "inteligencia de una persona se correlaciona con el tamaño de su cerebro", este es un criterio subjetivo (y falso). Pero si es un hecho objetivo que este libro contiene este criterio

(falso). Estamos interesados en las propiedades objetivas del documento. Las propiedades objetivas no son los criterios o evaluaciones subjetivos contenidas en los documentos; Las propiedades objetivas tienen un potencial cognitivo (o informativo) que asegura que el lector pueda diferenciar entre los planteamientos falsos y los verdaderos. Nuestra concepción de propiedades objetivas de los documentos es una reminiscencia del concepto del "World III" (24) de Karl Popper, en donde se refiere a los libros como "conocimiento objetivo", y opera con experimentos con los pensamientos muy similares a los míos. Sin embargo, mi concepto de la objetividad de los documentos no la tomé prestada de Popper, y existen grandes diferencia entre ellas, porque la base teórica de Popper es el dualismo y la mía es el monismo. No tenemos aquí espacio para evaluar la teoría de Popper en relación con la mía. Esto es controvertido y ha sido muy criticado tanto por la filosofía como por la ciencia de la información. En el último caso ver Rudd (25).

¿Que debemos entender por propiedades de un documento?

En el sentido amplio de la palabra, las propiedades de un documento son todas aquellas afirmaciones verdaderas que puedan hacerse de ese documento.

Un documento puede describir los logros de Christian IV, establecer los puntos de fusión de los metales, presentar información sobre la composición de aditivos alimentarios y sus consecuencias para la salud humana, investigar al unicornio como símbolo psicoanalítico, etc. Las propiedades aquí mencionadas puede decirse que tratan la reflexión del documento, la representación o tratamiento de una parte de la realidad (o de la conciencia y la imaginación humanas). El aspecto de la realidad que refleja (su cercaneidad) es una de las propiedades centrales del documento. También resulta significativo la forma en que se trata o refleja la realidad, por ejemplo, si sus planteamientos son ciertos o falsos, representativos, superficiales o fundamentales, etc. Hay una categoría de propiedades que puede denominarse relacionales: ¿cómo se relaciona este documento con otros documentos?¿El documento elabora, solapa, corrige o hace que otros documentos sean superfluos?

Los documentos pueden caracterizarse por el lenguaje, la forma, el tipo, etc., que frecuentemente representan propiedades menores. Hjörland (26). Y finalmente, los documentos pueden caracterizarse por el tipo de papel, la encuadernación, la topografía, etc., que en la mayoría de los casos serían insignificantes, pero que pueden ser centrales para fines específicos (la historia de un libro). Las propiedades de un documento surgen especialmente cuando este se utiliza, por ejemplo, leyendo un documento relacionado con una actividad particular (la investigación, la docencia u otra). La frecuencia y la estructura de las palabras utilizadas, es decir, el lenguaje expresado en el documento, también pertenecen a las propiedades del mismo. Estas últimas propiedades por lo general no aparecen directamente mediante la lectura del documento sino, por ejemplo, a través de su procesamiento para funciones de automatización, búsqueda o indización automatizada, clasificación, etc. Terminaré aquí el análisis de estas últimas propiedades, aún cuando no desempeñan naturalmente una gran función en la literatura de la ciencia de la información. El lenguaje en que está expresado el documento juega un gran papel práctico en la búsqueda de la información porque estos elementos con frecuencia están accesibles para la búsqueda ya sea en bases de texto completo (siguen siendo las excepciones), o en forma de representación de partes del texto en bases de datos,

usualmente los títulos y los resúmenes. Pasaré por alto este problema. Estoy de acuerdo con Spang-Hanssen (27, p. 20) en que el contenido de un documento no puede describirse de forma profunda simplemente por la formalización de su lenguaje.

He dado ya una breve definición de las propiedades de un documento. Ahora debemos tomar en consideración hasta qué punto las propiedades de un documento pueden describirse objetivamente.

Resulta muy curioso que la objetividad signifique dos cosas diferentes en relación con el enjuiciamiento de las propiedades de un libro(descritas aquí según la epistemología realista):

1. independiente del sujeto que percibe;
2. de acuerdo con la realidad. En el primero de estos sentidos, mientras más lectores identifiquen estas mismas propiedades en el libro, mayor objetividad habrá. En el sentido de: "de *acuerdo con la realidad"*, la relación es inversamente proporcional. Como para ser capaz de identificar las propiedades significativas en un libro científico se necesitan calificaciones especiales, quizás sólo un grupo limitado puede captar todo el potencial de una obra. En otras palabras, las propiedades fácilmente identificadas por la mayoría serán las menos significativas (o las más indiscriminadas), y por tanto menos objetivas en el segundo sentido de este término. (Esta situación es especialmente lo que sucede con la investigación básica, donde ocurre la reorientación teórica. En los contextos más cotidianos, el "proceso normal de investigación"(en el sentido kuniano), no es necesario obtener este contraste expreso entre los dos requisitos de la objetividad).

Para repetir: existe un contraste directo entre los dos conceptos de la objetividad en la evaluación de las propiedades más significativas de un libro y por ende sus materias. La solución de este problema es una decisión de la mayoría. La solución es una argumentación explícita, si no una prueba, por lo menos el establecimiento de una probabilidad. Hemos visto que la descripción de las propiedades de un documento en sí no es una cosa simple, susceptible a la automatización, pero es muy dependiente de condiciones particulares (que con frecuencia tienen un carácter teórico). Cuando sostenemos que las propiedades de un documento son objetivas, aún cuando su descripción requiere de prerrequisitos subjetivos especiales, esto implica que la realidad, la comprobación del documento en la práctica, en el análisis final decidirá su potencial informativo, sin importar cuántas malas concepciones se hayan hecho con anterioridad. La historia se convierte en el juicio final de la objetividad de los planteamientos acerca de las propiedades del documento. (Y aún cuando la historia raramente decidirá esto al final, mantenemos el concepto de las propiedades objetivas en los documentos que constituyen las bases de nuestros intentos por analizarlas.)

Las diferentes propiedades de los documentos pueden tener diferentes significados para diferentes objetivos o disciplinas científicas. Las disciplinas o teorías científicas pueden tener centros de atención disímiles o diferentes intereses epistemológicos. Por tal motivo pueden haber diferencias marcadas en la identificación de las propiedades de los documentos. La identificación de las propiedades a partir de un punto de vista teórico estrecho es más pragmática que una perspectiva más general. La identificación de las propiedades de los documentos desde un punto de vista superior o general presupone la habilidad de evaluar los potenciales de las diferentes teorías, lo que presupone una perspectiva más filosófica. El personal de la bibliotecología y la ciencia de la información con un profundo grado de conocimiento de la materia y con expectativa en la búsqueda en las bases de datos y en la evaluación de búsquedas realizadas a profesionales, cumple frecuentemente importantes prerrequisitos para identificar esas propiedades generales.

LAS MATERIAS Y LAS PROPIEDADES DE LOS DOCUMENTOS

En el uso filosófico los documentos representan la variable individual y sus propiedades y relaciones los predicados (las propiedades y las relaciones conjuntamente se denominan los atributos lógicos del documento).

Los ejemplos mencionados de las propiedades del documento (la parte de la realidad de que trata, su valor verdadero, su método, etc.) constituyen los predicados de primer grado (o predicados de primer orden), al igual que su estructura léxica, etc.

Cuando un bibliotecario o especialista de información categoriza documentos con una descripción temática, es con los predicados de primer grado con los que interactúa: ya sea leyendo el libro, o inspeccionando su estructura léxica(y en el caso extremo puede construir un programa de computadora que categorice los documentos a partir de esta estructura). Sobre las bases de este análisis de los predicados de primer grado del documento, le asigna un predicado de segundo grado, un predicado predicado. (Nota 4) *La asignación de una materia es, por tanto, una función de las propiedades del documento y en sí constituye un atributo del documento.* (Nota 5)

Ver de esta forma la materia como una función de las propiedades de un documento no nos dice en sí de que se trata la materia. A pesar de esto, el concepto de predicado esclarece la relación entre la materia de un documento y sus otros atributos. (Nota 6)

Para determinar el concepto de materia debemos preocuparnos por *qué* propiedades del documento forman parte de la descripción de la materia, y *en qué forma* estas propiedades juegan su papel. Resulta algo extremadamente fácil decir en la práctica qué es la materia (el concepto ingenuo de materia): la designación de una materia frecuentemente sólo requiere señalar una o algunas de las propiedades significativas del documento, en particular, las condiciones del mundo real que refleja el documento. Si el documento tiene la propiedad que trata sobre el estilo de construcción de Christian IV, entonces al documento podrá asignársele el predicado de materia " Estilo de construcción de Christian IV. "Hay en este ejemplo una identidad evidente entre lo que hemos definido como propiedad del documento y su materia, pero como hemos hecho una elección entre las muchas propiedades teóricas infinitas, la descripción de materia no es en principio idéntica a los predicados del primer orden del documento. No hay una explicación del porqué se ha escogido en este caso esta propiedad como materia. En otras palabras, debemos analizar más de cerca esta función de la materia. (Nota 7)

¿Qué propiedades del documento entran en la descripción de la materia?

Como se ha resaltado con anterioridad, con mucha frecuencia en la práctica las propiedades simples y fuertes forman las bases del análisis de la materia. Teóricamente, sin embargo, esto se vuelve extremadamente complicado, y tan pronto se intenta excluir una propiedad, surge un ejemplo hipotético en el cual esta propiedad formará parte importante de la determinación de la materia. ¿La autoría de un documento es parte del análisis de la materia? Si, en el caso de las autobiografías, y como Boserup (28) indica, también hipotéticamente en otras situaciones). No pretenderé demostrar aquí que todas las propiedades de los

documentos entran en la función de determinación de la materia, ni tampoco pretendo eliminar aquellas que no forman parte de esta función. Mi punto de partida es que no existe una parte bien definida o definible de las propiedades de los documentos que entra en el análisis de la materia (y que justamente esta situación lleva al concepto agnóstico de asunto de Patrick Wilson).

De la misma forma planteo que la función de definición de la materia no puede ser un procedimiento previamente determinado al analizar las propiedades, tal y como pretende establecer la fórmula PMEST de Ranganathan. Considero que exactamente la elección de determinadas propiedades de los documentos o de determinadas funciones de estas propiedades lleva inevitablemente al camino idealista. Como los bibliotecarios y los especialistas de la información gustan mucho de tener directivas y procedimientos claros y firmes, la tendencia idealista continuamente se esconde en las alas dentro de la propia concepción de materia. (Pero por supuesto, en el desarrollo concreto de los sistemas de información se deben describir los procedimientos, por ejemplo en el uso de los sistemas de clasificación y los tesauros, yo mismo he sido un vocero de los procedimientos definidos y explícitos (listas) en la descripción de las materias. (29)

Mi punto de partida para una teoría materialista de materia descansa en la concepción pragmática de la materia presentada con anterioridad. Las materias dan una evaluación de las propiedades del documento en relación con la optimización de la percepción potencial del documento. No está previamente definido qué propiedades del documento son pertinentes, y qué funciones analíticas tienen que instituirse en relación con estas propiedades, pero esto si depende del contexto. (Nota 10)

Por tal motivo las materias deben definirse en sí como los potenciales epistemológicos de los documentos. Un potencial es una propiedad intangible, de ahí el problema con la definición de las materias. Pero el potencial de algo no es una idea subjetiva u objetiva. El potencial es una posibilidad objetiva. El Uranio desarrolló sus potenciales como combustible atómico antes de que la ciencia conociera estas posibilidades, y muchos autores han sido enterrados antes de que el potencial significativo de su obra haya sido reconocido. La etapa actual del desarrollo social es la que determina qué cosas y qué obras tienen qué potenciales. En una etapa determinada el uranio era un metal que no era particularmente valioso y no tenía un potencial especial. En otra etapa es una importante fuente de energía, y en una tercera etapa quizás llegue a ser alguna otra cosa nuevamente. *Esto es como decir que es el nivel de desarrollo de la sociedad humana, la práctica humana, lo que constituye una materia* (Nota 8).

De ahí que la descripción temática de un documento es, de una u otra forma, una expresión de los potenciales epistemológicos de un documento, según lo percibe aquel que describe la materia. Mientras mejor prescriba la descripción los potenciales del documento, mejor, más correcta y más objetiva será la descripción de la materia. La comprensión de esto puede esclarecerse leyendo el ejemplo concreto analizado en el apéndice de este artículo. Sin embargo, la interpretación de una descripción dada de una materia debe comprender las calificaciones(e intereses) de la persona que ha realizado la descripción de la materia. Cuando Patrick Wilson (11, p. 92) escribió (en relación con lo que el usuario puede esperar encontrar en una ubicación particular en un sistema de clasificación de una biblioteca): “nada definitivo puede esperarse de las cosas que

encontramos en una posición dada", esto es sólo correcto a partir de este prerrequisito subjetivo. Podemos afirmar con la adherencia de la hermenéutica que la percepción del potencial de un documento depende del *preconocimiento* de la persona que determina la materia. En contraposición a muchos seguidores de la hermenéutica yo, sin embargo, deseo conservar el concepto del potencial objetivo o materias de los documentos.

La descripción de una materia es por tanto el pronóstico de los futuros potenciales. Este pronóstico puede basarse en criterios positivos o negativos. La descripción de la materia o descripción temática puede verse como una especie de visión o como una evaluación relacionada con la investigación actual. El prerrequisito más importante en la descripción de la materia no es el tipo específico de método, sino la madurez de los criterios.

El uso de los sistemas de materia asume, por tanto, también la interpretación. El usuario debe entrar en el universo del sistema y de sus medios. Esto es escasamente excepcional. En algunos casos los documentos se solicitan por el llamado "principio del origen o la fuente", lo que requiere que los documentos permanezcan en las colecciones y en el orden en que fueron originalmente organizados. Esto exige un análisis de la organización que existía cuando se creó la colección. La solicitud de los documentos y el conocimiento siempre están basados en premisas particulares, en criterios y suposiciones generales. Frecuentemente es necesario conocer estas premisas para obtener un resultado satisfactorio de la descripción de las materias. El grado necesario de interpretación depende de la medida en la cual la descripción se haya anticipado y satisfaga las necesidades del usuario. Con el principio de la fuente sólo se intenta un grado bajo de anticipación, ya que el principio no pretende considerar el contexto actual del usuario. Por el contrario, la base de datos de Ringdok sobre farmacología antes mencionada tiene un alto grado de satisfacción de las necesidades del usuario. Los sistemas de información que toman en consideración las necesidades de los usuarios son más caros de crear y de mantener, pero al mismo tiempo son más económicos en lo que se refiere a los recursos que utilizan.

Raramente se presenta una descripción de materia como afirmación directa acerca del potencial de un documento; aparece con mayor frecuencia en forma de referencia a una disciplina académica ("la materia es sicología"), es decir, un área de problema socialmente definida, dentro de la cual ejerce particularmente su contribución el documento para la solución de un problema. Como se mencionara con anterioridad, las materias pueden también expresarse indirectamente haciendo simplemente énfasis en la cualidades especiales ("trata la arquitectura de Christian IV"), que `puede también colocarse dentro de una disciplina (historia, historia del arte) o que sirve directamente como base desde donde el propio usuario evalúa la materia del documento (por ejemplo, "atracciones turísticas")

Los temas acerca de la expresión de las materias, de "los lenguajes de recuperación de la información" y de la representación en el texto van más allá del alcance de este artículo. Pero como estos temas presuponen un conocimiento de lo que son las materias, la teoría de la "materia" propuesta que

aquí se presenta constituye un prerrequisito para las teorías más profundas sobre estas cuestiones.

Podemos ahora volvernos al problema de Patrick Wilson relacionado con "los fenómenos mal definidos" de los autores. La designación de una materia refleja la claridad o imprecisión de un documento pero no de la forma mencionada por Wilson. El objetivo de analizar la materia es determinar si un documento tiene un potencial epistemológico relacionado con los futuros usuarios de una categoría dada o de un concepto dado, por ejemplo "la hostilidad". Si lo tiene se clasifica bajo ese concepto, si no lo tiene, no se clasifica bajo ese concepto. (Si se pone bajo ese concepto para esclarecer la terminología confusa en ese campo, puede también considerarse como una especie de potencial informativo, incluso de tipo más indirecto.) La asignación de una materia a un documento es de hecho, un criterio claro de que este documento "tiene un potencial epistemológico dentro del concepto de "hostilidad", aunque este criterio claro esté basado en muchas deliberaciones acerca de si el documento realmente contribuyó a esta materia, porque carecía de imprecisión en su uso de los conceptos. En la práctica real encontramos con frecuencia otras posibilidades, preferiblemente desde una perspectiva ideal, por ejemplo, la caracterización de los métodos o enfoque teórico del trabajo, que muy bien puede darle un perfil temático más elevado a la obra en dependencia de su estructura; en otras palabras, las decisiones acerca de la temática de un documento no son típicamente un criterio de "todo o nada". (Nota 10)

Materias y epistemología

Los documentos son fuentes para el proceso cognoscitivo al igual que las personas, las cosas, los procesos, los planteamientos etc. son también fuentes para la cognición humana. La forma en que el hombre adquiere el conocimiento es materia de interés de los epistemólogos. La cognición científica, que además de la epistemología también crea una teoría de la ciencia y de las metodologías de las disciplinas académicas, forma parte de la actividad cognoscitiva humana (un caso especial importante).

Existen varios tipos de epistemología, por ejemplo el idealismo (positivismo), el realismo científico y el materialismo. Está fuera del alcance de este artículo tratar la epistemología como tal. El objetivo de este artículo es esclarecer el concepto de materia, y con tal objetivo en mente resulta necesario mirar a la determinación de la materia desde un punto de vista epistemológico. Esto sigue, en particular, a la conclusión de la sección previa: que la determinación de una materia constituye una evaluación de prioridades de las propiedades de un documento en relación con la categorización y la descripción temática de ese documento. La forma en que avanzan esta categorización y esta descripción es un elemento decisivo para la "visibilidad" del documento en las bibliotecas y en las bases de datos, y por tanto para su función potencial en el desarrollo futuro del conocimiento.

La epistemología filosófica comprende el conocimiento más generalizado acerca de cómo la persona, por ejemplo un investigador, o toda una disciplina, debe examinar el mundo para ampliar el conocimiento humano. De ahí que concluya que en tanto la teoría sea capaz de producir resultados útiles, esta teoría será también la base de la determinación de las materias de los documentos.

Si un investigador tiene una pregunta particular, por ejemplo acerca de los simios, o los orígenes de la vida, la hipótesis y la formulación de la pregunta son primarias. Los métodos pueden utilizarse para investigar el tema,"métodos empíricos" o "el análisis teórico" o "la investigación bibliotecaria" (es decir la búsqueda de literatura) son secundarios. El esclarecimiento de la pregunta y los

conceptos centrales comprendidos, todos estarán de igual forma en un nivel central. La pregunta *determina* qué cosas, procesos, documentos etc. son pertinentes para estudiar, y en qué sentido lo son. Otro tema es, en qué medida pueden identificarse los documentos pertinentes. Yo sostengo que es extremadamente difícil identificar los documentos que son más pertinentes en la ciencia moderna. Hjörland (29) para el análisis de este problema en un estudio de caso). El efecto de esta identificación que resulta tan difícil es que la base de los sistemas de identificación asume la posición de un importante problema científico. La descripción de la materia de un documento (es decir, la evaluación, la asignación de las prioridades y la consecuente categorización de los potenciales del documento) comprende una visión o comprensión acerca de los problemas futuros que pueden demandar el uso del documento en cuestión. La causa de esto descansa en dos afirmaciones: 1. todo documento posee una cantidad infinita de propiedades (de manera tal que resulta imposible contarlas todas); 2. las propiedades que resultan esenciales para un contexto no necesariamente lo son para otro(de forma tal que puede determinarse una serie fija de prioridades para todos los casos, como lo ilustra el ejemplo del sistema de Ranganathan).

La epistemología tiene algo pertinente que decir sobre lo que significa "describir". ¿Qué significa describir, por ejemplo, el contenido de un libro? Abordaremos ligeramente los aspectos epistemológicos de esto, basándonos en Krober y Segeth (30). El concepto de descripción se utiliza más comúnmente en relación con las percepciones de los sentidos que están presentes de forma sistemática y ordenada a través de la deliberación y el lenguaje. Una descripción exitosa puede lograr una visión muy precisa del elementos descrito, pero sólo puede determinar la forma en que este objeto está constituido, no el porqué está constituido de esa forma. Por esa misma razón la descripción se mantiene en los aspectos superficiales del objeto, y no penetra en su esencia, incluyendo las razones de su existencia. Una descripción es, por tanto, el primer paso en la cognición, que posteriormente es sustituida por otros modos de cognición que penetran profundamente en la esencia de las cosas. El programa de la epistemología positivista de limitar el método científico a la simple descripción de los hechos es demasiado estrecho comparado con lo anteriormente expresado. El requisito del positivismo de una descripción completa de un fenómeno es imposible e innecesario. La descripción completa es imposible debido a que la cantidad infinita de propiedades de un fenómeno exigiría una descripción infinitamente extensa. La descripción completa es innecesaria porque tanto para el conocimiento científico como para los fines prácticos humanos, una descripción igualmente detallada de todas las propiedades y relaciones significativas e insignificantes, necesarias y casuales,

generales y particulares resulta inútil. Lo que se necesita es el conocimiento de lo significativo, lo general entre lo particular, lo necesario y lo típico. La descripción puede por tanto cumplir su función en el proceso de obtención de conocimiento sólo cuando no sea absoluta y distinta de otros medios de cognición, como la explicación, la hipótesis, el pronóstico, etc. La descripción debe, de hecho, verse en el contexto de esos otros modos de conocimiento.

No vemos razón para dudar que la misma situación se cumpla para la descripción de las materias de un documento: una descripción "pura" de los documentos sin conexión con otros modos de cognición como la hipótesis, el pronóstico etc. puede sólo obtener las propiedades más triviales y superficiales del documento. La comparación que de las descripciones de la materia hacen los bibliotecarios y los sociólogos de la literatura sociológica, por ejemplo, da cierta visión acerca de la situación (31): porque los documentos no se "describen" simplemente, sino que se evalúan en relación con su valor sociológico, los criterios de los sociólogos sobre materia eran los más precisos y útiles. Es banal descubrir que mientras mejores calificaciones tengamos en una disciplina académica, mejores serán los criterios acerca de las propiedades significativas de un libro de ese campo; y también se cumple lo contrario: mientras más pobres sean las calificaciones, más casual y superficial será la valoración y las propiedades que se resaltan.

En esta sección hemos visto un ejemplo de cómo dos teorías epistemológicas (el positivismo y el materialismo) ven el papel de la descripción en el desarrollo del conocimiento, y a partir de este ejemplo hemos visto el papel fundamental que juega la epistemología en la evaluación de las materias, y la forma en que los mismos problemas teóricos que ocurren en relación con los objetos materiales ocurren también en relación con el papel de los documentos en el desarrollo del conocimiento.

Es naturalmente decisivo para una teoría de la materia reconocer cómo distinguir entre las propiedades superficiales y las propiedades accidentales por una parte, y las propiedades significativas por la otra. Nuevamente esto se convierte en un problema básico para la epistemología(tanto como el problema del método científico). Así como resulta inútil describir la flora por sus características superficiales (como el color) en lugar de por sus características significativas (por ejemplo la categorización de las plantas con semillas o con esporas), resulta naturalmente necesario describir los documentos de acuerdo a las características significativas y no a las superficiales. De ahí que lo que se necesita es una teoría epistemológica que facilite el desarrollo del conocimiento encaminado a la esencia de las cosas. *Tal teoría se opone claramente a las concepciones que se basan en la investigación y el análisis de las materias como algoritmos, una "trampa" o método a priori.* Es el método el que debe ser una reflexión de la esencia del objeto.

La teoría materialista, en contraposición a la teoría pragmática, se caracteriza por un interés más amplio y más previsor en la epistemología. La teoría realista y materialista del concepto de materia no intenta simplemente resolver los problemas limitados del ahora y el aquí, sino que espera contribuir con la mayor

conciencia posible a las consecuencias de largo plazo. Las materias no deben estructurarse simplemente de forma estrechamente instrumental, sino que debe intentarse, por ejemplo, lograr contribuir a una penetración más profunda de las ciencias en la esencia íntima de la realidad. Las categorías de la materia deben mostrar esto de forma tal que reflejen los aspectos significativos y generales de la realidad. *En la práctica la teoría materialista de la materia frecuentemente opera con los conceptos de las ciencias, porque las ciencias son los órganos cognoscitivos de la sociedad.* (Nota 10) Por supuesto, las ciencias no son naturalmente ni controvertidas, ni objetivas, ni infalibles, pero como ideal, el debate acerca de la objetividad de la investigación científica es parte de la ciencia. (Nota 11) *Por tanto el análisis de la materia es en sí, y en su forma más profunda, parte del proceso científico de la recopilación del conocimiento.* Este análisis depende de factores contextuales, entre los que se incluyen el volumen de la literatura y el sistema de puntos de acceso. (Nota 12)

NOTAS

Nota 1:

¿Cuál es la ocurrencia del concepto “acerca de”? en la base de datos *Library and Information Science Abstracts (USA):*

S1 5504 ABOUT?= ocurrencias total en julio de 1989
S2 560 1/DE,TI= ocurrencias en los títulos y descriptores
S3 74 2/DE = ocurrencias como descriptores
S4 68 PY-1989
S5 2865 PY-1988
S6 5744 PY-1987(el chequeo del manual muestra que las ocurrencias de título crean ruido)
S7 5872 PY-1986
S8 5392 PY-1985
S9 5933 PY-1984
S10 5986 PY-1983
S11 5963 PT-1982
S12 5651 PY- 1981 La cifra total de referencias en LISA distribuida por año de impresión
S13 5469 PY-1980
S14 5388 PY-1979
S15 4506 PY-1978
S16 4171 PY-1977
S17 3790 PY-1976
S18 3681 PY-1975
S19 2695 PY-1974
S20 2978 PY-1973
S21 2985 PY-1972
S22 2576 PY-1971
S23 0 S4 y S3 1989 ¿El descriptor “acerca de”?
S24 1 S5 y S3 1988 distribuido por año de impresión
S25 1 S6 y S3 1987
S26 1 S7 y S3 1986
S27 0 S8 y S3 1985
S28 0 S9 y S3 1984
S29 0 S10 y S3 1983
S30 0 S11 y S3 1982
S31 0 S12 y S3 1981
S32 1 S13 y S3 1980
S33 1 S14 y S3 1979 Muestra que el uso del descriptor se concentra en los años
S34 2 S15 y S3 1978
S35 2 S16 y S3 1977
S36 11 S17 y S3 1976 interpretado como novedad

S37 30 S18 y S3 que no ha comprendido
S38 15 S19 y S3 1974
S39 3   S20 y S3 1973
S40 0   S21 y S3 1972
S41 0   S22 y S3 1971
S42 0   S3 y PY= 1970
S43 0   S3 y PY=1969

Nota 2:

La etimología del concepto “materia” (“emne” escandinavo).

*Nudansk prdbog (*13 udgave) sostiene que la palabra “emne” fue tomada prestada alrededor del 1760 del “emne” noruego o del sueco “amne”, la misma palabra que “evne”. Menciona tres significados de los cuales sólo los dos primeros son de interés en relación con lo que estamos tratando: 1.material para tratar en el discurso o la escritura; tema; motivo; 2. materia (“materia prima”), que está en parte elaborada, ej. acerca de las palabras claves antes del archivo final. *Nusvensk ordbok* menciona cuatro significados de los cuales el primero “materia prima”, “algo para producir a partir de”.

“Emne” puede traducirse como “materia” en inglés. El concepto de “materia” tiene el el *Oxford English Dictionary,* segunda edición, 18 significados principales. Resulta complicado que el término “materia” en inglés tenga tantos significados, entre ellos el “subjekt” danés (es decir, “sujeto” gramatica). De los 18 significados en el *OED* debe mencionarse el siguiente:

5.La sustancia de que está constituida una cosa o de la que está hecha.

7.Lógica a. Aquello que tiene atributos; aquello de lo cual se establece un criterio. b. El término o parte de una proposición de la cual se afirma o se niega un predicado.

8.Gram. El miembro o parte de una oración que denota aquello que está relacionado, que tiene predicado (es decir, de lo cual se afirma algo, se pregunta algo o se expresa un deseo); una palabra o grupo de palabras que denota(n) de lo que se está hablando y que constituye el “nominativo” para un verbo finito.

9. Filosofía moderna. Una conciencia o materia de pensamiento más completa: El pensamiento, como la “materia” a la cual son inherentes las ideas; aquello a lo que se le atribuyen todas las representaciones u operaciones mentales; el pensamiento o agente cognoscitivo; el yo o ego (correlativo al *objeto* sb.6).

(Los significados 5, 7, 8 y 9 se obtienen de “subjectum” latín del uso que hace Aristóteles con los significados: 1. materia de la cual están constituidas las cosas; 2. sujeto de atributos (cualidades); 3. sujeto de predicado (nombres).

10. El asunto o tema de un arte o una ciencia

12a. Aquello sobre lo cual se opera o actúa; una persona o cosa a la cual se dirige la acción o la influencia, o que es receptor de algún tratamiento

13a. En un sentido especializado: aquello que forma o que se escoge como tema del pensamiento, consideración o investigación; un tópico, un tema.

14a. El tema de una composición literaria; acerca de lo que trata un libro, un poema. etc.

18. atrib. y Comb.... (sentido 14, principalmente con referencia la catalogación de libros de acuerdo a sus materias) tarjeta de materia, catálogo, entrada, encabezamiento, índice, lista, referencia,...

El significado que tiene para nosotros un interés especial es, por supuesto, el 14 (y las combinaciones18), o sea que la “materia” es “de lo que se trata el libro”. Esta definición sin embargo, no resuelve el problema. ¿Qué quiere decir que el libro sea sobre una materia x? De acuerdo a los significados 12a y 13a y las definiciones arriba mencionadas del danés y el sueco tenemos evidencias de que nuestro concepto de “materia” o de “emne” es una “materia prima” sobre la cual actúan los humanos.

En la terminología alemana, verá que los índices de materia en las bibliotecas, libros, etc. reciben con frecuencia el nombre de “Sach-” o “Fachregister”. “Fach” es una referencia a profesiones o disciplinas científicas. Eso quiere decir que en alemán existe una conexión directa entre la terminología utilizada para nuestra “materia” y los grupos sociales que

puedan estar utilizando estos documentos. Es decir, el concepto de "materia" no tiene un equivalente preciso en alemán, pero los conceptos correspondientes apuntan a la función para referir los documentos a categorías de usuarios.
El significado etimológico de "materia prima" resalta el hecho que no son las propiedades innatas en las cosas en sí, sino sus funciones para el usuario humano las que constituyen las "materias" .
(En el artículo 1 he comparado el concepto de "materia" con el concepto de "valor". Esto permite una mejor captación del significado de "materia"; el oro tiene su valor, no por sus propiedades químicas en sí (son necesarias: que el oro sea "precioso" se debe en parte al hecho de que no se corrociona con facilidad por las influencias químicas), sino por las condiciones culturales especiales. El "valor" no es una propiedad innata en las cosas, sino una función de las propiedades de las cosas y de la cultura humana.)
Por tanto, hemos visto que nuestra concepción de "materia" en la Bibliotecología y la Ciencia de la Información no se contrapone a los significados importantes de la lengua en general. De haber existido dicha contraposición, nuestra posición se hubiera debilitado porque tendríamos que haber fundamentado un uso especial de la palabra. Por supuesto no estamos planteando que el concepto general "emne" o "materia" no pueda tener otros significados también, como se aprecia en el *OED*, pero hacemos énfasis en una cara del concepto que apoya nuestros puntos teóricos.
Nota 3:
No todos los investigadores modernos son de la opinión de que las cosas existen objetivamente y tienen propiedades objetivas. Por ejemplo el influyente libro *Understanding computers and cognition: a new foundation for design*, escrito por Terry Winograd y Fernando Flores /32,p.73/ asume una posición contraria.
Nota 4:
Un ejemplo de predicado es "F es simétrico", en donde la propiedad de la simetría es un predicado para partes de un cuerpo que tienen una relación particular entre sí/33/
Nota 5:
Hay otros predicados de segundo grado además de la asignación de la materia. Si, por ejemplo, se dice que un documento se caracteriza por pertenecer a la escuela estructuralista, (y este criterio se plantea directamente a partir de las propiedades del documento), esta es una meta descripción que no es idéntica a una descripción de la materia, pero que a veces puede ser parte de la descripción de la materia. (Si la asignación de la materia se basa en uno de los atributos secundarios, puede en sí misma convertirse en atributo de tercer grado, pero eso no tenemos que tratarlo ahora aquí).
Nota 6:
Otro concepto significativo para el concepto de materia es el propio concepto de "concepto". En los últimos veinte años han ocurrido cambios significativos en relación con los *conceptos* en la investigación dentro de la sicología, la filosofía, y la lingüística. Estos progresos no pueden resumirse aquí, pero son de gran importancia para el significado del concepto de materia. Uno de los resultados es que algunos conceptos deben verse hoy como el resultado de un argumento inductivo. Smith/34,p.518/ ofrece el siguiente ejemplo:

El animal tenía originalmente las propiedades típicas de un pájaro.
El animal adquirió accidentalmente las propiedades típicas de un insecto..
El animal produjo una descendencia con propiedades típicas de un pájaro.
Este animal probablemente sea un pájaro.

Es decir que los seres humanos, cuando se enfrentan a un problema de categorización, son capaces de elevar las similitudes y emplear las deducciones, lo que requiere una facilidad para suposiciones posteriores. Esto entra en contradicción directa con el criterio expresado por Beghtol (35, p. 95-96) de que el clasificador juzga la relación de clase sobre *las bases de las similitudes entre los documentos. Aquí proponemos el criterio de que así como la investigación moderna en conceptos ha ido tras la similitud como único*

criterio válido para el concepto, es correspondientemente necesario ir tras la similitud de los documentos como el único criterio de las relaciones de materia

Nota 7:

Según mi experiencia muchas personas ven esta discusión como algo innecesariamente complicado. ¿Por qué no es posible apreciar las materias como propiedades más tangibles de los documentos? Esto por supuesto se cumple en la mayoría de los casos. Pero pienso que en el trabajo particular con el concepto de materia en la sicología y en las ciencias sociales se necesita una concepción mucho más abstracta y más complicada de materia que la que previamente se ha discutido en la literatura de la LIS. En el apéndice se dan ejemplos que profundizan el conocimiento de los problemas del análisis de la materia en la sicología y las ciencias sociales. Merece señalarse que las críticas de los conceptos de materia (por ejemplo la "cercaneidad") con frecuencia provienen de personas que tienen antecedentes en las ciencias sociales. (cf Swift y otros/36/). Esto por supuesto no significa que el concepto de materia que aquí proponemos tenga sólo validez para las ciencias sociales. Por el contrario, las necesidades de las ciencias sociales contribuyen a una generalización del concepto de materia de forma tal que sea fructífero en otras áreas. Una teoría general de la ciencia de la información tiene que basarse en una generalización de la experiencia y de las teorías dentro de disciplinas específicas(como opuesto a lo opuesto: que una teoría terminada esté forzada a campos específicos).

Nota 8:

Debo la expresión "es la práctica humana la que constituye una materia", a mi colega Anders Orom, quien la acuñó en respuesta a una presentación oral de mi teoría de la materia.

Nota 9:

Esta relación nos lleva a una nueva interrogante: ¿Hay documentos sin materia? En teoría tenemos que responder no a esta pregunta; no podemos imaginarnos documentos sin ningún potencial cognoscitivo. Y resulta una rara experiencia considerar en la práctica no asignar alguna designación de materia. En casos específicos la falta de posibilidades claras para la clasificación usualmente refleja que el documento en cuestión no es apropiado para adquirirlo o para incluirlo en una base de datos particular. De ahí que la falta de una "materia" usualmente exprese una inconsistencia entre las políticas de acceso y de indización.

Desafortunadamente puede ocurrir una contradicción en las descripciones de la materia. Los documentos a los que le corresponde un sistema de clasificación (o lenguaje -IR- de recuperación) reciben clasificaciones simples o pocas, que se corresponden con la categoría respectiva en el sistema. Los documentos vagos o penetrantes reciben muchas más clasificaciones por lo que logran una visibilidad involuntaria. Este fenómeno tiene que contenerse. Los sistemas de información tienen que proporcionar un uso óptimo del conocimiento en el volumen de documentos recopilados. En el caso anterior el documento logra su visibilidad a costa de otros documentos: si todos los documentos se ubicaran en todas las categorías, todos los valores de categorización serían nulos e inútiles. También pueden ocurrir raras situaciones en donde la descripción temática de un documento cause más daño que beneficio, y esas descripciones deben evitarse.

Nota 10:

Existe, además del análisis temático para fines científicos e intelectuales, el análisis temático de una naturaleza más pragmática. El análisis temático de los documentos no siempre tiene que verse como un proceso científico de cognición, ya que con frecuencia la percepción/cognición también se extiende naturalmente y se hace pertinente a una percepción más ordinaria. Este criterio sobre el papel de las disciplinas científicas se opone a muchos científicos de la información quienes tratan de evitarlas y en su lugar describen los documentos de acuerdo a "categorías semánticas más fundamentales", por ejemplo el Classification Research Group.

Nota 11:
Este énfasis en las disciplinas y no en las "formas de conocimiento" o "tópicos" representa una alternativa para un criterio más difundido representado en la bibliotecología, por ejemplo en el reciente libro de Langridge *Subject analysis/37/.* Como este libro representa una teoría diferente acerca del análisis temático, yo debo hacer un breve comentario sobre el mismo.
Langridge analiza el concepto de materia en dos componentes principales:
(a) La tesis de que existen las categorías fundamentales del conocimiento resulta esencial para su libro. Estas son las categorías filosóficas que retoman a Platón y a Aristóteles y que han sido introducidas en la Ciencia de la Información y la Bibliotecología por S. R. Ranganathan. Langridge prefiere la expresión "formas del conocimiento" a estas categorías fundamentales.
Existen relativamente pocas "formas del conocimiento"; Langridge menciona doce, por ejemplo, la filosofía, las ciencias naturales, la tecnología, las ciencias humanísticas (conductuales o sociales), la historia, la religión, el arte, la crítica y la experiencia personal.
(b) Además de estas "formas de conocimiento" Langridge opera con "tópicos" que son "los fenómenos que percibimos". En donde las "ciencias humanas" son una "forma de conocimiento", la "conducta humana" un tópico.
Además de los dos componentes fundamentales existe un tercero:
(c) El concepto de disciplina (o "campo de aprendizaje") (p.31): desafortunadamente, esta extremadamente importante distinción ha sido borrada de la mente de muchas personas por la existencia de un tercer tipo que combina tanto la forma del conocimiento como el tópico. Por ejemplo, la ética es la filosofía (la forma) de la moral(el tópico); la zoología es la ciencia (la forma) de los animales (el tópico); la sicología es la ciencia (la forma) de la conducta humana (el tópico).
A Langridge no le gusta el concepto de disciplinas científicas como concepto en el análisis temático o de materia. Son inestables: "...las disciplinas que constituyen las especializaciones pueden resultar inestables, pero las disciplinas fundamentales, o las formas del conocimiento no lo son. Las especializaciones son conveniencias prácticas para compartir los trabajos intelectuales del mundo: las formas son permanentes, características inherentes del conocimiento"(p.32).
El concepto de Langridge de materia asume los antes mencionados "componentes fundamentales" como puntos de partida para el análisis temático o de materia. *No* hace referencia al contexto del usuario, al "criterio pragmático" del análisis temático.
En mi clasificación de las concepciones de "materia", la teoría de Langridge, en la tradición de Ranganathan, debe denominarse como "idealista objetiva".
Mi punto de vista difiere de otras formas:
Primero, en mi teoría las disciplinas son el punto central de partida. Con frecuencia son confusas e inestables, lo admito, pero son lo mejor que tenemos. Es trabajo de las propias disciplinas, conjuntamente con la filosofía (y espero que con la LIS), hacerlas más claras y mejor definidas.
Segundo, las categorías filosóficas fundamentales son importantes, pero epistemológicamente deben entenderse como generalizaciones de la investigación científica. La investigación científica, a propósito, no es sólo una investigación empírica, sino también teórica. No existen delimitaciones estrictas entre la ciencia y la filosofía. Estas categorías filosóficas son *relativamente* estables, pero *no* son "características permanentes e inherentes del conocimiento." (Yo interpreto esta afirmación como una clara posición idealista)
Tercero, el concepto de Langridge de "tópicos" como "fenómenos percibidos" representa la posición positivista, empírica e "idealista subjetiva" como punto de partida fundamental. Desde una posición "realista"(en el sentido platónico y escolástico) o "racionalista", es todo lo contrario: los fenómenos percibidos son incluidos en las "ideas inmortales."

Langridge al parecer asume una posición “racionalista” o “idealista objetiva”, en donde los “fenómenos percibidos” se incluyen dentro de “ideas inmortales”. Tanto los puntos de vista racionalista como empírico contienen parte de la verdad; es precisamente la inclinación hacia uno de estos dos criterios a expensas del otro lo que lleva al “idealismo subjetivo” o al “idealismo objetivos”.Las ciencias comienzan con fenómenos percibido como las flores(la botánica), las piedras (la geología), las estrellas (la astronomía), las sustancias químicas (la química), etc, pero con el desarrollo de las ciencias, los objetos percibidos se vuelven objetos más apercibidos. Las plantas, por ejemplo, se definen como organismos vivos con clorofila y la microbiología reconoce a organismos vivos que son tanto plantas como animales (porque tienen tanto boca como clorofila). Es decir: las cosas percibidas influyen en las ciencias y las “formas de conocimiento” (empiricismo) y el conocimiento teórico así obtenido cambian nuestras percepciones y nos permiten ver cosas nuevas (racionalismo).

Desde una posición realista materialista y moderna (”realismo calificado” frente al “realismo ingenuo”) las disciplinas científicas representan o reflejan el mundo, el mismo mundo que percibimos. Pero estas cuestiones son difíciles, y muchas ciencias tienen dificultades para determinar cuáles son sus materias. Esto tiene que esclarecerse, pero no es sensible a la ciencia de la información y la bibliotecología seguir su propio camino, tratar de avanzar sola y evitar las situaciones confusas mediante la elección de una teoría del conocimiento idealista en lugar de una materialista, basar su análisis de materia en “ideas inmortales” o en “fenómenos percibidos”.

Los “tópicos” o los “fenómenos percibidos” deben ser parte de la misma realidad del estudio de la ciencia. La percepción científica y la no científica deben incluirse en la dimensión teórica del análisis, en donde “la teoría de los niveles integrativos” constituye un buen punto de partida.

Langridge sigue la tradición de la Ciencia de la Información y la Bibliotecología, una línea más orientada hacia la biblioteca, con R:S:Ranganathan y el British Classification Research Group como figuras principales.

Esta tradición parece estar separada de otra vía de investigación, representada por ejemplo por *Language and representation in information retrieval* de Blair, podríamos decir, una línea más orientada hacia las bases de datos. Ambas tendencias están muy preocupadas por las cuestiones epistemológicas, y sus principales diferencias pudieran verse como diferentes posiciones epistemológicas, en donde la escuela de Ranganathan y sus seguidores representan una línea racionalista u “objetiva idealista”, mientras que Blair, siguiendo al último Wittgenstein representa un punto de vista pragmático.

En mi propia investigación trato de utilizar lo mejor de ambas tradiciones (y de otras también) e integrarlo en otra tradición epistemológica, la del materialismo/realismo. La selección de la posición epistemológica no constituye una “opción libre”. Una posición errada es científicamente improductiva y la investigación que se desarrolla en esa línea entrará en contradicción con la realidad y la investigación no florecerá, sino que representa un aliado ciego. Las posiciones epistemológicas por tanto no se escogen, sino que se trabajan en la investigación fundamental para resolver problemas teóricos. Una posición materialista o realista representa, por el contrario de lo que todo el mundo piensa, soluciones ya determinadas. Crea el camino para el trabajo teórico y empírico concreto.

Nota 12:

Principios de desarrollo para la descripción de la materia.

En la práctica, por supuesto, con frecuencia habrá varias descripciones de materia de un documento dado. Además de las descripciones de materia están también las propiedades del documento, por ejemplo, en forma de conceptos en las bases de datos (de título, texto completo u otras fuentes). La función de las descripciones de materia tienen que verse, naturalmente en relación con ese sistema de posibilidades. Esto pertenece a las cuestiones técnicas (“lenguajes de recuperación de la información”) que no trataremos

aquí. Lo que resulta significativo en relación con esto es que la explosión de la información (es decir, el crecimiento en la cantidad de documentos entre los cuales se debe discriminar) ha tenido consecuencias para los aspectos cualitativos dentro de la descripción de la materia. El usuario de los documentos llega a conocer, por supuesto en mayor o menor medida, las propiedades de los documentos. Sobre la base de esto, el propio usuario hace una evaluación de la materia. Mientras haya menos documentos involucrados, se podrán describir y analizar mejor las propiedades del documento, y más certera será la descripción de la materia. Muchos bibliotecarios y especialistas de la información mediante una comprensión implícita de esta situación podrán, por supuesto, brindar acceso a la mayor cantidad posible de propiedades del documento y responder por tantas propiedades como posibilidades prácticas le permita el sistema de materia. Mientras más amplia sea la cantidad de documentos en donde se realiza búsqueda, más difícil será localizar los documentos verdaderamente pertinentes. Por lo que será más conveniente que mientras más crezca la cantidad de documentos, más selectivas deben ser las descripciones de materia. En otras palabras: Mientras mayor sea la cantidad de documentos, mayor será la necesidad de realizar una verdadera descripción de materia y no un simple registro de las propiedades del documento.

Siempre que se descanse en la descripción de la materia y no se investiguen personalmente las propiedades primarias habrá una mayor incertidumbre en lo que respecta a que un predicado que predica sea más un producto indirecto que un predicado en sí. Por el contrario, descansar en la evaluación de la materia realizada por otros explota el servicio de valor añadido y ahorra tiempo. Los sistemas de información deben buscar una solución óptima a este dilema.

La hipótesis puede formularse de forma más precisa: *mientras mayor sea la cantidad de documentos, más necesario será describir sus materias sobre la base de las necesidades del usuario (en lugar de las propiedades de los documentos).* La multiplicidad de las propiedades y las relaciones entre ellas crea un exceso que deja al usuario sin poder determinar la pertinencia mediante el análisis de las propiedades. El peso es simplemente demasiado grande.

Un ejemplo que apoya esto es el desarrollo del índice de materia en *Kompas Danmark,* índice de productos sobre mercado danés. Mientras más productos aparezcan dentro de un campo, más se basan las descripciones en las necesidades de los usuarios. Hace treinta años, los productos químicos se describían principalmente por sus propiedades químicas, en la actualidad se describen más comúnmente por sus tipos de uso (por ejemplo, fertilizantes, químicos de fotografía, etc). El campo de la computación es una excepción de esta tendencia general, en él antes era común describir el hardware de acuerdo a un objetivo específico, pero en la actualidad la tendencia es hacer énfasis en la universalidad y describir las propiedades.

Otro ejemplo es la proposición de introducir el concepto de “funcionario civil político” en la administración central danesa. (Weekendavisen, 27.7.1990). Esto concuerda con nuestro criterio de que diferentes partidos políticos necesitarán “descripciones de materia diferentes” de la información existente, y mientras esta necesidad sea mayor, mayores serán las cantidades de información. La salvaguarda de los principios de la democracia descansa quizás no tanto en tener oficialmente funcionarios civiles “neutrales” y sistemas de información neutrales, sino en tener un análisis y sistemas de información consistentes que puedan garantizar verdaderas alternativas.

Los comentarios anteriores se incluyen aquí para demostrar que las materias no consisten en una función *a priori* de las propiedades de los documentos, sino que todo el contexto en el cual se lleva a cabo la descripción de la materia determina esta función, y que las regularidades pueden describirse aparentemente para determinar la dependencia de la descripción temática en los factores contextuales.

APÉNDICE

El análisis de materia: un ejemplo concreto

¿Cuál es la materia del libro de Robert A:Wicklund titulado: *Zero -variable theories and the psychology of the explainer (38)?*

De acuerdo al título del libro, este trata de ciertos tipos de teorías ("teorías de variable cero) y de "la sicología del que explica". La última materia está relacionada con la "sicología de la ciencia".

Si usted mira el libro, verá que "las teorías de variable cero no están evaluadas favorablemente; se explican como teorías simplistas y el libro trata de explicar el porqué estos tipos de teorías ocurren tanto en la sicología moderna. Por qué tantos psicólogos (o por qué tantos explicadores en general) tienden a utilizar estos tipos de teorías simplistas en nombre de teorías más variadas?

Usted puede leer en el prefacio del libro las siguientes oraciones: "El lector no debe suponer que este sea un libro sobre la filosofía de las ciencias sociales, o sobre un pronunciamiento moral sobre lo que es bueno o malo en la teorización antigua y en la moderna. Por el contrario, se invita al lector a considerar la parte psicológica del *que explica*".

Antes de presentar mi análisis de la materia del libro, miraremos el análisis de la Library of Congress (LC). En el *Cataloging-in-Publication Data* aparecen los siguientes términos: "1. Sicología. Filosofía, 2. Psicólogos, Sicología. 3. Explicación".

Esto quiere decir que la Library of Congress, en su primer campo de selección de términos de materia, no está en disposición de *seguir la afirmación de Wicklund en el prefacio,* mientras que los dos planteamientos siguientes puede decirse que están de acuerdo con la autocomprensión del libro. Esto se aplica especialmente a la última expresión temática.

Mi análisis de materia es el siguiente: Considero el libro importante porque trata sobre un tema olvidado en la investigación psicológica, o en la sicología como ciencia: la aparente decadencia del nivel teórico en la sicología. Diferentes análisis concretos de las teorías psicológicas ilustran esta condición, análisis que en la investigación psicológica posterior ha sido sustancialmente reducido. Un ejemplo de ellos es la casi clásica teoría de la personalidad por H. A. Murray de 1938.

En mi opinión, lo más esencial sobre el libro de Wicklund es en particular la documentación concreta de la evidente decadencia de la teoría psicológica. Hay muchos libros acerca de la filosofía y la metodología de la sicología que encaminan la sicología como ciencia, pero hay raros casos que documenten la evidente decadencia de la teoría. Parece como si la sicología no explotara lo mejor de su propia teoría y del conocimiento de la filosofía y de otras ciencias. ¿Cómo puede explicarse esto?.

La explicación que da Wicklund a esta condición evidente, según mi opinión, no es correcta. La explicación de Wicklund es diferente a mi forma de ver las cosas. Wicklund ve la fundamentación de la decadencia teórica como algo de poca importancia en este libro. Su interés principal es utilizar este material para explicar no sólo la condición de la sicología, sino la condición de la sicología de

los explicadores en general. El material que yo considero que tiene el mayor valor potencial es, para el autor del libro, solo un elemento simple.
Esto quiere decir que existe una marcada diferencia entre la opinión del autor y la mía acerca del valor potencial, el potencial epistemológico de este libro. Y por tanto acerca de cuál es su materia. El libro tiene, como cualquier libro, una ilimitada cantidad de propiedades. Analizar la materia de un libro es elegir las propiedades que tienen el mayor potencial para el conocimiento humano. De ahí que mi análisis de la materia sea uno distinto al del autor como lo indica el título y las palabras citadas en el prefacio. La razón por la cual el análisis de Wicklund y el mío sobre la materia central del libro sean tan diferentes descansa en mi evaluación profesional de la explicación de Wicklund, a la cual caracterizo como individualista: Wicklund busca una explicación para la decadencia de la teoría psicológica en los mecanismos psicológicos de las personas que producen estas teorías.
En realidad Wicklund en concordancia con su explicación, escribe acerca de fenómenos psicológicos interesantes y pertinentes(como es el caso del rumor y la competencia) que deben ser parte del patrón de explicación, pero en mi opinión, se necesita tomar como base una descripción cultural y social más amplia para poder entender estos mecanismos.
Pienso que los ejemplos documentados de la decadencia en la teoría psicológica puede en parte remontarse al mercado de los libros de sicología (y al mercado de los psicólogos). Durante un largo período después de la Segunda Guerra Mundial, el mercado de los libros psicológicos (y de los psicólogos) era el "mercado del vendedor", y era demasiado fácil vender incluso los libros de sicología más malos (y hacer investigación mala). Este fenómeno se describe en un artículo escrito por Jurgen Kagelmann, psicólogo consultante para la Psychologie Verlags Union, Munich, en la revista *Psychologie Heute,* octubre 1988. El criterio más importante de Kagelmann es que las posibilidades de ventas (demasiado) fáciles en los años 70 hicieron que ocurriera una abrumadora producción de libros de sicología de una calidad dudosa. Todo lo que pudiera imprimirse entre dos cubiertas era lanzado al mercado, y el mercado era insaciable. Este es un ejemplo de una explicación no individualista, que en mi opinión está más próxima a la verdad que la explicación de Wicklund, aún cuando esta no sea una explicación completa.
De ahí que yo considere que Wicklund tiene la tendencia de individualizar y sicologizar un problema social, y de esta forma su libro contiene una contradicción. Wicklund en este libro actúa además como un "explicador", y también se inclina a una teoría muy simplista y positivista la cual en realidad el libro debería atacar.
El potencial epistemológico del libro de Wicklund descansa, en mi opinión, especialmente en su documentación de ciertas condiciones de la ciencia psicológica que resulta importante establecer. De ahí que la *materia del libro sea la epistemología de la psycología,* la metodología, la teoría de la ciencia y la filosofía. En mi opinión, la LC estaba en lo cierto cuando hizo su primera selección de términos de materia (Psicología-filosofía), que como ya se mencionara estaba en contradicción con la afirmación de Wicklund en el prefacio.
Yo no_consideraría las "teorías de la variable cero" la materia del libro. Esto es escasamente un concepto con un futuro, ni siquiera una explicación de la decadencia en la teorización. Sigue siendo una pregunta abierta el hecho de que

lo que se ha dado en llamar “sicología variable” (39, p.522) sea o no un concepto valioso.

En lo que se refiere a la materia propuesta “sicología del explicador,” para mí constituye una cuestión teórica si la conducta de diferentes explicadores puede explicarse mediante los mismos mecanismos psicológicos sin importar lo que los explicadores están tratando de explicar. La cosa es si puede existir una “teoría del explicador”. Una teoría así tiene que incluir no sólo explicaciones de la conducta humana(es decir explicadores psicológicos, profesionales, así como legos), sino también otros tipos de explicación. Un libro así estaría realmente más cerca de una disciplina denominada “teoría de la decisión”, y de eso no se trata el libro de Wicklund. Yo llego a la conclusión de que tiendo a dudar del valor de la materia propuesta “sicología del explicador”. Esta duda incluye también el término de materia “Explicación” que da la LC. El libro de Wicklund es escasamente una contribución al concepto de explicación en general.

La última materia propuesta que quiero analizar es la “sicología de los psicólogos” (LC: ”Psychologists-Psychology”). Esa materia existe y se escriben libros sobre ella. Pueden describir, por ejemplo, la captación o reclutamiento de los psicólogos, la motivación que tienen para escoger la profesión, la socialización profesional y muchos otros temas. El libro de Wicklund no es, según mi opinión, de este tipo.

Considero, ya lo he señalado, que la materia del libro de Wicklund es la “filosofía y la epistemología de la sicología”. Mi juicio, por supuesto, es sujetivo y pudiera estar en general o en parte errado. La única forma de decidir esto es analizando los argumentos. Los argumentos acerca de la materia de un libro son fundamentalmente los mismos argumentos acerca del desarrollo del conocimiento.

La materia de un libro es su potencial epistemológico (objetivo). La descripción de materia que más se aproxime a la predicción del papel del documento en el desarrollo del documento es su más correcta descripción de materia. La evidencia de la verdad en la definición de la materia descansa en la argumentación. Si la argumentación antes expuesta no puede refutarse constituye entonces una sugerencia acerca de lo que trata la materia del libro de Wicklund mejor que la que ha ofrecido la LC. Si mi descripción de materia de determinado libro puede refutarse entonces está errada, pero esto no cambia mi teoría acerca de de lo que son las materias: el potencial de los documentos para el desarrollo del conocimiento.

REFERENCIAS

Artículo publicado en Journal of Documentation, vol 48, No. 2, june 1992, pp. 172-200. Traducción de Hilda Bello Quintana.

(1) Vygotsky, Lev Semenovich. Tankning og sprog. Bind 1-2.Kobenhavn: Hans Reitzel, 1982.

(2) Frohmann, Bernd. Rules of Indexing: Critique of mentalism in information retrieval theory, Journal of Documentation, 46(2),1990,81-101.

(3) Moller, Bente Ahlers. Vidensklassifikation. En Komparative analyse af Statsbibliuotekets systematiske katalog. Arthus: Statsbiblioteket, 1979.

(4) Mark Pejtersen, Annelise. The Meaning of “about” in fiction indexing and retrieval. Asñlib Proceedings, 31, 1979, 251-257.
(5) Mark Pejtersen, Annelise. Design of a classification, and use of the scheme for control of librarians´search strategies. En: Harbo O. y Kajberg, L. eds. Theory and application of information research: Proceedings of the Second Internatiponal Research Forum on Information Science, 3-6 agosto 1977, Royal School of Librarianship, Copenhagen, Londres, Mansell, 1980, 146-159.
(6) Belkin, Nicholas. The problem of “matching” in information retrieval. En: Harbo O y Kajberg, L eds. Theory and application of information research: Proceedings of the Second International Research Forum on Information Science, 3-6 agosto, Royal School of Librarianship, Copenhagen, Londres: Mansell, 1980, 187-197.
(7) Belkin, NJ. y Brooks, H:M. ASK for information retrieval: Part I. Backgorund and theory. Journal of Documentation, 38(2), 1982, 61-71.
(8) Belkin, N:J. ...ASK for information retrieval: Part II. Results of a design study. Journal of Documentation, 38(3), 1982, 145-164.
(9) Farradane, J:E. Fundamental fallacies and new needs in classification. En: The Sayers Memorial Volume. Londres. Lñibrary Association, 1961, 120-135.
(10) Farradane, J. E. Concept organization for information retrieval.InformationStorage and Retrieval, 3, 1967, 297-314.
(11) Wilson, Patrick.Two kinds of power: an essay on bibliographiical control. Berkeley: University of California Press, 1968.
(12) Gopinath, MA...Colon Classification. En: MALTBY, A. ed. Classification in the 1970s: a second look. Edición revisada. Londres. Clive Bingley, 1976, 51-80.
(13) Ranganathan, SR. Documentation and its facets. Londres: Asia Publishing House, 1963.
(14) Tranekjaer Rasmussen, Edgar. Bevidsthedsliv og erkendelse. Nogle psykologisk-erkendelsesteoretiske betragtninger. Festskrift udgivet af Kobenhavns Universitet I anledning af Hans Majestat Kongens Fodselsdag, 11 de marzo, 1956. Kobenhavn: Munksgaard, 1956.
(15) Johansen, Thomas. Indledende betragtninger over emners beslagtethed. Kobenhavn: Denmarks Bibliotesskole, 1975.
(16) Johansen, Thomas. An Outline of a non-linguistic approach to subject relations. International classification, 12(2), 73-79.
(17) Johansen Thomas. Elements of the non-linguistic approach to subject relationships. International Classification, 14(1), 1987, 11-18.
(18) Johansen Thomas. On the relationships of material subjects. International Classification, 14(3), 1987, 138-144.
(19) Johansen Thomas. Om sammensatte struktur. En: Orden I papirerneen hilsen til J:B: Friis Hansen. Redigeret af Ole Harbo og Lei Kajberg. Kobenhavn: Denmarks Biblioteksskole, 1989, 157-165.
(20) Steiger, Rolf. Zu philosophisch-weltanschaulichen Frageen der Informationssprachen.Informatik, 20, 1973, 52-55.
(21) Bookstein, Abraham....A decision theoretic fopundation for indexing. Journal of the American Society for Information science, 26(1), 1975, 45-50.
(22) Soergel, Dagobert.Organizing information: principles of database and retrieval systems. Londres: Academic Press,. 1985.
(23) Dahlberg, Ingetraut. Grundlagen universaler Wissensordnung. Problema und Moglichkeiten eines universalen Klassifikationssystem des Wissens. Munchen: Verlag Dokumentation, 1974.

(24) Popper, Karl.Objective knowledge: an evolutionary approach. Oxford: Clarendon Press, 1972.
(25) Rudd, David. Do we really need World III? Information Science with or without Popper. Journal of Information Science Principles and Practice, 7, 1983, 99-105.
(26) Hjörland, Birger. Indledende betraftninger over faglitteraturens typologi og udtrysformer.Blibliotеksarbejde, 29, 1990, 35-50.
(27) Spang-Hanssen, Henning. Kunnskapsorganisasjon, informationsgjenfinning, automatisering og sprak. En: Kunnskapsorganissasjon og informationsgjenfinning. Seminar arrangert 3-7, diciembre 1973 i samarbeid mellom Norsk hovedkomite for klassifikasjon, Statens Bibliotеksskole og Norsk Dokumentasjonsgruppe. Oslo: Riksbibliotekjenesten, 1974, 11-61. (Skrifter fra Riksbibliotektjenesten, Nr. 2)
(28) Boserup, Ivan. Hvad er emnedata? En: Emnedata I online-alderen. Under redaktion af Niels-Henrik Gylstorff, Niels C. Nielsen og Morten Laursen Vig. Danmarks ForskningsbiblioteksforeningsInternatmode Nyborg Strand 7-8. febrero 1984. Kobenhavn: Bibliotekscentralens Forlag, 1984, 31-42.
(29) Hjörland, Birger. Information Retrieval in psychology: implications of a case study. Behavioral and Social Sciences Librarian, 6(3/4), 1988, 39-64.
(30) Krober, Gunter...Beschreibung. En: Klaus, George...Marxistisch-Leninistisches Woterbuch der Philosophie I-IIIReinbeck bei Hamburg: Rowohlt, 1983, Band I, 214.
(31) Krarup, Karl..Reader oriented indexing: an investigation into the extent to which subject specialists should be used for the indexing of documents by and for professional readers, based on a sample of sociological documents indexed with the help of PRECIS indexing system. Copenjaguen: The Royal Library, 1982.
(32) Winograd, Terry...Understanding computers and cognition: a new foundation for design. Nueva York: Addison-Wseley, 1987.
(33) Segeth, Wolfgang En Klaus, Georg...Marxistisch-Leninistisches Worterbuch der philosophie I-III. Reinbeck bei Hamburg:Rowohlt, 1983, Band III, 961-962.
(34) Smith, Edward. Concepts and induction En: Posner, Michael, ed. Foundations of cognitive science. Cambridge, Mass, Londres. MIT, 1989, 501-526.
(35) Beghtol, Clare.Bibliographic classification theory and text linguistic: aboutness analysis, intertextality and the cognitive act of classifying documents. Journal of Documentation, 42, 1986, 84-113.
(36) Swift, D. F. ”Aboutness” as strategy for retrieval in the social sciences. Aslib Proceedings, 30, 1978, 182-187.
(37) Langridge, DW. Subject analysis: principles and procedures. Londres: Bowker Saur, 1989.
(38) Wicklund, Robert. Zero-variable theories and the psychology of the explainer. Berlin: Springer, 1990.
(39) Holzkamp, Klaus. Grundlegung der Psychologie. Frankfurt, 1983.

LA IDENTIFICACIÓN DE CONCEPTOS EN EL PROCESO DE ANÁLISIS DE MATERIAS PARA LA INDIZACIÓN

Mariâgela Spotti Lopes Fujita

Universidad Nacional del Estado de São Paulo (Brasil)

INTRODUCCIÓN

La indización como acto de construir índices es práctica bastante antigua en el tratamiento de los documentos. Sabemos que en las bibliotecas de la antigüedad ya existían listas de los documentos almacenados. A partir del momento que la ordenación de esas listas necesitó de una organización por materias se establecieron profundos cambios en el acercamiento al acto mecánico de construir índices, o sea, introducirse en el proceso de análisis del contenido de los documentos.

La indización, como proceso de análisis documental, es realizada más intensamente desde el aumento de las publicaciones periódicas y de la literatura científico-técnica en general, que impulsaron la necesidad de crear mecanismos de control bibliográfico en centros de documentación especializados. Así, las bibliografías como mecanismo de control bibliográfico surgen fuera del ámbito de las bibliotecas tradicionales y representan una evolución en las técnicas de tratamiento de la información, dando un impulso teórico-práctico, en aquella ocasión, a una nueva área: la Documentación.

La esencia de la evolución de las técnicas de tratamiento temático de la información está ligada al análisis documental como extensión del tratamiento temático que abarca la creación de resúmenes y la indización. La indización, según Chaumier (1980, p. 42), es la "parte más importante del análisis documental", es una combinación metodológica altamente estratégica entre el tratamiento del contenido de los documentos y su recuperación por parte del usuario. Además de estratégica, demuestra una estrecha relación entre el proceso y la finalidad de la indización.

Según el World Information System for Science and Technology (1) (1981, p. 84) hay que considerar la indización desde dos puntos de vista: en cuanto proceso, que consiste en describir e identificar un documento, y en cuanto a su finalidad, permitiendo buscar y acceder a la información almacenada.

La indización en el análisis documental, desde el punto de vista de los sistemas de información, también es reconocida como la parte más importante porque condiciona los resultados de una estrategia de búsqueda. Un buen o mal desempeño de la indización se refleja en la recuperación de la información a través de los índices.

Esto nos lleva a considerar que la recuperación del documento más pertinente es una cuestión de buscar a aquel cuya indización proporciona la identificación de conceptos más pertinentes a su

contenido, produciendo una correspondencia con la materia buscada en los índices.

En la identificación de conceptos, el indizador, después del examen del texto, pasa a abordarlo de una forma más lógica a fin de seleccionar los conceptos que mejor representen su contenido.

Entretanto, la identificación de conceptos, analizada desde los aspectos de la lectura, de la búsqueda por la tematicidad y de las concepciones de la lectura, agrega un considerable grado de complejidad que, ciertamente, acarrea dificultades al indizador, como verifica un estudio de observación de la lectura de cuatro indizadores del Centro Coordinador Nacional del Sistema Especializado en el Área de la Odontología, antigua Subred Nacional de Información en Ciencias de la Salud Bucal del convenio BIREME/KELLOGG/USP (Fujita, 1998), buscando analizar la importancia de la identificación de conceptos como estrategia de lectura.

Considerando que la identificación de conceptos, realizada durante la lectura documental, es el objetivo del análisis de contenido en la indización, las dificultades observadas posibilitaron la motivación para investigar, por medio de la revisión de la literatura, la identificación de conceptos a partir de la lectura documental, de la tematicidad y de las concepciones del análisis de materias.

IDENTIFICACIÓN DE CONCEPTOS EN LA INDIZACIÓN: CONCEPTUALIZACIÓN Y PROCESO

De acuerdo con los "Principios de indización" formulados por el World Information System for Science and Technology (1981, p.84), la indización es definida como "el acto de describir o identificar un documento en términos de su contenido", y para la norma 5693 de la International Standardization for Organization (1985, p. 2), la indización es "la representación del contenido de los documentos por medio de símbolos especiales, que extraídos del texto original, están recogidos en un lenguaje de información o de indización"

La indización, como operación de representación documental, se desarrolla de acuerdo con un proceso compuesto de operaciones básicas. Según los "Principios de indización" del World Information System for Science and Technology (1981, p.84), "durante la indización, los conceptos son extraídos del documento a través de un proceso de análisis, y después

traducidos para los términos de los instrumentos de indización (tales como tesauros, listas de encabezamientos de materias, esquemas de clasificación, etc.)

El proceso de indización, por tanto, comprende dos etapas: la analítica, en la que se realiza la comprensión del texto como un todo, la identificación y la selección de conceptos válidos para la indización, y la etapa de traducción, que consiste en la representación de conceptos en términos de un lenguaje de indización:

- **Determinación de la materia: establecimiento de los conceptos tratados en el documento.**
- **Representación de conceptos en términos de un lenguaje de indización: traducción de los conceptos en términos del lenguaje de indización.**

Según Vickery (1980), el proceso de indización se compone solo de una etapa, de sumarización, entre lo analítico y la traducción. Esta diferencia de etapas o estadíos se explica por el hecho de que Vickery desdobló la fase analítica en dos, de análisis y de sumarización, que puede ser entendida como síntesis. En los "Principios de indización" la fase de determinación de la materia, o análisis de materias, engloba la operación de síntesis, como se verá a continuación.

La primera fase, el análisis de materias, razón de nuestro estudio por abordar el proceso de lectura, se subdivide en otras tres:

- **comprensión del contenido del documento,**
- **identificación de los conceptos que representan ese contenido,**
- **selección de los conceptos válidos para la recuperación.**

En observación aparte, el texto de "Principios de indización" llama la atención sobre el hecho de que "en la práctica estos tres estadíos se superponen" (World Information System for Science and Technology,1981, p. 86), pero no explicita en qué momento. Más adelante observaremos que, de hecho, estos tres sub-estadíos se superponen durante la lectura del documento.

Para la comprensión del contenido del documento, el texto se refiere a documentos gráficos y no gráficos. Para los gráficos (libros, monografías, periódicos, informes, tesis) apunta la impracticabilidad de una lectura extensiva del texto, porque la considera ideal.

Así mismo, para que el indizador no obvie ninguna información relevante, apunta una serie de partes importantes del texto que merecen especial atención durante su lectura: título, introducción

y las primeras frases de capítulos y párrafos; ilustraciones, tablas, diagramas y sus explicaciones; palabras o grupos de palabras subrayadas o impresas con tipos diferentes. Después, advierte que los primeros epígrafes del texto presentan generalmente las intenciones del autor, y que las partes finales comunican el alcance de esas intenciones. Por eso, no recomienda la indización solamente por el título o por el resumen.

En la identificación de conceptos (segunda fase del establecimiento de conceptos), el indizador, después del examen del texto, pasa a abordarlo de una forma más lógica a fin de seleccionar los conceptos que mejor representen su contenido. Para eso, recomienda que la identificación de conceptos se haga obedeciendo a un esquema de categorías existente en el área cubierta por el documento, como por ejemplo: el fenómeno, el proceso, las propiedades, las operaciones, el material, el equipamiento, etc.

Aunque el texto "Principios de indización" no se refiere a la lectura durante las fases de identificación y selección de conceptos, es posible observar que está implícita en la frase citada anteriormente "después del examen del texto, pasa a abordarlo de una forma más lógica".

En la selección de conceptos es necesario tener en cuenta los objetivos para los cuales se indiza la información. Por tanto, no todos los conceptos identificados serán necesariamente seleccionados.

La publicación de los "Principios de indización" por el World Information System for Science and Technology (1981) generó la primera norma para el análisis, identificación de materias y selección de términos para la indización publicada por la ISO, con el número 5963 en 1985, con el título "Documentation-methods for examing documents, determining their subjects, and selecting indexing terms". En 1992 la Associação Brasileira de Normas Técnicas (ABNT) publicó la traducción de esa misma norma con el número 12676, titulada "Métodos para el análisis de documentos-determinación de sus materias y selección de términos de indización".

La Norma 12676 de la ABNT (1992, p.2) indica para la indización tres fases:

a) examen del documento y establecimiento de la materia de su contenido

b) identificación de los conceptos presentes en la materia

c) traducción de esos conceptos en los términos de un lenguaje de indización.

En el acápite "examen del documento", la Norma 12676, al mismo tiempo que considera ideal la lectura total del documento apunta su impracticabilidad operacional, ofreciendo al indizador la posibilidad de analizar el texto a través del examen cuidadoso de sus partes, como título, resumen, etc.

Al recomendar el examen del documento a través de una lectura de sus partes no especifica qué tipo de documento tiene esas partes, o si todos los documentos, indistintamente, la poseen. Advierte en nota al pie que "no se recomienda indizar por cualquiera de estos elementos aisladamente." (Associação Brasileira de Normas Técnicas, 1992, p. 2)

Después del examen del documento, la Norma 12676 define la fase de Identificación de Conceptos como: "un acercamiento sistemático para identificar aquellos conceptos que son los elementos esenciales en la descripción de la materia." (Associação Brasileira de Normas Técnicas, 1992, p. 2)

El acercamiento sistemático de la norma para identificación de conceptos, por tanto, va a ser como el "esquema de categorías existente en el área cubierta por el documento" propuesto en los "Principios de indización" (World Information System for Science and Technology, 1981, p. 87) porque recomienda un cuestionamiento del texto a través de preguntas preparadas para identificar determinados conceptos esenciales.

La identificación de conceptos, según la norma (Associação Brasileira de Normas Técnicas, 1992, p. 2), se realiza después del examen del documento cuando el indizador deberá seguir un acercamiento sistemático para la identificación de aquellos conceptos que son esenciales en la descripción de la materia. Obsérvese que la norma no se refiere a una interrupción de la lectura, pero sí al examen que corresponde a la exploración de partes del texto. La lectura, aquí, parece estar implícita porque la identificación de conceptos se hace por un cuestionamiento.

Se puede suponer, por tanto, que el acercamiento sistemático es un cuestionamiento que el indizador realiza para extraer mejor los conceptos cuando estuviera haciendo la lectura de las partes del texto, pero la norma brasileña no explicita qué cuestiones serían las más indicadas para cada parte del texto. Además de eso, no hace ninguna mención respecto a que la lectura pueda ser realizada mejor por el acercamiento sistémico del cuestionamiento y ser esta considerada como estrategia de lectura, recomendando que "después de examinar el documento" el indizador debe abordarlo sistemáticamente. ¿Eso podría significar que la identificación de conceptos es independiente de la lectura? Supongo que la norma brasileña considerase la identificación de conceptos independiente de la lectura, lo que ciertamente sería lo indicado, en la fase de "examen del documento", la necesidad de extraer, por lo menos, un enunciado de materia con el cual sería realizada la identificación de conceptos. Así, se presume que la identificación y selección de conceptos deba ser realizada durante la lectura.

La selección de conceptos, diferente de los "Principios de indización" (World Information System for Science and Technology, 1981), está incluida en la Norma 12676 en el acápite "Identificación de conceptos", recomendando que el indizador no necesita representar como términos de indización todos los conceptos identificados, sino que seleccionará los que se adecuen a los objetivos de uso de los términos.

Más adelante, la Norma incluye el acápite "Selección de términos de indización" relativo a la traducción de los conceptos en términos de indización y aquí queda establecida la diferencia entre la selección de los conceptos identificados durante la lectura y la selección de los términos de indización en los lenguajes documentales para la representación de los conceptos seleccionados.

En la etapa de "Traducción" de esos conceptos en el lenguaje de indización del sistema, la norma recomienda procedimientos para la verificación de los descriptores controlados y la preparación de una lista de aquellos términos para los cuales no hubiera, en el tesauro, una exacta representación de las materias encontradas en el documento.

La misma norma, en el acápite "control de calidad", apunta la necesidad de la calidad y consistencia de la indización y además relaciona los factores que la garantizan: la imparcialidad del indizador, el conocimiento del indizador sobre el campo que engloba los documentos a ser indizados, las ventajeas del contacto directo con el usuario y la receptividad de los lenguajes documentales para los nuevos términos.

Considerando aclarada la función de cada una de las etapas de la indización, es posible afirmar que una de las etapas consideradas más importantes del trabajo del indizador es el análisis de materias, que tiene como objetivo identificar y seleccionar los conceptos que representan la esencia de un documento. Se trata de un proceso complejo, pues exige esfuerzos del profesional (indizador) para seguir una metodología adecuada a fin de obtener resultados satisfactorios. La eficacia de ese trabajo puede ser avalada por los resultados obtenidos por los usuarios en el momento de la recuperación de la información.

ANÁLISIS DE MATERIAS PARA LA IDENTIFICACIÓN DE CONCEPTOS: UN ANÁLISIS DE CONCEPCIONES

En la literatura se verifica que la palabra materia tiene varias interpretaciones. En vista de eso, el proceso también puede ser denominado Análisis temático, Análisis documental, Análisis conceptual o Análisis de contenido.

Para Langridge (1977), citado por Albrechtsen (1993), el término Análisis de materia abarca el conocimiento del contenido de los documentos y la determinación de sus características significantes.

Chu y O´Brien (1993, p. 439) consideran el análisis de materia como la fase inicial del proceso de indización, el cual decidirá sobre los principales tópicos de la materia de un documento, precediendo a la fase de traducción de esos tópicos de acuerdo con el lenguaje documental adoptado por el sistema.

El término análisis de materia es el más utilizado, pero gran parte de los autores lo considera como la etapa de traducción de los conceptos extraídos de los

documentos para un vocabulario controlado en el proceso de indización como un todo.

Según Vickery (1968, p. 356), el análisis de materia es visto por su producto, o sea: “Análisis de información significa derivar de un documento el conjunto de palabras que sirven como una representación condensada de ese documento. Esta representación puede ser usada para identificar un documento, para proveer los puntos de acceso en la búsqueda, para indicar su contenido, o como sustituto del documento”.

Según Naves (2000), al análisis de materia lo envuelve una gran complejidad, pues, además del problema de la terminología, existe la influencia directa de las personas que ejecutan, conocida como subjetividad, por la cual el indizador interpone sus propios valores en su actuación de intermediario entre los autores y los usuarios, por lo que la tarea del indizador sea determinar, de forma precisa, el contenido del documento.

Para Cesarino y Pinto (1980), existen dos situaciones en las cuales los profesionales hacen análisis de materia. La primera sería al recibir un documento e insertarlo en el sistema de información. Hacen un análisis para determinar su contenido informativo, observando los objetivos del sistema y las necesidades de los usuarios. La segunda ocurre cuando, al recibir la solicitud de información hacen un análisis de esta con el objetivo de comprender la necesidad de información requerida por el usuario, identifican los conceptos existentes en la solicitud y traducen los mismos para el lenguaje del sistema. Estas dos situaciones tienen por objetivo identificar la necesidad informacional del usuario.

Se destaca, por tanto, que el proceso de análisis de materia reviste una subjetividad característica, dadas las circunstancias y elementos que intervienen, pues, a partir de la lectura del documento hecha por el indizador, se realiza un proceso de comunicación interactivo entre tres variables: lector, texto y contexto. Cada una de esas variables estará sujeta a diferentes condiciones, pero es el indizador como lector la variable más influyente en esa interacción para el análisis de materia, porque necesita realizar la comprensión de la lectura mediante su cognición.

Destacamos que, como el indizador tiene el objetivo de hacer la materia conocida para los usuarios interesados, la función de ese profesional es “aumentar la visión de lo que otros pueden leer en un texto”. (Hutchins, 1977, p. 19)

Para algunos autores, como veremos más adelante, el análisis de materia implica determinar la tematicidad del documento mediante la identificación y selección de los conceptos que componen el asunto o tema principal y los secundarios.

Siendo la materia la información relevante abordada en el texto, es preciso resaltar que la selección de la materia o información relevante sufre la influencia de la política de indización del sistema de información en el cual se inserta el indizador. La institución decidirá si el tema extraído del documento será más o menos específico, o si considerará un nivel más genérico, basándose en el perfil del usuario que establezca atender.

Conforme a Albrechtsen (1993, p. 221), dependiendo de los objetivos institucionales se considerará la concepción de análisis de materia que el sistema de información seguirá y, consecuentemente, el indizador atenderá ese aspecto en cuestión. Se consideran, así, diferentes concepciones de análisis que, ciertamente afectan el desempeño del indizador en tanto lector. Al respecto,

Albrechtsen (1993, p. 220), clasifica los diferentes puntos de vista en tres tipos de concepciones:

- Concepción simplista: Considera las materias como entidades objetivas absolutas, que pueden derivar de una abstracción lingüística del documento o de sumas, usando métodos estadísticos de indización. De acuerdo con esta concepción la indización puede ser totalmente automática.
- Concepción orientada al contenido: Abarca una interpretación del contenido del documento que va hacia los límites de la estructura superficial léxica o gramatical. El análisis de materias del contenido de los documentos abarca la identificación de los tópicos o materias que no están explícitamente colocados en la estructura textual superficial del documento, pero que son fácilmente percibidos por un indizador humano. Abarca, por tanto, una abstracción indirecta del documento.
- Concepción orientada por la demanda: Considera las materias como instrumentos para la transferencia del conocimiento, por tanto, direccionada por una finalidad pragmática de información y conocimiento. Conforme esta concepción, los documentos se crean para la comunicación del conocimiento, y las materias deben, por tanto, ser ajustadas para funcionar como instrumentos de mediación y transmisión de ese conocimiento para cualquier persona interesada. De esta forma, cuando el indizador analiza un documento no se concentra en representar o resumir la información implícita o explícita sino en cuestionarse cómo podría hacer ese contenido o parte de él, visible para el usuario potencial, qué términos debería utilizar para llevar ese contenido hasta el lector interesado.

Las tres concepciones tienen ventajas y desventajas. La principal ventaja de adoptar una concepción simplista se refiere al abaratamiento de computadoras y softwares, lo que significa una indización automática a bajo costo. Su principal inconveniente se refiere a la imposibilidad de la transferencia del conocimiento desde el punto de vista social.

La concepción orientada al contenido ayuda al mejoramiento de las técnicas y el trabajo del indizador, y puede ser simple al focalizar apenas la representación de los documentos sin considerar sus posibles usos.

La concepción orientada a la demanda tiene la ventaja de permitir la transferencia y la diseminación del conocimiento, pero conlleva un alto grado de responsabilidad al distinguir la prioridad de determinadas materias en usuarios potenciales para asegurar su utilización. En este caso se hace necesario enfatizar el examen de cómo se analiza el texto para la definición de una materia, teniendo en cuenta el contexto donde esté insertado.

Para Naves (1996), las dos últimas concepciones –indización orientada al contenido y la orientada a la demanda- son complementarias, y más que complementarias son intrínsecas, porque en el momento que el indizador está leyendo y procurando identificar y seleccionar los conceptos para la determinación de las materias del documento, está objetivando encontrar la materia que le es familiar debido a su práctica de indización y también definir lo que puede interesar al usuario del sistema de información.

La cuestión de la indización de materias del documento nos hace reafirmar que la actividad está vinculada a la lectura, aclarando que el indizador realiza dos operaciones, identificación y selección de conceptos, durante la misma y que la traducción de los términos que representan conceptos en descriptores del

lenguaje del sistema se debe hacer después de esto para que el análisis sea conceptual y marcado por la demanda. La preservación del contenido del documento es una garantía de la relevancia en la recuperación, objetivo de toda indización de contenido.

Esto nos remite a los principios de indización del sistema PRECIS, que en los inicios de la década del 70, preconizaba la "preservación del contexto" basado principalmente en un análisis cuya proposición era la identificación de conceptos a partir de la investigación de la estructura profunda del texto, usando un principio de organización de las materias en facetas de acuerdo con un orden de citación estructurado en una entrada de dos líneas y tres posiciones que revelaban la estructura superficial (Fujita, 1989). Este principio distinguía muy bien las dos partes para el funcionamiento del sistema PRECIS: la parte sintáctica, formada por la estructura de las entradas y la gramática compuesta de operadores de función atribuidos a los conceptos durante el análisis conceptual, y la parte semántica en que los términos identificados por los operadores eran traducidos por términos de un vocabulario controlado. El análisis conceptual del PRECIS se basaba en la idea que el autor utilizaba la terminología sin ningún tipo de asociación con instrumentos de control del vocabulario.

Si el principio de "preservación del contexto" se basa principalmente en el análisis conceptual y tenía como proposición la identificación de conceptos a partir de la investigación de la estructura profunda del texto, la contribución de la lingüística textual, fundada en la gramática transformacional, se hace decisiva para la difícil trascripción entre forma y contenido (Pinto Molina, 1994). La base teórica del PRECIS es la gramática transformacional, el uso de su análisis conceptual garantiza que el contenido será representado en la identificación de conceptos.

En este momento conviene esclarecer la doble función de la selección de conceptos que ocurre en dos momentos diferentes del análisis de materias: durante la identificación de conceptos para la determinación de la materia y, después, la identificación de conceptos durante la traducción de los términos que representan los conceptos para los términos del lenguaje documental adoptado por el servicio de análisis. Este esclarecimiento respaldará e iluminará la continuidad de nuevos estudios en análisis documental, porque confirma que las concepciones de la lectura orientada al contenido y a las demandas deben ser intrínsecas, y deben caracterizar la lectura documental. La concepción de lectura orientada al contenido debe servir para la identificación de conceptos y la concepción de lectura orientada a la demanda debe servir para la selección de conceptos.

Regresando al proceso de análisis de documentos de la Norma 12676 de la Associação Brasileira de Normas Técnicas ya mencionada, verificamos que la identificación y la selección de conceptos son etapas del análisis de materias. Al recomendar un acercamiento sistemático para la identificación de conceptos, la Norma 12676 revela un fuerte indicativo de los principios de la concepción orientada al contenido, porque según Albrechtsen (1993, p. 220) "abarca la identificación de tópicos o materias que no están explícitamente colocados en la estructura textual superficial del documento".

Se puede observar también que la identificación de conceptos hecha por el acercamiento sistemático es también un análisis orientado al contenido.
Por tanto, los “Principios de UNISIST” y la Norma 12676 contienen las dos concepciones, la orientada al contenido, porque propone el acercamiento sistemático para la identificación de conceptos y la orientada a la demanda, porque orienta la selección de los contenidos conforme los objetivos de uso de los términos, confirmando el aspecto intrínseco de las dos concepciones de análisis de materias durante la lectura documental.

IDENTIFICACIÓN DE CONCEPTOS Y LA BÚSQUEDA POR LA TEMATICIDAD

Con la evidencia de la identificación de conceptos en el análisis de materias orientada al contenido y a la demanda, existe necesidad de orientación sobre ese proceso. Como ya nos habíamos referido a la sistemática identificación de conceptos por cuestionamiento, indicada por la Norma 12676 y ya habíamos comparado el análisis conceptual del PRECIS, considerando esta una buena orientación, no cuestionamos el orden y la relevancia de los conceptos conforme a la tematicidad intrínseca (relevancia autor) y a la tematicidad extrínseca (salida lector) del contenido del documento.
El sistema PRECIS fue influido por el “análisis de facetas” ideado por Ranganathan (1960), un bibliotecario indio que concebía el contenido de un documento como un conjunto de materias específicas relacionadas entre sí desde una perspectiva particular. Dentro de esa perspectiva facetada, Derek Austin, como creador del PRECIS, propone el análisis conceptual del PRECIS con elementos del análisis en facetas.
Al demostrar la concepción del sistema de indización, Fujita (1989, p. 5) relata una síntesis evolutiva de doscientos años de estudios realizados en torno a la indización alfabética de materias, desde la publicación en 1876 de la obra básica de Charles Ammi Cutter, “Rules for dictionary catalog” hasta la creación del sistema de indización PRECIS por Derek Austin en 1974.
Por esta síntesis evolutiva es posible observar que la preocupación principal de los estudios es el producto final, o sea la generación del índice, el análisis que abarca la transformación del contenido en el índice se expresa tanto por la proposición de categorías como por los sistemas de indización, como se observa en la evolución de los principales estudios teóricos:

- J. Kaiser (1911): Con la publicación del trabajo “Systematic Indexing” propone el análisis de materias por la combinación de tres categorías: “concreto”, “proceso” y “lugar”.
- S. R. Ranganathan (1933): Desarrolla un esquema basado en el análisis de facetas y en el uso de cinco categorías: personalidad, materia, energía, espacio y tiempo.
- E. J. Coates (1960): En su libro “Subject Catalogues” presenta la formulación de encabezamientos de materias específicos por categorías: cosa-parte-materia-acción.
- J. W. Metcalfe (1959): Admite que la entrada debe ser directa y discute el propósito de la catalogación de materia en el sentido de indicar solamente la clase de materia en la que está insertado.
- M. F. Lynch (1973): Creó y desarrolló los índices articulados de materias para el Chemical Abstracts.

- J. E. L. Farradane (1977): Creó un sistema de indización que adopta nueve operadores relacionales para indicar las relaciones entre los términos en etapas de discriminación en el espacio y en el tiempo.
- POPSI (Postulated based Permuted Subject Indexing Language), creado por Neelameghan y Gopinath (1975), es un sistema basado en principios clasificatorios que utiliza encabezamientos de clasificación como términos de entrada en la producción de los índices cuyos patrones se derivan de la clasificación de dos puntos de Ranganathan.
- T. C. Craven (1978): Creó inicialmente el sistema NEPHIS (Nested Phrase Indexing System) y después, como consecuencia de una evolución experimental, el sistema LIPHIS (Linked Phrase Indexing System). Ambos son sistemas de indización automática.
- Derek Austin (1974): Creó para la British National Bibliography (BNB) el PRECIS, cuyo funcionamiento se basa en estructuras semánticas y sintácticas y en un esquema de operadores lógicos.

La noción de índice siempre estuvo muy ligada al proceso de indización. Los índices que antes existían en los sistemas de recuperación de la información, tales como los antiguos catálogos de fichas de las bibliotecas, fueron considerados dentro de una perspectiva clasificatoria, porque los llamados encabezamientos de materias se habían creado con una influencia de la terminología clasificatoria y no del texto y su contenido.

El gran elemento transformador dentro de la indización alfabética en los estudios teóricos fue el análisis en facetas propuesta por Kaiser, Ranganathan y sus seguidores, marcando la posibilidad de mayor especificidad y uniformidad con el uso de conceptos esenciales: espacio, tiempo, proceso, cosa, acción, etc.

Después de Ranganathan, el Classification Research Group desarrolló la aplicación de los estudios del análisis en facetas asumiendo la influencia de la clasificación facetada y pasó a utilizar y desarrollar una metodología facetada (Piedade, 1983, p. 79-80). Vickery (1975, p. 181-189), citado por Esteban Navarro (1999, p. 74), por ejemplo, amplió la cantidad de facetas propuestas por Ranganathan: personalidad, materia, energía, espacio y tiempo (PMEST) a tipo, estructura, constituyentes, propiedades, procesos, operaciones, técnicas, generalidades.

Según Esteban Navarro (1999, p. 73), “la faceta permite describir las relaciones que mantienen entre sí los conceptos mediante la formulación de una serie de preguntas peculiares para el dominio disciplinar en que se sitúa la materia del documento.”

En ese sentido, las facetas relacionadas con la materia “materiales dentales”, por ejemplo, serían reveladas a partir de los siguientes conceptos:

- tipo de materiales dentales: materiales dentales metálicos y materiales dentales no metálicos
- constituyentes: oro, aluminio, porcelana, plata
- propiedades: resistencia a la fractura, fotoelasticidad, rigidez
- procesos: amalgamación, polimerización
- técnicas de laboratorio: Fase Gama, etc.

En la visión de Esteban Navarro (1999, p. 79), la identificación de conceptos en la indización debe utilizar preguntas construidas sobre la base del “análisis de facetas que caracterizan un

conjunto de materias relacionadas entre sí". Si retomamos la recomendación de los "Principios de indización" del World Information System for Science and Technology y de la Norma12676 para la identificación de conceptos (2), veremos que la base es la misma, o sea, la influencia del análisis de facetas permea el proceso de análisis en la indización. Los autores citados mencionan la identificación del tema refiriéndose a conceptos, categorías y facetas, que podemos entender como la misma cosa, porque el tema está constituido por la presencia de conceptos.

En la metodología propuesta por Tálamo (1987), el proceso de indización consiste en identificar el tema de un documento por medio de un mecanismo de preguntas y respuestas agrupadas por generalidades y que responden a cada una de las cuestiones fundamentales siguientes: Quién? (ser), Qué? (tema), Cómo? (modo), Dónde? (lugar) y Cuándo? (tiempo). Según la autora, identificando esta estructura temática se encontrará el objetivo principal del texto, esto es, las informaciones relevantes separándolas de las accesorias.

Kobashi (1994) se basa en el mecanismo de preguntas conceptuales de Lasswell (1971), utilizadas por García Gutiérrez y Lucas (1987), un método analítico con fines de indización que abarca: Who, What, When, Where, Why. Se destaca que la categoría Quién? no fue identificada en textos científico-técnicos, en tanto la categoría Qué? es esencial por ser un elemento nuclear de la estructura temática. Según la autora, las categorías Cuándo?, Dónde? y Cómo? son categorías accesorias a la principal, Qué?, pudiendo aparecer o no en el texto, independiente del "orden de procedencia entre ellas."

El tema, por tanto, posee una estructura temática compuesta de conceptos o categorías o facetas cuya identificación develará el análisis conceptual del documento. La composición de las categorías identificadas formulará el tema del documento en cuestión.

Respecto de "dónde" localizar conceptos, eso depende de la identificación de la estructura temática. Conforme a la legibilidad y a la estructura textual del documento, el tema podrá estar formulado de forma clara en el "objetivo" del trabajo y, cuando eso no suceda, será necesaria la identificación de los conceptos dentro de la estructura textual del documento.

En el punto de la determinación del tema, es necesario aclarar que se refiere a la determinación de la materia, de forma idéntica. Consideramos el término tema y no materia porque así aparece en la bibliografía.

Para algunos investigadores del área, es relevante, nos referimos a la tematicidad (aboutness) del documento cuando se busca investigar sobre la problemática de la identificación del tema. La tematicidad es pertinente al análisis de materias porque estamos tratando de su objetivo principal que es la identificación de la materia o tema mediante análisis conceptual compuesto de la identificación y la selección de conceptos. Podemos decir que la materia del documento o tematicidad del documento es el centro principal y el más carente de esclarecimientos en los estudios de análisis documental.

Conforme aclara Albrechtsen (1993), el concepto “aboutness” pasa a ser investigado en lugar del concepto “subject” que, más recientemente regresa en otras investigaciones.

Todd (1992, p, 101) afirma que el término “aboutness” es usado en la literatura más reciente como sinónimo de materia y relata, desde Cutter hasta Borko y Bernier, las nociones acerca del término materia:

- materia, desde el punto de vista de Cutter, es tema o tópico de investigación, estando o no determinado en el título.
- Kaiser describe materia como cosas en general, reales o imaginarias, y las condiciones que están ligadas a ellas; nociones establecidas por las categorías concreto y proceso.
- Ranganathan habla de contenido de un documento como un término asumido o aislado.
- Coates identifica la materia como una abstracción de la idea general que está contenida en el contexto de una unidad de la literatura.
- Vickery habla del tema de los cuales libros, partes de los libros, partes de artículos se escriben; como un agregado complejo de aspectos específicos; compuesto de términos elementales.
- Borko y Bernier definen la materia como “the foci of work”, o tema central para el cual la atención y los esfuerzos del autor fueron direccionados. Son aquellos aspectos del trabajo que contienen las ideas del asunto, explicaciones o interpretaciones, las cuales pueden ser indizadas.

Clarificando la definición anterior, afirmamos que tematicidad siempre será el contenido relevante del documento, aunque algunas variantes influyan en la determinación de ese contenido como los intereses informacionales de los usuarios del sistema de información, entre otras. Por tanto, escoger el tema de un documento siempre estará relacionado con los intereses de tales usuarios, independiente de la cantidad de informaciones referentes al tema seleccionado.

Todd (1992, p. 102) dice que, según la opinión de Wilson (1985), podemos entender también que el grado de relación entre tematicidad y significado es variable porque depende del “uso que la persona puede encontrar de la tematicidad del documento en una cierta época, y el mismo documento puede tener diferentes significados para el mismo lector en diferentes épocas, en tanto el documento posea una pertinencia fundamental”

El término “aboutness”, originario de la lengua inglesa y usado por Fairthorne (1969) y otros, puede significar “lo que trata un texto” en portugués. En lengua portuguesa hay divergencias entre los investigadores para traducir “aboutness”: para algunos puede ser tematicidad (3), por considerarlo un sustantivo ligado al término temático, en tanto otros adoptan “atinência” (pertinencia).(4)

El concepto de materia (subject), según Albrechtsen (1993, p. 219), pasa actualmente a constituirse como área central de estudio, puesto que el tratamiento realizado en torno al concepto de “aboutness” o tematicidad tendía a manejar documentos como fuentes aisladas del conocimiento, a excepción de Hutchins (1977), y Beghtol (1986). La reinserción del concepto de materia por Blair (1990), Hjörland (1992), Weinberg (1988) y Soergel (1985) enfatizó la función primaria de la indización para servir a la búsqueda por el conocimiento y recomendar que el indizador no debe direccionar exclusivamente sobre el contenido, sino anticipar el impacto y el valor de un determinado documento para uso potencial.

Bajo esta influencia, Begthol (1986) distingue entre “aboutness” y “meanings”; “aboutness” es el contenido intrínseco del documento, que independiente del uso temporal que un individuo pueda hacer del mismo en el análisis y que le hace posible una tematicidad relativamente permanente y un número variable de “meanings” (significados), puede ser medido de acuerdo con el uso particular del documento teniendo en cuenta los usuarios. En definitiva, por “aboutness” debe entenderse el contenido relativamente permanente del documento y por “meanings” el significado comprendido por el usuario.

Aclarando la definición de Beghtol (1986), afirmamos que la tematicidad siempre será el contenido relevante del documento, por eso algunas variables influirán en la determinación del contenido, como los intereses informacionales de los usuarios del sistema de información, entre otras. Por tanto, escoger la materia de un documento estará relacionado con los intereses de tales usuarios, independientemente de la cantidad de informaciones referentes a la materia seleccionada.

Otra forma de referirse a esta cuestión la presenta Cavalcanti (1989), investigadora del área de lectura en Lingüística Aplicada. Para la autora, la tematicidad intrínseca es tema importante para el autor y la tematicadad extrínseca es tema importante desde el punto de vista del lector. La autora denomina la tematicidad intrínseca “salida autor“ y a la extrínseca “relevancia lector”.

Hutchins (1977, p.33) destaca que el indizador, durante la búsqueda de la comprensión del texto, procura identificar materias familiares contrariamente al lector común que, durante la lectura normal, va a encontrar informaciones nuevas para ampliar su conocimiento sobre el tema. En opinión de la autora, el indizador, en el momento de la indización, procura identificar primeramente, de manera automática, temas familiares al conocimiento previo que adquirió sobre el área del tema en el cual trabaja pero es preciso también apuntar a los temas nuevos que puedan interesar a los usuarios del documento al mismo tiempo que posibilitan al indizador ampliar su vocabulario sobre la terminología del área.

Lo ideal es que haya una equivalencia entre la relevancia de la materia del documento tanto para el indizador como para el usuario. Para cumplir ese objetivo contribuirá elaborar informaciones documentales (índices y resúmenes) consistentes.

Considerando que la determinación de la tematicidad intrínseca y extrínseca es parte del análisis de materias, entendemos que el indizador, durante la lectura documental, realiza un análisis de materias en la fase inicial del proceso de indización.

Se entiende, hasta aquí, que la identificación y selección de conceptos son operaciones características de un análisis de materias cuya concepción está orientada al contenido y a la demanda y, por tanto, volcada a la preservación del contexto del documento antes de acometer la operación de traducir los términos para un lenguaje documental. Ocurre que la identificación y la selección de conceptos sirven para la composición de un enunciado temático que identifica la materia del texto, siendo esa una condición necesaria para la comprensión global del texto.

LA LECTURA EN EL PROCESO DE IDENTIFICACIÓN DE LOS CONCEPTOS

Considerando el hecho de que el contenido del documento estará mejor representado si la identificación y la selección de conceptos fueron realizadas dentro de la concepción orientada al contenido y a la demanda, eso nos lleva a la necesidad de la comprensión de la lectura por parte del indizador porque el análisis orientado al contenido supone la explicitación del significado del texto, una situación que no se resuelve sin que haya comprensión de la lectura.

Para analizar la importancia de la lectura para el análisis de materias en la indización, consideramos que la concepción orientada al contendido, está, de hecho, comprometida con una comprensión de la lectura para la identificación y selección de conceptos, toda vez que "abarca una interpretación del contenido del documento que va hacia los límites de la estructura superficial léxica y gramatical." (Albrechtsen 1993, p. 220)

En el análisis de la literatura, vemos que Foskett (1973), antes de los "Principios de indización" del World Information System for Science and Technology (1981) y de la Norma 12676, acreditaba que la determinación de la materia de un documento debería ser hecha por la lectura íntegra del documento, por eso, como consideraba que no había tiempo disponible para la lectura de todo el documento sugiere partes del texto a ser leídas, alertando que el título es, muchas veces escogido para llamar la atención y no para indicar la materia abordada.

De la misma forma, Chaumier (1980, p.43) consideró que, para los fines de la indización, el conocimiento del contenido del documento "se hace a través de una lectura rápida, o lectura en diagonal del documento" enumerando, a continuación, las partes para una lectura más precisa. Cuando aborda la selección de conceptos, no hace ninguna referencia a la lectura, afirmando que la selección depende de un "verdadero análisis conceptual del documento".

Van Slype (1991, p. 116), al estructurar el proceso de indización humana en cuatro etapas (examen del contenido del documento, selección de los conceptos, traducción de los conceptos en descriptores y establecimiento de las relaciones sintácticas entre los descriptores), recomienda al documentalista, en la primera fase, una lectura rápida "en diagonal" de las partes del documento para solamente ver de lo que trata; una selección de conceptos indica claramente que "en la medida que realiza la lectura, el documentalista identifica los conceptos" y, más adelante, esclarece cuáles nociones deben ser extraídas o seleccionadas.

La lectura para la indización, dentro de la Norma 12676, de los "Principios" del World Information System for Science and Technology y de los teóricos del área, exceptuando a Van Slype,

aparece completamente desvinculada del proceso de identificación y selección de conceptos, y es considerada como mero examen de las partes del documento.

Lo que es evidente, en la postura de Foskett (1973), Chaumier (1980) y de la Norma 127676, es el aspecto más técnico del examen del documento sin cuestionarse o identificar al indizador como lector.

En relación con la comprensión, podemos considerarla como condición de la lectura, o sea, no existe lectura sin comprensión. Cuando hablamos de lectura para la indización, podemos decir que el indizador necesita comprender el texto para identificar y seleccionar conceptos, pues solamente lo hará si hubiera comprensión. También en la visión de Farow (1995, p. 243): "el proceso de indización consiste en la comprensión del documento que va a ser indizado, seguido por la producción de un conjunto de términos de indización".

Pensando en términos de una lectura documental, Farrow (1991, p. 151) considera "razonable asumir que los indizadores comprenden el texto esencialmente del mismo modo que los lectores frecuentes", por lo que aborda la influencia de las condiciones de tiempo, objetivo definido, modelo a ser producido y áreas de materias específicas con estructura textual normalizada de los documentos inducidos por un proceso repetitivo y automático.

Lancaster (1993, p. 20-21) aborda la cuestión de la lectura en la indización cuando examina "La práctica de la indización" en la que se preocupa por las restricciones del tiempo y de la cantidad de documentos del servicio de indización al ponderar que "al indizador raramente le es dado el lujo de poder leer un documento de principio a fin". Luego se refiere a las partes más interesantes y oportunas en la lectura del indizador, saltando el examen de las partes del documento recomendado por la Norma 12676. De manera mucho más apropiada, considera como presupuesto el hecho "de que es posible leer el documento a indizar" y que "el motivo es la decisión sobre lo que se debe incluir en la indización" basándose en los intereses de la comunidad a la que sirve.

Lo más importante es la advertencia que Lancaster (1993, p. 22) hace a los indizadores que realizan "análisis conceptual" influidos por el vocabulario del sistema, pues juzga que eso compromete la representación del contenido del documento. Por eso critica la Norma 12676 (1992, p. 2) cuando afirma que "tanto el

análisis como la traducción deben ser realizadas con el auxilio de los instrumentos de indización".

De la misma forma creemos que el uso del vocabulario controlado durante la lectura para la identificación y selección de conceptos podría impedir la comprensión del contexto del documento y su adecuada representación, pues el hecho de que términos no asociados al vocabulario no serán considerados significativos establece que el indizador realice un análisis bastante superficial del contenido del documento, o sea, no está representando realmente las ideas del autor, sino apenas ajustando palabras, de una forma muy simplista.

Es necesario esclarecer que, cuando la Norma 12676 (1992, p. 2) en el acápite "Examen del documento", aborda la cuestión de que una lectura completa del documento es "impracticable y no siempre necesaria" para la indización, posiblemente está distinguiendo, en ese momento, la lectura documental del proceso global de lectura. Pero la norma no comenta los motivos por los cuales admite ser impracticable la lectura completa del documento, es razonable suponer que eso se debe al hecho de que el trabajo de un indizador no se restringe a pocos documentos si consideramos la totalidad del acervo de una biblioteca.

La actividad de indización se inicia con la lectura del documento a analizar. Esa lectura, la documental, difiere de la lectura común porque exige otros procedimientos, aunque los conocimientos necesarios para el buen entendimiento de un texto sean comunes.

En la lectura documental el lector es caracterizado como idizador, pudiendo ser un bibliotecario o un individuo con formación superior en el área de la materia con la cual trabaja (lector-especialista). Este lector-indizador tiene un objetivo definido: identificación y selección de conceptos que representen el contenido del texto y que coincidan con las necesidades de la comunidad usuaria del sistema de información, no siéndole posible realizar una "crítica de la ciencia y la validación de las contribuciones vinculadas por los textos" (Ginez de Lara, 1993, p. 50), considerando que, en la mayoría de las veces, no presenta un conocimiento especializado sobre la materia que indiza. Eso no significa que el indizador bibliotecario no pueda ser un especialista por no poseer formación superior en el área de la materia en que trabaja, por el contrario, puede convertirse en un

especialista con los años de práctica en la actividad de indización y cursos especializados.

La lectura realizada por el indizador es también caracterizada como una lectura dinámica, no habiendo necesidad de hacer una lectura del documento completo, excepto en la indización para la elaboración de índices internos o alfabéticos-remisivos (aquellos que están al final de un libro) porque, como bien afirma Collinson (1971, p. 18), "en primer lugar el libro deberá ser leído completamente, dos o tres veces, como un todo", demostrando que ese tipo de indización, exige del indizador una lectura completa de la obra.

Como se ha visto anteriormente, el indizador, por realizar una lectura con objetivos profesionales, sufrirá la presión de la falta de tiempo debido a la gran cantidad de material que necesita leer para indizar. En la lectura para los fines de la indización, por tanto, el lector-indizador deberá utilizar estrategias propias de la lectura documental que le faciliten lograr su objetivo.

CONSIDERACIONES FINALES

El análisis de materias es la etapa más importante del trabajo del indizador. Tiene como objetivo identificar y seleccionar los conceptos que representan la esencia de un documento. El proceso de identificación de conceptos tiene cierto grado de complejidad por exigir del indizador el uso de una metodología adecuada para garantizar buenos resultados en la recuperación, lo que presupone el conocimiento de acercamientos sistematizados en el texto. En el análisis de la lectura, la identificación de conceptos depende de la tematicidad del texto y está relacionada a la lectura del indizador y a sus concepciones de análisis de materias adquiridas por su formación, objetivos y políticas de indización.

La identificación y la selección de conceptos son operaciones características de un análisis de materias cuya concepción esté orientada al contenido y a la demanda. Por tanto, destinada a la preservación del contexto del documento antes de la traducción. Considerando el hecho de que el contenido del documento estará mejor representado si la identificación y la selección de conceptos fuera realizada dentro de la concepción orientada al contenido y a la demanda, eso nos lleva a la necesidad de la comprensión de la lectura por el indizador porque el análisis orientado al contenido presupone la explicitación del significado

del texto, una situación que no se resuelve sin que haya comprensión de la lectura.

Los aspectos teóricos que fundamentan los estudios sobre la lectura documental indican la importancia de la identificación de los conceptos en el análisis de materia. Por ello, se da especial importancia a los estudios sobre tematicidad que abarcan la determinación de las materias dentro del análisis y las propuestas metodológicas de la indización indicadas por la normalización y por los estudios teóricos. Los estudios sobre tematicidad revelan la necesidad de la distinción entre tematicidad intrínseca (aboutness- inherente al contenido del documento) y extrínseca (meanings- significado para el usuario del sistema) y comprueba la concepción del análisis de materia orientado a la demanda.

La selección de conceptos es parte integrante de la identificación de conceptos realizada durante el análisis de materia y existe para que el indizador pueda prever la adecuación de los conceptos representados en la recuperación conforme la demanda de los usuarios.

Se considera que la concepción de análisis está directamente vinculada con su formación educacional (concepción orientada al contenido) y con la postura del sistema de información (concepción orientada a la demanda) y no por el hecho de ser un lector menos o más hábil. Por eso se recomienda que la formación del indizador sea direccionada por la importancia de la identificación y la selección de conceptos hecha durante el análisis de materias conforme al uso de una metodología adecuada.

REFERENCIAS

Artículo publicado en: Revista Digital de Biblioteconomía y Ciencia de la Información, v.1, n.1, p. 60-90, jul/dic. Brasil, 2003
(Traducción no oficial de la compiladora.)

(1) Sistema Internacional vinculado a la UNESCO y conocido por las siglas UNISIST.

(2) Escoger los conceptos puede obedecer a un esquema de categorías reconocidas como importantes en el campo cubierto por el documento, ejemplo: el fenómeno o proceso, las propiedades, las operaciones, el material o el equipamiento, etc." (World Information System for Science and Technology, 1981, p. 87)

(3) Preferimos usar tematicidad por considerar que este término está más relacionado con la noción de tema del documento.

(4) El término "atinência" (pertinencia) fue empleado en la traducción del libro "Indización y resumen: teoría y práctica" de Lancaster y es utilizado por Naves en sus trabajos (1996, 2000).

BIBLIOGRAFÍA

ALBRETCHTSEN, H. Subject analysis and indexing: from automated indexing to domain analysis. The Indexer, London, v.18, n. 4, p. 219-24, 1993.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. NBR 12676: Métodos para análise de documentos - determinação de seus assuntos e seleção de termos de indexação. Rio de Janeiro, 1992. 4 p.

AUSTIN, D. PRECIS: a manual of concept analysis and subject indexing. London: Council of the British National Bibliography, 1974. 551 p.

BEGHTOL, C. Bibliographic classification theory and text linguistics: aboutness analysis, intertextuality and the cognitive act of classifying documents. Jornal of Documentation, London, v. 42, n. 2, p. 84-113, 1986.

BLAIR, D. C. Language and representation in information retrieval. Amsterdan: Elsevier Science Publisher, 1990.

CAVALCANTI, M. C. Interação leitortexto: aspectos de interpretação pragmática. Campinas: UNICAMP, 1989. 271 p.

CESARINO, M. A. N.; PINTO, M. C. M. F. Análise de assunto. Revista de Biblioteconomia de Brasília, Brasília, v. 8, n. 1, p. 32-43, jan./jun. 1980.

CHAUMIER, J. Travail et methodes du/de la documentaliste: connaissance du problème. Paris : ESF/Libraries Techniques, 1980. Exposé 3, Chap.3: L'indexation, p.42-7.

CHU, C. M.; O'BRIEN, A. Subject analysis: the critical first stage in indexing. Journal of Information Science, London, v. 19, n. 6, p. 439-54, 1993.

COLLINSON, R. L. Índices e indexação: guia para indexação de livros, e coleções de livros, periódicos, e coleções de livros, periódicos, partituras musicais, com uma
seção de referência e sugestões para leitura adicional. Tradução de Antônio Agenor Brinquet de Lemos. São Paulo: Polígono, [1971]. 223 p.

CRAVEN, T. C. Linked phrase indexing. Information Processing and Management, New York, v. 14, p. 469, 1978.

ESTEBAN NAVARRO, M. A. E. Elementos, actividades y criterios para la identificación, compreensión y selección de conceptos en la indización analítica. In: GARCIA MARCO, F. J. G. M. Organización del conocimiento en sistemas de información y documentación. Zaragoza: Capítulo Español de la ISKO, Universidad Carlos III de Madrid, 1999. v. 3, p. 69-93.

FAIRTHORNE, R. A. Content analysis, specification, and control. Annual Review of Information Science and Technology, Medford, NJ, v. 4, p. 73-109, 1969.

FARRADANE, J. A. A comparison of some computer produced permuted alphabetical subject indexes. International Classification, Munich, v. 4, n. 2, p. 94-101, 1977.

FARROW, J. All in the mind: concept analyis in indexing. The Indexer, v. 19, n.4, p.243-7, 1995.

FARROW, J. A cognitive process model of document indexing. Jornal of Documentation, London, v. 47, n. 2, p. 149-166, 1991.

FOSKET, A. C. A abordagem temática da informação. Tradução de Agenor de Briquet de Lemos. São Paulo: Polígono, 1973. Tradução de: Subject approach to subject information.

FUJITA, M. S. L. PRECIS na língua portuguesa: teoria e prática de indexação. Brasília: UnB/ABDF, 1989.
_____. A leitura em análise documentária. 1998. 184 f. Relatório final (Projeto Integrado de Pesquisa) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista; CNPq, Marília.
GARCIA GUTIERREZ, A.; LUCAS, R. Documentación automatizada de los medios informativos. Madrid: Paraninfo, 1987.
GINEZ DE LARA, M. L. A representação documentária: em jogo a significação. São Paulo, 1993. 133 f. Dissertação (Mestrado em Ciência da Comunicação)– Escola de Comunicação e Artes, Universidade de São Paulo.
HJÖRLAND, B. The concept of subject in information science. J. Doc., London, v. 48, n. 2, p.172-200, 1992.
HUTCHINS, W. K. On the problem of aboutness in document analysis. Journal of Informatics, v. 1, p. 17-35, 1977.
INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. Documentation - methods for examining documents, determining their subjects, and selecting indexing terms. Suíça: ISO, 1985. 5p. (ISO 5963-1985 (E))
KAISER, J. O. Systematic indexing. London: Pitman, 1911.
KOBASHI, N. Y. A elaboração de informações documentárias: em busca de uma metodologia. 1994. 195 f. Tese (Doutorado em Ciências da Comunicação) – Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo.
LANCASTER, F. W. Indexação e resumos: teoria e prática. Trad. de Antonio Agenor Briquet de Lemos. Brasília: Briquet de Lemos/Livros, 1993.
LANGRIDGE, D. Classificação: abordagem para estudantes de biblioteconomia. Tradução de Rosali P. Fernandes. Rio de Janeiro: Interciência,1977. 120 p.
LASSWELL, H. D. A estrutura e a função da comunicação na sociedade. In: COHN, G. Comunicação e indústria cultural. São Paulo: Nacional/EDUSP, 1971.
LYNCH, M. F.; PETRIE, J. H. A program suite for the production of articulated subject indexes. Computer Journal, Oxford, v. 16, p. 46-51, 1973.
METCALFE , J. Subject classifing and indexing of libraries and literature. New York: Scarecrow, 1959.
NAVES, M. M. L. Fatores interferentes no processo de análise de assunto: estudo de caso de indexadores. 2000. 257 f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte.
NAVES, M. M. L. Análise de assunto: concepções. Revista de Biblioteconomia de Brasília, Brasília, v. 20, n. 2, p. 215-226, jul./dez, 1996.
NEELAMEGHAM, A.; GOPINATH, M. A. Postulated-based permuted subject indexing (POPSI). Library Science with a slant to documentation, v. 12, n. 3, p. 79-87, 1975.
PIEDADE, M. A. R. Introdução à teoria da classificação. 2. ed. rev. aum. Rio de Janeiro: Interciência, 1983. 221 p.
PINTO MOLINA, M. Interdisciplinry approaches to the concept and practice of written text documentary content analysis (WTDCA). Jornal of Documentation, London, v. 50, n. 2, p. 111-133, jun. 1994.
RANGANATHAN, S. R. Colon Classification. E. Goldston: London, 1933.
SOERGEL, D. Organizing information –principles of database and retrieval systems. New York: Academic Press, 1985.

TÁLAMO, M. F. G. M. Elaboração de resumos. Escola de Comunicação e Artes, 1987. 14 f. Datilografado.
TODD, R. T. Academic indexing: what`s it all about? The Indexer, London, v. 18, n. 2, p. 101-104, apr. 1992.
VAN SLYPE, G. Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales. Trad. Pedro Hípola e Félix de Moya. Madrid: Fundación Germán Sánchez Ruipérez; Pirámide, 1991. 200p. Tradução de: Les languages d'indexation: conception, construction et utilisation dans les systèmes documentaires.
VICKERY, B. C. Analysis of information. In: KENT, A., LANCOUR, H. (Ed.). Encyclopedia of library and information science. New York: Decker, 1968. v. 1, p. 355-384.
_____. Classification and indexing in science. Londres: Butterworths Scientific Publications, 1975
_____. Classificação e indexação nas ciências. Tradução de Maria Christina Girão Pirolla. Rio de Janeiro: BNG/Brasilart, 1980. 274 p.
WEINBERG, B. H. Why indexing fails the researcher. The Indexer, London, v.16, n. 1, p. 3-6, 1988.
WILSON, P. Subject and the sense of position. In: CHAN, C. et al. Theory of subject analysis: a manual. Littleton, Colorado: Libraries Unlimited, 1985. p. 306-23.
WORLD INFORMATION SYSTEM FOR SCIENCE AND TECHNOLOGY. Princípios de indexação. R. Esc. Bibliotecon. UFMG, v. 10, n. 1, p. 83-94, 1981.

INDIZACIÓN

Rosa Giráldez Rodríguez
Universidad de La Habana (Cuba)

1. CONCEPTOS FUNDAMENTALES

Generalidades

La principal función de un sistema de información es poner a la disposición de los usuarios la información relevante a sus intereses para lo cual tiene que realizar diferentes procesos, uno de los más importantes que se realiza es la indización.

Con frecuencia la indización y la búsqueda se consideran como operaciones paralelas. Es decir, la indización como un proceso que se realiza con los documentos, y la búsqueda como un proceso que se realiza con las solicitudes. Esta consideración no es correcta. Por eso es necesario enfatizar que la indización es un proceso que se aplica tanto a los documentos que van a formar parte de la colección del sistema, como a las solicitudes de búsqueda que formulan los usuarios para recuperar determinada información relevante a sus intereses.

También es conveniente aclarar que el proceso de indización con frecuencia se utiliza en los diferentes departamentos de un centro de información como un medio de control operativo. Por ejemplo se pueden indizar los catálogos de editoras de libros en una cataloteca o sección de adquisición.

2. INDIZACIÓN DE LOS DOCUMENTOS

Aspectos básicos

La indización de los documentos es un proceso complejo, que forma parte del procesamiento de la información, por medio del cual se representan en algún portador material características esenciales de los documentos que permiten su posterior recuperación sin tener que revisar toda la colección.

Indizar es una forma de clasificar. También se podría decir que clasificar es una forma de indizar. Ahora bien, de forma convencional, en la asignatura indización se estudiarán otras variantes para asignar clases a los documentos, generalmente utilizando palabras del lenguaje natural.

Existen diferentes formas de indización de acuerdo con la característica esencial del documento que se utilice como clave de búsqueda para su posterior recuperación. Cada forma de indización dará lugar a un tipo de índice. La indización de materia o temática es la más importante y la que se estudiará con más detalle en este texto.

Lenguajes de recuperación de la información

Los lenguajes de recuperación de la información (LRI) son los lenguajes artificiales que se utilizan para indizar los documentos y las solicitudes de información y tienen tres componentes fundamentales:

VOCABULARIO
SINTAXIS
REGLAS PARA SU USO

Fases fundamentales de la indización de materia

La indización de materia se realizará a través de una serie de tareas que se agrupan en tres fases:
1) Análisis de contenido
- se revisa el documento para determinar su contenido.
- se toma la decisión sobre qué conceptos claves del contenido se van a extraer.
- se expresan los conceptos claves extraídos en los términos del autor o del propio indizador.
2) Traducción de los términos asignados en el análisis de contenido a los términos índices del vocabulario del lenguaje de la indización del sistema.
- se consulta el vocabulario controlado del sistema.
3) Organización del índice.
- se organizan, de acuerdo con la forma que se haya establecido, los términos utilizados para indizar los documentos de la colección y se obtiene el índice de materia.
Cada una de estas fases puede ser un proceso intelectual realizado por el hombre, o puede ser un proceso total o parcialmente automatizado.

La indización de materia como un proceso de comunicación
Un análisis teórico de la indización como un proceso de comunicación permite señalar algunos aspectos importantes:
a) Es necesario entender el significado del texto que se indiza y tomar decisiones válidas sobre su contenido teniendo en cuenta los intereses de los posibles usuarios.
b) Los términos índices sustituyen al texto completo del documento en el proceso de la búsqueda. Es decir, se asume que el acceso posterior a la información del documento, por medio de los términos índices, elimina la necesidad de tener que revisar los propios documentos para realizar la búsqueda.
c) Existe una interacción entre:
- los objetivos que se propone la información contenida en el documento (se reflejan de forma explícita o implícita los objetivos y criterios del autor)
- la información trasmitida por los términos índices y los usuarios de la información con sus sistemas de juicio de valores
d) El significado de los términos índices está relacionado con el contexto histórico social del usuario, su educación, experiencia y perfil profesional.
La fuente emisora original es el autor que trasmite un mensaje en el documento que escribe. En el contenido del documento quedan reflejados sus criterios, valoraciones, conocimientos e ideología. Generalmente también se escribe para determinado tipo de usuario con propósitos preestablecidos.
El indizador es al mismo tiempo, receptor y trasmisor. Primeramente recibe el documento, lo analiza y lo indiza. Ha sido receptor del mensaje original, lo ha transformado en los términos índices, y posteriormente lo trasmite al usuario. El usuario como receptor recibe este mensaje transformado, lo analiza lo interpreta de acuerdo a su sistema de juicio y valores. Posteriormente podrá tomar la decisión sobre si el documento original responde a sus necesidades de información.

La indización de la solicitud de búsqueda

La operación de búsqueda-recuperación para responder a una solicitud temática comprende las cuatro fases siguientes:
1ra. Fase: Indizar la solicitud y formular la prescripción de búsqueda
2da. Fase: Trazar la estrategia de búsqueda
3ra. Fase: Confrontar los términos índices de la prescripción de búsqueda con los términos del índice de materia de los documentos de la colección
4ta. Fase: Predecir la relevancia del documento

El proceso de indización de la solicitud generalmente es un proceso menos complejo que el correspondiente a la indización del documento. Si el contenido está explícitamente expuesto en la solicitud del usuario la indización se limita a la segunda fase de la indización de materia, es decir, se traduce el contenido conceptual a los términos del lenguaje del sistema.

Con frecuencia ocurre que los usuarios al formular su solicitud no expresan con toda claridad y precisión sus verdaderas necesidades de información. Para evitar las consecuencias negativas que esto pueda traer es conveniente que el trabajador de la información que va a realizar la búsqueda, tenga primeramente, un intercambio con el usuario. Esta entrevista le permitirá apreciar si la solicitud está correctamente expresada o sí por el contrario, es necesario redefinirla.
Seguidamente se explicará con más detalle solamente la primera fase de la operación de búsqueda-recuperación. La primera fase de la operación de búsqueda-recuperación puede desglosarse en los siguientes pasos:
a) Se traduce el contenido temático de la solicitud a los términos del lenguaje de indización
b) Se precisan otros detalles de la búsqueda
c) Con el resultado de las operaciones a) y b) se formula la prescripción de búsqueda, la cual puede contener otras indicaciones adicionales.

Variables asociadas con el proceso de indización

Existen diversas formas para realizar el proceso de indización, cada una de las cuales da lugar a un tipo de índice con características especiales en su construcción y en su aplicación como dispositivo de recuperación de la información.
Independientemente del sistema de indización que se utilice hay una serie de variables que inciden en el proceso y que, en gran medida, definen su calidad. Entre estas variables las más importantes son las siguientes:

- el indizador
- la colección de documentos
- la política y las reglas de indización

- grado de exhaustividad
- profundidad
- especificidad
- el lenguaje de indización

El indizador

El indizador es la persona que realiza el trabajo intelectual de la indización y puede considerarse el factor de mayor importancia de todos los que afectan la calidad de este proceso.

En el trabajo del indizador influyen, además de una serie de rasgos personales, su dominio de la actividad científico informativa y sus conocimientos sobre idiomas extranjeros y sobre la materia o materias de la colección de documentos que tiene que analizar.
En muchos sistemas se requieren especialistas de materia (químicos, biólogos, etc.) con conocimientos de ciencia de la información para que puedan indizar con calidad determinadas colecciones especializadas. Ahora bien, en la mayoría de las bibliotecas para clasificar y asignarle epígrafes de materia a los libros generalmente no se emplean especialistas de materia, sino trabajadores de la información con nivel universitario, los cuales en la mayoría de los casos producen índices de calidad. No obstante, en algunas ocasiones tienen que enfrentarse a situaciones a las que no pueden darle la mejor solución.
Un aspecto primordial para que la labor de indización sea fructífera es que el indizador conozca los intereses y las posibilidades reales de búsqueda de los usuarios que van a utilizar el índice. Es necesario que trate de penetrar en los procesos cognoscitivos que se desarrollan cuando un usuario utiliza el índice. Tiene que intentar hallar las respuestas a las siguientes interrogantes:
¿Cómo el usuario determina las palabras claves de su perfil de búsqueda?
¿Qué hace cuando no encuentra registros de su interés bajo una palabra clave?
¿Cómo modifica las palabras claves a medida que desarrolla la búsqueda y se enfrenta con los términos de entrada del índice?

El indizador, al estudiar la forma en que los usuarios utilizan el índice y valorar la efectividad del sistema para recuperar o no recuperar un documento determinado en respuesta a una solicitud dada, hace un estimado de los parámetros relacionados con el futuro uso del sistema e inserta esos parámetros en las reglas de decisión para asignar los términos índices a los documentos.

La colección de documentos
Por muchos esfuerzos que se desplieguen para realizar una indización de calidad no se podrá lograr ofrecer un buen servicio de información dentro de una rama del conocimiento si la colección no es adecuada o es insuficiente.
También es necesario considerar que en las condiciones actuales los servicios de información no pueden ser satisfactorios si no procesan las fuentes no publicadas (flujo ascendente de información) tales como tesis, informes de investigaciones parciales, ponencias, patentes, normas y otros.

La política de indización. Reglas de indización
El sistema de información traza la política de indización, la cual se traduce en una serie de lineamientos para guiar el trabajo del indizador con la finalidad de lograr elaborar índices que funcionen como dispositivos, lo más efectivos posibles en situaciones determinadas, para recuperar la información. Una parte de los lineamientos que emanan de la política de indización se convierten en reglas de indización, o sea en disposiciones concretas que deben cumplirse con exactitud.
La política de indización se traza teniendo en cuenta los intereses de los usuarios y el tipo y volumen de la colección de documentos.

La política de indización establece las pautas para determinar la exhaustividad, profundidad y especificidad de la indización.

Exhaustividad

La materia que abarca el contenido de un documento es la totalidad de tópicos que se tratan en el mismo. La exhaustividad en la indización de un documento se define como el número máximo de diferentes tópicos indizados. Por ejemplo, un documento trata sobre el tópico central A y tres tópicos colaterales B, C y D. Si se indizan los cuatro tópicos el grado de exhaustividad empleado para indizar este documento será máximo.

Profundidad

La profundidad de la indización se define como el número de diferentes términos seleccionados para indizar el documento. Esta variable también se denomina densidad de indización.

No existe necesariamente, una relación de igualdad entre la exhaustividad y la profundidad. Un término índice puede comprender varios tópicos o un solo tópico puede requerir varios términos. Así, por ejemplo, se podría haber decidido solamente indizar el asunto central A del documento. Al hacer el análisis de contenido se consideran todos los aspectos que se tratan sobre el tópico A, lo cual requiere que se asignen un total de 10 términos índices al documento. En este caso la exhaustividad no sería máxima, ya que no se indizaron los tópicos colaterales (B, C y D) del documento, pero la profundidad si sería máxima con respecto al asunto A, pues se indizaron todos los aspectos de ese tema.

Especificidad

La especificidad es una propiedad semántica de los términos, es el nivel de detalle y exactitud con que se representa un concepto dado. Para apreciar el verdadero significado de la especificidad es necesario tener en cuenta uno de los tipos más importantes de relación que existe entre los conceptos, es decir, la relación género/especie.

Por ejemplo, si BIBLIOTECAS representa el género, entonces los diferentes tipos de bibliotecas serán las especies:

BIBLIOTECAS ESCOLARES
BIBLIOTECAS NACIONALES
BIBLIOTECAS PÚBLICAS
BIBLIOTECAS UNIVERSITARIAS

La relación de una especie con su género es una relación de subordinación. Una especie está en un nivel genérico inferior que su género. Por tanto, si se nombra una especie en la indización se es más específico que si se nombra su género.

Estas características de la indización afectan medidas importantes de la efectividad de un sistema de información: el recobrado y la precisión. **Recobrado** es la capacidad del sistema para recuperar los documentos relevantes de un fondo en respuesta a una solicitud de información. **Precisión** es la capacidad del sistema para retener los documentos no relevantes en respuesta a una solicitud de información

No siempre los términos específicos están comprendidos en la relación género/especie. Por ejemplo, si se utilizan los términos VAGONES y LOCOMOTORAS en lugar de TRENES, se están usando términos más

específicos, pero los términos VAGONES y LOCOMOTORAS no son una especie de TRENES, sino que son partes de los trenes. En este caso la especificidad está comprendida dentro de la relación parte/todo.
En otros casos los términos más específicos representan conceptos más limitados que no caen exactamente dentro de la relación género/especie ni todo/parte, como en el caso que se utilice VOLTAJE en lugar de ELECTRICIDAD. El voltaje no es ni una especie ni una parte de la electricidad, es simplemente un aspecto más limitado dentro del estudio más amplio de la electricidad.

El lenguaje de indización

El lenguaje de indización influye de forma decisiva en los resultados de la indización. En los lenguajes se pueden desarrollar determinados mecanismos para elevar la efectividad del sistema.
En términos generales se puede decir que el vocabulario del lenguaje proporciona los términos que se pueden usar en la indización. Si es muy específico facilita la indización específica, si por el contrario, carece de especificidad se convertirá en un freno para la indización específica. Así por ejemplo, un indizador decide utilizar el término específico BIBLIOTECAS UNIVERSITARIAS al analizar un documento, pero al consultar el vocabulario constata que sólo aparece el término genérico BIBLIOTECAS. En este caso el lenguaje ha impedido que se indice el documento en forma específica.

Fundamentos metodológicos del proceso de indización

La indización es un proceso que comprende dos fases fundamentales, el cual se puede realizar siguiendo una metodología de trabajo que comprende varios pasos. No se puede establecer una guía de trabajo única, inflexible. Hay una serie de variantes que será necesario introducir acorde con el sistema de indización que se esté aplicando, con el lenguaje que se utilice. Por tanto, la metodología de trabajo que se aplicará en este texto puede servir de guía general para realizar el proceso de indización, pero será necesario tener en cuenta que en cada caso particular habrá que hacerle algunas modificaciones en correspondencia con los principios, objetivos y características del sistema de indización que se vaya a aplicar. En este caso la guía se ha elaborado suponiendo que se va a aplicar un sistema de indización que se compone de:
- un lenguaje de indización con un vocabulario autorizado formado por una lista alfabética de términos autorizados y los no autorizados (sinónimos, casi-sinónimos y otros) . Los términos no autorizados se presentan con una referencia cruzada de USE para indicar el término que debe usarse. Por ejemplo:

ENSEÑANZA A DISTANCIA Término autorizado

Estudios dirigidos Término no autorizado

USE
ENSEÑANZA A DISTANCIA Referencia cruzada

- una política de indización que ha trazado una serie de pautas de modo que la indización se realice de acuerdo con los intereses de los usuarios, con el tipo de documentos que se van a indizar, y con una profundidad tal que permite que a

cada documento analizado se asigne, en caso necesario, hasta un máximo de 8 términos índices como promedio.

Guía metodológica de trabajo

1. Se revisa el documento.
2. Se formula la interrogante ¿es valioso para la colección?

Hay que tomar la decisión de si se debe o no analizar el documento para indizarlo e incluirlo en la colección. Esta decisión se tomará considerando los intereses de los usuarios. Claro está que si la política de selección y adquisición ha sido adecuada los documentos que lleguen a la etapa del procesamiento analítico - sintético es porque son de interés para el sistema. De todos modos este paso es necesario ya que muchos de los trabajos que se van a analizar son artículos de revista. Una revista puede ser importante para el sistema, pero, no obstante, es posible que determinados artículos no respondan a los intereses de los usuarios.

Si el documento no es valioso no se analiza, es decir no se sigue el proceso . Se desvía a otro destino donde puede tener mayor utilidad o simplemente se elimina. Por supuesto que si el documento es un artículo de una revista, la cual tiene otros artículos que sí son de interés, no pueden ser desviados ya que sería absurdo mutilar la revista.

En los grandes sistemas integrales el personal que hace la selección desvía los documentos, de acuerdo con la rama del conocimiento, hacia los especialistas calificados para que los analicen.

3. Si el documento es de interés para la colección se anotan los datos bibliográficos en la hoja de trabajo (registro bibliográfico) de acuerdo con las reglamentaciones establecidas por el sistema.
4. Se analiza el contenido del documento y se asignan los términos para expresar los conceptos claves, utilizando las propias palabras del autor o del indizador. Este es el paso más importante y complejo de todo el proceso.
5. Se consultarán los términos asignados (TA) con el vocabulario autorizado (VA).
6. Con cada término asignado se plantea la pregunta: ¿Está el TA en el VA?
7. Si el término TA está en el VA se utiliza como término índice (TI) y se escribe en la hoja de trabajo .
8. Si el TA no está en el VA hay que plantearse la pregunta: ¿Es un identificador? si es un identificador, o sea un nombre propio de personas, institución, organización, se utiliza cómo TI y se escribe en la hoja de trabajo.
9. Si el TA no es un identificador hay que hacerse la pregunta: ¿Tiene una referencia de USE?
10. Si el TA no es un identificador, pero tiene una referencia de USE se busca el término autorizado correspondiente y se utiliza como TI añadiéndolo a la hoja de trabajo.
11. Si el TA no tiene referencia de USE se buscan posibles sinónimos en diccionarios, glosarios u otro tipo de repertorio.
12. ¿Se encuentra algún sinónimo?
13. Si se encuentra algún sinónimo hay que averiguar si está ó no en el VA. Si está en el VA se utiliza como TI y se anota en la hoja de trabajo.
14. Si no se encuentra un sinónimo (o casi-sinónimo) (o si el sinónimo encontrado no está en el VA) se estudia la posibilidad de incluir en el VA el TA en primera instancia (o el sinónimo encontrado que no está en el VA).
15. Hay que tomar la decisión si debe o no incluirse en el VA.

16. Si se toma la decisión de incluir el término en el VA se llena la tarjeta que ordena que sea incorporado el vocabulario y se utiliza como TI adicionándolo a la hoja.
17. Si se toma la decisión de no incluirlo en el VA no se utiliza como TI y se sigue el proceso con otro TA (paso 5).

3. ANÁLISIS DE CONTENIDO. FUNDAMENTOS METODOLÓGICOS

El proceso de indización de un documento es parte del mecanismo de recuperación de un sistema de información y determina, o por lo menos influye de modo notable, en la habilidad del sistema para responder a las solicitudes de los usuarios. La indización es una labor intelectual que comprende dos fases fundamentales. En la primera el indizador somete el contenido del texto en lenguaje natural a un análisis conceptual. Seguidamente, en la segunda fase, representa los conceptos claves extraídos en el lenguaje del sistema. Es decir, traduce los términos asignados en el análisis de contenido a los términos autorizados por el sistema.
El análisis de contenido comprende dos partes:
- análisis semántico: se identifican los conceptos
- análisis sintáctico: se establecen las relaciones entre los conceptos
No siempre se establecen las relaciones sintácticas entre los términos. En estos casos el análisis de contenido se reduce al análisis semántico o de identificación de los conceptos que deben ser indizados.
Para hacer el análisis de contenido se leen algunas partes del texto, los aspectos más destacados del documento: título, introducción, epígrafes o partes del trabajo, términos o expresiones en letras destacadas, conclusiones, encabezamientos de las Tablas y los diagramas y otras partes que se consideren de importancia. Generalmente no es necesario leer el texto línea por línea, aunque en algunos casos excepcionales de artículos de revistas será preciso leer una gran parte del material.
Es importante que el indizador entienda el contenido, el significado del texto que está analizando. De esto se deriva, que en la mayoría de los casos para indizar documentos en una rama especializada del conocimiento se requieran especialistas de la materia de que se trate. Además, el indizador tiene que tener la habilidad de hacer juicios críticos valiosos sobre los conceptos claves que debe indizar en relación con los intereses de los usuarios.
El indizador tiene que guiarse para su trabajo por la política y las reglas de indización del sistema:

- ¿Con qué exhaustividad debe analizar?

Tiene que determinar qué tópicos debe indizar.
- ¿Con qué profundidad debe analizar?

Tiene que determinar si sobre un tópico dado sólo se indiza el asunto central o también los asuntos secundarios. Asigna el número de términos que sea necesario sin pasarse del número máximo que esté establecido por el sistema.

- ¿Con qué especificidad debe analizar?

Tiene que determinar si debe utilizar el término más genérico, o el más específico, o un grado intermedio de especificidad, o si

utiliza el concepto genérico más el específico. También es posible que la decisión sobre la especificidad no sea igual para todos los tópicos, sino que se acondicione, en algunos casos, a la costumbre de los usuarios.

Con todas estas premisas, el indizador para tomar la decisión sobre qué conceptos clave debe extraer para su indización se formulará ante cada concepto X que identifica, las siguientes interrogantes:

- ¿El concepto X debe ser indizado?
- ¿El concepto X se encuentra dentro de los marcos de intereses de los usuarios?
- ¿El concepto X se trata con suficiente profundidad en este documento para que merezca ser indizado, o simplemente se menciona como algo secundario?
- ¿Cuántos aspectos del concepto X hay que señalar en la indización?

A medida que se toma la decisión de qué conceptos clave se deben extraer también se le asignan términos para nombrarlos, utilizando las palabras del propio indizador o las del autor del documento. Después, como se ha dicho anteriormente, el indizador tiene que seleccionar los términos índices del vocabulario autorizado para expresar los conceptos claves, es decir, tiene que traducir los términos asignados a los términos índices del lenguaje del sistema. Se formula esta interrogante:

¿El término índice debe asignarse al documento analizado?

En algunas ocasiones se puede tomar la decisión de asignar términos índices para expresar un juicio de valor sobre el documento en relación con aspectos que no están explícitamente expresados.

Diferentes tipos de conceptos en el análisis de contenido

En el análisis de contenido de los documentos se pueden identificar diferentes tipos de conceptos.

En un documento titulado “Diccionario químico de ácidos y bases” se pueden distinguir dos tipos de conceptos:

- conceptos de materia
  - Química - la disciplina
  - Acidos y bases - los aspectos que se estudian dentro de la disciplina
- conceptos de forma
  - Diccionario - la forma en que se presenta el documento

Los conceptos de materia indican sobre qué trata un documento. Estos tipos de conceptos son esenciales para el análisis conceptual.

Al analizar un documento lo primero que hay que hacer es decidir a qué área o rama del conocimiento pertenece, es decir, a qué disciplina o subdisciplina corresponde. Después hay que identificar los aspectos de la disciplina que se estudian en el documento. Por ejemplo, en un documento que trata sobre “la psicología del adolescente”, “psicología” es el concepto de disciplina y “adolescente” es el aspecto que se estudia en esa disciplina.

El mismo aspecto o fenómeno puede estudiarse por diferentes disciplinas. Por ejemplo el concepto adolescente puede estudiarse en medicina, educación, sociología, etc.

Los fenómenos estudiados por las disciplinas pueden ser entidades o aspectos concretos, tales como automóviles, esmeraldas, bibliotecas, centrales azucareros; pueden ser aspectos que expresen una acción, como formación, transmisión, obtención, recolección; o pueden ser ideas abstractas, tales como belleza, odio, amor, ternura.

También dentro de los conceptos de materia se pueden distinguir los de lugar y los de tiempo. Por ejemplo, si se analizan los títulos "La enseñanza programada en Cuba" y "Las revoluciones burguesas del siglo XVIII", se puede apreciar que en el primer caso "Cuba" es un concepto de materia de lugar, y en el segundo caso el siglo XVIII es un concepto de materia de tiempo. Estos conceptos de materia de lugar y tiempo en muchos casos no tienen nada que ver con el lugar o la fecha en que se publica el documento.

Algunos autores consideran la disciplina como un concepto de forma intelectual, es decir un concepto de forma relacionado con la temática. Por ejemplo, en un libro sobre Historia de Cuba se identificará "Cuba" como el concepto de materia, ya que el libro trata sobre Cuba, e Historia sería un concepto de forma intelectual, ya que el libro es de Historia. En este texto no se seguirá ese criterio, sino que se identificará la disciplina como un concepto de contenido y no como un concepto de forma intelectual.

Los conceptos de forma expresan lo que es un documento y no la temática o materia de que trata el documento. En la tabla 1 se relacionan los tipos fundamentales de estos conceptos. En algunos casos pueden surgir confusiones al analizar un documento y hay que tener cierto cuidado para distinguir si un término representa un concepto de materia o de forma.

Tabla 1. Conceptos de forma

1. FORMA FÍSICA
Libros, folletos, discos, películas, cintas magnéticas, diapositivas, microfilms, etc.
2. FORMA DE PRESENTACIÓN
2.1 FORMAS DE SÍMBOLOS UTILIZADOS PARA LA PRESENTACIÓN
Lenguaje. Ejemplos: español, ruso, inglés, etc.
Matemática. Ejemplos: estadísticas, fórmulas, etc.
Gráficos. Ejemplos: diagramas, dibujos, etc.
2.2 FORMAS DE ORDENAMIENTO, EXPOSICIÓN O SELECCIÓN
Orden. Ejemplos: alfabético, cronológico, etc.
Forma literaria. Ejemplos: ensayos, conferencias, cartas
Colecciones. Ejemplos: antologías, enciclopedias
Reglamentaciones. Ejemplos: códigos, normas, leyes, constituciones.
Información secundaria. Ejemplos: resúmenes, índices, bibliografías, sinopsis, reseñas, concordancias, extractos.
2.3 FORMAS PARA LECTORES DETERMINADOS
Ejemplos: Estadísticas para dirigentes, psicología para los padres, lecturas para los niños.

La clasificación de los conceptos que se presenta no es una clasificación rígida, ni está establecida como norma general, pero puede servir como orientación y apoyo al trabajo del indizador. Se pueden considerar diferentes modificaciones, pero

después que el indizador o grupo de indizadores que trabajan con un mismo tipo de colección de documentos, llega a acuerdos concretos sobre los tipos de conceptos que van a indizar es necesario que los cumplan como una de las reglamentaciones del trabajo.
En algunas disciplinas ciertos conceptos son especialmente importantes. Por ejemplo, en Historia los conceptos de materia de lugar y tiempo (fecha). En Química las fórmulas de los compuestos como conceptos de materia y los números de patentes como formas numéricas de presentación.

Consistencia en la indización
La consistencia en la indización representa el grado de acuerdo en la asignación de términos índices a un documento o a una solicitud de información entre dos indizadores (interconsistencia) o en un mismo indizador en diferentes momentos (intraconsistencia).
El círculo A representa el conjunto de términos índices asignado a un documento dado por el indizador 1 y el círculo B el conjunto asignado al mismo documento por el indizador 2. La proporción de términos iguales asignados en el proceso de indización expresa la consistencia de los indizadores 1 y 2.
En el caso que haya consistencia total los dos círculos se superponen, es decir, ambos indizadores seleccionan los mismo términos para indizar el documento.

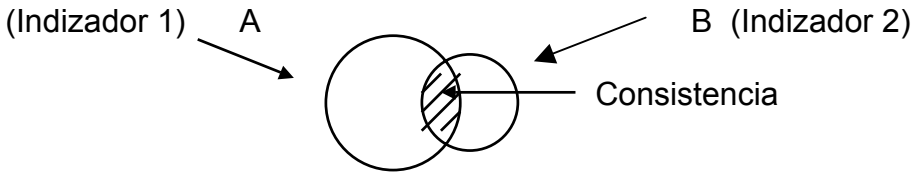


Fig. 1 Representación gráfica de la consistencia entre los indizadores 1 y 2 al indizar un mismo documento

La consistencia interindizador depende fundamentalmente de tres factores:
- la experiencia en el trabajo de indización
- los conocimientos sobre el tema que se indiza
- la longitud y complejidad del documento

Es necesario señalar que en la literatura aparecen diferentes criterios al analizar los factores o aspectos que influyen en la consistencia en la indización.
La importancia de este concepto se deriva del hecho que existe una elevada correlación entre una recuperación efectiva de la información y la consistencia. Esto justifica que se realicen grandes esfuerzos por mantener una elevada consistencia y poder contribuir a asegurar la calidad en el sistema de recuperación.

4. LENGUAJES DE RECUPERACIÓN DE LA INFORMACIÓN (LRI)
Definición, componentes, rasgos y funciones de los LRI
Los LRI son lenguajes artificiales, es decir, lenguajes creados por el hombre, que se utilizan para indizar los documentos y las solicitudes con la finalidad de recuperar la información almacenada y satisfacer las demandas de los usuarios. Cumplen, pues, la función comunicativa de una forma muy especial. Una forma

que posibilita la comunicación entre los autores, los indizadores y los usuarios, utilizando como canales de comunicación los documentos y los índices.
Los LRI tienen, al igual que las lenguas naturales, como mínimo tres componentes fundamentales:
VOCABULARIO
SINTAXIS
REGLAS PARA SU USO
Se han diseñado muchos LRI con diferentes características, pero todos poseen una serie de rasgos comunes que los diferencian de las lenguas naturales. De estos rasgos los más importantes son los siguientes:

1. Se mantiene una relación unívoca entre los términos y los conceptos que ellos expresan. Es decir, un término expresa un solo concepto, refleja de forma veraz un objeto o fenómeno de la realidad objetiva.
2. Sus componentes (vocabulario, sintaxis y reglas de uso) forman un sistema que tiene que estar en correspondencia con el objetivo para el cual se creó. El LRI tiene que ser capaz de expresar de forma adecuada los contenidos de los documentos y de las solicitudes de información considerando los intereses de los usuarios del sistema de información.

Los rasgos esenciales contribuyen a que, conjuntamente con otros factores, los LRI cumplan las funciones básicas siguientes:
- Eliminan la ambigüedad en su poder expresivo al establecer una relación unívoca entre los términos del vocabulario y los conceptos que estos expresan.
- Facilitan la labor de indización al relacionar los términos autorizados y mostrar (en muchos casos) sus relaciones lógicas, lo cual determina con qué precisión el indizador puede expresar el contenido del documento.
- Mejoran la consistencia de la indización.
- Sirven de apoyo a la operación de búsqueda - recuperación al establecer con qué precisión se pueden expresar los intereses de los usuarios y facilitar la confrontación de los términos índices de la prescripción de búsqueda con los de los registros de datos de los documentos, lo cual permite determinar si existe o no concordancia semántica entre la solicitud y el documento.

Vocabulario
El vocabulario o léxico de un LRI es el conjunto de términos que se utiliza para expresar el contenido informacional de un documento (libro, folleto, informe, artículo de una revista, tesis, diapositiva, microfilm, etc.) o una solicitud de información. Los términos del LRI pueden estar representados en diferentes formas:
- palabras aisladas o combinación de palabras
- códigos numéricos, alfabéticos o alfa-numéricos
- códigos en combinación con palabras del lenguaje natural.

Homonimia. Sinonimia. Relaciones paradigmáticas. Sistema sindético
Para garantizar que se cumpla el rasgo esencial de todo LRI, o sea que entre los términos y los conceptos exista una relación unívoca, es necesario eliminar de su vocabulario la homonimia y la sinonimia.
La **homonimia** se elimina con aclaraciones sobre los significados del término, las cuales se colocan entre paréntesis y a continuación de los vocablos que sean **homógrafos**, es decir una misma palabra con dos o más significados.

Los siguientes grupos de palabras son ejemplos de homógrafos:

PLANTA (INSTALACION INDUSTRIAL)
PLANTA (PARTE INFERIOR DEL PIE)
PLANTA (PISO DE UN EDIFICIO)
PLANTA (EN BOTÁNICA)

TANQUE (RECIPIENTE)
TANQUE (ARMAMENTO)

MERCURIO (PLANETA)
MERCURIO (METAL)

La **sinonimia** se elimina estableciendo un conjunto de clases equivalentes entre los sinónimos, o casi-sinónimos, o sea entre dos o más palabras diferentes que tienen significados iguales o parecidos. Después se selecciona una de estas palabras que represente el conjunto, y se establecen referencias de USE (o VÉASE) desde las otras hacia la seleccionada.
Los siguientes sinónimos forma un grupo de clases equivalentes:
CARBOHIDRATOS
HIDRATOS DE CARBONO
GLUCIDOS
De este grupo de clases equivalentes se puede seleccionar el término CARBOHIDRATOS para que represente el conjunto. Después se coloca la referencia cruzada USE desde Glúcidos e Hidratos de carbono hacia CARBOHIDRATOS:
CARBOHIDRATOS

-
-
-
-

Glúcidos
USE CARBOHIDRATOS

-
-

Hidratos de carbono
USE CARBOHIDRATOS

Las relaciones que se acaban de explicar son ejemplos de **relaciones paradigmáticas**, relaciones lógicas de carácter extralingüístico que se establecen entre los términos por algún rasgo común de tipo semántico o morfológico. A continuación se exponen otros ejemplos de relaciones paradigmáticas.

Ejemplos de paradigmas (conjunto de palabras con relaciones paradigmáticas)

Paradigma morfológico

<u>INFORM</u>ACIÓN
<u>INFORM</u>ACIONAL } El rasgo común es la raíz <u>INFORM</u>
<u>INFORM</u>ATIVO

Paradigma semántico

BIBLIOTECAS UNIVERSITARIAS
BIBLIOTECAS PÚBLICAS
BIBLIOTECAS ESCOLARES

} El rasgo común es que todos representan tipos de bibliotecas

En los vocabularios de los LRI generalmente se expresan relaciones paradigmáticas del tipo semántico.
Todas estas relaciones y aclaraciones sobre los términos, que sirven de guía para el mejor uso del vocabulario conforman el llamado **sistema sindético**. En la tabla 2 se resumen los principales componentes de un sistema sindético.

Tabla 2. Componentes de un sistema sindético

	PRINCIPALES TIPOS DE RELACIONES	**SÍMBOLOS UTILIZADOS**
RELACIONES **PARADIGMÁTICAS** (del tipo semántico)	**Relaciones de equivalencia** (Referencias cruzadas) - de los términos no autorizados a los autorizados - de los términos autorizados a los no autorizados	USE o VÉASE **UP (USADO POR)**
	Relaciones jerárquicas - para señalar el término más amplio o genérico - para señalar el término más limitado o específico	**TA (T. más amplio)** **TG (T. genérico)** **TL (T. limitado)** **TE (T. específico)**
	Relaciones asociativas - para señalar los términos relacionados	**TR (T. relacionado)**
NOTAS DE ALCANCE Y ACLARATORIAS		

Sintaxis. Relaciones sintagmáticas
La sintaxis de un LRI es el conjunto de reglas para combinar los términos del vocabulario en cadenas, frases o unidades sintácticas capaces de expresar conceptos o significados más complejos o más abarcadores, que no podrían ser expresados si se utilizaran los términos del vocabulario de forma aislada; y se llaman relaciones sintagmáticas a esas relaciones lingüísticas entre los términos para formar las cadenas, frases o unidades sintácticas.
La sintagmática requiere determinadas reglas de asociación para representar las relaciones mutuas entre las palabras para formar una cadena sintáctica.
Ejemplos de sintagmas (conjunto de palabras con relaciones sintagmáticas):
- La Sexta Cumbre de los Países No Alineados se celebró en La Habana
- Formación de profesores. Química. Cuba

Clasificación de los lenguajes

Se pueden establecer diferentes clasificaciones de los LRI, pero en este texto se clasificarán, de acuerdo con el rasgo diferencial correspondiente la coordinación de los términos del vocabulario, en dos grandes grupos:
- lenguajes precoordinados: los que realizan la coordinación antes de la indización o durante la indización
- lenguajes poscoordinados: los que realizan la coordinación de términos en el momento de la búsqueda, es decir después de la indización

Se ha seleccionado esta clasificación entre las distintas que aparecen en la literatura porque es la que se puede vincular más directamente con la indización y, además, para tener un marco de referencia que facilite estudiar los lenguajes. Ahora bien, debe analizarse con una óptica flexible ya que no existe una línea fija de separación entre los dos grupos de lenguajes. Así, por ejemplo, en los lenguajes precoordinados también aparecen palabras aisladas, términos que no son el producto lógico de dos o más términos simples. A su vez en los lenguajes poscoordinados aparecen términos compuestos, lo cual significa que se ha realizado esta precoordinación en el momento de diseñar el vocabulario. Estos términos pueden, a su vez, coordinarse con otros en el momento de la búsqueda ya que pertenecen a un lenguaje que funcionará en un sistema poscoordinado.

Lenguajes precoordinados predominantemente enumerativos

Los lenguajes precoordinados son predominantemente enumerativos y al confeccionar su vocabulario se relacionan todos los términos que se pueden utilizar en la indización de los documentos o las solicitudes, sin permitir una ulterior combinación de términos para formar clases más complejas. Todas las coordinaciones de términos que se consideren necesarias se hacen al desarrollar el vocabulario, es decir se precoordina antes de la indización. Por eso algunos autores incluyen a los lenguajes precoordinados dentro de los lenguajes no manipulativos, ya que no se pueden manipular los términos en el momento de la búsqueda.

Lenguajes precoordinados enumerativos con síntesis

En este tipo de lenguajes las clases principales también se enumeran en su vocabulario, pero se han previsto los medios para que durante la indización se pueda aplicar en alguna medida la síntesis. Esto significa que se pueden construir clases más complejas (no enumeradas previamente) a partir de las clases contenidas en el vocabulario.

Lenguajes precoordinados predominantemente sintéticos

Estos lenguajes se diseñan de modo tal que al aplicarlos al proceso de indización permitan sintetizar (relacionar o coordinar) el contenido del documento mediante la relación de los conceptos claves extraídos en el análisis conceptual del texto.

Lenguajes poscoordinados

Son los que utilizan vocabularios formados por clases básicas, que en muchos casos están formados por una sola palabra, es decir libre de precoordinación. Se utilizan en los sistemas poscoordinados, los cuales al indizar utilizan los términos que sean necesarios del vocabulario, o sea, es una indización coordinada para cada documento.

En el momento de la búsqueda coordinan los términos de la solicitud de información de modo de localizar todos los documentos que tengan el producto lógico de los términos que representan la solicitud.

Tabla 3. Principales tipos de LRI

LENGUAJES PRECOORDINADOS	Clasificaciones jerárquicas	Rubricadores Clasificaciones universales (CDU)
	Lenguajes alfabéticos de materia	Epigrafiarios Lenguajes de relación o articulados
	Clasificaciones facéticas	
LENGUAJES POSCOORDINADOS	Lenguajes de descriptores	Kristal, Pusto, Nepusto

Clasificaciones jerárquicas

Las clasificaciones jerárquicas se elaboran partiendo del principio que establece que las materias se pueden dividir en submaterias o materias más específicas. Este proceso de subdivisión se puede repetir tantas veces como sea necesario. De este modo se elabora una estructura jerárquica de tipo ramificada o de árbol. Dentro de las clasificaciones jerárquicas se incluyen muchas de las clasificaciones temáticas o rubricadores. Estas clasificaciones generalmente se hacen dentro de una rama del conocimiento. Las clases se subdividen en subclases, pero estas subdivisiones, en la mayoría de los casos, no se continúan hasta muchos subniveles.

Como ejemplos de clasificaciones temáticas o rubricadores se pueden señalar las clasificaciones que presentan muchas revistas referativas (revistas de resúmenes). Su objetivo primordial es ordenar los resúmenes por grupos temáticos o clases genéricas. Al aplicar estas clasificaciones se le asigna a cada resumen el código que representa la clase o subclase que describe del modo adecuado el tema central del documento.

Las clasificaciones temáticas o Rubricadores facilitan la localización de las fichas con resúmenes y permiten su rápida revisión dentro de una temática de interés para el usuario. Este tipo de clasificación requiere como complemento un índice analítico, otra vía para buscar en caso de necesidad las materias más específicas.

Además, en muchos casos hay que tomar una decisión sobre a qué grupo temático se asigna un documento determinado.

El otro tipo de clasificaciones jerárquicas son las universales, que abarcan todo el universo de conocimientos. Se pueden mencionar como ejemplos la Clasificación Decimal Universal (CDU), la Clasificación Decimal de Dewey, la Clasificación Biblioteco-Bibliográfica de la URSS, la Clasificación de la Biblioteca del Congreso (LC) de Washington. Estas clasificaciones eran, en sus primeras ediciones, predominantemente enumerativas. Las últimas ediciones han incrementado las posibilidades de síntesis, de crear clases más complejas durante el proceso de indización uniendo algunas de las clases previamente enumeradas.

La CDU previó esto desde sus inicios y por eso está considerada como una clasificación semifacética y utiliza los dos puntos (colon) para expresar la síntesis, tal como se ilustra en el ejemplo siguiente:

341.67 Desarme. Reducción de armamentos. Prohibición de las armas nucleares, químicas y bacteriológicas
623.454.8 Radiación penetrante. Armas atómicas
094.2 Acuerdos internacionales y tratados (determinante de forma)

El documento "Tratado internacional para la no proliferación de armas nucleares" se indizará asignándole el siguiente código: 341.67:623.454.8(094.2)

Epigrafiarios o listas de encabezamientos de materia

Los epigrafiarios o listas de encabezamientos de materias son lenguajes alfabéticos de materia y caen dentro de los lenguajes precoordinados enumerativos, que ordenan alfabéticamente los términos y no ofrecen posibilidades para coordinar clases más complejas en el momento de la indización. Los lenguajes de epígrafes clásicos eran de este tipo. En la actualidad estos lenguajes prevén mecanismos para asignar epígrafes complejos en el momento de la indización. Presentan, en estos casos, cierta posibilidad de síntesis. Además, con el empleo de los subepígrafes se aumenta la especificidad del lenguaje.

Los catálogos de materia de muchas bibliotecas son ficheros en los cuales los RB de los documentos (las fichas catalográficas en este caso) se ordenan en una secuencia alfabética del epígrafe de materia. Esto significa que la clave de orden del catálogo de materia es el epígrafe.

Por tanto estos ficheros son índices alfabéticos de materia que han utilizado para el proceso de indización una lista de epígrafes.

Otros lenguajes en los que predomina la síntesis son los que tienen reglas de uso para utilizar frases modificadoras en el proceso de indización. En un primer paso se asignan los términos del vocabulario autorizado, que serán los puntos de acceso a la entrada del índice. En un segundo paso se añade una frase modificadora que representa el contexto en que se encuentra el término índice en el documento.

Clasificaciones facéticas

Las clasificaciones facéticas se basan en la síntesis, o sea la coordinación de términos durante el proceso de indización. Una de estas clasificaciones fue desarrollada en toda su plenitud por el científico hindú S. R. Ranganathan, uno de los bibliotecarios más notables de todos los tiempos.

Ranganathan planteó que cualquier tema podría considerarse como una combinación de uno o más "conceptos básicos" o "categorías fundamentales" que designó con el nombre de facetas. Postuló que existían cinco facetas o categorías fundamentales aplicables a todas las áreas de conocimientos a las cuales nombró del siguiente modo:

PERSONALIDAD
MATERIA
ENERGÍA
ESPACIO

TIEMPO
En una disciplina o área de conocimiento determinado, una faceta agrupa a un conjunto de términos que representa un número de fenómenos o aspectos que comparten alguna característica común dentro de ese campo temático. Así, cada disciplina tendrá sus facetas respectivas. Por ejemplo, en medicina hay dos facetas esenciales:
ÓRGANOS (PERSONALIDAD)
PROBLEMAS (ENERGÍA) (incluye métodos de estudio, tratamientos, etc.)
Las clasificaciones facéticas se desarrollan a partir de los fundamentos científicos del análisis facético, que ha sido aplicado por diversos grupos de investigadores, entre los cuales se ha destacado el "Classification Research Group" (CRG) de Londres.
El método del análisis facético puede resumirse en los siguientes pasos:
1. Se analiza una muestra de la literatura dentro de la rama de la especialidad que se proyecta clasificar. Este paso es la base para desarrollar los pasos 2 y 3.
2. Se determinan las facetas, que son los aspectos más generales y fundamentales de la materia. A veces, si es necesario, las facetas se subdividen en subfacetas.
3. Se seleccionan los términos que corresponden a cada faceta o subfaceta. El ordenamiento de los términos dentro de cada faceta puede ser jerárquico o alfabético, o una combinación de ambos tipos de ordenamiento.
4. Se asignan códigos para representar las facetas y los términos dentro de cada faceta.
5. Se establece una fórmula facética para indicar la secuencia de las facetas al indizar los documentos.

En estas clasificaciones las clases no están previamente formuladas (por eso no son enumerativas). En el momento de la indización es cuando se formulan las clases escogiendo los términos, que sean necesarios para expresar el contenido del documento de las distintas facetas.
Estos términos se combinan en una secuencia determinada, que se llama fórmula facética. Es decir, se sintetiza una clase más compleja no previamente enumerada en la clasificación. Generalmente en el lugar de los términos en el lenguaje natural se utilizan los códigos que los representan.
Las facetas representan categorías semánticas en contraposición con las categorías gramaticales del lenguaje natural de las frases del lenguaje natural.
Un análisis más detallado del significado del papel de las categorías en las clasificaciones facéticas permite llegar a la conclusión de que en realidad tienen una función dual:
- En la expresión paradigmática (vertical) sirven para agrupar los términos por características semánticas comunes dentro de una misma faceta.
- En la expresión sintagmática (horizontal) sirven para situar los términos en un orden preferido (de acuerdo con la fórmula facética).

Lenguajes de descriptores
Los lenguajes de descriptores se utilizan en los llamados sistemas poscoordinados. Estos sistemas empezaron a desarrollarse y a extenderse a partir de 1945 como una respuesta a la urgente necesidad de disponer de sistemas que permitieran no sólo la recuperación de la información por múltiples aspectos y por cualquier grado de complejidad, sino que también posibilitaran y facilitaran la

utilización de procedimientos mecánicos y automáticos en los sistemas de información.

A continuación se relacionan algunos de los principales aportes que sirvieron de base al pleno desarrollo de los sistemas poscoordinados y los lenguajes de descriptores:

- En 1915 H. Taylor (EE.UU) utilizó tarjetas perforadas de superposición para realizar la búsqueda poscoordinada en ornitología.
- En 1939 W. E. Batten (RU) diseñó un sistema de indización coordinada para patentes.
- En 1946 G. Cordonnier (Francia) diseñó un sistema de indización con tarjetas de coincidencia óptica, también llamadas “peck a book”
- En 1946 Calvin Mooers (EE.UU) elaboró un sistema de búsqueda mecánica que llamó Zatacoding con tarjetas con bordes perforados y un selector mecánico. Mooers introdujo el término “descriptor” para designar las unidades léxicas del vocabulario. También introdujo el término “lenguaje de recuperación” (retrieval language)
- En 1953 Mortimer Taube (EE.UU) elaboró el sistema unitérmino, que operaba con palabras extraídas del texto del documento. No se ejercía, por tanto, un control del vocabulario.

El sistema unitérmino fue el que prácticamente inició la fase del desarrollo acelerado de los sistemas poscoordinados. Este sistema tal como fue concebido tenía una serie de fallas. La indización resultante no expresaba de manera unívoca y precisa el contenido de los textos debido a la falta de control del vocabulario, presentándose problemas de sinónimos, homógrafos, dificultades para la búsqueda genérica y falsas coordinaciones.

En un intento de contrarrestar estas dificultades se introdujeron algunas modificaciones. Se empezaron a utilizar términos formados por más de una palabra para identificar sin ambigüedades determinados conceptos, como por ejemplo “partículas subatómicas”, “física nuclear”, “carbón activado”, “intercambiadores de calor”.

Posteriormente se analizó la conveniencia de establecer un vocabulario controlado y surgieron los “tesauros”. En fechas anteriores ya habían surgido otros Tesauros con diferentes propósitos. El más conocido es el “Thesaurus of English words and phrases” del británico P.M. Roget en 1852. El principio básico de este Tesauro es agrupar las palabras de acuerdo con las ideas; este sigue siendo uno de los principios básicos de los nuevos Tesauros. Una diferencia fundamental es que el Dr. Roget trataba principalmente con palabras sencillas y los nuevos Tesauros trabajan con conceptos y los nombres de estos conceptos son las entradas de los Tesauros, que en algunos casos son palabras aisladas, pero en otros son combinaciones de dos o más palabras.

En los sistemas poscoordinados, al indizar los documentos asignan un conjunto de términos del lenguaje de descriptores. Esos términos no están relacionados entre sí y deben representar todos los aspectos del contenido temático del documento.

Los sistemas poscoordinados están estructurados de forma tal que permiten recuperar los documentos coordinando, en el momento de la búsqueda, los términos índices de la prescripción de búsqueda. Estos sistemas ofrecen muchas posibilidades de formar conceptos complejos a partir de conceptos simples.

Se puede señalar que aún con un vocabulario no muy amplio se pueden expresar muchos conceptos, ya que los términos simples pueden utilizarse para expresar diversos conceptos complejos.
No se deben precoordinar los términos en el vocabulario si no es realmente necesario. Ahora bien, en determinados casos sí es necesario usar términos precoordinados para asegurar la precisión del sistema. Por tanto hay que mantener un balance adecuado al analizar, dentro de la totalidad de los términos del vocabulario, la cantidad de términos simples (unitérminos) y de términos compuestos (formados por más de una palabra).
Los lenguajes de descriptores ofrecen la posibilidad de multiplicación lógica de los sistemas poscoordinados. Es decir, con la coordinación de los términos se logra expresar de una manera muy específica los conceptos. Esto aumenta la capacidad del sistema de retener los documentos no relevantes en respuesta a la necesidad real de información del usuario, lo que equivale a afirmar que se aumenta la precisión del sistema.

Estudio comparativo de los lenguajes

El estudio comparativo de los lenguajes se resume en la Tabla 4. Este estudio se basa en los siguientes rasgos esenciales que caracterizan a los lenguajes:

1. SINTESIS
2. RELACIONES PARADIGMÁTICAS
3. RELACIONES SINTAGMÁTICAS
4. FUERZA SEMÁNTICA
5. EFECTIVIDAD PARA LA BÚSQUEDA (GENÉRICA Y/O ESPECÍFICA)
6. RECUPERACIÓN MULTIFACÉTICA

La síntesis expresa la capacidad del lenguaje para construir clases más complejas a partir de clases simples.
Las relaciones paradigmáticas son las relaciones lógicas extralingüísticas entre los términos que facilitan expresar con más precisión los conceptos.
Las relaciones sintagmáticas son relaciones lingüísticas entre los términos para expresar conceptos con significado más complejo o abarcador.
La fuerza semántica es la capacidad del lenguaje para expresar del modo más preciso y exacto un mensaje informativo. La fuerza semántica está en dependencia de los rasgos anteriores.
La efectividad para la búsqueda se correlaciona con la estructura del lenguaje con todos los rasgos anteriores. Además se consideran los dos tipos extremos de búsqueda: la genérica y la específica.
La recuperación multifacética es la recuperación de la información por múltiples aspectos y por cualquier grado de complejidad.

Tabla 4. Estudio comparativo de los principales tipos de lenguajes

	1	2	3	4	5		6
	SINTESIS	Rel. PARAD.	Rel. SINTAG.	Fuerza semántica	EFECT. GEN.	BUSQ. ESPC.	Recuperación multifacética
A. CLASIFICACIONES JERÁRQUICAS							
Rubricadores	no	no	no	débil	si	no	no

C. Universal	cierto grado de síntesis	no detalladas	no detalladas	regular	si	reg	no
B. EPIGRAFIARIOS O LISTAS DE ENCABEZAMIENTOS DE MATERIA	poca	no detalladas	no detalladas	débil	reg	reg	no
C. CLASIFICACIONES FACÉTICAS	sintét.	no detalladas	no detalladas	fuerte	reg	si	reg
D. LENGUAJES DE DESCRIPTORES	sintét.	detalladas, en general usan tesauros	no detalladas	fuerte	reg	si	si

Comparación de epígrafes y descriptores

Es conveniente empezar por señalar que los epígrafes, al igual que los descriptores, son términos que se asignan para expresar el contenido esencial de los documentos. La diferencia está en que los epígrafes forman parte de un vocabulario de un lenguaje precoordinado y los descriptores pertenecen generalmente a tesauros y se utilizan en sistemas con indización coordinada (sistemas poscoordinados). Algunos sistemas no establecen esta diferencia. Por ejemplo el MEDLARS (Medical Literature Analysis and Retrieval System) de la Biblioteca Nacional de Medicina de EUA tiene uno de los más grandes sistemas automatizados postcoordinados y denomina a su lenguaje MESH (Medical Subject Headings) que significa "encabezamientos de materia de medicina" o "epígrafes de medicina".

5. INDIZACIÓN CON EPÍGRAFES

Los epígrafes generalmente se utilizan en las bibliotecas para preparar los índices de materia (catálogos de materia) de las colecciones de libros. En el análisis de contenido de los libros, el título generalmente es el elemento más importante, pero también es necesario consultar otras partes que brindan información esencial para garantizar una mejor indización, tales como las siguientes:

-Título y subtítulo
-Tabla de contenido
- Prefacio
- Información de las contraportadas

Ejemplos de los principales tipos de epígrafes

Al estudiar los epígrafes hay que considerar dos aspectos esenciales:

a) La forma de presentación
b) El contenido

a) Por la forma de presentación los epígrafes pueden ser:

Epígrafes simples: formados por una sola palabra.

Ejemplos:

EDUCACIÓN ELECTRONES
DERECHO RELOJERÍA
EXISTENCIALISMO SOCIALISMO

Epígrafes compuestos: formados por más de una palabra, o una palabra con una aclaración entre paréntesis, o dos o más palabras separadas por una coma para indicar una inversión.

Ejemplos:

MEDICIONES DEL APRENDIZAJE
POLÍTICA CIENTIFICA
COLUMNA JUVENIL DEL CENTENARIO
ARTE Y SOCIEDAD
PLANIFICACIÓN (ECONOMÍA)

PLANIFICACIÓN (URBANISMO)

MONTECARLO, MÉTODO DE
SOGAMOSO, VALLE
MARTÍ, JOSÉ, 1853-1895

b) El contenido es el aspecto primordial de los epígrafes, su razón de ser. Lógicamente se comprenderá que existen múltiples variantes de encabezamientos de contenido. Seguidamente se brindarán algunas aclaraciones y ejemplos con respecto a los principales tipos de epígrafes, considerando su contenido.

Epígrafes biográficos. El nombre de la persona se pone en forma invertida, seguida de la fecha de nacimiento y muerte:

Ejemplos:

MARTÍ, JOSÉ, 1853-1895
LUZ Y CABALLERO, JOSÉ DE LA, 1800-1862
CASTRO RUZ, FIDEL, 1926-

Epígrafes históricos o cronológicos. Designan a nombres propios de épocas o etapas históricas y/o geológicas, de acontecimientos relevantes, tratados, convenios, alianzas.

Ejemplos:

EDAD MEDIA
EDAD DE PIEDRA
HISTORIA ANTIGUA
PAÍSES NO ALINEADOS
PLAYA GIRÓN, BATALLA DE, 1961
POLONIA, OCUPACIÓN, 1939-1945
VERSALLES, TRATADO DE, 28 de junio de 1919 (Alemania)

Epígrafes étnicos. Se utilizan para indizar trabajos referentes a pueblos nómadas, tribus, razas o grupos humanos con características especiales.

Ejemplos:

AZTECAS
INDIOS DE NORTEAMÉRICA
INDIOS DE SUR AMÉRICA
INDIOS DE BOLIVIA
INCAS
MAYAS

Subepígrafes

Los subepígrafes son las palabras que se adicionan al epígrafe, después de un guión, para representar un concepto más específico, más detallado.

A continuación se pondrán algunos ejemplos de los diferentes tipos de subepígrafes.

Subepígrafes temáticos o de materia

CIENCIA-ENSEÑANZA
AERONÁUTICA-VUELOS
EDUCACIÓN-HISTORIA
QUÍMICA ANALÍTICA-ANÁLISIS POR MICROPRUEBAS
ARTE-HISTORIA-SIGLO XX

Subepígrafes formales o de forma

EDUCACIÓN-DISCURSOS, ENSAYOS, CONFERENCIAS
QUÍMICA-BIBLIOGRAFÍA
AUTORES CUBANOS-DICCIONARIOS
MARTÍ, JOSÉ-MANUSCRITOS
FÍSICA-MANUALES

Subepígrafes de ubicación geográfica

EDUCACIÓN SUPERIOR-ESPAÑA
INDUSTRIA MINERA-SAN SALVADOR
ARTE-POLONIA
EDUCACIÓN-CUBA
TELEVISIÓN-LEGISLACIÓN-GRAN BRETAÑA

Al indizar documentos de naturaleza histórica o descriptiva el nombre del país se asignará como epígrafe y el asunto como subepígrafe.

Ejemplo:

HABANA-DESCRIPCIÓN
PERÚ-HISTORIA
PARÍS-CALLES
CUBA-HISTORIA-PERÍODO COLONIAL, 1514-1898
CUBA-HISTORIA-GUERRA DE LOS DIEZ AÑOS, 1868-1878
CUBA-HISTORIA-REVOLUCIÓN, 1959-

Es importante aclarar que no en todos los casos hay una plena coincidencia en el uso de los epígrafes entre los colectivos de trabajo de diferentes centros. Esto se

debe, entre otras causas, a que a veces se toman decisiones que se fundamentan en los intereses y costumbres de los usuarios del centro, o a los criterios o puntos de vista de los responsables del colectivo. Lógicamente al hacer una revisión de las mismas obras en los catálogos de los diferentes centros se aprecian algunas inconsistencias, es decir habrá ciertas discrepancias en los encabezamientos de materia asignados en la indización.
También es un deber señalar que todos los ejemplos y ejercicios utilizados en este texto han sido tomados de los ficheros confeccionados por especialistas en información de la Biblioteca Central Rubén Martínez Villena de la Universidad de La Habana, de la Biblioteca Nacional José Martí o del Centro de Documentación de la Oficina Regional de la UNESCO de La Habana.

Lineamientos generales para la asignación de epígrafes

Las primeras reglas de importancia para la asignación de epígrafes fueron elaboradas por Charles Cutter en 1876 y, a pesar de que ha transcurrido más de un siglo, los principios en que se sustentan siguen teniendo validez.
En este texto se ofrecerán algunos lineamientos generales para realizar este tipo de indización, los cuales no difieren en lo esencial de las reglas de Cutter. Por eso es necesario destacar que lo importante para desarrollar un trabajo con un enfoque nuevo y moderno es utilizar términos que respondan a las necesidades de los usuarios en correspondencia con los avances científico-técnicos, políticos y sociales del momento actual. Esto exige descartar términos obsoletos e introducir nuevos términos, las listas de epígrafes tienen que crecer y renovarse. Ahora bien, esto no significa que hay que cambiar cada cinco minutos, sino solamente cuando lo requieran las circunstancias. Los cambios o la creación de los nuevos términos deben ser analizados por personas con deseos de renovación, pero al mismo tiempo con conocimientos, autoridad y experiencia.

Pasos de los lineamientos generales para indizar con epígrafes

1. Se asignará el epígrafe que defina de modo más preciso y específico el asunto que se desea indizar.

Ejemplos: Si un documento trata sobre electrones se le asignará el epígrafe ELECTRONES y no FÍSICA ATÓMICA, QUÍMICA ATÓMICA o ÁTOMOS; si un documento trata sobre perros se indiza con el término PERROS y no con ZOOLOGÍA, MAMÍFEROS o ANIMALES DOMÉSTICOS.

2. Se añadirá a cada epígrafe los subepígrafes que sean necesarios para expresar del modo más preciso el asunto.

Ejemplo: CUBA-HISTORIA-DESCUBRIMIENTO, EXPLORACIÓN Y CONQUISTA, 1492-1519

3. Se utilizarán, siempre que sea posible, palabras en el idioma español.

Ejemplos: Se utiliza RETROALIMENTACIÓN y no FEEDBACK; TRATAMIENTO EN TANDAS y no BATCH PROCESSING; SISTEMAS EN LINEA o SISTEMAS DIRECTOS y no SISTEMAS ON-LINE; ACTAS y no PROCEEDINGS; NORMAS y no STANDARDS; NORMALIZACIÓN y no STANDARIZATION.
Por excepción se emplean ciertos términos o expresiones en otros idiomas cuando son de uso muy frecuente o no tienen equivalentes en español. Por ejemplo, muchos autores consideran que es conveniente seguir utilizando SOFTWARE, HARDWARE, KWIC (Key Word In Context), etc.

4. Es necesario considerar dos cuestiones en contradicción. Por un lado, es conveniente asignar tantos epígrafes como sean necesarios para representar todos los aspectos esenciales del contenido del documento. Por otro lado, hay que analizar que en los catálogos de materia con cada epígrafe aparece la ficha bibliográfica del documento. Así, si se asignan cuatro epígrafes hay que introducir cuatro fichas en el catálogo. Esto es una limitante seria ya que implica un rápido y exagerado crecimiento del tamaño del catálogo y un aumento del tiempo y el trabajo para confeccionar las fichas y para realizar la búsqueda. Por tanto, hay que conciliar estas dos cuestiones contradictorias y adaptar una línea intermedia, poniendo un límite al número máximo de epígrafes y tratando de limitar, en la mayoría de los casos, el número asignado a uno o dos.

5. Se seleccionará, entre los posibles sinónimos y casi-sinónimos, un término como epígrafe representativo y se hará referencia de los otros términos equivalentes.

Ejemplo: Cuando se utilice el epígrafe MEDICIONES DEL APRENDIZAJE en el índice debe aparecer también la siguiente referencia cruzada:

TESTS
véase MEDICIONES DEL APRENDIZAJE

6. No se debe dar entrada por la temática y la forma de presentación, a un mismo documento.

Ejemplo: Un documento titulado “Manual de Laboratorio de Química” se entra por QUÍMICA-MANUAL DE LABORATORIO y no se entra por MANUAL DE LABORATORIO-QUÍMICA.

7. No se debe, en general, dar entrada por el asunto y por el país, sino que hay que seleccionar la entrada de acuerdo con el caso de que se trate. En las temáticas de ciencias exactas y naturales, técnicas, artes y muchas de las ciencias sociales (aunque no todas) se entrará por el asunto. En las temáticas históricas o descriptivas se entra por el país o el nombre de la ciudad.

Ejemplos:

ARTE-ESPAÑA
EDUCACIÓN-CUBA
CUBA-HISTORIA-REVOLUCIÓN, 1959-
PARÍS-CALLES

Esta regla tiene sus excepciones. Así, en casos especiales que se estime necesario se entrará por el asunto y el país.

8. Se consultarán las listas y los repertorios antes de crear nuevos términos.

6. INDIZACIÓN CON DESCRIPTORES

La indización con descriptores corresponde a la indización coordinada, es decir a sistemas poscoordinados. Este tipo de indización permite la recuperación multifacética de la información almacenada, lo cual significa que un documento dado puede indizarse asignándole tantos descriptores como sea necesario para describir todas las facetas y subfacetas expresadas en su contenido. La

recuperación se realiza coordinando en la búsqueda los descriptores que representen el producto lógico de la información solicitada y seleccionando los documentos que respondan a ese producto lógico y, por tanto, a la demanda formulada.

Reglas sobre la presentación de los descriptores
Se pueden consultar otras obras para ampliar sobre este asunto. En este texto solamente se brindarán las reglas más importantes y necesarias.

1. Forma del término
Utilizar, siempre que sea posible, los sustantivos.
Ejemplos: EVALUACIÓN en lugar de EVALUATIVO
No utilizar verbos
Ejemplos: PROGRAMACIÓN en lugar de PROGRAMAR; COMPATIBILIDAD en lugar de COMPATIBILIZAR; AUTOMATIZACIÓN en lugar de AUTOMATIZAR

2. Número del término
Utilizar el singular para términos que expresan conceptos que no pueden contarse por unidades.
Ejemplos: Nombres de disciplinas (QUÍMICA, FÍSICA, etc.); procesos (CAPACITACIÓN, POLÍTICA CIENTÍFICA); materiales y propiedades específicas (UREA, FUERZA); nombres propios (LEY DE NEWTON)
Utilizar el plural para términos que expresan conceptos que pueden contarse por unidades.
Ejemplos: CENTROS DE INFORMACIÓN, BIBLIOTECAS, CRÉDITOS ACADÉMICOS, MIMEÓGRAFOS, COMPUTADORAS, LENGUAJES, MEDIOS DE DIFUSIÓN MASIVA.

3. Forma de la entrada
Se utilizará la entrada directa. Únicamente en casos muy excepcionales se hará la inversión de la entrada.

4. Se debe evitar la utilización de:
Signos de puntuación
Abreviaturas

Tabla 5. Recomendaciones para el uso del número gramatical de los descriptores.

TIPO DE TÉRMINO	USO DEL SINGULAR EJEMPLOS	USO DEL PLURAL EJEMPLOS
Procesos	CAPACITACIÓN CONSTRUCCIÓN	-
Nombres propios	LEY DE NEWTON PLUTÓN	-
Disciplinas	QUÍMICA INGENIERÍA	-
Equipos, aparatos, objetos,		PULVERIZADORES

partículas, locales, otros	-	TRACTORES BALANZAS MIMEÓGRAFOS MESONES BIBLIOTECAS LENGUAJES
Sucesos o eventos	-	EXPLOSIONES HURACANES EMBOSCADAS CONFERENCIAS COLOQUIOS
Sustancias, materiales propiedades	si el término es específico	si el término es genérico
	ETANOL CELOFÁN CERA VISCOSIDAD COMBUSTIÓN	ALCOHOLES PLÁSTICOS CATALIZADORES PROPIEDADES FÍSICAS PROPIEDADES QUÍMICAS

Clasificación de descriptores

Los descriptores al igual que los epígrafes, pueden por su forma de presentación ser simples o compuestos

Por su contenido también pueden ser muy diversos según la rama del conocimiento. Precisamente los lenguajes de descriptores suelen utilizarse para indizar colecciones especializadas, ya que es posible indizar por aspectos muy específicos. En muchos vocabularios de descriptores (Tesauros) aparecen índices auxiliares con clasificaciones jerárquicas que agrupan a los descriptores dentro de las categorías de la rama del conocimiento de que se trate.

Lineamientos generales para la asignación de descriptores

Los lineamientos 1, 3, 5 y 8 de la asignación de epígrafes también tienen validez en el proceso de asignación de descriptores, por tanto hay que considerarlos conjuntamente con los siguientes lineamientos:

1. Hacer un análisis de contenido profundo del documento que se va a indizar.
2. Asignar en la mayoría de los casos un máximo de ocho descriptores para expresar todos los conceptos importantes del contenido. En caso necesario se podrán asignar más de ocho descriptores.
3. Si el nombre del país se considera un elemento necesario para la recuperación se añadirá al conjunto de descriptores del documento para que aparezca como una entrada en el índice.
4. Se podrán utilizar cuando se estime necesario descriptores de forma como por ejemplo BIBLIOGRAFÍAS, ESTUDIOS CRÍTICOS, INVESTIGACIONES EDUCACIONALES.
5. Se utilizarán las siglas o abreviaturas previamente autorizadas.

Ejemplos de indización con descriptores

Los ejemplos que se ofrecen se basan en fichas con resúmenes informativos y con los elementos bibliográficos de los documentos. Al final del resumen se presentan los descriptores con letras mayúsculas.

SETIEN Q., Emilio y Lilia F. Pérez. Vías de formación del sistema de conocimientos bibliológico-informativo. *Ciencias de la información* (La Habana) 26 (2) jun. 1995: 42-46 (e)
Se explica cómo las conclusiones de los estudios realizados en la Biblioteca Nacional "José Martí" sobre el carácter, contenido y estructura de las disciplinas relacionadas con la profesión condujeron a la identificación de las vías de formación del sistema de conocimientos de referencia. Se exponen las razones por las cuales se denomina a ese sistema bibliológico-informativo y las que permiten reconocer disciplinas rectoras, complementarias y específicas. Se concluye que el sistema es una expresión de la tendencia del movimiento contradictorio de la ciencia a la especialización y a la integración. Se ejemplifica el caso de la bibliotecología. 13 refs.
BIBLIOTECOLOGÍA
SISTEMA BIBLIOLÓGICO-INFORMATIVO

BLISS Nonie J. The emergence of International Librarianship as a field (El surgimiento de la Bibliotecología Internacional como un campo) *Libri* (Copenague) 43 (1) Jan.-Mar. 1993: 39-52 (i)
La bibliotecología es por naturaleza internacional en alcance y propósito. En los últimos años esta dimensión ha crecido en diversas direcciones. Esto se analiza en cuatro áreas: contexto técnico; práctica profesional; educación; y el control y la normalización de la información y los formatos de información. Se señala que ha habido pocos trabajos críticos sobre la bibliotecología internacional. La literatura sobre este tema ha sido narrativa y descriptiva, trabajos de opinión, conjunto de datos o encuestas.
BIBLIOTECOLOGÍA

LANCASTER, F.W./y otros/ Ranganathan´s influence examined bibliometrically (La influencia de Ranganathan examinada bibliométricamente) *Libri* (Copenague) 42 (3) jul.-sep. 1992: 268-281 (i)
Se hace un estudio bibliométrico de las citaciones de los trabajos de Ranganathan en el período de 1959-1990, utilizando el Social Sciences Citation Index. Este análisis demuestra que su influencia no ha disminuido. Sus libros se citan con más frecuencia que sus artículos. Se ha citado en un amplio rango de contexto. Las "cinco reglas" son consideradas por varios autores como la base filosófica de la bibliotecología. Se hace referencia a sus trabajos en el análisis facético, la estructura temática, generación de tesauros por computadora, sistemas de indización y diseño de sistemas de expertos.
BIBLIOTECOLOGÍA
BIBLIOMETRÍA
ANÁLISIS FACÉTICO
CIENTÍFICOS NOTABLES: RANGANATHAN

LANCASTER, F.W. *Thesaurus construction and use: a condensed couse*. General Information Programme and UNISIST (PGI-85/WS/II). París: UNESCO, 1985, 89 p. (i)
Contenido de un curso que se basa en el Seminario Regional sobre lenguajes de indización organizado bajo los auspicios de la UNESCO por el Centro Argentino

de Información de Ciencia y Tecnología, celebrado en Buenos Aires en 1978. Contiene dos componentes principales: un conjunto de 84 ilustraciones que pueden convertirse en slides o transparencias; y un texto para explicar y ampliar las ilustraciones. Está dividido en 14 unidades temáticas. Tiene un componente práctico. Cada estudiante (o un pequeño grupo) debe completar un pequeño tesauro de alrededor de 200 términos en alguna área temática. Las instrucciones de este curso se corresponden generalmente con la guía de la Unesco para elaborar tesauros. 10 refs.
CURSOS DE ESTUDIOS
TERMINOLOGÍA
ELABORACIÓN DE VOCABULARIOS
TESAUROS
LENGUAJES DE INDIZACIÓN

7. SISTEMAS DE INDIZACIÓN. TIPOS DE ÍNDICES

Introducción. Componentes de un sistema de indización

Un sistema, en un sentido amplio, es un conjunto de componentes interrelacionados con el propósito de cumplir una serie de objetivos determinados. En su forma más general y sencilla se puede representar con el siguiente esquema:

Tanto para hacer el análisis de un sistema que esté funcionando, como para realizar el diseño de un nuevo sistema, es necesario, primeramente, identificar o definir de modo claro y preciso los siguientes aspectos:

- Objetivos
- Funciones
- Componentes
- Requisitos y limitantes para su óptimo funcionamiento
- Medio ambiente

Esta breve introducción permite definir un sistema de indización como un conjunto de partes interrelacionadas que forman un todo integral con la finalidad de cumplir el siguiente objetivo: Elaborar guías efectivas (índices) para conducir con calidad y eficiencia la búsqueda de información o documentos, de modo de satisfacer con prontitud las demandas de los usuarios. El sistema de indización tendrá que realizar una serie de funciones para cumplir el objetivo propuesto y obtener como principal producto final uno o varios índices. Las funciones se desarrollarán a través de los diversos pasos que conforman el proceso de indización.
Al estudiar la estructura de un sistema de indización hay que considerar los principales elementos que componen el sistema. Si el sistema opera con un vocabulario controlado, el LRI es uno de los elementos que juega un papel central. Esto llega a un extremo tal que, con mucha frecuencia, se identifica el lenguaje de indización (o de recuperación de la información) con el propio sistema.

Los recursos humanos y materiales son aspectos de primera importancia para el proceso operativo del sistema. Estos aspectos, conjuntamente con los costos y beneficios, hay que considerarlos detenidamente al analizar los requisitos y limitantes para el óptimo funcionamiento del sistema.
Es importante no perder de vista el medio ambiente, ya que un sistema de indización es un subsistema dentro de un sistema mucho más amplio al que se ha denominado sistema de información. Las variables del medio ambiente actúan, influyen y en gran medida determinan cómo opera el sistema, pero no pueden ser controladas por éste.

Tabla 6. Principales componentes de un sistema de indización

ELEMENTOS SUPERESTRUCTURALES

1. Objetivos del sistema
2. Directrices y reglas de indización
3. Metodología de trabajo

ELEMENTOS BÁSICOS

ENTIDADES (Componentes materiales)

4. Entradas
   - Documentos
   - Solicitudes de información
5. LRI
6. Salidas
   - Índices de documentos
   - Solicitudes de información indizadas
7. Otros (personal, materiales, equipos, etc.)

PROCESOS (componentes operacionales)

Es conveniente aclarar que aunque las entradas proceden del medio ambiente sí se han considerado como elementos básicos del sistema, ya que son los materiales a partir de los cuales se desarrolla todo el proceso de indización y sin ellos el sistema no existiría. Las salidas, por otro lado, son el producto final del proceso y se revierten en el ambiente del sistema. Tanto las entradas como las salidas están vinculadas al medio ambiente, pero pueden ser controladas por el sistema y forman parte de su estructura básica.

Principales sistemas de indización

Existen múltiples variantes de sistemas de indización, pero todos tienen como objetivo central el que se plantea en la introducción de este capítulo con un enfoque general. La diferencia entre los diversos sistemas depende de las vías que se utilicen para lograr los objetivos específicos que se propongan, de las técnicas que se apliquen para desarrollar el proceso de indización. Consecuentemente también diferirán sus productos finales, es decir, diferentes sistemas elaborarán índices con distintas estructuras, formatos y eficiencias.

Este trabajo estará centrado en el estudio de los sistemas de indización de materia por lo cual a continuación se describen las

características más representativas de los principales tipos de estos sistemas:

1. Indización con epígrafes
2. Indización coordinada (con descriptores)
3. Indización en cadena
4. Indización por rotación (indización permutada)
5. Indización de relación o articulada
6. Indización de citación

Es necesario aclarar que los sistemas de indización de la mayoría de los servicios informativos elaborarán dentro del mismo sistema varios índices. A modo de ejemplo vamos a citar el Chemical Abstracts Service que produce múltiples índices.

- Indización con epígrafes

Estos sistemas de indización emplean epigrafiarios, que son lenguajes alfabéticos de materia del tipo precoordinado con vocabularios enumerativos, que también se conocen como listas de epígrafes o listas de encabezamientos de materia. En estos sistemas los términos compuestos se crean en el momento de la indización, adicionando al epígrafe de entrada los subepígrafes necesarios, por tanto permiten un cierto grado de síntesis. El resultado es un índice alfabético de materia que constituye el catálogo de materia de muchas bibliotecas.

Estos índices se aprenden a manejar sin dificultad alguna. Su principal desventaja es que su elaboración requiere un serio esfuerzo intelectual, no posibilitan la recuperación multifacética y son muy voluminosos y, por tanto, retardan la búsqueda.

- Indización coordinada (con descriptores)

Los sistemas de indización coordinada utilizan lenguajes de descriptores. Con frecuencia estos sistemas operan con un fichero dual, el cual está formado por un fichero directo y un fichero inverso.

El fichero directo contiene los registros bibliográficos del documento que ofrecen toda la información bibliográfica y en muchas ocasiones también otra información, como por ejemplo un resumen del contenido y un código que posibilita su localización física.

El fichero inverso funciona como un índice que conduce al fichero directo. Los registros del fichero inverso tienen los términos índices (descriptores) y los números que refieren a los registros del fichero directo a los que se les ha asignado el descriptor. Por ejemplo, las tarjetas unitérmino ordenadas alfabéticamente por el descriptor constituyen un fichero inverso.

Estos sistemas tienen la ventaja de que posibilitan la recuperación multifacética de la información y que facilitan el almacenamiento y recuperación automática (son muy apropiados para los sistemas automatizados).

La mayoría de los lenguajes de descriptores carecen prácticamente de gramática, pues al indizar un documento simplemente se yuxtaponen los descriptores. Esto, que por un lado simplifica las cosas, también tiene la desventaja que puede producir falsas e incorrectas coordinaciones.

- Indización en cadena

El sistema de indización en cadena generalmente se apoya en una clasificación facética. En el análisis de los documentos los conceptos compuestos que representan su contenido temático se sintetizan a partir de las notaciones de los términos extraídos de las correspondientes facetas. Por ejemplo la revista referativa británica LISA (Library Information Science Abstracts) utiliza el sistema de indización en cadena. Para representar un tema compuesto extrae los distintos conceptos claves y forma con los términos correspondientes una cadena con el siguiente orden:

- Procesos técnicos
- Operaciones y agentes de los procesos, equipos y promoción de uso
- Fondos y materiales de bibliotecas. Uso de los materiales
- Tipos de bibliotecas. Usuarios
- Subdivisiones de lugar, tiempo y forma

Ejemplo 1:
Para indizar un documento que trata sobre:
“Los edificios de bibliotecas universitarias en Cuba” se combina la notación correspondiente a “Edificios” con la de “Bibliotecas” y con la de la Subdivisión de Lugar siguiendo reglas definidas sobre el orden de citación. El resultado es la cadena compuesta:

Edificios. Bibliotecas universitarias. Cuba

A partir de esta cadena se confeccionan las entradas al índice en el orden reverso, eliminando de modo sucesivo eslabones en la cadena. En el índice aparecerán las siguientes entradas:
Cuba: Bibliotecas universitarias. Edificios
Bibliotecas universitarias. Edificios
Edificios. Bibliotecas

Ejemplo 2:
Un documento se analiza y se describe su contenido temático por la siguiente cadena temática:

A B C D

A partir de esta cadena se derivan las siguientes entradas al índice

D C B A
C B A
B A
A

En los ejemplos 1 y 2 se puede apreciar que el último término de la cadena temática es el primero en la entrada más completa al índice. Así, si la cadena comprende cuatro términos, generalmente se hacen cuatro entradas al índice. La primera tendrá 4 términos, la segunda 3 y así sucesivamente.
También está establecido que puedan hacerse entradas adicionales para ampliar las posibilidades de la recuperación, tal como se muestra en el ejemplo a continuación.

Ejemplo 3:

Un documento se analiza y se escribe la frase temática que expresa su contenido:
“Los curricula en las escuelas bibliotecarias en Australia”

A partir de esta frase se construye la cadena temática, la cual se ordena desde el término más genérico hasta el más específico:
Australia. Escuelas bibliotecarias. Curricula
Escuelas bibliotecarias. Curricula
Curricula. Educación (Profesional). Bibliotecología
Educación (Profesional). Bibliotecología
En este caso hay una entrada adicional por educación. Además en el índice aparecerán las siguientes referencias:

Bibliotecología
  véase Escuelas bibliotecarias
        Educación (Profesional). Bibliotecología
Cursos
  véase Curricula

Profesional. Educación. Bibliotecología
  véase Educación (Profesional). Bibliotecología

La indización en cadena se puede definir como un método para elaborar un índice alfabético de materia de una forma semiautomática, de acuerdo con un proceso que comprende dos fases:
1ra. Fase: El indizador construye la cadena temática que conduce del término de nivel más genérico hacia el término de nivel más específico siguiendo los pasos siguientes:
- hace el análisis de contenido del documento, extrae los conceptos claves y construye una frase temática
- consulta la clasificación facética y selecciona las notaciones con los términos correspondientes que representen los conceptos claves contenidos en la frase temática
- ordena los términos (con sus notaciones), según el orden de citación establecido, construyendo la cadena básica temática
2da. Fase: A partir de la cadena básica temática se confeccionan las entradas al índice, eliminando de modo sucesivo eslabones en la cadena. Un buen mecanógrafo y con experiencia en este trabajo puede realizar esta fase.

- Indización por rotación (Permutada)
La indización por rotación generalmente se conoce en la literatura como indización permutada, generando los llamados índices permutados entre los que se encuentran los índices KWIC (Key Word in Context) y KWOC (Key Word Out of Context).
Estos sistemas no ofrecen control del vocabulario, sino que utilizan el lenguaje natural libre. Se basan en la rotación de las palabras significativas de los títulos de los documentos o de frases o de términos compuestos. Los índices que se generan son listas alfabéticas de las palabras clave presentadas en su contexto.

Ejemplo:

Un título de un documento se expresa por la siguiente secuencia de palabras.

P. Q. R. S. T.

En el índice KWIC aparecerán cinco puntos de acceso en una columna central los que mantendrán un orden alfabético. Cada punto de acceso se encuentra rodeado de su contexto:

				P	Q	R	S	T
			P	Q	R	S	T	
		P	Q	R	S	T		
	P	Q	R	S	T			
P	Q	R	S	T				

Estos sistemas se pueden aplicar de modo manual para indizar colecciones pequeñas, pero su creador, Peter Luhn de Estados Unidos, los concibió para ser utilizados en la indización automática.

La indización automática procede del siguiente modo. Primeramente se almacena en la computadora una lista negativa (stop-list) que contiene todas las palabras no significativas, las cuales no se deben seleccionar para indizar. Por ejemplo, artículos, conjunciones, preposiciones, verbos auxiliares y algunos adjetivos y nombres. Después el título del documento se introduce en la computadora y las palabras se seleccionan indirectamente con la lista negativa. Las palabras que no aparezcan en la lista se utilizan como entradas al índice. Para cada documento habrá tantas entradas en el índice como palabras clave contenga su título.

Ejemplo:
Se va a indizar un documento con el siguiente título:
“Los curricula de las escuelas bibliotecarias en Australia”
Se rota por cada palabra clave y se generan las siguientes entradas al índice:

en Australia. Los **curricula** de las escuelas bibliotecarias
Los curricula de las **escuelas** bibliotecarias en Australia
de las escuelas **bibliotecarias** en Australia. Los curricula
bibliotecarias en **Australia**. Los curricula de las escuelas

Después se ordenan estas frases alfabéticamente por la palabra clave.

- Indización de relación o articulada. Roles. Conectores

Los sistemas de relación se apoyan en una serie de principios lógicos con la finalidad de elaborar índices con entradas que se fundamentan en estructuras sintácticas. El enfoque de relación cobró especial importancia en el sistema de sintaxis de J. E. L. Farradane. Un aspecto esencial en este sistema son los llamados roles, que son indicadores representados mediante códigos numéricos, alfabéticos o alfanuméricos, que se utilizan para expresar la función de cada término en una cadena.
También es importante establecer las relaciones entre las palabras. Tres mecanismos se pueden utilizar para simbolizar estas relaciones:
- conectores de relación o enlaces (links), que pueden indicarse por códigos numéricos, alfabéticos o alfanuméricos

- preposiciones, conjunciones u otras palabras para especificar las relaciones entre los términos
- convenciones sobre el orden de las palabras

En los últimos años han surgido diferentes sistemas de indización de este tipo, en la mayoría de los cuales la labor intelectual la realiza el hombre y las tareas de rutina la computadora. A modo de ejemplo se explicarán dos de estos sistemas: el ASI, que corresponden a un índice articulado y el PRECIS que corresponde a un índice de relación.

ASI. El ASI (Articulated Subject Index) es un sistema de indización que elabora índices de materia articulados y que se diseñó por la Escuela de Postgrado de Bibliotecología y Ciencia de la Información en la Universidad de Sheffield en el Reino Unido. Se aplica en los "World Textile Abstracts" y "Rubber Plastics Research Association Abstracts".
En este sistema el indizador formula una frase que expresa el contenido esencial del documento. Los términos de esta frase que deben aparecer como entradas al índice los señala colocando los símbolos < > antes y después de cada término.
Las frases marcadas se introducen en la computadora que las procesa para crear una modificación para cada término de entrada.

Ejemplo:
El contenido del documento identificado con el número 1234 se puede expresar por la siguiente frase:

Silicosis en mineros de cobre 1234

Se señalan los términos que deben aparecer como entradas al índice:

<Silicosis> en <mineros> de <cobre> 1234

Las modificaciones por cada término de entrada las crea la computadora con el reordenamiento de la frase mediante un algoritmo que comprende seis reglas, el cual controla el orden de las palabras en las modificaciones:

Silicosis en mineros de cobre, 1234
Cobre, mineros de, silicosis en, 1234
Mineros de cobre, silicosis en, 1234

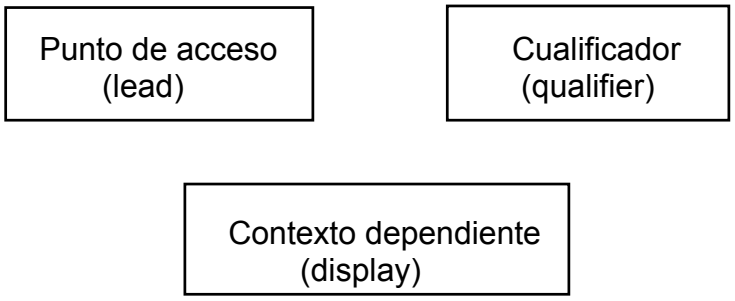
Las entradas del índice simulan las creadas por los indizadores que trabajan sin la computadora. Por supuesto que los indizadores son más flexibles y tienen otras posibilidades que no tiene la computadora, ya que pueden utilizar, si lo consideran conveniente, palabras no incluidas en la frase. No obstante, aunque la computadora pueda tener ciertas limitantes en ese sentido, compensa con creces esas limitantes con el trabajo y el tiempo que ahorra cuando se trata de indizar grandes colecciones.

PRECIS. El acrónimo PRECIS (**PRE**serve **C**ontext **I**ndexation **S**ystem) se refiere a un sistema de indización de materia que organiza los conceptos según el principio de "dependencia del contexto", es decir establece que el significado de un término está en dependencia de las palabras que lo rodean.
Este sistema se desarrolló en el Reino Unido por Austin Dereck y sus colegas y se aplicó en 1971 en la Bibliografía Nacional Británica (BNB) para proporcionar un índice alfabético de materia a los registros UK/MARC.

PRECIS evolucionó durante cuatro años a partir del sistema de indización en cadena que se aplicaba con anterioridad.
El proceso de indización se divide entre el indizador y la computadora. El indizador realiza la labor intelectual, todas las tareas que requieren el juicio humano. La computadora realiza las operaciones de rutina, implementa las decisiones del indizador y las referencias cruzadas y ordena alfabéticamente las entradas.
El resultado es una cinta magnética con el índice alfabético de materia de la BNB, que puede ser utilizado para la búsqueda y como punto de partida para tirar un índice impreso.
En este sistema hay dos aspectos esenciales:
- la sintaxis y
- la semántica

La sintaxis se ocupa de las relaciones entre los conceptos indizables del documento. Consiste en un esquema de operadores de rol que guían al indizador en el análisis del texto de un modo lógico y le ofrecen una base paso a paso para la identificación y el ordenamiento de los conceptos en una cadena de términos que expresa el contenido esencial del documento que se desea indizar.
La semántica se refiere al significado de los conceptos que son independientes de un determinado documento. Los términos que representan los conceptos conforman un tesauro abierto con un sistema sindético que contiene tres tipos principales de relaciones:
- relaciones de equivalencia (sinónimos y casi-sinónimos)
- relaciones jerárquicas (genéricas-específicas y parte-todo)
- relaciones asociativas

A los términos utilizados en la indización se le asignan números indicadores de referencia (RIN - Reference Indicator Number). Estos números conducen a la extracción automática, a partir del tesauro almacenado en la máquina, de la referencia cruzada apropiada véase o véase también.
De este modo se autogenera y desarrolla automáticamente el tesauro o sistema especial de clasificación, que se caracteriza por ser flexible y tener hospitalidad, ya que se puede, con relativa facilidad, hacer adiciones, eliminaciones o correcciones.
El PRECIS se basa en una serie de principios lógicos que son independientes de la rama del conocimiento y que se pueden resumir en los siguientes puntos:
1) A partir del análisis de contenido de los documentos se forman las cadenas de términos. De estas cadenas se seleccionan las entradas principales que servirán de puntos de acceso al índice.
2) Cada término de entrada debe ser coextensivo con el texto que indiza (Esto significa que aparece conjuntamente con el contexto, no se pierden eslabones). No hay pérdida de especificidad, ya que todos los componentes de la cadena original están presentes en cada entrada.
3) El orden de los términos en la cadena se establece mediante los operadores de rol.
4) Cada entrada al índice debe tener sentido cuando se lea y ser similar en forma al lenguaje natural.
5) Las entradas deben estar apoyadas por referencias cruzadas para establecer los vínculos entre las palabras semánticamente relacionadas.
6) Se utiliza un formato de entrada que contiene tres posiciones básicas, que esquemáticamente se representa así:



- El punto de acceso (lead) lo ocupa el término que sirve de acceso al índice.
- El cualificador (qualifier) aparece en la misma línea que el punto de acceso y expresa el contexto más amplio que el término situado en el punto de acceso.
- El contexto dependiente (display) aparece en una segunda línea, un poco hacia dentro, y expresa un contexto dependiente, más limitado que el del término situado en el punto de acceso.

Los términos se pueden ir corriendo y situarse en las diferentes posiciones. El hecho de que un término pueda o no aparecer en cualesquiera de las posiciones básicas está bajo el control del indizador.

El sistema PRECIS permite al usuario entrar al índice por cualesquiera de los términos significativos que conjuntamente representan el planteamiento de una temática compuesta.

Para lograr sus propósitos este sistema se compone de:

1) Un conjunto de 26 operadores de rol que indican la función de los términos y determinan su posición en la cadena de conceptos. Son una guía para el trabajo del indizador y dan las instrucciones a la computadora, pero no aparecen en el índice impreso.
2) Un tesauro abierto o vocabulario autorizado con una estructura de árbol modificada para evitar la dispersión terminológica. Este tesauro controla la forma de cada entrada en el índice y las referencias cruzadas.
3) Reglas que se basan en la gramática inglesa (se han aplicado a otras lenguas, por ejemplo danés, francés y alemán) expresadas mediante los operadores de rol para asegurar que las entradas sobre un tema dado se organicen consistentemente.

En el proceso de indización se va formando un fichero que almacena todos los registros de los documentos indizados los cuales se identifican por un número SIN (Subject Indicator Number). Con el SIN se puede localizar el registro de datos del documento que forma un paquete que se identifica con la etiqueta MARC. En este sistema se entra por el índice el cual conduce a la unidad con el SIN y al registro con toda la descripción bibliográfica.

En el proceso de indización el indizador tiene que realizar los siguientes pasos:

1) Hace el análisis de contenido del documento que sumariza en una o más frases.
2) A partir de esta frase escribe la cadena de términos:
   - primeramente busca un término que denote acción; después busca el objeto de la acción (o key system), que puede tener un número de elementos dependientes.
   - asigna a los términos los operadores de rol que expresen su función.
   - organiza los términos índices en una cadena de acuerdo con el valor de los operadores.

3) Asigna el número de Clasificación Decimal Dewey
4) Asigna el número LC (Library of Congress)
5) Asigna el encabezamiento de la LC
6) Asigna el RIN a todos los nuevos términos
7) Asigna el SIN

Existen 26 operadores divididos en cinco grupos:
7 operadores principales (números del 0 al 6)
6 operadores interpuestos (letras minúsculas de la **p** a la **t** y la letra **g**)
8 operadores diferenciantes (letras minúsculas **h**, **i**, **j**, **k**, **m**, **n**, **o**, **d**)
2 operadores conectivos (**v**, **w**)
3 operadores de interrelación (**x**, **y**, **z**)

Ejemplo:
Se desea indizar con el método PRECIS un documento titulado "La administración en los centrales azucareros en Cuba"

1) Se hace el análisis de contenido del documento que se sumariza en la siguiente frase:
   Administración de los centrales azucareros en Cuba
2) Se analiza esta frase en sus diferentes componentes que después se organizan en una cadena de términos:
   a) se identifica "administración" como el término que denota acción
   b) después se establece que el objeto de la acción (o key system) es "centrales azucareros", es decir es la entidad que se administra
   c) el medio ambiente es la localidad "Cuba"
   d) se asignan a los términos los operadores de rol que expresan su función y se marcan con el símbolo ✓ los términos que se desea que funcionen como puntos de acceso (lead)

      ✓
      Administración (2)
      ✓
      Centrales Azucareros (1)
      ✓
      Cuba (0)

   e) se organizan los términos índices en una cadena de acuerdo con el valor de los operadores

      Cuba (0)
      Centrales azucareros (1)
      Administración (2)

Tabla 7 Operadores de rol del PRECIS.

Operadores principales	
Entorno de los sistemas observados	0 Localización
Sistema observado	1 Sistema observado
	2 Acción
	3 Agente de la acción
Datos en relación al observador	4 Puntos de vista como forma
Sustancia seleccionada	5 Muestreo poblacional región de estudio
Presentación de los datos	6 Blanco / Forma

Operadores interpuestos

Elementos dependientes	p	Parte (propiedad)
	q	Miembro de un grupo
	r	Agregado
Elementos enlazantes	s	Definidor de rol
	t	Asociación
Coordinación	g	Concepto de coordinación
Operadores diferenciantes	(h, i, j, k, m, n, o, d)	
Operadores conectivos	(v, w)	
Operadores de interrelación temática	(x, y, z)	

- Indización de citación

La indización de citación se considera un sistema de indización de materia, aunque el proceso que se aplica y la estructura y características de los índices que resultan se apartan de todos los sistemas explicados. Además, las palabras que se utilizan como claves de búsqueda son los nombres de determinados autores.

Este sistema no se fundamenta en la asignación o extracción de términos para expresar el contenido del documento, sino en el hecho de que los autores al publicar sus trabajos suelen presentar un conjunto de referencias bibliográficas de los documentos consultados, los cuales tratan sobre temáticas iguales o afines a los asuntos por ellos tratados.

El estudio de las referencias bibliográficas demostró que éstas forman una red de documentos con temáticas afines y, por tanto, pueden utilizarse como claves para la localización del contenido de materia de los documentos.

Este sistema de indización es el que utiliza el "Institute for Scientific Information" de Filadelfia para publicar el "Science Citation Index" (Índice de Citación de Ciencias) que es una publicación trimestral con alcance internacional y multidisciplinario, que abarca un número considerable de las publicaciones científicas y técnicas más importantes del mundo.

Un índice de citación consiste en dos partes fundamentales:

1) índice de fuentes, que es un índice de todos los artículos publicados en un grupo selecto de revistas
2) índice de citación propiamente, que es un índice ordenado por autor de todos los artículos del grupo de fuentes analizadas

Para usar estos índices primeramente se localiza un autor conocido en el índice de citación y después se buscan las fuentes que citan sus artículos en el índice de fuentes. Si no se conoce algún autor hay un índice de materia ("permuterm") que conduce a los nombres de los autores.

En el Índice de Citación las entradas se ordenan alfabéticamente por el apellido del autor citado. Debajo de cada autor citado se ordenan cronológicamente las referencias citadas. A su vez bajo cada referencia citada aparecen, en orden alfabético del primer autor, las fuentes que citan.

De la publicación de estos índices se han derivado algunas investigaciones:

1. Estadísticas: Se ha contado la frecuencia de citación de las revistas de los autores y de los artículos.
2. Patrones de comunicación entre los investigadores: Relaciones interdisciplinarias, por ejemplo entre biología y medicina.

3. Evaluación de revistas: Basada en los datos estadísticos de la frecuencia de citación.

Un aspecto muy importante de estos índices es que no requieren del esfuerzo intelectual del indizador; se confeccionan mediante procedimientos de rutina que se realizan con computadoras digitales. No existe la posibilidad de elaborar estos índices manualmente, ya que es necesario manejar un excesivo número de datos. Los índices de citación se han calificado como efectivos, pero resultan voluminosos y caros.

CLASIFICACIÓN TIPOLÓGICA DE LOS ÍNDICES

Los índices pueden ser muy variados en dependencia del sistema de indización que se aplique para su elaboración, y la clave de búsqueda que interese destacar en el índice. En la Tabla 8 se relacionan los principales tipos de índices de acuerdo con la característica esencial del documento que se utilice como clave de orden para organizar el índice. La **clave de orden** es el punto de acceso al índice y funciona a su vez como clave de búsqueda-recuperación. Esta Tabla por supuesto que no es exhaustiva; pero sí incluye los índices más conocidos y de más aplicación.

En los índices de nombres de personas se puede señalar como ejemplo el índice de autores en el cual el apellido del primer autor es la clave de orden del índice y es al mismo tiempo el elemento que se utiliza para la búsqueda-recuperación.

Los índices de citación se explicaron aclarando que son índices de materia, pero se han incluido en los índices de nombre de personas porque en su estructura la clave de orden principal son los nombres de los autores citados.

Los índices cronológicos son los que se utilizan en casos que la fecha de publicación del documento sea un dato muy importante que es necesario utilizar para la búsqueda. Por ejemplo, en el caso de los libros raros la fecha de publicación es la primera clave de orden del índice.

Los índices del número de la norma (standard) se utilizan como complemento de otra serie de índices para tener acceso a las normas por el número con que han sido registradas en los organismos de normalización.

En los índices topográficos la clave de orden es el código numérico que corresponde al ordenamiento de los documentos en los estantes. Por ejemplo, en muchas bibliotecas clasifican los documentos con la Clasificación Decimal de Dewey. En estos casos el código de la clasificación que se le asigna al documento se escribe en el registro de datos (ficha catalográfica) y en el marbete que se pega al propio documento. Este código permite la localización física de los documentos, los cuales estarán agrupados por temáticas en el almacén, ya que el código de clasificación representa su contenido temático.

Tabla 8. Principales tipos de índices

1. ÍNDICES DE NOMBRE DE PERSONAS, INSTITUCIONES Y OTRAS ENTIDADES

1.1 Índices de autores

1.1 Índices de nombres citados en obras

1.2 Índices de citación
1.3 Índices de nombres de instituciones
1.4 Índices de otras entidades (Catálogos de universidades, colegios profesionales, editoriales, equipos de laboratorio, reactivos químicos, piezas de repuestos, etc.)

2. ÍNDICES DE MATERIA

2.1 Índices generales de materia (alfabéticos)
2.1.1 Índices de epígrafes
2.1.2 Índices de descriptores
2.1.3 Índices en cadena
2.1.4 Índices por rotación (permutados, KWIC, KWOC)
2.1.5 Índices de relación o articulados
2.2 Índices especiales de materia
2.2.1 Índices de nombres de sustancias químicas
2.2.2 Índices de nombres científicos (animales, plantas, medicamentos, enfermedades, virus, espectros, etc.)

3. ÍNDICES DE FÓRMULAS QUÍMICAS

3.1 Índices de fórmulas moleculares
3.2 Índices de anillos (sistemas cíclicos en química)

4. ÍNDICES NUMÉRICOS

4.1 Índices cronológicos (libros raros)
4.2 Índices de números de patentes
4.3 Índices de concordancia de patentes
4.4 Índices del número de registro de sustancias químicas
4.5 Índices del número de la norma (standard)
4.6 Índices topográficos

Estructura del índice de materia

Al analizar la estructura de un índice de materia se pueden considerar dos componentes esenciales que caracterizan a cualquier tipo de índice independientemente de otros rasgos que puedan distinguirlos. Estos componentes son:

- ENTRADA
- SISTEMA SINDÉTICO

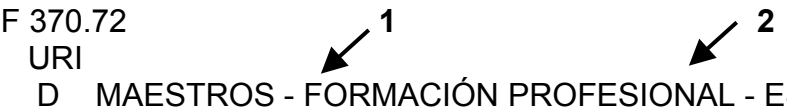
La entrada, a su vez, generalmente contiene las tres partes siguientes:

1 Término o punto de acceso
2 Complemento del término de acceso (subepígrafe, contexto o frase modificadora)
3 Referencia

El aspecto 2 no se presenta siempre en todos los índices. A continuación se ilustran varios ejemplos de entradas en distintos tipos de índices, señalando con los números 1, 2 y 3 las tres partes componentes.

Ejemplo:

Entrada en un catálogo de materia de una biblioteca que utiliza un sistema de indización con epígrafes



↑ Uria Rodríguez, Ma. Esther
3 Dimensiones personales de los profesores de E.G.B. en formación. } **2**
Barcelona, 1979, 36 p. 22 cm.

El **sistema sindético** contiene las referencias cruzadas, es una parte muy importante de los índices y guía al usuario de los términos no autorizados a los que se han utilizado en el índice con la palabra **véase**. Además envían, mediante la referencia de **véase también**, desde algunas entradas a otras conceptualmente relacionadas y que probablemente sean de interés del usuario. Su propósito es evitar que se pierdan entradas y ahorrar el tiempo del usuario.
El sistema sindético del vocabulario se corresponde en muchos casos, por lo menos en gran medida, con el sistema sindético del índice. También muchas veces cuando el vocabulario tiene un sistema sindético muy desarrollado se recomienda, para utilizar con el máximo aprovechamiento los índices, consultar el vocabulario antes de iniciar la búsqueda en el índice. Por ejemplo, así ocurre en los tesauros.
Algunos sistemas que tienen índices muy voluminosos publican, de forma separada, una guía sobre cómo utilizar el índice, en la cual aparecen todas las referencias cruzadas e indicaciones especiales para el mejor uso del índice. Este es el caso de la revista referativa "Chemical Abstracts" que publica el "Index Guide" para orientar al usuario sobre el uso adecuado de sus múltiples índices.

8. LOS TESAUROS: PROPÓSITO Y SURGIMIENTO DE LOS TESAUROS

Los Tesauros son vocabularios que se presentan en una nueva modalidad y se utilizan para el análisis y la recuperación de la información.
El propósito esencial de un Tesauro es ayudar al usuario a encontrar el término adecuado para un determinado concepto, para un significado dado. Esto contrasta con el propósito principal de un diccionario o un glosario que es el de definir o expresar el significado de una determinada palabra o término.

Definición y rasgos esenciales

La palabra "Tesauro" viene del latín "Thesaurus", que en plural es "Thesauri", y el diccionario de la Real Academia Española de la Lengua lo define como "Tesoro", que equivale a "tesoro de palabras".

Según los lineamientos del UNISIST un Tesauro puede definirse de acuerdo a su función y a su estructura; esto sirvió para elaborar de una forma más sencilla y global la siguiente definición: "Un Tesauro es un vocabulario controlado, estructurado y dinámico para utilizar en el proceso de indización de una rama determinada del conocimiento."
En esta definición están contenidos los tres rasgos esenciales que caracterizan y distinguen a un Tesauro de otros vocabularios, los cuales se analizarán a continuación:

1) Es un vocabulario controlado, porque es una lista de términos (o unidades lógicas) autorizados; un dispositivo de control terminológico. Indica los términos que se pueden utilizar para indizar (DESCRIPTORES) y señala los que no

están permitidos utilizar (NO DESCRIPTORES), los cuales envían con una referencia de USE al término autorizado o DESCRIPTOR.

2) Es un vocabulario estructurado, porque los términos no se presentan en forma aislada, sino que están relacionados con otros descriptores por su significación conceptual. Es decir, se presentan en su contexto semántico conformando lo que algunos autores denominan el "artículo léxico". Esto se puede señalar como el rasgo más característico de un Tesauro y que lo distingue, de modo significativo, de otros tipos de vocabularios.
3) Es un vocabulario dinámico, porque desde su concepción inicial se deben establecer normas sencillas para su mantenimiento y actualización, de modo de permitir eliminar, adicionar o modificar términos en correspondencia a los dictados de los avances que marque el progreso científico técnico.

Objetivos y funciones

El objetivo de un Tesauro es proporcionar un vocabulario que contenga los términos adecuados para la indización de los documentos y las solicitudes de información en una rama del conocimiento.

Sus funciones se desprenden de su objetivo. Esto significa que es un dispositivo de trabajo para el indizador y la persona que busca la información (el trabajador de la información o el usuario) que realiza las siguientes funciones:

- Permite traducir el vocabulario del lenguaje natural de los documentos al lenguaje artificial del sistema (el vocabulario del Tesauro)
- Posibilita encontrar el término preciso para un significado dado
- Facilita la recuperación de la información relevante ante una solicitud determinada
- Ayuda a mantener la uniformidad terminológica dentro de la rama del conocimiento para que haya sido elaborado.

Componentes de un tesauro. Formas de presentación

Hay tres componentes fundamentales de un Tesauro:

- Introducción
- Vocabulario Estructurado
- Clasificación Temática

En la INTRODUCCIÓN se explica la estructura del Tesauro y su finalidad. Se hacen aclaraciones sobre la forma en que se elaboró, la temática que abarca, los idiomas utilizados y otros rasgos que puedan ayudar a su uso más eficiente.

El VOCABULARIO ESTRUCTURADO está constituido por los descriptores con sus artículos léxicos y los no descriptores con su referencia de USE. Estos términos se pueden ordenar alfabéticamente o de acuerdo con el orden de la clasificación temática.

La parte de la CLASIFICACIÓN TEMÁTICA presenta, primeramente, la clasificación con sus códigos numéricos. Después, dentro de cada clase (o subclase) se relacionan todos los términos pertenecientes a esa clase en orden alfabético.

Muchos Tesauros presentan otras partes. Estas partes son índices auxiliares que ayudan a loca-lizar el término que se busca. De estos índices auxiliares, el de mayor utilidad es el índice rotativo de los descriptores. Es decir, un índice del TIPO KWOC (Key Word Out of Context. Palabra clave fuera del contexto) en el cual las

entradas corresponden a las palabras significativas del descriptor; cada término aparecerá tantas veces como palabras significativas contenga.

Tesauros monolingües y plurilingües. Recuperación multifacética

Los Tesauros monolingües están expresados en un solo idioma. Los plurilingües están expresados en dos o más idiomas, lo cual ofrece grandes ventajas para el intercambio de información y para mantener la uniformidad del léxico en el flujo informativo.

Los Tesauros, independientemente del idioma o idiomas que utilicen, tienen como finalidad esencial servir de apoyo a la recuperación multifacética de la información. Esto significa que son un complemento de sistemas que están diseñados para recuperar la información, por los distintos aspectos, que están contenidos en los documentos. Estos sistemas son fundamentalmente los sistemas postcoordinados y los sistemas articulados que ya se explicaron.

Elaboración de un tesauro

Análisis previo

Se parte del hecho que para estar en condiciones de analizar la posibilidad de elaborar un Tesauro se requiere de una experiencia de no menos de cinco años (preferiblemente más de cinco) en el trabajo de indización de los documentos y una base sólida de conocimientos acerca del proceso de indización, los lenguajes de recuperación de la información y de la actividad científica informativa en general.

Partiendo de esta premisa se formularán una serie de interrogantes que pueden servir de pauta para el análisis previo que debe hacer un Centro de Información para determinar si realmente es necesario elaborar un Tesauro.

- ¿Se ha establecido la conveniencia de usar un vocabulario controlado?

Hay que pensar que existen autores de mucho prestigio que son partidarios de la indización libre.

- ¿Si hay un volumen grande de documentos indizados con otro tipo de vocabulario, qué complicaciones traería hacer un cambio?

Hay que considerar, las complicaciones de tipo económico, técnico y material.

- ¿Es realmente necesario elaborar un Tesauro?

Hay que indagar si existe ya otro Tesauro de la misma temática que pueda utilizarse o adaptarse con pequeñas modificaciones.

Después de estas consideraciones y otras más que surgirán de acuerdo con las características del Centro y del Sistema de Indización utilizado, es posible que se llegue a la conclusión que es conveniente elaborar un Tesauro, entonces será necesario plantear y despejar una segunda serie de interrogantes.

- ¿Se cuenta con personal para desarrollar la labor?
- ¿Será necesario buscar asesoramiento y colaboración con otros Organismos?
- ¿Se dispone del tiempo suficiente?
- ¿Se cuenta con la colaboración del centro de cálculo para utilizar la computadora?

Si se encuentran respuestas objetivas y lógicas a todas las interrogantes que surjan al analizar la magnitud y complejidad de la tarea y si queda bien fundamentada la necesidad, conveniencia y posibilidad de elaborar un Tesauro, se puede y debe iniciar el trabajo.

Guía metodológica de trabajo

Lo primero es crear un equipo para desarrollar el trabajo y tener en cuenta que en la mayoría de los casos, será necesario utilizar como colaboradores a diferentes especialistas en los subcampos temáticos del Tesauro.

En el procedimiento para elaborar un Tesauro hay que contemplar dos aspectos esenciales:

- Estudios básicos, que brindarán un marco conceptual y una base teórica para fundamentar el trabajo.
- Tareas operativas, que señalan los pasos que hay que realizar para desarrollar las tareas concretas.

Teniendo en cuenta estos aspectos esenciales se expondrán unos lineamientos generales que pueden servir a modo tentativo, de guía metodológica de trabajo. Esta guía consta de cuatro fases, las cuales no hay que realizar estrictamente de forma sucesiva, sino que por lo contrario, en determinados momentos, es conveniente y necesario trabajar paralelamente en dos o más fases. De modo aproximado se puede decir lo mismo con respecto a los pasos que comprende cada fase.

En diferentes casos habrá que, necesariamente, hacer modificaciones y precisar detalles para adecuar la guía a la realidad concreta que se presente.

FASE PRIMERA

1-a. Estudios básicos:

- Estudiar la introducción, estructura y formato de otros Tesauros.
- Analizar las normas, lineamientos y reglas internacionales y nacionales para el establecimiento y desarrollo de Tesauros.
- Estudiar distintos tipos de documentos sobre la aplicación, uso y confección de Tesauros.

Al finalizar este paso se podrá:

- Determinar la estructura y formato del Tesauro que se va a elaborar.
- Tomar decisiones sobre las partes componentes del artículo léxico.
- Mandar a reproducir los modelos para el artículo léxico.
- Si se va a utilizar la computadora establecer los contactos con el Centro de Cálculo.
- Puntualizar los compromisos con la imprenta si se va a editar.

1-b. Determinar los subcampos de conocimientos que abarcará el Tesauro dentro de su temática principal.

En este caso será necesario consultar con diversos especialistas de la rama. Además, se debe analizar la estructura orgánica del Organismo que dirige la rama, los planes de estudio, las clasificaciones especializadas y las tablas de contenido de revistas y libros de la rama.

1-c. Diseñar una clasificación temática:

Esta clasificación comprenderá como clases principales todos los subcampos que abarque el tema del Tesauro. Las clases principales se pueden dividir en subclases (no más de 2 subniveles jerárquicos), o simplemente indicar los aspectos esenciales que comprende la clase. También se puede hacer una clasificación facética. Cada clase o faceta debe ser identificada con un código numérico.

Este paso es la culminación del paso 1-b, así que las consultas y análisis que se hicieron en ese paso se continúan y profundizan en este.

1-d. Tomar los datos bibliográficos de todos los documentos consultados:

Este paso queda abierto. Es decir, se seguirá añadiendo tarjetas con los datos bibliográficos de los documentos que se consultan en el transcurso del trabajo.

FASE SEGUNDA

2-a. Recopilar los términos que servirán de base de datos inicial para hacer el Tesauro.

Es conveniente copiar los términos en tarjetas (de papel o cartulina) de 12,5 cm x 7,5 cm. Lo más conveniente es aplicar el método empírico, o sea, tomar los términos que se hayan utilizado en el proceso de indización. También se pueden tomar de otros Tesauros, glosarios, listas, diccionarios, libros, etc.

2-b. Someter a una revisión crítica los términos recopilados.

Se debe analizar la utilidad de cada término desde el punto de vista de la indización y considerando como aspecto principal su papel en la recuperación de la información. Con este análisis surgirán nuevos términos, se modificarán otros y se eliminarán algunos.

2-c. A cada término se le asignará el código de la clase temática correspondiente.

Esta tarea se puede hacer de modo individual o en equipo, pero siempre se harán discusiones colectivas para analizar si el código ha sido bien asignado.

2-d. Hacer listas con todos los términos y sus códigos, ordenados alfabéticamente (una de control y una para cada miembro del equipo).

Es conveniente que estas listas se hagan a dos columnas y a cuatro espacios, lo cual permitirá que sin utilizar demasiado papel se puedan añadir los nuevos términos que se incorporen en el desarrollo del trabajo.

FASE TERCERA

3-a. Agrupar los términos escritos en las tarjetas por clases temáticas y distribuir los grupos entre los especialistas que integran el equipo.

A cada especialista le tocará uno o más grupos de acuerdo con los subcampos afines a su especialidad.

3-b. Pasar los términos y sus códigos al modelo del artículo léxico. En caso que sea una referencia de USE pasarlo al modelo del término NO DESCRIPTOR .

Estos modelos se deben llenar con lápiz, ya que hay que borrar con bastante frecuencia.

3-c. Confeccionar el artículo léxico de cada término.

Esta es la tarea más importante y trabajosa en la elaboración de un Tesauro. En el epígrafe 7.8.3 se dan algunas indicaciones para realizar esta tarea.

3-d. Revisar y discutir los artículos léxicos confeccionados.

Cada especialista entregará, en fecha previamente acordada, los artículos léxicos terminados. El responsable del equipo revisará cuidadosamente y hará las sugerencias de modificaciones en el propio modelo, pero sin borrar nada de lo que está escrito. Posteriormente el responsable del equipo (o el equipo completo) analizará el trabajo con la persona que confeccionó el artículo léxico. Se discutirán las discrepancias y se tratará de llegar a acuerdos o a soluciones de compromiso.

FASE CUARTA

4-a. Revisión final de los artículos léxicos.

Se verificará que se cumplan las relaciones recíprocas entre los descriptores, las reglas en cuanto a la forma, la exactitud y la veracidad de las notas de alcance y la correcta ubicación de cada descriptor dentro de su clase temática.

4-b. Pasar a los modelos que servirán para el procesamiento automatizado por la computadora.

4-c. Perfilar y precisar los últimos detalles.

Lineamientos para confeccionar los artículos léxicos

1) Se analiza el término y su clasificación (representada por el código numérico). De acuerdo con este análisis se podrá llegar a una de las tres conclusiones siguientes:

CONCLUSIÓN	OPERACIÓN A REALIZAR
El término debe ser eliminado	Se pone la palabra NO en la parte superior del modelo
El término no corresponde a la clase asignada	Al lado del código asignado se pone una interrogante o una flecha que señale el código que se considere más adecuado
El término corresponde a la clase asignada	Se procede a desarrollar el paso 2

2) Se analiza con profundidad el significado y alcance del término para determinar sus posibles relaciones semánticas. Para seleccionar los términos que estén semánticamente relacionados con el descriptor cabecera del artículo léxico se puede (y se debe) consultar nuevamente la lista de trabajo con todos los términos candidatos, otros tesauros, glosarios, clasificaciones, documentos sobre la temática, con compañeros especialistas, etc.

INTERROGANTES QUE SE PUEDEN FORMULAR	SI LA RESPUESTA ES AFIRMATIVA HAY QUE LLENAR EL CAMPO
¿Es necesario incluir una nota de alcance?	3- NA
¿Tiene uno o más términos con los que guarde una relación de equivalencia?	4- UP
¿Tiene uno o más términos genéricos?	5- TG
¿Tiene uno o más términos específicos?	6- TE
¿Tiene uno o más términos relacionados?	7- TR

3) Se revisan las relaciones recíprocas. Este paso se comienza a realizar con el grupo de términos de una clase temática después que se hayan completado todos los modelos de artículo léxico de los términos correspondientes a esa clase.

En cada modelo de artículo léxico que se haya completado se revisan las relaciones recíprocas de todos los términos con el término cabecera.

A cada término que pertenezca a la misma clase que el término cabecera se le hace una marca (✓) y se verifica si existe otro modelo de artículo léxico que contenga:

- Ese término en el campo 1- TÉRMINO
- El código correspondiente en el campo CÓDIGO
- El término cabecera del artículo léxico que se está revisando en el campo recíproco; a este último término recíproco también se le hace una marca (✓) para indicar que ya tiene su modelo de artículo léxico.

A cada término del artículo léxico que no pertenezca a la misma clase que el término cabecera se le pondrá a la izquierda el número de la clase a la que

pertenece. Posteriormente se hará el chequeo de las relaciones recíprocas con esos términos.

NOTA

Extracto de los tomos I y II de Indización, texto publicado por el Departamento de Textos y Materiales Didácticos del Ministerio de Educación Superior en la década del 80 que sirvió de base para la Asignatura de Indización en la carrera de Información Científico-Técnica y Bibliotecología de la Universidad de La Habana.

INDIZACIÓN DE DOCUMENTOS CIENTÍFICOS

Wilfrid Lancaster
Universidad de Illinois, Chicago (EE.UU)

El término «indización» se refiere a la asignación a un documento de una o más etiquetas que sirven para identificarlo y/o describirlo y para facilitar su posterior recuperación de algún tipo de base de datos. Aunque el término base de datos se aplica generalmente a una colección de registros en formato electrónico, que pueden ser procesados mediante ordenadores, su significado no tiene por qué ser tan restringido. También puede aplicarse a colecciones de registros (que representan documentos) en forma impresa. Así, pueden considerarse como bases de datos herramientas impresas tan importantes como Biological Abstracts y Chemical Abstracts.

Las etiquetas aplicadas a un documento en la indización pueden ser de varias clases, incluyendo nombres de autores, instituciones en las que trabajan, y números de los documentos (por ejemplo, de informes técnicos). La aplicación de tales etiquetas es relativamente fácil porque tienden a ser inequívocas: en la mayoría de los casos resultará obvio determinar quiénes son los autores, a qué instituciones representan, etcétera. Por el contrario, la indicación de materias -es decir, la asignación de etiquetas que representan el contenido de un documento (aquello «sobre lo que trata»)- resulta más difícil porque no es probable que dos personas se muestren totalmente de acuerdo «sobre lo que trata» un documento o sobre las etiquetas que mejor representan su contenido.

Este capítulo se centra exclusivamente en la indicación de materias. Además, trata sólo de la indicación de documentos para su inclusión en bases de datos electrónicas, puesto que en la actualidad ésta es su aplicación más importante; realmente está muy extendida la creencia de que los índices y las bases de datos impresas habrán desaparecido por completo muy pronto.

FASES EN LA INDIZACIÓN DE MATERIA

La indicación de materias comprende dos fases principales:

1. Análisis conceptual
2. Traducción.

Ambas fases están bastante separadas desde el punto de vista intelectual, aunque no siempre están claramente diferenciadas e incluso pueden darse de forma simultánea en la práctica.

El análisis conceptual, ante todo, implica determinar de qué trata un documento -es decir, cuál es su contenido- y «traducción» se refiere a la selección de un determinado término o grupo de términos para representar el contenido del documento.

Esta afirmación es un poco simple. La indización de materias se lleva a cabo generalmente para satisfacer las necesidades de una audiencia particular -los usuarios de un servicio concreto de información o de una base de datos específica. Una indización de materias eficaz implica decidir no solamente sobre el contenido de un documento, sino también sobre las razones que hacen probable que ese documento resulte de interés para un grupo concreto de usuarios. En otras palabras, no existe un conjunto «correcto» de términos de indización válido para cualquier publicación. Un mismo documento podría ser indizado de forma

muy diferente en distintas bases de datos, y debería ser así, si los grupos de usuarios están interesados en ese documento por razones asimismo diferentes.
Por tanto, el indizador tiene que hacerse varias preguntas sobre un documento:

1. ¿De qué trata?
2. ¿Por qué hay que incluirlo en la base de datos?
3. ¿En qué aspectos estarán interesados los usuarios de la base de datos?

Considérese, por ejemplo, un detallado informe técnico en el que se describen experimentos con varios cultivos agrícolas utilizando diferentes formas de riego. Si este documento se indiza para una base de datos de agricultura, es probable que el interés principal esté en el aspecto del cultivo -los propios cultivos, el efecto del riego sobre su crecimiento, su vitalidad y productividad; suelo, clima y otros factores medioambientales; y aspectos económicos, si fueran objeto de discusión. Por otro lado, este informe podría resultar de interés para varias bases de datos de ingeniería. En tal caso, los aspectos del cultivo pueden que resulten irrelevantes, frente al interés principal con relación al equipamiento para riego que se describe, los aspectos hidráulicos, o incluso los materiales usados (en tuberías, mangueras de riego, etc.), si se incluye algo inusual al respecto. Por tanto, para la base de datos de agricultura, los aspectos relativos a los cultivos han, de ser indizados con mayor detalle y los de.ingeniería con menor detalle (si son indizados), mientras que para la base de datos de ingeniería lo que se requiere es justo lo contrario. Para una tercera base de datos que trate, por ejemplo, de economía, también el énfasis será diferente.
Y así es como tiene que ser. Cuanto más especializada sea la base de datos, más probable será que la índización pueda y deba ser ajustada a la medida de los intereses precisos de sus usuarios. Fidel (1) utiliza la expresión «indización centrada en el usuario» para referirse al principio de indización basada en las necesidades de información de una audiencia concreta. De ello se desprende que los indizadores tienen que saber mucho más que los principios de la indización. En particular, tienen que estar absolutamente familiarizados con los intereses del colectivo al que se presta el servicio y con las necesidades de información de sus miembros.
El principio de indización centrada en el usuario podría incluso llevarse más lejos afirmando que, en relación con una determinada colección de documentos y un grupo concreto de usuarios, cualquier conjunto óptimo de términos de indización lo sería sólo en un determinado momento. Unos pocos años después, el mismo grupo de usuarios puede que necesite acceder a la misma colección (o a otra muy parecida), pero desde perspectivas diferentes. Un ejemplo obvio de ello podría ser una colección de informes técnicos en una organización dedicada a la investigación: al cambiar las prioridades de la organización y de sus intereses de investigación, puede cambiar también la forma en que la colección resulte útil para el colectivo. Esto puede ser especialmente cierto en el caso de la investigación interdisciplinar.
A lo largo de esta discusión se ha supuesto que un indizador humano puede tomar decisiones inteligentes sobre el «contenido» de un documento. Pero no todo el mundo está de acuerdo con eso. Algunos autores, incluso han llegado a afirmar que la indización de materias es una tarea virtualmente imposible porque no hay dos individuos que estén totalmente de acuerdo sobre el contenido de un documento. Se han publicado muchos trabajos sobre esta cuestión con puntos de vista teóricos o filosóficos –Wilson (2), Maron (3), Hutchins (4), y Swift et al. (5)

constituyen buenos ejemplos. Estas discusiones resultan de gran valor al recordarnos los problemas que conlleva el logro de una indización consistente dentro de una base de datos. Sin embargo, muchas son extraordinariamente pesimistas. La realidad es que los indizadores humanos experimentados son capaces de describir la materia de los documentos de formas más útiles para grupos específicos de usuarios, o de otro modo, las bases de datos no podrían justificar su existencia ni los servicios de información podrían funcionar eficazmente. La indización orientada al usuario no se centra sobre el «contenido» de un documento en un sentido teórico, sino en los rasgos o características del documento que lo hacen interesante para un grupo particular de usuarios; lo cual es un planteamiento muy práctico.

La traducción, el segundo paso de la indización de materias, implica la conversión del análisis conceptual de un documento en un conjunto concreto de términos de indización. Con respecto a esto, se puede establecer una distinción entre indización por extracción (indización derivada) e indicación por asignación. En la primera, se seleccionan palabras o frases presentes en el documento para representar su materia.

La figura 1 ofrece un ejemplo simple. La materia del documento (en realidad, título y resumen solamente) que aparece en la primera parte de la figura puede ser representada por completo mediante las palabras o frases que aparecen bajo el epígrafe de «Palabras clave». Estas han sido simplemente extraídas del título y del resumen y tienen exactamente la forma en que aparecen en el texto. Una extracción de este tipo puede ser hecha por humanos o, en ciertas circunstancias, mediante ordenadores.

La indización por asignación supone asignar términos a un documento a partir de una fuente distinta al propio documento. Los términos podrían proceder del propio indizador; por ejemplo, un indizador podría decidir que el término aumento de escala (comercial) que no aparece explícitamente en el resumen, es un buen término para utilizar con el documento de la figura 1.

La indización por asignación implica, más generalmente, el intentar la representación de la esencia del análisis conceptual mediante el uso de términos de algún tipo de vocabulario controlado. En la parte inferior de la figura 1 se puede ver la traducción que un indizador ha efectuado del análisis conceptual (y también de las palabras clave) con términos obtenidos de un vocabulario controlado.

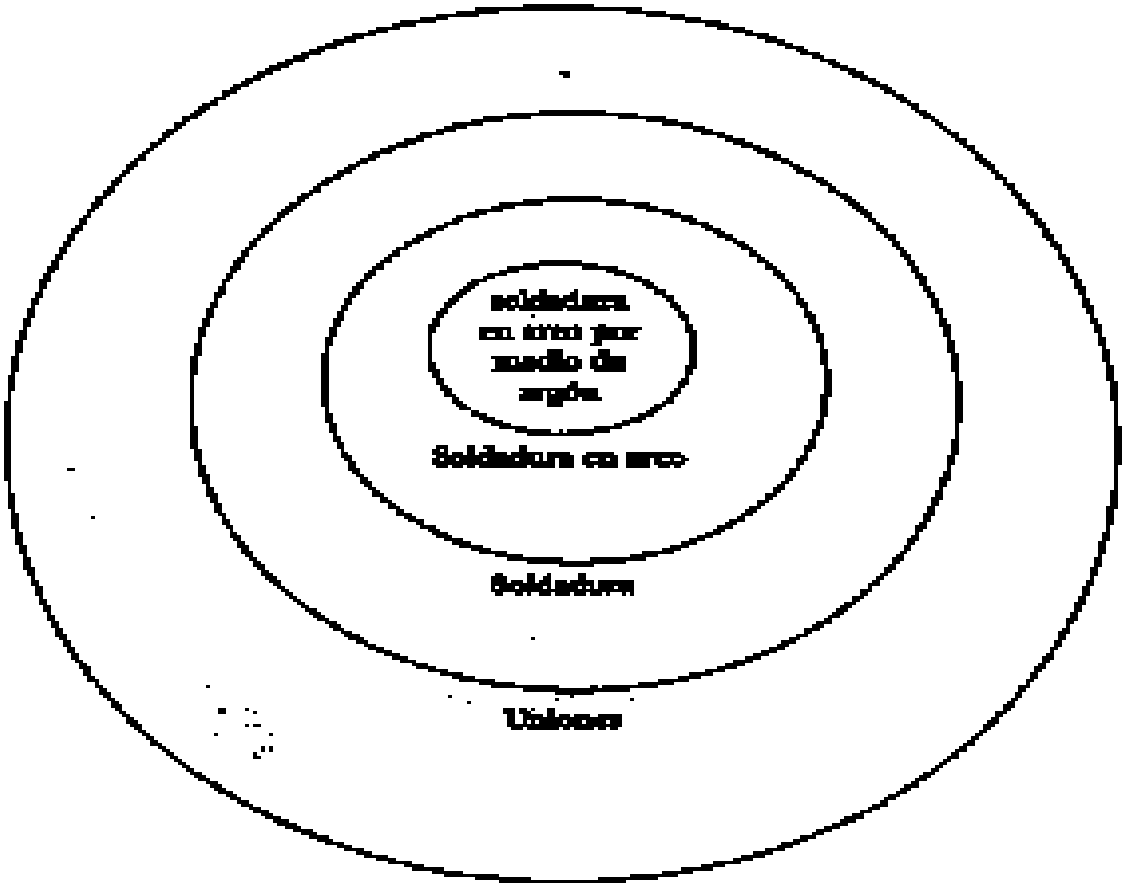


Fig. 1 Muestra de resumen de documento indizado con palabras clave y vocabulario controlado.

TÍTULO

El combustible de alcohol en la actualidad

RESUMEN

Describe las diversas fuentes a partir de las que se puede obtener etanol por destilación, incluyendo cultivos de varios tipos, residuos agrícolas, residuos municipales y lodos industriales. Compara los costes de producción del etanol con los de la gasolina y trata los problemas que plantea la transformación de la producción de etanol en una planta piloto en una producción

comercial a gran escala. Discute las ventajas e inconvenientes del gasohol, una mezcla de gasolina y de combustible de alcohol, y explora los problemas que hay que solventar antes de que los automóviles impulsados por alcohol resulten prácticos.

PALABRAS CLAVE

Combustible de alcohol, Etanol, Cultivos, Residuos agrícolas, Residuos municipales, Lodos industriales, Costes de producción, Plantas piloto, Producción comercial, Gasohol, Coches.

TÉRMINOS CONTROLADOS

Combustibles de alcohol, Gasohol, Costes de producción, Gasolina, Cultivos, Residuos agrícolas, Combustibles obtenidos de desechos, Residuos domésticos, Residuos industriales, Proyectos piloto, Utilización de residuos, Automóviles.

VOCABULARIOS CONTROLADOS

Un vocabulario controlado es básicamente una lista de autoridades. En general, los indizadores pueden asignar a un documento sólo aquellos términos que aparecen en la lista adoptada por la entidad para la que trabajan.

Sin embargo, normalmente el vocabulario controlado es más que una mera lista. Generalmente llevará incorporada alguna forma de estructura semántica. En concreto, esta estructura está diseñada para:

1. Controlar los sinónimos, eligiendo una de las formas como la aceptada y reenviando a ella desde todas las demás.
2. Distinguir entre homógrafos. Por ejemplo, Mercurio (metal) es un término totalmente diferente de Mercurio (planeta).
3. Juntar o enlazar aquellos términos cuyos significados muestren una relación más estrecha. Se pueden identificar explícitamente dos tipos de relaciones: la jerárquica y la no jerárquica (o asociativa). Por ejemplo, el término residuos domésticos está en relación jerárquica con residuos (como una especie de este término) y con residuos agrícolas (que también es una especie de residuos), al tiempo qué está asociado con términos como eliminación de residuos y saneamiento, que forman parte de jerarquías totalmente diferentes.

Pueden identificarse tres clases principales de vocabularios controlados: sistemas de clasificación bibliográfica (como la Clasificación Decimal de Dewey), listas de encabezamientos de materia, y los tesauros. En todos ellos se intenta presentar los términos tanto en orden alfabético como de forma «sistemática». En las clasificaciones bibliográficas, la parte alfabética tiene un carácter secundario y aparece en forma de índice de la ordenación principal, que es la jerárquica. En el tesauro, a simple vista la disposición de las entradas es alfabética, pero hay una estructura subyacente a la lista alfabética, establecida con referencias cruzadas. La lista tradicional de encabezamientos de materias es similar al tesauro en el uso de la ordenación alfabética. Se diferencia de él en que incorpora una estructura jerárquica imperfecta y no logra distinguir claramente entre la relación jerárquica y la asociativa. Estos tres tipos de vocabulario controlan sinónimos, distinguen homografías y agrupan los términos relacionados, pero aplican métodos algo diferentes para conseguir tales propósitos. El tipo de vocabulario controlado más utilizado en la actualidad, el tesauro, es tratado con más detalle en otro capítulo.

NÚMERO DE PUNTOS DE ACCESO

Un registro será recuperado cuando algún «elemento» utilizado para consultar la base de datos aparezca en el registro. Normalmente ese elemento de búsqueda será una palabra o una frase, pero podría ser otra cosa, como un valor numérico o algún tipo de código. Generalmente, los elementos son combinados de algún modo, en vez de ser usados de manera aislada. Por ejemplo, una búsqueda con el término «carreteras» recuperará solamente registros en los que esta palabra (es decir, esta cadena de caracteres) aparezca; «carreteras OR pistas de aterrizaje» recuperaría solamente aquellos registros en los que aparezcan una de las palabras (o ambas); y «carreteras AND reparaciones» recuperará sólo registros en los que aparezcan ambas palabras. Las palabras utilizadas en estos ejemplos son puntos de acceso, y se les llama así porque hacen que los registros sean accesibles (recuperables). Está claro que, cuantos más puntos de acceso tenga un registro, más probabilidades habrá de que sea recuperado.

En la figura 2 se muestra claramente la relación entre el número de puntos de acceso y la recuperabilidad. En ella, un indizador ha intentado representar el contenido de un artículo del Newsweek (21 de Abril de 1997) en tres niveles diferentes. El artículo trata fundamentalmente de la ciencia y de cómo la investigación científica y la interpretación de los resultados de investigación pueden verse influidos por los intereses políticos y sociales y por modas sociales. Este tema central ha sido cubierto por el indizador por medio de los cuatro términos elegidos para el nivel 1 de indización.

Fig. 2 Tres niveles de acceso a las materias tratadas "The Science Wars".

NIVEL 1	NIVEL 2	NIVEL 3
Ciencia	Ciencia	Ciencia
Política	Política	Política
Influencias sociales	Influencias sociales	Influencias sociales
Modas	Modas	Modas
	Constructivismo	Constructivismo
	Estudios de la mujer	Estudios de la mujer
	Crítica literaria	Crítica literaria
	Expansión del universo	Expansión del universo
	Cosmología	Cosmología
	Conducta animal	Conducta animal
	Heredabilidad	Heredabilidad
		Constante de Hubbe
		Babuinos
		Ornamentación
		Aborto
		Cáncer

Para ilustrar esta materia central, el autor del artículo ofrece varios ejemplos. En el nivel 2, el indicador intenta representar los temas adicionales más importantes junto a la materia central introduciendo más términos; y este mismo proceso se amplía todavía más en el nivel 3.

¿Qué efecto pueden tener estos diferentes niveles de indización sobre la recuperabilidad de un registro de este artículo incluido en una base de datos -por ejemplo, una base de datos que cubra los artículos sobre temas científicos en revistas de divulgación? Es obvio que un registro del nivel 1 podrá ser recuperado solo cuando un usuario utilice alguno de los cuatro puntos de acceso o una combinación de ellos, como, «ciencia AND política.» Y, sin embargo, este artículo también puede resultar de interés para alguien que busque información sobre un tema algo diferente. Por ejemplo, el artículo tiene cierta relevancia con relación a la influencia de los estudios de la mujer sobre la interpretacíón de los resultados de la investigación científica. En este caso el registro debería ser fácilmente recuperable con la indicación de nivel 2 y de nivel 3, puesto que los términos «ciencia» y «estudios de la mujer» están presentes. Es concebible que también pueda ser recuperado con la indización de nivel 1, pero ello exigiría mucho mayor ingenio y perseverancia por parte del usuario. Por ejemplo, el usuario tendría que darse cuenta de que los «estudios de la mujer» pueden ser considerados como «influencia social» sobre la ciencia. Más importante aún es que la búsqueda podría no limitarse al tema específico de interés, con lo que sería recuperado todo lo que trata de la influencia social en la ciencia (probablemente un gran número de elementos en el caso de una base de datos voluminosa), y la mayor parte de lo recuperado no tendría la más mínima relevancia, o interés.

La indización de nivel 3 permitiría que este artículo fuera recuperado por otros usuarios de la base de datos, que pueden encontrarlo relevante para otros intereses diferentes. Por ejemplo, alguien puede que quiera encontrar todos los artículos en que los puntos de vista sobre el aborto puedan haber influido sobre la interpretación de los resultados de la investigación científica. En este caso, el

artículo podría ser recuperado con la indización de nivel 3, pero no con la de nivel 1 o 2 -al menos no fácilmente y no sin recuperar al mismo tiempo una gran cantidad de material irrelevante.
Este artículo no se centra de manera significativa en la relación entre la cuestión del aborto y la interpretación de los resultados de investigación. Y, sin embargo, podría merecer la pena recuperarlo, especialmente si se trata de un tema sobre el que se ha escrito muy poco o tiene poca presencia en la prensa de divulgación. Además, incluso si la referencia a este tema es demasiado escasa, el artículo puede conducir al usuario a información adicional. Por ejemplo, el articulo menciona otros dos artículos relevantes aparecidos en revistas médicas y también da el nombre y la procedencia institucional de un investigador académico cuyo trabajo ha sido central en este tema concreto.

NIVELES ÓPTIMOS DE INDIZACIÓN

De todo lo que antecede se podría llegar a la conclusión de que en una base de datos se debería indizar todo tan completamente como fuera posible -al nivel 3 del ejemplo e incluso más allá. Pero no tiene que ser necesariamente así, por dos razones. En primer lugar, cuando se utiliza el esfuerzo intelectual humano, la indización exhaustiva requerirá más tiempo y resultará más cara que la indización a un nivel más superficial (se mencionan más adelante algunas alternativas a la indización humana). En segundo lugar, cuanto más completamente se indice un documento, más probable resultará su recuperación en casos en los que sería juzgado irrelevante. Volviendo al ejemplo de la figura 2, la indización de nivel 3 (en particular) cubre temas cuyo tratamiento en el artículo es muy marginal. Ciertamente el artículo se refiere a la constante de Hubble (una medida de la tasa de expansión del universo), pero realmente no da mucha información al respecto. Alguien que se plantee una búsqueda rigurosa sobre este tema (o sobre la conducta de los babuinos, el cáncer, la crítica literaria u otros temas tratados tangencialmente) puede decir que este artículo trata el tema tan brevemente que resulta inútil. Si se recuperan muchos registros como éste, puede que el usuario termine desconfiando de la base de datos.
La indización a nivel 3 también crea otros problemas: cuantos más términos se usen para indizar un documento, mayor probabilidad habrá de recuperar registros completamente irrelevantes, puesto que los términos sugerirán relaciones que son completamente falsas en lo que concierne a este artículo. Por ejemplo, el artículo podría ser recuperado en una búsqueda de información sobre el uso de los babuinos en la investigación sobre el cáncer (porque «babuino» y «cáncer» son puntos de acceso en el nivel 3), pero no tiene ninguna relevancia para el tema. Aunque las relaciones falsas pueden darse en cualquier nivel de indización («modas en política», «modas en crítica literarias, etcétera), es obvio que serán más frecuentes cuando se usan muchos términos que cuando se usan pocos.
Todo esto sugiere que es probable que exista un cierto nivel «óptimo» de representación de un documento dentro de una base de datos. Excepto en algunas situaciones inusuales, ese óptimo no es cuantificable con exactitud: nunca se podrá decir, por ejemplo, que diez términos es lo «correcto» y que once son demasiados y nueve muy pocos. Por otra parte, basándose en la experiencia práctica, se puede llegar a la conclusión de que una indización con unos diez términos (de promedio) por registro parece dar mejores resultados que un promedio de cinco o de veinte términos por registro.

Cuál es el mejor nivel en cualquier situación dependerá de las características de los documentos representados en la base de datos y de la forma en que ésta es utilizada. Si los usuarios de una determinada base de datos siempre necesitan realizar búsquedas exhaustivas -no se pueden permitir el lujo de perder nada- será necesaria una indización detallada, incluso si, por las razones aducidas anteriormente, ello puede producir, a menudo, resultados de poca relevancia. Aunque exista la necesidad de búsquedas exhaustivas, posiblemente en ciertos casos de asistencia sanitaria o de atención a los clientes (servicio de ayuda), es poco usual. Lo más común es que un usuario de una base de datos esté buscando aquellos documentos que ofrezcan la mayor información sobre algún tema más que cualquier posible referencia al mismo.

Las características de los documentos incluidos en la base de datos también influirán mucho sobre el número de puntos de acceso requeridos. Cuanto más complejos y multifacéticos sean los documentos cubiertos, más necesaria será la indización de forma exhaustiva. En una empresa de ingeniería, por ejemplo, puede que merezca la pena una indización muy detallada de los archivos de los contratos de la compañía, con términos que representen a todos los materiales usados, las condiciones de funcionamiento (como temperaturas y presión), dimensiones de los productos, etcétera. El detalle de la indización no estará justificada por el tamaño y complejidad de los archivos, sino por la gran importancia de esos recursos para la compañía. El coste de una indización detallada está más que justificado si sirve para evitar que la compañía tenga que invertir en volver a diseñar un componente ya diseñado anteriormente o para evitar que tenga que hacer de nuevo un diseño costoso o un error de instalación.

Diferentes categorías de materiales pueden ser indizados a diferentes niveles, incluso dentro de la misma base de datos, debido a las diferencias en su complejidad o valor para la organización. Por ejemplo, en la base de datos de una organización industrial, los informes técnicos de la propia compañía probablemente habrán de ser indizados con mayor detalle que los informes adquiridos de fuentes externas; y los artículos de interés para la compañía, obtenidos de las revistas técnicas, pueden ser indizados incluso con menos términos.

ESPECIFICIDAD DEL VOCABULARIO

La longitud de un registro de una base de datos (número de puntos de acceso proporcionados) constituye un factor determinante para su recuperabilidad. Visto de forma algo diferente, si hay que realizar búsquedas razonablemente exhaustivas en una base de datos, los registros deben tener la suficiente longitud. Al mismo tiempo, se quiere evitar una situación en que las búsquedas en la base de datos recuperen con frecuencia un gran número de registros irrelevantes. Para reducir esa posibilidad, la terminología utilizada para representar a los documentos en la base de datos tiene que ser suficientemente específica.

Los términos utilizados para el acceso por materias tienen que ser lo bastante específicos como para permitir que las búsquedas se realicen a un nivel de detalle adecuado a los intereses de los usuarios de la base de, datos. Por ejemplo, si un usuario quiere información sobre soldadura en arco por medio de argón, puede que no quiera recuperar todo lo que haya sobre soldadura en arco y seguro que no querrá todo sobre soldadura en general. Para este usuario, el vocabulario tiene que ser más específico que «soldadura» o incluso que «soldadura en arco»; el

término preciso que necesita es el de «soldadura en arco por medio de argón». Está claro que el productor de una base de datos debe tener unos buenos conocimientos sobre las necesidades e intereses de las personas que probablemente la utilizarán.
En la figura 3. se muestra el efecto de la especificidad sobre las posibilidades de recuperación. Piénsese en un usuario que quiere utilizar una base de datos de ingeniería en busca de documentos que traten de la «soldadura en arco por medio de argón». Si los registros son totalmente específicos, el término aparecerá en todos los registros relevantes: el usuario debería tener la posibilidad de recuperar todos los registros relevantes y sólo los relevantes. Sin embargo, a medida que el vocabulario se va haciendo menos específico, la búsqueda podría obtener cada vez más registros completamente irrelevantes: todos los que tratan de soldadura en arco, de soldadura, o incluso de ¡uniones! Un nivel apropiado de especificidad es el requisito más importante de un vocabulario controlado utilizado en la indización.

Fig. 3 Efecto de la especificidad del vocabulario sobre la capacidad de recuperación.

MÉTODOS AUTOMÁTICOS

La indización humana suele requerir mano de obra intensiva y resultar costosa, de manera que hay un gran interés en encontrar procedimientos alternativos que hagan accesibles los documentos mediante alguna base de datos. Son posibles dos alternativas: la indización automática y la búsqueda textual. Si se tiene un texto en formato electrónico, un ordenador es capaz de extraer palabras o frases que puedan ser buenos indicadores del contenido del texto. Las palabras o frases así seleccionadas se convierten en puntos de acceso de búsqueda (términos de indización) del texto, en lugar de los puntos de acceso elegidos por humanos. Esta es la indización automática por extracción.
La forma más simple de indización por extracción se basa exclusivamente en la frecuencia de las palabras. Es decir, las palabras comunes (artículos, preposiciones, conjunciones) se ignoran, pero todas las demás se ordenan por su frecuencia de aparición. Aquellas palabras o frases que aparecen en el texto con más frecuencia son seleccionadas. Puesto que la frecuencia de aparición es a menudo una buena medida de relevancia, la indización automática de este tipo puede tener un buen éxito. El ordenador puede extraer aquellas palabras y frases que un humano hubiera seleccionado del texto.
Pero la indización por extracción puede ser más sofisticada. Por ejemplo, se pueden hacer programas de extracción que ignoren palabras o frases que aparezcan con frecuencia en el texto pero que también aparezcan frecuentemente en toda la base de datos. Así, sí la base de datos tratara de irrigación, este término no debería ser seleccionado nunca, incluso aunque aparezca con mucha frecuencia en gran cantidad de documentos. Por otra parte, palabras o frases que aparezcan raramente en la base de datos serán seleccionadas incluso aunque no aparezcan con frecuencia en el texto tratado. Así, el término «geotérmico» puede ser seleccionado incluso si aparece pocas veces en el documento, porque es un término de rara aparición en la literatura sobre irrigación.

El software para indización por extracción puede también tener en cuenta otros criterios como la posición dentro del texto: a las palabras en títulos, encabezamientos de secciones y quizás en otros lugares, en el proceso de selección se les puede asignar más peso que el dado a otras palabras.
En algunas variantes especiales de indicación por extracción, los programas de selección buscan y extraen textos de un tipo determinado, como nombres de individuos o de organizaciones.
Los ordenadores también pueden ser utilizados para incluir textos dentro de categorías concretas preseleccionadas. Esta es una forma de indización por asignación porque los términos de indización (o los códigos de las categorías) son asignados al texto por el ordenador. Los programas que llevan a cabo este tipo de operación lo hacen comparando las palabras del texto con las palabras asociadas a las categorías que hay que asignar. Se asigna una categoría (un término) cuando las palabras del texto se ajustan suficientemente bien con el perfil de las palabras de la categoría. Por ejemplo, el perfil para la categoría «aerodinámica» puede incluir palabras o frases como «resistencia», «aleteo», «capa límite», «fuerza de sustentación» y «flujo de fluidos», además de la propia «aerodinámica».
En general, este tipo de indización se realiza con la ayuda del ordenador en vez de estar completamente automatizada. Como los programas, a veces, asignarán las categorías de forma incorrecta e incluso fallarán en la asignación de alguna categoría que debería ser identificada, normalmente la categorización automática es validada por humanos.
Los procesos automáticos más simples, basados en la extracción a partir del texto, son poco costosos y capaces de producir representaciones aceptables del texto. Pero la asignación automática de categorías preestablecidas es algo más costoso, porque los perfiles de las categorías deben actualizarse con frecuencia, lo que puede ser muy costoso en el caso de contar con muchos cientos o miles de categorías. Es muy poco probable que este procedimiento más costoso sea completamente automático, porque generalmente se hará necesaria la revisión por humanos. Aunque los procedimientos de indización automática pueden resultar satisfactorios para muchos fines, es improbable que produzcan representaciones de la misma calidad que las debidas a indizadores expertos.

BÚSQUEDA TEXTUAL

La búsqueda textual supone que la base de datos contiene texto almacenado en formato electrónico y que el texto se puede buscar. Es decir, quien consulta la base de datos puede utilizar los programas de búsqueda para encontrar textos en los que aparezca una determinada palabra o combinación de palabras. El texto almacenado puede ser incluso el documento entero -un artículo completo de una revista o diario, la totalidad de un informe técnico, un fragmento de cartas, o cualquier otra cosa- o puede ser más pequeño que el texto completo -un extracto de algún tipo, un resumen, o quizás simplemente el título del documento.
Desde luego, actualmente, la mayor parte de los textos impresos en papel existen primero en formato electrónico. Esto, junto al hecho de que el almacenamiento electrónico se haya abaratado tanto, ha llevado a la creación de numerosas bases de datos de texto completo de gran tamaño. Casi todas las búsquedas en Internet implican la búsqueda de texto completo o parcial.

Es muy conveniente tener texto completo en formato electrónico de forma que pueda ser visualizado en pantalla o impreso cuando convenga. Sin embargo, la posibilidad de buscar el texto completo no siempre es mejor que poder buscar algo menor, como un extracto o un resumen.
La búsqueda de texto puede tener un éxito completo cuando pueda ser llevada a cabo utilizando términos muy específicos y/o inusuales, sobre todo en el caso de distintos tipos de nombres. Pero se logra bastante menos éxito cuando se trata de «conceptos» amplios. Volviendo al ejemplo de la figura 1, debería resultar relativamente fácil hacer una búsqueda sobre «etanol» en una base de datos textual, pero resultaría extremadamente difícil buscar sobre «cultivos», porque hay cientos de maneras diferentes para expresar esta idea (es decir, nombres de cultivos concretos). En una búsqueda textual sobre etanol obtenido a partir de cultivos, puede ser más fácil recuperar todos los documentos en los que aparece el término «etanol», y eliminar manualmente todos los irrelevantes, antes que intentar pensar en todos los términos posibles de cultivos.
Para este tipo de búsquedas amplias y genéricas, del tipo de «cultivos», es para lo que resulta más útil el vocabulario controlado, porque si está construido de forma apropiada debería enlazar todos los términos de cultivos e incluso permitir la realización de una búsqueda de todo el grupo de términos de cultivos («cultivos» y todo lo que se encuentre por debajo en la jerarquía) con una única orden.
Aunque la indización por humanos bien preparados resulta cara, sin embargo, tiende a hacer más eficiente y menos costosa la búsqueda en las bases de datos. En otras palabras, una base de datos sin indización humana y sin control del vocabulario descarga el peso (y el coste) sobre los usuarios.

INDIZADORES

El nivel y el tipo de formación exigida a los indizadores varían considerablemente en función del tipo de indización y de la materia a cubrir. La indización por extracción, como se ha visto, puede hacerse con éxito por medio de ordenadores; de modo que, las personas también pueden hacer lo mismo con éxito, con muy pocos conocimientos o experiencia. Por ejemplo, se podría formar a alumnos de bachillerato para llevar a cabo este tipo de indización con muy pocos conocimientos o experiencia. Podrían ser preparados para trabajar casi como los ordenadores -buscando palabras o frases que aparezcan con frecuencia, dándoles más peso si aparecen en determinados lugares (por ejemplo, títulos, subtítulos y pies de figuras), etcétera.
Por otro lado, la indización por asignación exigirá mucha más formación y preparación, especialmente si el indizador debe elegir los términos más adecuados de un vocabulario extenso y cuidadosamente controlado. Los indizadores deberían tener una cierta familiaridad con la materia a tratar, y entender su terminología, aunque no tengan por qué ser necesariamente unos expertos en la materia. Algunas organizaciones han tenido problemas con indizadores demasiado “expertos” -que tienden a interpretar demasiado y quizás a ir más allá de lo que el propio autor quiere expresar (por ejemplo, indizando una posible aplicación no específicamente identificada en el artículo), o incluso a mostrar prejuicios no indizando cuestiones del autor que para ellos resulten inaceptables. Sin embargo, la falta de conocimiento de la materia puede conducir a una sobreindización. Incapaz de distinguir entre dos términos, puede que el indizador asigne ambos, cuando sólo es necesario uno o sólo uno es el correcto.

Como se ha subrayado con anterioridad, conocer los intereses de los usuarios de una base de datos es particularmente importante porque la «buena» indización debería ajustarse a las necesidades específicas de un determinado colectivo, siempre que sea posible. Los años de experiencia también deberían constituir un factor que afecte a la calidad de la indización, lo mismo que determinadas características del individuo como la capacidad de concentración, de lectura y de comprensión rápidas. Por último, y posiblemente lo más importante de todo, un buen indizador debe disfrutar con su trabajo. Resulta poco probable que se obtenga una buena indización de personas que odian lo que están haciendo.

CONSISTENCIA Y CALIDAD DE LA INDIZACIÓN

Generalmente se usa la palabra «consistencia» para referirse al grado en que dos o más indizadores están de acuerdo sobre los términos que hay que asignar a un documento. De nuevo, es probable que el grado de consistencia dependa en gran medida del tipo de indización. Está claro que la indización por extracción mediante ordenador será plenamente consistente; de igual modo, puede esperarse un alto nivel de consistencia entre humanos cuando la extracción sea hecha por humanos.

Por otro lado, como se sugirió en la discusión sobre el «contenido», es muy poco probable que se obtenga un nivel de consistencia muy alto en una indización por asignación que utilice un vocabulario amplio de términos muy específicos. No obstante, incluso en esta situación, se pueden conseguir niveles aceptables de consistencia gracias a buenos programas de formación de indizadores, a unas directrices claras, a distintas ayudas de indización (por ejemplo, herramientas accesibles en línea o ayudas en línea), y a procedimientos de control de calidad (por ejemplo, el trabajo de una indizador novato revisado por otro con más experiencia).

«Calidad» es un concepto más impreciso cuando se aplica a la indización. Puesto que ésta es algo esencialmente subjetivo, y puesto que diferentes colectivos de usuarios pueden tener diferentes intereses al respecto, no puede existir una única indización «correcta» para cada documento. Una «buena» indización para una base de datos concreta supone que los indizadores asignen términos que ofrezcan puntos de acceso útiles a los usuarios de esa base de datos. «Útil» aquí significa que los puntos de acceso proporcionados permitan a los usuarios recuperar la mayor parte de los documentos que quieren recuperar, pero evitando, al mismo tiempo, la recuperación de una gran cantidad de documentos no deseados.

Para una discusión detallada sobre consistencia y calidad de indización, así como sobre la relación entre ambas, véase Lancaster (6).

EL FUTURO DE LA INDIZACIÓN

A pesar de la proliferación de bases de datos textuales y del hecho de que cada vez son más accesibles por Internet, parece poco probable que la necesidad de indizadores experimentados desaparezca en el futuro más inmediato. La facilidad con que se puede hacer una base de datos accesible a través de Internet anima cada vez más a las organizaciones a desarrollar las suyas propias -por ejemplo, a las bibliotecas para producir bases de datos de recursos importantes a nivel local. La indización humana, con alguna forma de control de vocabulario, puede aumentar considerablemente la utilidad de tales recursos. Además, las

organizaciones pueden construir bases de datos útiles para ellas mismas descargando documentos de diversas fuentes de Internet.
Puede que resulte necesaria una indización local para aumentar el valor de tales bases de datos. Del mismo modo, algunos bibliotecarios están comenzando a darse cuenta de que una función importante de la biblioteca en un entorno digital puede ser la de construir recursos en red relevantes a nivel local. Por último, los desarrollos tecnológicos han creado nuevos retos, como los asociados con la indización de bases de datos de imágenes y sonidos. Puede que pase mucho tiempo antes de que los ordenadores puedan reemplazar totalmente a los humanos en la indización y en las demás tareas de tipo intelectual, necesarias para la recuperación de la información.

REFERENCIAS

Capítulo 6 de: Procesamiento de la información científica. Madrid: Arco/Libros, 2001, pp. 164-181.
(1) FIDEL, R., “User-centered indexing”. En: Journal of the Ameiican Society for Information Science, 1994, 45, 572-576.
(2) WILSON, P., Two Grands of Power: an Essay on Bibliographical Control. Berkeley: University of California Press, 1968.
(3) MARON, M. E., “On indexing, retrieval and the meaning of about”. En: Journal of the American Society for Information Science, 28,1977, 38-43.
(4) HUTCHINS, W. J., “The concept of «aboutness» in subject indexing”. En: Aslib Proceedings, 1978, 30,172-181.
(5) SWIFT, D. E. et at., “Aboutness' as a strategy for retrieval in the social sciences”. En: Aslib Proceedings, 1978, 30,182-187.
(6) LANCASTER, F. W., Indexíng and Abstracting in Theory and Practice. 21 ed. Urbana-Champaign: University of Illinois, Graduate School of Library and Information Science, 1998.

LENGUAJE NATURAL E INDIZACIÓN AUTOMATIZADA

Eva María Méndez Rodríguez
José Antonio Moreiro González
Universidad Carlos III de Madrid (España)

INTRODUCCIÓN

Vivimos en un mundo esencialmente lingüístico en el que las cosas son lenguaje y el lenguaje es una cosa. La cultura, la producción científica, y en definitiva, el conocimiento que aporta al ser humano el dominio de la realidad, se conforma, se construye y difunde a través del lenguaje. El hombre piensa, lee, y escribe gracias al lenguaje (al lenguaje natural) de tal suerte que su código se erige como un potencial comunicativo.

En este contexto de la comunicación humana, la Documentación presenta una estructura lingüística(1) ya que el discurso sobre el que se emiten los datos se ejecuta en lenguaje natural, como *un aluvión de estructuras cognitivas en lenguaje natural.* Si bien es cierto que el lenguaje natural es aquel conjunto de signos y símbolos orales y escritos por medio de los cuales los seres humanos se comunican entre sí, dentro de este trabajo definiremos lenguaje natural como aquel conjunto de palabras utilizadas por un autor para expresar sus ideas en un documento.

Es evidente pues, que existe una estrecha relación entre la Lingüística y la Gestión de la información, que podríamos explicar haciendo una extrapolación del concepto saussuriano de *signo lingüístico*, compuesto por significante (plano de la expresión, esto es, los grafemas que componen los términos de los documentos científico-técnicos) y significado (plano del contenido, o de la esencia semántica de los conceptos sobre los que se realiza el análisis de contenido en Documentación):

Significante → análisis formal	Significado → análisis de contenido
• Análisis morfológico • Análisis sintáctico • Análisis morfosintáctico • Análisis fonológico	• Análisis semántico (de base textual)
Descripción física o catalogación	**Resumen e indización**
ANÁLISIS DOCUMENTAL = Referencia del documento fruto del análisis	

A pesar de la omnímoda implicación de la Lingüística en el Análisis Documental, en esta aproximación nos centraremos en la semántica, y concretamente en la

semántica informática de cuyo desarrollo depende en gran medida la indización automatizada.
Como venimos diciendo, la comunicación científica se establece en lenguaje natural, un lenguaje que en su expresión escrita adolece de serias ambigüedades e imprecisiones derivadas precisamente de la falta de significado unívoco y preciso de las palabras que lo componen; presenta múltiples dificultades para el tratamiento de la información al estar compuesto por decenas de miles de palabras, y estar sujeto a diferentes accidentes léxico-semánticos (como la homonimia, polisemia, sinonimia, y figuras retóricas como anfibología, metáfora, símil, metonimia, anáfora, sinécdoque, etc.) que impiden la univocidad del signo lingüístico, y por ende, la comunicación exacta.
Pese a ello, hoy, el tratamiento y la recuperación de información en Lenguaje Natural es posible gracias a la intervención del ordenador. Cada vez son más abundantes los *software* documentales basados en el lenguaje natural que se destinan a interrogar bases textuales constituidas tanto en lenguaje cotidiano como en una terminología especializada.

La trascendencia de estos programas para el tratamiento y la indización del lenguaje natural aumenta en el contexto en que nos encontramos: la explosión de la información textual posibilitada por ordenador, donde la edición electrónica a finales del siglo XX se ha convertido en un hecho a la vez que un problema para la recuperación de información. Por ello, a través de este trabajo, proponemos mostrar cómo ha evolucionado la indización automática en la gestión de las palabras desde los inicios en lenguajes absolutamente libres, hasta el momento presente determinado por la regularización de las palabras en términos contrastados mediante tesauros y bases de conocimiento.

DE LA INDIZACIÓN A LA INDIZACIÓN AUTOMATIZADA: JUSTIFICACIÓN

La indización ha sido tradicionalmente uno de los temas más importantes de investigación en Documentación, ya que los índices han facilitado la recuperación de información tanto en los sistemas manuales tradicionales como en los nuevos sistemas informatizados. La indización *per se,* está abocada a la recuperación de información.
Con las oportunas salvedades históricas, podríamos decir que, el concepto de recuperación de información es tan antiguo como el mundo escrito, y se magnifica su importancia cuando hablamos de un *mundo informativo digital*, en el que numerosas representaciones del conocimiento humano se hacen en formato electrónico.
La indización es uno de los procesos fundamentales del análisis de contenido, y son muchas las definiciones que se han dado pero todas ellas la definen como una técnica, la de caracterizar el contenido tanto del documento como de las consultas de los usuarios, reteniendo las ideas más representativas para vincularlas a unos términos de indización, bien extraídos del lenguaje natural empleado por los autores, o de un vocabulario controlado o lenguaje documental

seleccionado *a priori*. Hoy en día es posible vincular el proceso de indización al lenguaje natural del documento gracias a los computadores; para hacerlo debemos discriminar la información aprovechando las estrategias utilizadas por los propios autores para presentar sus publicaciones, pues destacan la información esencial en títulos, resúmenes, y en los párrafos iniciales de las diferentes partes de los textos. También nos valemos de otras estrategias sintácticas y semánticas, como las que se derivan de la función que cumplen las palabras en las oraciones y del peso semántico que tienen las palabras en los textos. Si optamos por manejar el texto completo, sólo será posible una recuperación eficaz en aquellos lenguajes cuyos términos gocen de gran estabilidad. Tal sucede en los propios de las ciencias aplicadas y de la tecnología, donde la búsqueda se podría hacer en las mismas expresiones usadas por el autor.

Lo más frecuente es que el texto original y su *traducción* documental se den dentro de los dominios propios de las distintas áreas del saber. En este caso la amplitud de uso de los términos, de la expresión y del estilo que es propia del lenguaje natural, se ve limitada por las características fundamentales del discurso científico, lo que favorece la pertinencia de uso del lenguaje natural con fines documentales (2):

- Recepción y emisión cualificada (competencia),
- Vocabulario especializado,
- Organización estructural útil a la ciencia,
- Modelado lógico-formal,
- Determinación más sistemática que el lenguaje común.

De igual forma que en la indización manual, el principio de indización automatizada es identificar un documento por un conjunto de palabras claves representativas de su contenido, que pertenezcan a un conjunto abierto de términos, — indización libre—, o que pertenezcan a un conjunto cerrado y referenciado en una lista de autoridad o en un tesauro —indización controlada—. Así pues, podemos definir la indización automatizada como el uso de máquinas para extraer o asignar términos de indización sin intervención humana, una vez se han establecido programas o normas relativas al procedimiento.

Los factores que hacen posible pensar en el paso de una indización manual a una indización automatizada son, los siguientes:

a) Alto coste de la indización humana (tiempo)

b) Aumento exponencial de la información electrónica y la proliferación del *full-text*

c) La Gestión Electrónica de Documentos (GED) y a la informatización de los procesos documentales

d) Automatización de los procesos cognitivos y la investigación creciente y los avances en el Procesamiento del Lenguaje Natural (PLN)

a) Alto coste de la indización humana en términos de tiempo es uno de los argumentos más sólidos que se ostentan para justificar el desarrollo de Sistemas de Indización Automatizada (3). Cómo explotar de manera pertinente con un coste y tiempo reducidos, el volumen siempre creciente de información textual, se ha convertido en un tema recurrente y obsesivo en todos los estudios de análisis documental de contenido, dando lugar a múltiples trabajos destinados a evaluar la coherencia y la pertinencia de indización automática frente a la humana (4).

Otros autores (5) encuentran la justificación de las investigaciones en indización automatizada, partiendo de la base que la indización humana es inadecuada para minimizar la subjetividad inherente a la indización, ya que el grado de consistencia

alcanzado, depende no sólo del conocimiento de técnicas de abstracción conceptual, ni del conocimiento y manejo de lenguajes documentales, depende también del grado de conocimiento que el analista tenga sobre el tema que se trata, exigiéndole que esté siempre actualizado en esa materia. Es importante señalar también la inconsistencia entre los indizadores e incluso de un mismo indizador en distintos momentos anímicos, ya que la indización es algo subjetivo; el ser humano utiliza el lenguaje en función de múltiples condicionamientos, parcialidades y sesgos personales y culturales involuntarios.

La exacerbación de *lo humano* como sinónimo de *lo racional* y lo perfecto es fruto del conservadurismo y de la fidelidad a la idea de ser humano, pero objetivamente desde el punto de vista de la indización o descripción característica del contenido de un documento, hay muchos casos de malos ejemplos en que la indización manual, es a todas luces, deficiente. Por tanto, todas estas argumentaciones nos han llevado a pensar que la indización automática es la formalización y/o automatización de la indización, con el objetivo de reducir la subjetividad del proceso, y el alto coste en tiempo de la indización manual.

b) El aumento exponencial de la información electrónica y la proliferación del *full-text*. En este sentido es interesante evocar la afirmación que hacía Jones6 en los años 80:

El valor de la indización automática se incrementará cuando la literatura de forma legible a máquina sea más importante que la producida por medios tradicionales. Entre tanto, el ordenador será de importante ayuda para el indizador en la elaboración de los índices, aliviándole de tareas rutinarias como la ordenación, clasificación e impresión. No obstante, por el momento, las acciones específicas de determinar lo que constituye la materia indizable del texto, y cómo se debe expresar, son funciones todavía de la inteligencia y creatividad humanas.

Esta afirmación que Jones hacía en 1986 como futurible, parece que es una situación del presente, no porque la literatura producida en forma legible por máquina sea más importante que la producción impresa, pero sí hay que tener en cuenta que la propia naturaleza de la información ha cambiado y cada vez más se presenta en formato electrónico. El crecimiento exponencial de cantidades de información producidas y/o reproducidas en redes Internet e Intranet es hoy ya una realidad; por ello parece inevitable que el valor de la indización automatizada se incremente y tienda a dominar con respecto a la indización tradicional humana. El incremento de la ciencia y de la comunicación electrónica, crece de manera imparable; cada vez son más las bases de datos que se pueden consultar a texto completo, al mismo tiempo que la vida media de la información tiende a disminuir, todo ello contribuye a que no exista un paradigma unificado para la recuperación de información. La tarea de convertir en accesibles todas estas informaciones relevantes requiere una serie de actividades que componen el ciclo documental, entre las cuales, el análisis de contenido tiene

un papel fundamental, con lo cual es lógico que las investigaciones en documentación busquen nuevas alternativas para optimizar la recuperación de información. Una de estas alternativas es la indización automatizada donde, acudiendo a otras disciplinas como la lingüística o la estadística, se pretende dar solución al problema de la caracterización del contenido documental, y con ello, de la recuperación de información.

c) La Gestión Electrónica de Documentos (GED) y a la informatización de los procesos documentales. Las organizaciones están asumiendo en la actualidad una tendencia incipiente de conversión de los archivos basados en papel a los Sistemas de Gestión Electrónica de Documentos (EDMS: *Electronic Data Management Systems)*. Esta tendencia supone una nueva filosofía en el tratamiento de la documentación, combinando la imagen con la información textual asociada a ella, que requiere una planificación exhaustiva, donde la indización de documentos digitales insta un proceso informatizado de comprensión e inferencia del contenido para su posterior integración y recuperación en los procesos.

La automatización de los procesos documentales — almacenamiento, recuperación y reproducción de los documentos— mediante herramientas y aplicaciones informáticas, está estrechamente ligado a la indización automatizada, ya que la mayoría de los sistemas GED incluyen un motor de indización y búsqueda para procesar el lenguaje natural y efectuar la recuperación por contenido.

d) La automatización de los procesos cognitivos y la investigación creciente y los avances en el Procesamiento del Lenguaje Natural (PLN). Existen numerosas metáforas antropomórficas aplicadas a las máquinas en el sentido de que la eficacia en el procesamiento de la información es la característica esencial que comparten el ordenador y la mente humana.

La mente humana posee una eficacia cualitativa en sus procesos cognitivos (percepción, decisión, planificación y lenguaje). Existen distintas teorías que avalan que el Lenguaje Natural, lenguaje de comunicación humana, no es un lenguaje interno de pensamiento sino que es un lenguaje fruto del aprendizaje. De esta afirmación, podemos deducir que las máquinas también pueden aprender el procesamiento del lenguaje natural, máxime si tenemos en cuenta que se pueden automatizar, con un relativo margen de adecuación o calidad, aquellos procesos o tareas en que se den dos condiciones: 1) que las tareas se puedan describir por una secuencia perfectamente definida de acciones elementales y 2) cuando esas tareas se deban repetir muchas veces; ambas condiciones se dan en los procesos de indización, por ello, son perfectamente automatizables. El lenguaje refleja y contiene infinitas posibilidades del pensamiento humano, mientras que las estructuras formales que son los modelos con los que puede operar el ordenador son de naturaleza finita. Una palabra es más que la secuencia de las letras de su significante, a causa del significado que se asocia a éstas y de su relación con otras palabras y con el contexto que las rodea. Podríamos explicarlo de una manera un tanto metafórica, que las relaciones que contiene un significante con su significado denotativo y connotativo en cada hablante, son como una *nube* que

cuelga de cada elemento del texto y que le parece distinta a cada persona, y el ordenador, no procesa esa *nube*, lo que hace es transformar las cadenas de caracteres.

Con todo lo indicado hasta ahora, podemos decir que nos encontramos en un momento de transición, donde la indización tradicional realizada manualmente para el análisis de contenido de documentos en formato impreso, convive con la indización automatizada destinada al análisis masivo de información textual en formato electrónico.

La indización consiste pues, en recorrer el documento para comprender y abstraer su magnitud significativa, de tal forma que dé como resultado una representación sintética de su contenido. Esta tarea compleja, exige conocimientos científicos, la comprensión del lenguaje natural y de la lengua del texto y un dominio práctico de un lenguaje documental (sea tesauro, sea lista de encabezamientos o lista de descriptores), además de una capacidad de análisis y síntesis. Todas estas exigencias que podemos estimar para una buena indización pueden concurrir o no en un indizador humano, pero son las que debemos exigirle a un sistema de indización automatizada.

Todo análisis semántico de un texto científico es una operación eminentemente intelectual que exige una doble competencia, primero en el plano de la lengua y también en el plano del pensamiento científico, y la máquina debe ser instruida de la misma manera en ambos órdenes de competencia.

Los distintos modelos de indización automatizada irán, como veremos a continuación, de una mera extracción en lenguaje natural, donde la palabra se entiende como objeto, pasando por una indización por tratamiento lingüístico sobre un vocabulario abierto, a una indización "inteligente" por conceptos, donde los sistemas de indización y búsqueda se erigen como una verdadera herramienta de búsqueda y recuperación documental.

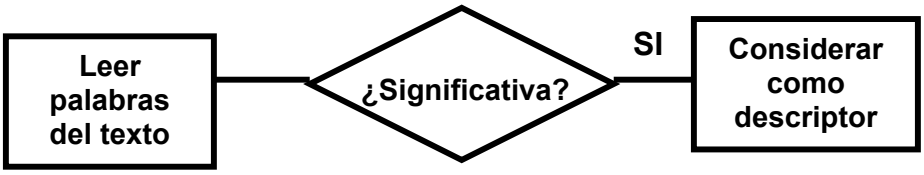
MODELOS DE INDIZACIÓN AUTOMATIZADA Y LENGUAJE NATURAL

En todos los estudios genéricos —como éste— sobre indización automatizada se realizan distintas aproximaciones para caracterizar o tipificar los modelos de indización automatizada, atendiendo a diversos criterios: uno de los más habituales es el criterio evolutivo (7), en tanto que al ser la indización automatizada un campo de investigación creciente se trata de primar más los avances de esta técnica informatizada de análisis de contenido que la tendencia profética que trate de discernir el futuro de estos sistemas; otro de los criterios más seguidos es el que se fundamenta en método de extracción terminológica, que distingue fundamentalmente los métodos de extracción lingüísticos de los no lingüísticos, donde los métodos lingüísticos abarcan todas las técnicas derivadas del PLN y los no lingüísticos el resto de las formas de extracción del vocabulario de corte estadístico, probabilístico e incluso, bibliométrico (8) o informétrico; otro de los parámetros que se tienen en cuenta para estudiar los sistemas de indización automatizada es la parte del documento que indizan, distinguiendo esencialmente, los sistemas que indizan las partes principales del documento (título, resumen) (9) de los que se destinan a indizar el texto completo; finalmente,

señalamos un criterio fundamental que aparece en múltiples trabajos: el control del vocabulario, que trata de hacer hincapié en la presencia de lenguajes controlados (tesauros o listas de materias) como elemento de control semántico del sistema de indización automatizada frente a una indización exclusivamente *full-text* (10).
Todos estos criterios utilizados para establecer una clasificación de los Sistemas de Indización Automatizada no son excluyentes, más bien responden a un c*ontinuum* de evolución. Lo más habitual es que a tenor de los cambios y de los avances, los modelos no se suplanten, sino que convivan (11) y se aúnen en un fin común, en este caso, conseguir una indización totalmente automatizada. Por ello, trataremos de incluir todos ellos en lo que hemos decidido llamar *generaciones de indización automatizada*, donde parece primar un criterio evolutivo, por razones de claridad expositiva, pero en realidad no queremos revelar sólo la evolución de los sistemas, sino el papel que ha desarrollado en Lenguaje Natural en cada uno de ellos. Así distinguiremos:
- Una primera generación de la indización automatizada, donde las palabras se entendían como objetos;
- Una segunda generación donde lo que prima es el análisis lingüístico para la desambiguación de conceptos;
- Y finalmente, distinguimos una tercera generación a la que hemos denominado indización "inteligente" en tanto que trata de abstraer no sólo conceptos sino modelos conceptuales fundamentados en bases de conocimiento.

IDENTIFICACIÓN AUTOMÁTICA DE LAS ENTRADAS: LA PALABRA COMO OBJETO

Los primeros índices automáticos, construidos por permutación de los elementos que componen las unidades susceptibles de indización (hasta entonces, sólo palabras) fueron los de tipo KWIC-KWOC. En los años 60, Luhn (12), conseguía aplicar la capacidad electrónica de los ordenadores a un campo ajeno al de las matemáticas. Pasó así el ordenador a ser considerado capaz de hacer análisis del contenido de los textos. Pero en realidad comenzaba una larga evolución que se desarrollaría entre la capacidad contable inicial y la reflexión cognitiva a la que aspiran las aplicaciones actuales. Desde el comienzo, los ordenadores se utilizaron para procesar textos, en especial para realizar traducciones automáticas (13), lo que está muy cerca de los usos documentales.
Estos primeros intentos se basaron en la identificación de las palabras que aparecían en títulos (14) de artículos científicos. Para hacerlo se utilizaba una base técnica muy sencilla: las palabras se consideraban como objetos exclusivamente y por tanto, desde su significante. Para llegar a ser una entrada del índice las palabras pasaban primero por el filtro de un antidiccionario, cualquier palabra que constase en éste (palabra vacía) y en la unidad que se debía indizar, se eliminaba, y así, las que permanecían se consideraban significativas y pasaban a ser elementos de indización. En la base de cualquier proceso de indización automática se iba a situar desde entonces un algoritmo, cuyo funcionamiento se puede explicar en tres pasos, según muestra en la figura:



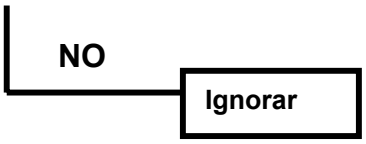


Fig. 1. Esquema del funcionamiento del algoritmo. Robredo (15)

La obtención por este medio de palabras claves daba como resultado innumerables referencias cuando se manipulaba el texto completo, ya que se alcanzaba una indización no selectiva e indiscriminada, incapaz de diferenciar, para el resultado final, las formas flexionadas de una misma palabra por género y número. Y mucho menos aún de reconocer los sinónimos (de tal forma que se podían dar varias entradas para un mismo significado) ni los homónimos (sumando significados distintos al mismo significante). La única posibilidad de orientación hacia el contenido que cada palabra quería representar venía a través de su presentación en contexto. Determinación ésta utilizada desde antiguo en la confección de los denominados índices de concordancias (16). Cuyo establecimiento se hacía sabiendo que la posible ambigüedad producida cuando las palabras se presentan aisladas quedaba limitada por un contexto que las definía y explicaba.

Los índices permutados tienen una entrada por cada palabra no vacía del documento o fragmento a indizar. Descomponen, por tanto, en elementos simples las expresiones sintagmáticas. La candidatura a ser palabra de indización se originaba exclusivamente en no haber sido eliminada por la lista negativa y en aparecer como caracteres de estructura independiente entre dos espacios del texto en blanco. El texto en ningún caso es tomado como una composición macroestructural, si no como una sucesión de símbolos.

Una consideración que aminora la diferencia entre la utilización del lenguaje natural sin limitaciones y la deseable regulación se establece al observar que muchos de los intentos hechos para indizar mediante ordenadores se han valido de la información presentada en los registros bibliográficos para facilitar su tratamiento. Partir de títulos y resúmenes ofrece como ventajas tener que procesar un menor volumen, hacerlo sobre la expresión de las ideas sustanciales y encontrar un vocabulario más representativo y, por tanto, más idóneo. Se utiliza así un recurso heurístico de interpretación sumaria del texto completo, aprovechando estrategias que ofrece el propio texto.

Un paso más en la representación automatizada consistió en hacer cálculo de la frecuencia estadística con que aparecían las palabras. Ya no bastaba simplemente con aparecer en la unidad documental que se indizaría para ser considerado candidato, ahora los términos se seleccionaban si su tasa se situaba próxima a una frecuencia de aparición media, quedando fuera las palabras cuyo umbral era muy alto y también aquellas que lo era muy escaso (17). La utilización del método cuantitativo es la única manera que permite generar algoritmos que haga a las máquinas entender la lengua (18). Aún así continuaba siendo una indización morfológica, aunque corregida hacia la pertinencia mediante la limitación de aquellas palabras cuya aparición fuera excesivamente abundante o rara dentro de un texto (19). Sin embargo, el texto seguía siendo considerado una sucesión de símbolos o caracteres, sin prestar atención a la composición

macroestructural. Y por ello, al situarnos aún dentro de una indización por palabras, lo implícito, las materias no nombradas, quedaban sin poderse recoger en los índices.
Podemos decir, no obstante, que esta primera generación de modelos para la indización automatizada, basada en criterios meramente estadísticos o probabilísticos, tiene una importancia significativa: por un lado desde el punto de vista de que son los primeros modelos que surgen como alternativa a la tediosa operación documental de la indización aprovechando el desarrollo de la informática, y por otro, porque son métodos que siguen usándose (bien combinados con otros modelos de base más lingüística para la indización o bien, como herramienta para la extracción de palabras en los procesos de elaboración de lenguajes controlados –tesauros-) en áreas específicas del conocimiento.

PROGRESOS HACIA LA DESAMBIGUACIÓN: LA FUNCIÓN DE LAS PALABRAS

Ya en los primeros intentos de los años 50 estaba latente un largo proceso para conocer la estructura sintáctica de las oraciones textuales. A principios de los 70 se iniciaban los modelos de análisis lingüístico que se han perpetuado en la mayoría de los sistemas actuales. Esta nueva generación de sistemas de indización automática, deberían de valerse del Procesamiento del Lenguaje Natural (PLN) cuyos primeros conatos surgían en aquella época, y que en la actualidad ha conseguido unos resultados *que sitúan al PLN en posición para liderar una nueva dimensión en las aplicaciones informáticas del futuro: los medios de comunicación del usuario con el ordenador pueden ser más flexibles y el acceso a la información almacenada más eficiente* (20).
El objetivo era eliminar la ambigüedad de las palabras filtrándolas a través de cuatro procesamientos, análisis o etapas sucesivas —*parsers* lingüísticos— (fig.2) de menor a mayor complejidad. Con ellas se busca comprender realmente el significado de los documentos:

a) morfológico-léxico;
b) sintáctico;
c) semántico y
d) pragmático.

a) *Procesamiento morfológico-léxico*: En primer lugar, se realiza una segmentación del *corpus* de textos en unidades menores, procediendo a una verticalización de las oraciones y asignándoles una serie de identificadores que serán utilizados como puntos de referencia en los diferentes análisis posteriores. Se trata no sólo de identificar las palabras, si no también las formas sintagmáticas, las siglas y las locuciones. Los elementos delimitados se contrastan con los dos diccionarios con los que el sistema trabaja (un diccionario que contiene todas las entradas de una lengua; otro con las locuciones e idiotismos), incluso en los sistemas más actuales, las palabras identificadas son sometidas a un proceso de lematización para alcanzar su forma canónica (21). Debe advertirse que presenta gran dificultad la captación de los conceptos del texto desde el léxico: en primer lugar, porque las asociaciones de palabras se alejan a veces mucho del sentido que tenían sus componentes originales, lo mismo que sucede con los términos polisémicos donde sólo el contexto determina el significado concreto.
Esta etapa tiene como función principal la de obtener el léxico, componente básico

de los posteriores análisis sintáctico y semántico; gracias al analizador morfológico, el análisis estadístico de frecuencias se realizará sobre datos formalizados y unívocos semánticamente.

b) *Procesamiento sintáctico*: utilizando una gramática y/o diccionarios, se analizan las palabras sintácticamente y se describe la estructura de las oraciones. El análisis sintáctico tiene un doble objetivo: por un lado, permite separar las unidades lingüísticas con sentido simples o compuestas, y por otro, permite desambiguar las categorías gramaticales asignadas por el analizador morfológico (22) y al mismo tiempo enriquecer y autogenerar los diccionarios de aplicación.
Los analizadores sintácticos determinan la construcción de las oraciones localizando la función que cumplen las palabras como sujeto, verbo, complemento (y tipos de complementos) (23).

c) *Procesamiento semántico*: Su objetivo es alcanzar el conocimiento temático de los textos, el significado, por tanto, de sus oraciones. Esta es la etapa se fundamentará, normalmente, bien en un análisis semántico-léxico (estudio de las relaciones paradigmáticas de significado: este análisis permite agrupar y jerarquizar el contenido del texto a través del reconocimiento nuevamente morfológico y del reconocimiento de sinónimos e hiperónimos), o/y en un análisis semántico-gramatical —estudio de las relaciones sintagmáticas, en el plano de la frase o, y su significado concreto en el contexto del documento— todo ello con la finalidad de Fig. 2 (24) reducir y homogeneizar la información léxica del texto que se pretende indizar.

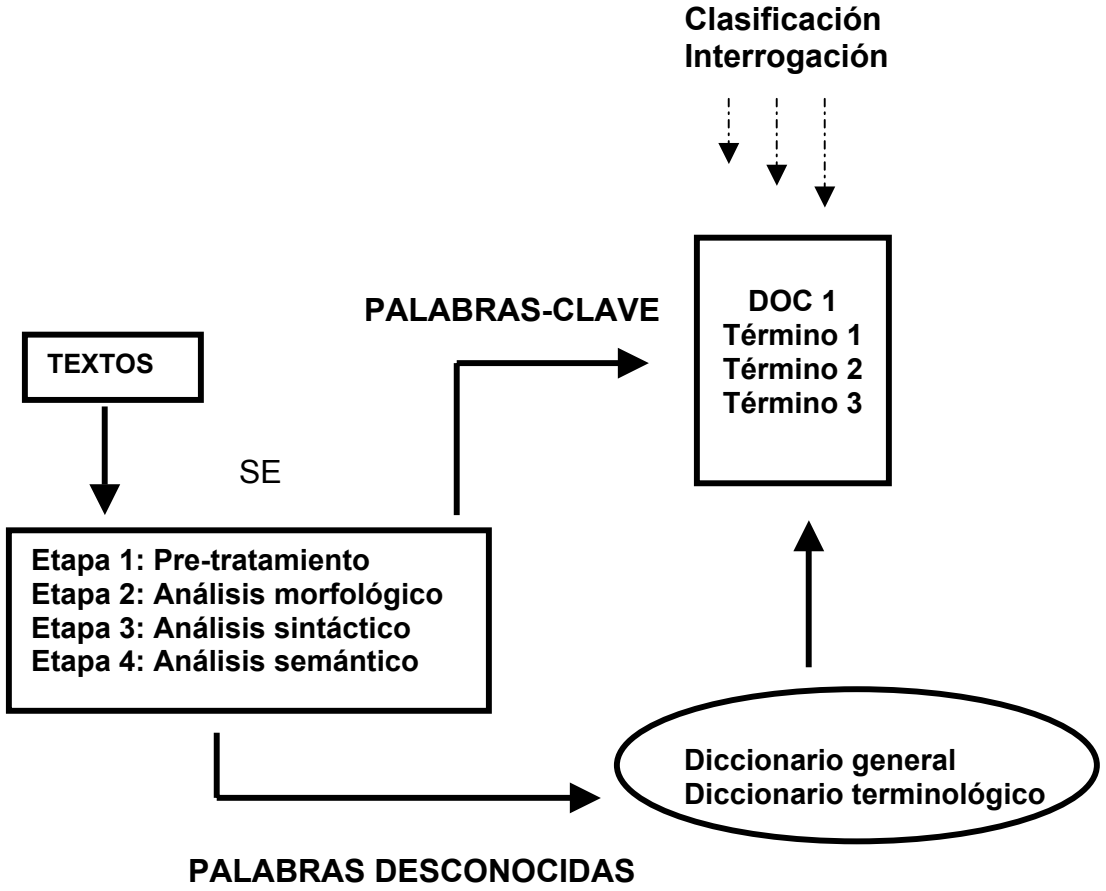


Fig. 2 Fases de la indización automatizada en un modelo de base lingüística de segunda generación, basado en Isabel Gachot.

Los enlaces dentro de esos esquemas pueden representarse gráficamente mediante estructuras arborescentes que permiten refinar las búsquedas ascendiendo hacia los genéricos descendiendo por los específicos. La base de este análisis se encuentra en los procesos deductivos por los que se establecen inclusiones conjuntivas, llegándose a representar los diferentes dominios conceptuales de un texto.

Para efectuar este nivel del análisis se emplean auténticos tesauros de términos. Los enlaces que éstos establecen, ya sea por jerarquías o por asociaciones, permiten precisar o ampliar cada búsqueda dentro de los textos de un campo especializado. No olvidemos que un tesauro contiene los conceptos (y las relaciones que existen entre ellos) mediante los que se representa el conocimiento de un campo científico-técnico.

Precisamente la utilización de los mismos tesauros supuso un avance que consistió en que, una vez procesado el texto y extraídos los términos preferentes, pasaron éstos a asociarse con dos descriptores de un tesauro. Fue éste el inicio de los mapas léxicos donde se representaban los términos del texto y una o varias parejas de términos del tesauro. El ejemplo clásico ha sido el definido por el programa PASSAT *(Programm zur automatischen Selektion vo Stichwörtern aus Texten)* que es el módulo de análisis de textos del software de recuperación de información GOLEM de la empresa informática Siemens.

d) *Procesamiento pragmático*: El análisis pragmático del texto es el más difícil de automatizar ya que implica un conocimiento del mundo real o *semántica de mundo*. Se trata de analizar las relaciones contextuales haciendo uso de algoritmos que permiten comprender el contexto del discurso (25).

Grishman (26), por ejemplo, advierte en su *Introducción a la lingüística computacional,* que una de las mayores dificultades para analizar el contenido del los textos en lenguaje natural es que gran parte de lo significativo está implícito en el discurso. Por eso, algunos de los estudios más avanzados en el desarrollo de *software* para el análisis de contenido, que por ello podríamos incluir en la generación siguiente abocada a una indización *inteligente* se basan, además de en un análisis puramente semántico, en un Análisis Cognitivo Discursivo (27) (ACD) y extraen, lo que se denomina *Estructura Fundamental del Significado* (SFS), además de otras técnicas como la constitución de Redes Semánticas, que veremos en el apartado siguiente.

HACIA UNA INDIZACIÓN *INTELIGENTE*

Las últimas tendencias, que nos permiten hablar de una nueva generación de sistemas de indización automatizada, giran en torno al acceso directo a los documentos a través del procesamiento lingüístico automático y la utilización del lenguaje natural, combinando otras técnicas como el análisis estadístico o la ponderación terminológica.

Se busca asegurar la coherencia a la vez que, al utilizar el lenguaje natural, permitir el acceso a los documentos sin formación previa en lenguajes documentales y sin conocer el vocabulario terminológico específico del campo interrogado, esto es, sistemas funcionales que permitan incluir interfaces

inteligentes que posibiliten la utilización del lenguaje natural como lenguaje de intercambio de conocimiento entre el documentalista o el usuario final y el sistema. Se trata de integrar todos los modelos y de aprovechar la modularidad en los sistemas para imprimir al ordenador una especie de competencia lingüística y/o cognitiva, teniendo como soporte no sólo bases lingüísticas, sino *bases de conocimiento.*

Podemos decir que en la evolución del procesamiento lingüístico de los documentos ha habido tres momentos marcados por la utilización de otros tantos instrumentos de análisis.

1. *Diccionarios:* que guiaron el análisis morfológico y el sintáctico utilizando reglas lingüísticas (gramática).
2. *Tesauros,* que permitieron explicitar las unidades semánticas mediante los enlaces de equivalencia, jerarquía y asociación que existían entre ellos, al aplicar reglas documentales.
3. *Bases de conocimiento,* que incluso indican los tipos de relaciones que se dan entre los conceptos y desambiguan el contenido del documento.

La gestión del conocimiento, que es la tendencia de todos los sistemas de información actuales, no tratan de crear un simple almacenamiento y acceso a la información, sino todo un proceso de manipulación, selección, mejora y preparación de la información, para dotarla de un valor añadido.

En este sentido la indización automatizada (genéricamente motor de indexación y búsqueda) serán un elemento fundamental para la recuperación de información en los nuevos sistemas de gestión del conocimiento, y por ello se conciben como sistemas de extracción de conceptos, construyendo Redes semánticas *input-output* (fig.3) basadas en bases de conocimiento. Podemos definir un concepto como una representación general y abstracta de un objeto, que permite la recuperación de información por ideas, definidas éstas como representaciones distintivas y detalladas de los objetos contenidos en los textos. En estos nuevos motores de indización y búsqueda -*v.gr.* Spirit, y su módulo de análisis semántico Spirit Sense (28) o Tropes (29)- incluidos dentro de software documentales destinados a la GED o a la Gestión del Conocimiento, podemos atisbar un influjo de las teorías lingüísticas de Saussure y una utilización de la lógica universal aristotélica para construir la semántica del texto y asociar las relaciones del contexto.

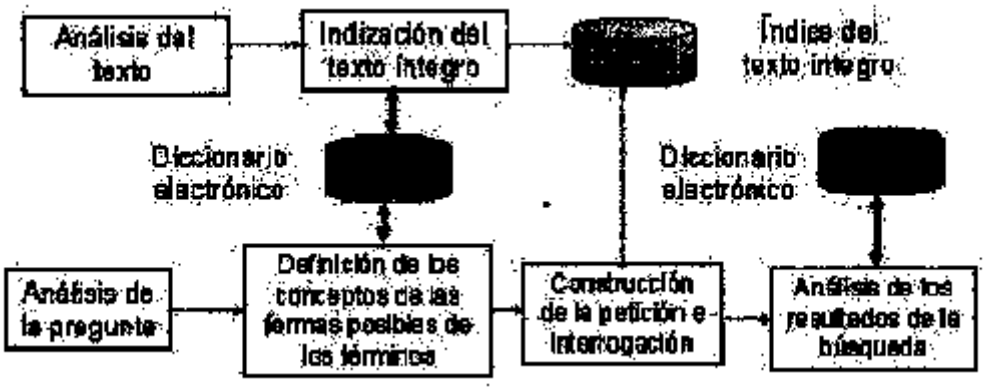


Fig. 3 Utilización de una red semántica.

Las *bases de conocimiento,* traducción forzada del término inglés *knowledge bases* según Leloup (30), aparecen pues, en estos sistemas, como un tesauro enriquecido con información morfológica, sintáctica y semántica, cuyo vocabulario se obtiene del *corpus* de documento de un área del saber. Los textos especializados presentan términos enlazados. Se trata de identificarlos tal como están en los textos, incluso nominalizando los verbos. Como los autores de un campo científico-técnico están al frente de la investigación, su lenguaje está por encima de los controlados y, por tanto, de los que poseen los analistas (31). Este análisis se fundamenta en el conocimiento que los expertos han depositado en los documentos, es decir, un conocimiento pragmático a través de la aprehensión de su realidad (*semántica de mundo*). Su aplicación precisa la intervención de la estadística, la informática, la lingüística y la Inteligencia Artificial.

En estos sistemas de indización de última generación, se trata pues, además de asimilar el PLN, de establecer relaciones semánticas desde un hecho con sus causas y consecuencias. Los tesauros ya tenían relaciones de asociación, pero las bases de conocimientos especifican cómo es esa asociación, la representan mediante estructuras arborescentes (generalmente *B-tree*) o en planos. Los términos existen en el texto igual que en los bancos de datos terminológicos, lo que ofrece más posibilidades que el uso de los tesauros que funcionan realmente como diccionarios. El tratamiento lingüístico permite recuperar palabras tanto en su forma canónica como flexionada. Precisamente, al tratar las palabras desde el nivel léxico, su procesamiento se complica, ya que las variaciones terminológicas son innumerables en los textos científicos debido a la inserción de unos términos en otros, a las coordinaciones entre términos, a las variaciones coordinadas y a la morfología derivacional.

La última generación de sistemas de indización, busca la representación del contenido utilizando conceptos y algoritmos que dan lugar a nuevas herramientas de software más complejas y dirigidas a la gestión del conocimiento. Están dirigidas a la indización de textos electrónicos digitalizados; responden a una arquitectura cliente-servidor y a entornos Internet/Intranet; permiten la indización e interrogación en lenguaje natural; combinan tanto el modelo estadístico (ponderación) con el lingüístico y suelen estar formados por 4 módulos: un módulo de construcción de reglas (canonización), un motor de indización; módulo de cálculo estadístico y un diccionario electrónico o base de conocimiento. Podemos decir con todo, que estos sistemas suponen la asunción del contexto informacional y la solución integrada para indizar el conocimiento electrónico.

CONCLUSIONES

A pesar de que a lo largo de toda la exposición venimos introduciendo algunos puntos de vista sobre el tema, de forma recopilatoria, podemos concluir lo siguiente:

- Las investigaciones en torno a la Indización Automatizada se deben al alto coste de la indización humana (tiempo), al aumento exponencial de la información electrónica, a la proliferación del *full-text*, a la GED, a la informatización de los procesos documentales, a la posibilidad de automatizar los procesos cognitivos y, sobre todo, a la investigación creciente y a los avances PLN. Fruto de estas

investigaciones podemos hablar de distintas generaciones de indización automatizada, según el modelo seguido.

- La tendencia que siguen las investigaciones en indización automatizada es a integrar todos los modelos y a la modularidad en procesos más simples —análisis estadístico + análisis lingüístico (análisis sintáctico, morfológico y semántico)— de un proceso complejo como es la indización. Aunque son muchos los autores, a los cuales nos adscribimos, que anuncian que el éxito de la indización automatizada vendrá de la mano del desarrollo de las técnicas de Procesamiento del Lenguaje Natural y en el desarrollo de sistemas híbridos y de la Inteligencia Artificial, esta modularidad en la que creemos para el desarrollo de la indización automatizada, puede reflejarse también en la necesidad de crear sistemas mixtos que conjuguen el *software* para el tratamiento del texto completo y la GED, con el *software* para el PLN.
- Las últimas tendencias en indización automatizada han dado lugar a programas específicos para la indización automatizada, pero dentro de *software* que se destinan a la gestión, almacenamiento y recuperación de información —verdaderos Sistemas de Gestión Electrónica de Documentos o Sistemas de Gestión del conocimiento— donde el módulo de procesamiento/indización (motor de indización) constituye una parte fundamental del sistema (tales programas son por ejemplo: Search'97, ZyIndex, Excalibur, entre otros). Se tiende pues a indizar los documentos en formato digital, por medios electrónicos y al acceso directo a los documentos a por su contenido a través del procesamiento lingüístico automático a fin de alcanzar una indización coherente. Al utilizar lenguaje natural, se accedería a los documentos sin formación previa en lenguajes documentales, donde —creemos— el papel del tesauro, como herramienta fundamental para la recuperación de información, no desaparecerá con el desarrollo de las bases de conocimiento, sino que reconvertirá su utilidad más, transparente para el usuario, en los momentos *input-output* del sistema.

El campo de investigación de la indización automatizada y de la recuperación de información es inagotable y se ve magnificado al introducir en él el fenómeno de la gestión de la información en Red (Internet/Intranet). Se trata pues, de ser receptivos y coherentes con el desarrollo tecnológico de nuestro tiempo, ya que en todo lo que implica extracción de datos *(data mining),* la gestión y la búsqueda del contenido son la próxima etapa, por ello los sistemas de indización "inteligentes" serán el futuro para una verdadera gestión del conocimiento (estructurado o no).

REFERENCIAS

Material bajado de Internet. Publicado en: Ciencias de la Información, vol. 30, No. 3, septiembre 1999, p. 11-24.

(1) Antonio GARCÍA GUTIÉRREZ. Estructura lingüística de la documentación, teoría y método. Murcia: Universidad, Secretariado de Publicaciones, 1990. p. 18.

(2) L. BLOOMFIELD. Aspectos lingüísticos de la ciencia. Madrid: Taller de ediciones, 1973. p. 105.

(3) Este argumento aparece desde las primeras investigaciones sobre indización automatizada llevadas a cabo en los años 50 (Cfr. E. GARFIELD. The relationship between mechanical indexing, structural linguistics and information retrieval. Journal of Information Science, (18):343-354. 1992) hasta las más recientes investigaciones de la década de los 90 llevadas a cabo en el INIST (Cfr. J.

CHAUMIER et M. DEJEAN. L'indexation documentaire: de l'analyse conceptuelle humaine à l'analyse automatique morphosyntaxique. Documentaliste-Sciences de l'Information, 27(6): 275-279. 1990)

(4) Tal es el caso del trabajo de Plaunt y Norgard, que describen la evaluación de dos algoritmos basados en la técnica de disposición léxica aplicados a 4626 documentos de la base de datos INSPEC, para crear un diccionario de asociaciones entre los ítems léxicos que contienen los títulos, autores y resúmenes y los términos controlados asignados a esos documentos por indizadores humanos, que servirá, en un primer estadio de aplicación del algoritmo, para comparar los encabezamientos de materia asignados de forma automática con los asignados por un catalogador. Christian PLAUNT and Barbara A. NORGARD. An Association-Based Method for Automatic Indexing with a Controlled Vocabulary. Journal of the American Society for Information Science, 49(10): 888-902. 1998.

(5) V. gr. Ghislaine CHARTON. Indexation manuelle et indexation automatique: dépasser les oppositions. Documentaliste-Sciences de l'information, 26(4-5): 181-187. Juillet-octobre 1989; Isidoro GIL LEIVA, José Vicente Rodríguez Muñoz. De la indización humana a la indización automática. En: Organización del conocimiento en Sistemas de Información y Documentación. Zaragoza: Fco. Javier García Marco, ed., 1997, p. 201-215.

(6) Kevin P. JONES. Getting Started in Computerized Indexing. The Indexer, 15(1): 12. 1986.

(7) Este es el enfoque del estudio, por ejemplo, de Isidoro GIL LEIVA y José Vicente RODRÍGUEZ MUÑOZ. Tendencias en los sistemas de indización automática. Estudio evolutivo. Revista Española de Documentación Científica, 19(3): 273-291. 1996.

(8) Vânia Lisbôa DA SILVEIRA GUEDES, es una representante de la corriente brasileña (Río de Janeiro) de aplicación de criterios estadísticos y de leyes bibliométricas —concretamente las leyes de Zipf y la Ley del punto T de Goffman— a la indización automatizada. Vid. Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos. Ciencias da Informação, 23(3): 318-326. 1994. Concretamente, es este artículo, realiza una aplicación de la bibliometría para la indización de un conjunto de textos sobre la mecánica de suelos.

(9) Según Garfield, en facetas del conocimiento muy especializadas (como la Química), un 60% de los términos pertinentes para la indización, están de forma explícita en el título, un 30% está implicado en alguna palabra del título, y sólo el 10% restante se extraía propiamente del texto del artículo. Cfr. E. GARFIELD. Op. cit, p. 344.

(10) Michel REMIZE. Le thésaurus face au texte intégral: une évolution tournée vers l'utilisateur. Archimag, (112): 40-41. Mars 1998.

(11) Un indicio de esto, puede ser el título tan sugerente de un artículo de 1995 de Isabelle GACHOT. Linguistique+statistiques+informatique = indexation automatique. O por ejemplo el sistema SMART emprendido por Salton en 1961, que intentaba procesar documentos de forma automática fundamentado en principios estadísticos, pero tomando como base principios lingüísticos utilizando tanto la morfología de las palabras como la sintaxis de las frases. Vid. G. Salton. The SMART system 1961-1976. Experiments in dynamic document processing.

Encyclopedia of Library and Information Science, vol. 28, 1980, pp. 1-28 (citado por Gil Leiva y Rodríguez Muñoz. Tendencias. Op. cit., p. 290)

(12) Hans Peter Luhn, especialista de IBM, fue el pionero en aplicar el análisis estadístico del vocabulario para efectuar una indización automatizada, constituyó un gran paso en la automatización o, más bien entonces, mecanización del análisis de contenido gracias a la autocodificación de los textos y la constitución de índices KWIC (Key Word In Context) que aún hoy se siguen utilizando para la localización de términos en algunos vocabularios controlados (tesauros).

(13) William LOCKE y Donald BOOTH. Machine translation of languages. Cambridge: MIT Press, 1955.

(14) Recordemos lo que apuntaba Garfield al respecto de la relevancia de los términos del título en disciplinas muy especializadas. E. Garfield, Loc.cit. (nota 9).

(15) Jaime ROBREDO. Indexação automática de textos. Uma abordagem otimizada e simples. Ciência da Informação, 20(2):131. 1991. La figura muestra el algoritmo de trabajo de los sistemas que extraían el lenguaje natural fundamentándolo en un antidiccionario, según muestra fig. 1, el algoritmo se desarrolla en tres pasos: 1) Las palabras del texto son comparadas con las del antidiccionario; 2) se desprecian aquellas que aparezcan a la par en el texto y en la lista y 3) las que permanecen son consideradas palabras-clave.

(16) Jennifer E. ROWLEY. Abstracting and Indexing. 2nd ed. London: Clive Bingley, 1988, p. 46.

(17) J. CHAUMIER et M. DEJEAN. Loc. cit.

(18) G. SALTON, L. ALLAN, y C. BUCKLEY. Automatic Structuring and Retrieval of Large Text Files. Communications of the ACM, 37(2): 97-108. 1994.

(19) La utilización de la frecuencia estadística de aparición de las palabras en la representación automática fue ampliamente tratada por V. ROSENBERG. A study of statistical measures for predicting terms used to index documents. Journal of the American Society for Information Science, 22(1): 41-50. 1971.

(20) Eduardo SOSA. “Procesamiento del lenguaje natural: revisión del estado actual, bases teóricas y aplicaciones”. En : Information World en Español, vol. 6, nº 12, enero-febrero 1997, p. 26.

(21) Por forma canónica entendemos la transformación de las formas conjugadas y flexivas en entradas de un diccionario.

(22) Por esta proximidad en el análisis, algunos modelos de indización de segunda generación, prefieren hablar de analizadores morfosintácticos, tratando de realizar un analizador con una gramática particular gobernada por la naturaleza de los textos que se indizan, y cuyo cometido será constituir una serie de modelos que constituyan un repertorio con todas las formas posibles para, a través del análisis flexional y de la lematización, reducirlos a su forma canónica. Esto demuestra que la serie de principios lingüísticos que operan en este tipo de modelos, es constante, pero su orden o fundamentación teórica es aleatoria.

(23) William WOODS. Transition network grammars for natural language analysis. Communications of the AMC, 13(10): 591-606. 1970.

(24) Fases de la indización automatizada en un modelo de base lingüística de segunda generación, basado en Isabelle GACHOT. Linguistique +statistiques+informatique=indexation automatique. Archimag, 84: 34-37.

Mai 1995;. y Michelle LUBKOV. L’abc du langage naturel. Archimag, (103): 24-25, abril 1997.

(25) H. KAMP. Discourse representation theory: What It is and Where It Ought to go?. En: Natural Language at the Computer, 1988, p. 95.
(26) R. GRISHMAN. Introducción a la lingüística computacional. Madrid: Visor, 1991.
(27) Sobre este aspecto, Vid. Rodolphe GHIGLIONE, et al. L'analyse automatique des contenus. Paris: Dunod, 1998. Donde se describen las técnicas lingüísticas e informáticas del software francés para el procesamiento del contenido textual y la recuperación de información: Tropes de Acetic. Información relativa a este programa, se puede recabar también en la web en: http://www.acetic.fr/prsentat.htm
(28) Sobre este programa de la empresa T-Gid, vid. Catherine LELOUP. Motores de búsqueda e indexación: entornos cliente servidor, Internet e Intranet. Barcelona: Ediciones Gestión 2000, 1998, p. 251-257. O la homepage de la empresa en: http://www.technologies-gid.com
(29) Sobre el funcionamiento y arquitectura del software Tropes, resulta muy interesante el libro de: Rodolphe GHIGLIONE, et al. Op. cit.
(30) Catherine LELOUP. Op. cit., p. 146.
(31) Xavier POLANCO. Infométrie et ingénierie de la connaisance. Nancy: INIST-CNRS, 1995.

BIBLIOGRAFÍA

BLOOMFIELD, L. Aspectos lingüísticos de la ciencia. Madrid: Taller de ediciones, 1973.

CHARTON, Ghislaine. Indexation manuelle et indexation automatique: dépasser les oppositions. Documentaliste-Sciences de l'Information, 26(4-5): 181-187. Juillet-octobre1989.

CHAUMIER, Jaques et Martine DEJEAN. L'indexation documentaire: de l'analyse conceptuelle humaine à l'analyse automatique morphosyntaxique. Documentaliste Sciences de l'Information, 27(6): 275-279. 1990.

COULON, Daniel, Daniel KAYSER. Informatique et langage naturel: présentation générale des méthodes d'interpretation des textes écrits. Technique et science informatique, 5(2): 103-128. 1986.

GACHOT, Isabelle. Linguistique+statistiques+informatique=indexation automatique. Archimag, 84: 34-37. Mai 1995.

GARFIELD, E. The relationship between mechanical indexing, structural linguistics and information retrieval. *Journal of Information Science*, (18): 343-354. 1992.

GHIGLIONE, Rodolphe, et al. L'analyse automatique des contenus. Paris: Dunod, 1998.

GIL LEIVA, Isidoro. La automatización de la indización de documentos. Gijón: Trea, 1999.

GIL LEIVA, Isidoro, José Vicente RODRÍGUEZ MUÑOZ. De la indización humana a la indización automática. En: Organización del conocimiento en Sistemas de Información y Documentación. Zaragoza: Fco. Javier García Marco, ed., 1997, p. 201-215.

GIL LEIVA, Isidoro y José Vicente RODRÍGUEZ MUÑOZ. Tendencias en los sistemas de indización automática. Estudio evolutivo. Revista Española de Documentación Científica, 19(3): 273-291.1996.

IBEKWE, Fidelia. Traitement linguistique des données textueles pour la recherche des tendances thématiques [documento www]. Grenoble: Université Stendhal, 1995.
Disponible en: http://atlas.irit.fr/vsst95/vsst95p8M2.html (consultado el 11 de mayo de 1999)
INDEXING Digital Documents it's NOT an Option [documento www]. University of Texas, rev. 27 de julio de 1997. Disponible en:
http://fiat.gslis.utexas.edu/~scisco/inel.html (consultado el 11de mayo de 1999)
JONES, Kevin P. Getting Started in Computerized Indexing. *The Indexer,* 15 (1): 9-13. April 1986.
KAMP, H. Discourse representation theory: What It is and Where It Ought to go? En: Natural Language at the computer, 1988.
LELOUP, Catherine. Motores de búsqueda e indexación: entornos cliente servidor, Internet e Intranet. Barcelona: Ediciones Gestión 2000, 1998.
LISBÔA DA SILVEIRA GUEDES, Vânia. Estudo de um critério para indexãçao automática derivativa de textos científicos e tecnológicos. Ciencias da Informação, 23 (3): 318-326. set-dez 1994.
LUBKOV, Michel. L'abc du langage naturel. Archimag, (103): 24-25, abril 1997
MOREIRO GONZÁLEZ, José Antonio. Implicaciones documentales en el procesamiento del lenguaje natural. Ciencias de la Información, 24(1):48-54. Marzo1993.
PLAUNT, Christian and Barbara A. NORGARD. An Association-Based Method for Automatic Indexing with a Controlled Vocabulary. Journal of the American Society for Information Science, 49(10): 888-902. 1998. [También accesibe en la web en: Papers on Information Retrieval and Autonomous Agents. Berkeley: University of California, Chris Plaunt's UC Berkeley Web Page, 25 de agosto de 1997, rev. 20 de diciembre de 1995. Disponible en: http://bliss.berkeley.edu/papers/assoc/assoc.html]
POLANCO, Xavier. Infométrie et ingénierie de la connaisance. Nancy: INIST-CNRS, 1995.
REMIZE, Michel. Le thesaurus face au texte intégral: une évolution tournée vers l'utilisateur. Archimag, (112): 40-41. Mars 1998.
ROBREDO, Jaime. Indexação automatica de textos: uma abordagem otimizada e simple. Ciencia da Informação, 20(2): 130-136. Jul/Dez 1991.
ROSENBERG, V. A study of statistical measures for predicting terms used to index documents. Journal of the American Society for Information Science, 22(1): 41-50. 1971.
ROWLEY, Jennifer E. Abstracting and Indexing. 2nd ed. London: Clive Bingley, 1988.
SALTON, G. Automatic text processing: The transformation, analysis, and retrieval of information by computer. Boston: Addison-Wesley, 1989.
SALTON, G., L. ALLAN and C. BUCKLEY. Automatic structuring and retrieval of large text files. Communications of the ACM, 37(2): 97-108. 1994.
SOSA, Eduardo. Procesamiento del lenguaje natural: revisión y estado actual, bases teóricas y aplicaciones. Information World en Español, 6(12): 26-29. Enero-febrero 1997.
SLYPE, Georges van. Los lenguajes documentales de indización: concepción, construcción y utilización en los sistemas documentales. Madrid: Fundación Germán Sánchez Ruipérez, 1991.

VERDEJO MAILLO, M. F. Comprensión del lenguaje natural: avances, aplicaciones y tendencias. Procesamiento del lenguaje natural: 5-29. 1994.
WOODS, William. Transition network grammars for natural language analysis. Communications of the AMC, 13(10): 591-606. 1970.

ELABORACIÓN Y MANTENIMIENTO DE TESAUROS

Wilfrid Lancaster
Universidad de Illinois, Chicago (EE.UU.)

Como se ha mencionado, la indización supone, con frecuencia, que indizadores bien preparados establezcan la representación de un tema mediante términos seleccionados de algún tipo de vocabulario. Un vocabulario controlado es, sobre todo, una lista de autoridades. En su forma más simple y básica, se trata de una mera lista de términos que el indizador debe utilizar. La figura 1 ofrece un ejemplo bastante simple. Se trata de una lista (muy incompleta) de la que un indizador debe seleccionar a la hora de indizar temas relacionados con los combustibles. Nótese que se trata de una simple lista alfabética, sin estructura, y que incluye términos de diferentes tipos -términos de combustibles; términos de actividades (minería de carbón), términos de procesos (combustión) y términos de propiedades (velocidad de combustión). Sin embargo, si el único grupo de términos relacionados con los combustibles que se permite utilizar al indizador, se trata de un vocabulario controlado totalmente legítimo, aunque poco elaborado.

Fig. 1 Lista parcial de autoridades de términos relacionados con los combustibles.

- *Carbón*
- *Carbón bituminoso*
- *Combustibles fósiles*
- *Combustibles nucleares*
- *Combustibles sintéticos*
- *Combustión*
- *Gas natural*
- *Ignición*
- *Minería del carbón*
- *Minería a cielo abierto*
- *Minería de excavación*
- *Oleoductos*
- *Petróleo*
- *Velocidad de combustión*

ESTRUCTURA DEL TESAURO

En la figura 2 se presenta una estructura típica de tesauro, basándose solamente en el limitado conjunto de entradas de la figura l. Los términos siguen estando en orden alfabético, pero se han añadido varios símbolos para hacer evidente la estructura.

La principal relación mostrada es jerárquica -la relación entre un género y, sus especies, y, de esa forma, también entre miembros iguales de la jerarquía (hermanos). Por ejemplo, el término combustibles tiene tres términos más específicos (TE): combustibles fósiles, combustibles nucleares y combustibles sintéticos. A éstos se les considera especies (es decir, «tipos»)

del género combustibles, y por tanto son miembros iguales (miembros al mismo nivel) de la jerarquía de los combustibles. Nótese que la relación jerárquica es totalmente recíproca; puesto que combustibles abarca a los combustibles fósiles como uno de sus específicos, combustibles fósiles debe mostrar combustibles como su término más genérico (TG).

Los términos también pueden estar relacionados entre sí de forma no jerárquica, reflejando, por ejemplo, el hecho de que los materiales estén en relación con sus propiedades -o con las actividades que se ejecutan sobre ellos. Esta relación menos formal (a veces conocida como relación asociativa) se indica con la referencia TR («término relacionado»). Así, los términos velocidad de combustión e ignición aparecen como relacionados (TR) con el término combustión. Aunque la condición de término relacionado no tiene por qué ser recíproca, generalmente lo es.

El tesauro también controla los sinónimos, o los términos que son casi sinónimos, eligiendo uno de ellos y estableciendo reenvíos desde los demás. Por ejemplo, se advierte a los indizadores que el término quema no está aceptado y que tienen que utilizar combustión. Nótese que la relación *use* es recíproca. Por ejemplo, bajo combustión se ve que este término se usa en vez de (UP) quema.

Es importante darse cuenta que un tesauro para la recuperación de información, si está construido adecuadamente, debería ser un esquema de clasificación perfecto. Incluso el ejemplo simple de la figura 2 muestra varias jerarquías perfectas, aunque sean relativamente pequeñas. A partir de la figura 2 puede obtenerse de forma más completa la estructura jerárquica y presentarla en cualquiera de las formas mostradas en la figura 3.

Nótese que cualquiera de los dos árboles jerárquicos de la figura 3 podría ser generado por ordenador a partir de los datos de la figura 2.

Fig. 2 Visualización de los términos de la fig. 1 en un tesauro típico.

Carbón

TG Combustibles fósiles
TE Carbón bituminoso
TR Minería del carbón

Carbón bituminoso
TG Carbón

Combustibles
TE Combustibles fósiles
Combustibles nucleares
Petróleo

Combustibles nucleares
TG Combustibles

Combustibles sintéticos
TG Combustibles

Combustión
UP Quema

Gas natural

TE Combustibles fósiles

Ignición
TR Combustión

Minería del carbón
TE Minería a cielo abierto
Minería de excavación

Minería a cielo abierto
TG Minería del carbón

Minería de excavación
TG Minería del carbón

Oleoductos

TG **Equipo de**

TR Ignición
Velocidad de combustión

distribución

Petróleo

TG Combustibles fósiles
TR Oleoductos
Quema *use* Combustión
Velocidad de combustión
TR Combustión

Y, al contrario, la jerarquía de términos genéricos y específicos de la figura 2 podría generarse por ordenador a partir de ambas alternativas de la figura 3, una vez que se añada una notación adecuada que refleje los niveles en la jerarquía.

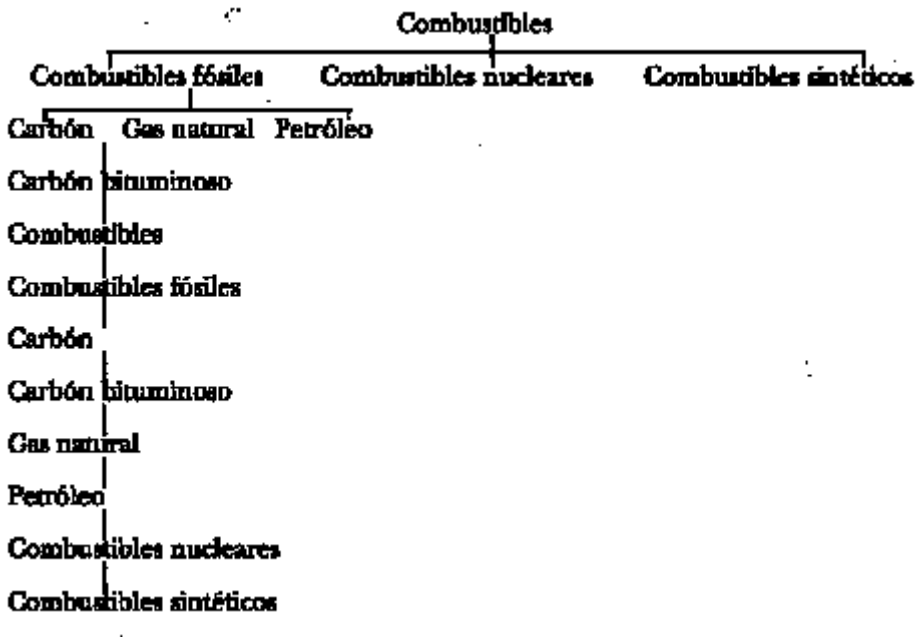


Fig. 3 Dos modos de presentación de una de las jerarquías incluidas en la fig. 2.

La presentación principal de los términos en un tesauro para recuperación de información es alfabética, con la estructura de la clasificación en forma de referencias cruzadas, como en el ejemplo de la figura 2. Sin embargo, pueden aparecer formas alternativas de presentación en un tesauro impreso o en línea, incluyendo una presentación jerárquica (probablemente parecida a la presentación inferior de la figura 3) y una presentación de palabras permutadas.

FINALIDAD DEL TESAURO

El tesauro es utilizado por los indizadores y por los usuarios de la base de datos. Una función obvia es la de controlar los sinónimos, de forma que documentos sobre temas esencialmente iguales no sean indizados bajo términos completamente diferentes. Por ejemplo, sin control del vocabulario, algunos indizadores podrían utilizar el término energía atómica, otros utilizar energía nuclear, e incluso otros, poder nuclear, ya que todos ellos se refieren a un

fenómeno esencialmente idéntico. En consecuencia, un usuario que busca información sobre este tema y que utilice sólo el término energía atómica podría no encontrar en la base de datos todo lo que sea relevante e incluso ni siquiera los documentos que posiblemente pudieran resultar de más valor. Controlando los sinónimos, el tesauro impide la dispersión de documentos similares.
La estructura del tesauro también ayuda, tanto al indizador como al usuario, a elegir los términos que parecen más apropiados para un documento determinado o una necesidad de información concreta. Por ejemplo, un indizador puede decidir en primera instancia que un documento trata de la «quema». Pero el tesauro puede llevarle a considerar que velocidad de combustión es un término mucho más apropiado. Lo mismo ocurriría en el caso de alguien que consultara la base de datos en busca de información sobre ese tema. Un ejemplo especial de esta situación es la forma en que la estructura de términos genéricos y específicos puede llevar a los indizadores o a los usuarios hasta el término más específico que resulte adecuado para la finalidad de la búsqueda -por ejemplo, conduciéndoles desde el término minería del carbón al término minería de excavación.
Otra función del tesauro, pero quizás de menor importancia, es la de clarificar el significado de ciertos términos y distinguir entre palabras homógrafas. Los homógrafos se distinguen utilizando calificadores entre paréntesis, fundamentalmente para conseguir dos palabras a partir de una sola cadena de caracteres, como por ejemplo pena (condena) y pena (tristeza). A algunos términos cuyo significado puede resultar oscuro se les puede añadir notas de aplicación para aclarar la forma en que deben ser usados.
Pero quizás la función más importante del tesauro es la de ayudar a quienes quieren llevar a cabo búsquedas realmente exhaustivas. El entramado de referencias incluidas en el tesauro así lo permite. Para poner un ejemplo obvio, es improbable que el término de indización combustibles recupere todo lo que hay en una base de datos sobre el tema. Normalmente, el usuario de la base de datos debe utilizar todos los términos de combustibles permitidos en el vocabulario controlado. Como muestra la figura 2, el usuario que busque bajo el término combustibles será enviado hasta los términos más específicos, a varios niveles, de forma que le mostrará el conjunto completo de términos necesarios, para una búsqueda exhaustiva sobre el tema. Si el sistema de recuperación ha sido bien diseñado, debería ser posible para el usuario realizar una búsqueda genérica sobre el término combustibles -una búsqueda que recupere todo lo que haya sido indizado bajo este término o con cualquier otro término que se encuentre por debajo de él en la jerarquía.
La estructura de términos relacionados puede servir también de ayuda en el caso de una búsqueda exhaustiva. Por ejemplo, un usuario que busque bajo el término refinado de petróleo será reenviado a términos relacionados como refinerías de petróleo, capacidad de las refinerías, y destilación, conceptos que pueden tener una cierta relevancia para el objeto de la búsqueda.

MÉTODOS DE ELABORACIÓN

Los tesauros se construyen, normalmente, mediante un proceso intelectual humano, aplicando métodos de "arriba a abajo" o "de abajo a arriba". Los grandes tesauros creados por las agencias gubernamentales y otras organizaciones de los Estados Unidos en los años sesenta utilizaron el planteamiento de arriba a abajo. Este es un método esencialmente de comité. Grupos de expertos en la materia se

reúnen para organizar la terminología en sus campos de especialización. Volviendo al ejemplo utilizado anteriormente, un grupo de expertos podría empezar con el término combustibles y decidir cómo habría que subdividirlo, en los distintos niveles, hasta establecer la jerarquía completa de combustibles. Luego centrarían su atención en las jerarquías relacionadas -propiedades de los combustibles, utilización, etc.

Este planteamiento de arriba a abajo es pesado y costoso. Además, es más teórico que empírico. El planteamiento de abajo a arriba se inicia con los propios términos, tal y como aparecen en los documentos a indizar, y luego se organizan éstos en las jerarquías adecuadas. La justificación de este planteamiento estriba en que tiene «autoridad» literaria (a veces llamada «autoridad bibliográfica»), lo que simplemente significa que un término está justificado (autorizado) si se sabe que se utiliza en la literatura, o, al menos, que aparece con suficiente frecuencia.

El procedimiento más eficiente para la elaboración de un tesauro basado en la autoridad literaria es el de reunir los términos a partir de fuentes ricas en terminología. Los glosarios y diccionarios especializados son candidatos obvios para ello, pero raras veces están completamente actualizados. Los documentos que se están publicando en un determinado momento en un área de conocimiento, especialmente las revistas científicas y los informes técnicos, son los que mejor reflejan el uso actual de los términos. Si existe una revista de resúmenes para el campo concreto, será una fuente excelente de terminología. Incluso, mejor sería una base de datos de resúmenes: se pueden elaborar programas para manipular la base de datos y generar listados de términos ordenados por la frecuencia de aparición.

Una vez obtenidos los términos, deben ser organizados en una estructura coherente y cohesiva. Supóngase que se está elaborando un tesauro en el campo de la biblioteconomía, que los términos han sido obtenidos del Library and Information Science Abstracts y que han sido registrados en fichas (*). Después de terminar la obtención (porque parece haberse alcanzado un punto de decreciente rendimiento), las fichas pueden ser agrupadas según términos «semejantes». Por ejemplo, un montón de fichas que representen tipos de bibliotecas, otro que represente los tipos de materiales que manejan las bibliotecas, un tercero con tipos de servicios que proporcionan las bibliotecas; etcétera.

Este proceso se muestra en la figura 4. Lo que ha sucedido es la división de la terminología de la biblioteconomía en una serie de aspectos o facetas. Alguno de los montones será relativamente amplio porque la faceta es amplia; otros montones pueden ser bastante pequeños. De hecho, puede ser necesario crear un (afortunado) pequeño montón misceláneo para incluir los términos que no parecen encajar bien dentro de ninguna de las facetas principales.

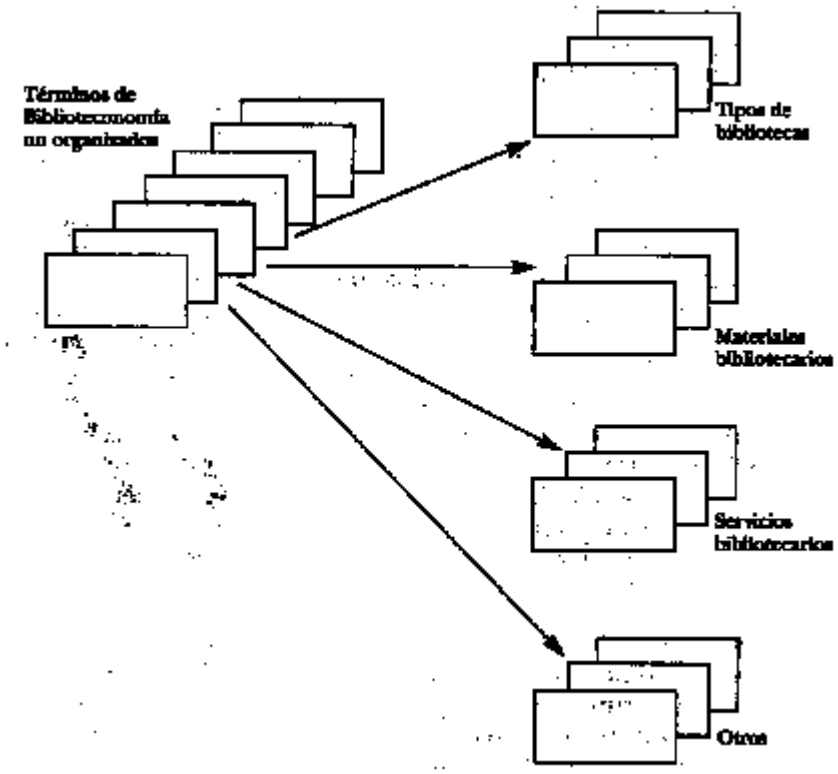


Fig. 4 Aplicación del análisis de facetas a los términos.

Después de haber identificado las facetas de este modo, cada una debe ser organizada en jerarquías del tipo mostrado en la figura 3. En Lancaster (1) puede encontrarse un estudio más completo del proceso de obtención de términos y de su organización.

MANTENIMIENTO DEL TESAURO

Excepto en ocasiones muy poco usuales, hay que mantener actualizados los tesauros para reflejar los cambios terminológicos en el campo, incluyendo términos que nunca han aparecido anteriormente. Las personas que con más probabilidad encontrarán términos nuevos serán los propios indizadores. Se les debe ofrecer la posibilidad de registrar cualquier término nuevo que no esté en el tesauro todavía pero que resulte necesario para poder indizar adecuadamente un documento. A veces, tales términos son registrados como términos «provisionales» y se les marca de alguna forma para distinguirlos de los que ya están en el tesauro. Estos términos provisionales serán revisados periódicamente para comprobar la frecuencia con que han sido utilizados. Si parece probable que su importancia persista, serán aprobados y añadidos en los lugares apropiados dentro de la estructura jerárquica. Las organizaciones implicadas en operaciones de indización a gran escala, como las agencias gubernamentales y los productores de bases de datos comerciales, probablemente tendrás uno o más lexicógrafos responsables de revisar los términos nuevos sugeridos, por los indizadores y demás usuarios de la base de datos, que tomarán la decisión sobre su inclusión definitiva y en que momento se añaden a la estructura del tesauro.

DIRECTRICES Y NORMAS

El primer tesauro para recuperación de la información se desarrolló en 1959, dando lugar a una intensa actividad en la construcción de tesauros por parte de agencias gubernamentales y asociaciones profesionales de los Estados Unidos. Esta experiencia, continuada a lo largo de los años sesenta, favoreció el desarrollo de directrices para la construcción de tesauros que culminó con la publicación de normas nacionales (en los Estados Unidos, en el Reino Unido, y en otros países) y, eventualmente, internacionales. La norma internacional (2) constituye una guía esencial para quien se plantee construir un tesauro, ya que se ocupa de la forma de los términos, características estructurales, métodos de presentación y demás elementos. Para una historia del desarrollo de las normas para la construcción de tesauros, véase Lancaster y Krooks (3).

MÉTODOS AUTOMÁTICOS

Hasta aquí, sólo se ha tomado en consideración la elaboración de tesauros por medio de procesos intelectuales humanos. Desde luego, se utilizan ordenadores para procesar los datos del tesauro, para ejecutar aplicaciones de mantenimiento, para imprimir y visualizar los tesauros, etcétera. También pueden ser utilizados para tareas de la construcción del tesauro, como generar listas de los términos candidatos que aparecen con mayor frecuencia en las bases de datos.

A lo largo de los años, varios investigadores, sobre todo Salton (4), han producido «tesauros» por procedimientos completamente automatizados. Las herramientas que han elaborado son muy diferentes de los tesauros altamente estructurados descritos aquí. Están construidos a partir de asociaciones estadísticas entre los términos que aparecen en el texto. Así, un «grupo del tesauro» podría consistir en términos (normalmente palabras simples o raíces de palabras) que tienden a aparecer juntas en el texto con frecuencia. Para poner un ejemplo completamente hipotético, un grupo del tesauro de carbón podría incluir, además de la propia palabra carbón, palabras como minería, excavación, bituminoso, antracita, quema, combustible y combustión. Es evidente que semejante grupo muestra una gran heterogeneidad -términos que indican tipos de carbón, procesos, actividades, etcétera. Es bastante diferente de la categoría que probablemente formaría un humano. Sin embargo, grupos de términos asociados de esta manera pueden seguir siendo de utilidad para las operaciones de proceso automático de textos. La figura 5. proporciona otro hipotético ejemplo basado en un término médico.

Fig. 5 "Grupo de tesauro" hipotético que puede formarse con programas de ordenador en base a datos de co-aparición de términos.

MIGRAÑA
Migraña
Pacientes
Dolor de cabeza
Ataques
Dolor
Síntomas
Sumatriptan

Cerebral
Sangre
Vascular
Ergotamina

Las asociaciones estadísticas también pueden usarse para visualizar palabras en línea que pueden ser de utilidad para quien haga una búsqueda en una base de datos. Para poner un ejemplo hipotético, cuando se teclea la palabra «carbón», un programa de asociación estadística podría generar una visualización de la palabra «carbón» y otras intensamente asociadas a ella, como:

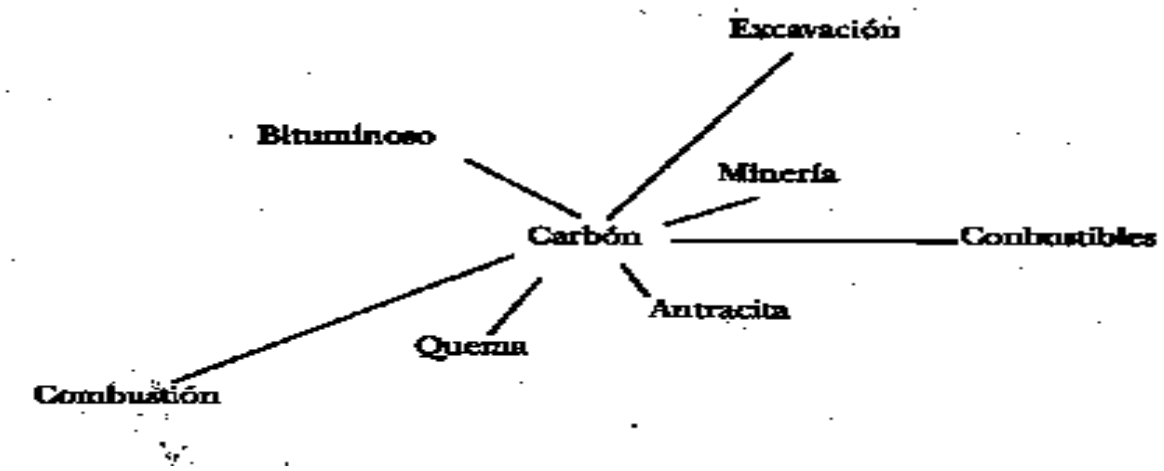


Fig. 6 Asociación estadística.

En este ejemplo, realmente simple, la distancia entre carbón y las demás palabras refleja lo estrechamente asociada que están por la frecuencia de aparición: quema aparece junto a carbón con mucha más frecuencia que combustión. Parece que las visualizaciones de este tipo han sido inventadas por Doyle (5), quien se refirió, a ellas como «mapas semánticos». Y han sido descubiertas de nuevo, más recientemente, por investigadores que se han planteado experimentos con las ayudas de búsqueda en el contexto de la WWW (6,7).

QUÉ ES LO QUE HACE QUE UN TESAURO SEA «BUENO»

Es posible analizar un tesauro impreso y establecer una valoración de su calidad: ¿sigue lo establecido en las normas?, ¿están adecuadamente construidas las jerarquías?, ¿se establece correctamente la reciprocidad de los términos?, etcétera. En la práctica, sin embargo, un tesauro sólo puede ser juzgado en relación con el uso que se hace de él -es decir, su validez para fines de recuperación. Como se ha mencionado, 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.

CONTROL DE VOCABULARIO EN EL CONTEXTO DE LA NFORMACIÓN DIGITAL

La World Wide Web ha hecho que las fuentes de información sean ampliamente accesibles en cantidades inimaginables hace incluso una década. La mayoría de esas fuentes existen en forma de texto. El hecho de que distintos motores de búsqueda permitan acceder a textos de páginas Web mediante la combinación de palabras (o partes de palabras) sugiere que el control del vocabulario es innecesario en el contexto de Internet. Sin embargo, en realidad Internet ha dado lugar a un creciente interés por los principios del control del vocabulario.

Diversas organizaciones, la mayoría de instituciones académicas, están intentando mejorar el acceso a aquellas fuentes de Internet consideradas más relevantes y valiosas para determinados ámbitos. Lo hacen identificando las fuentes que quieren hacer accesibles (una forma de control de calidad) e indizando estos recursos de algún modo. En esencia, lo que están haciendo es proporcionar indicadores a recursos evaluados de interés potencial a su ámbito (ingenieros, educadores, usuarios de bibliotecas públicas, o cualquier otro tipo de usuarios). Este trabajo está bien explicado en el libro de Wells (8).

Las organizaciones que hacen este trabajo proporcionan "portales" para seleccionar las fuentes de Internet, utilizando procedimientos de indización convencionales. Casi sin excepción, la indización se basa en alguna forma de vocabulario controlado -encabezamientos de materias, clasificaciones o tesauros.

La culminación de este desarrollo es la reciente propuesta de que las bibliotecas académicas deberían asumir la responsabilidad de desarrollar un «portal de expertos», que podría promover el desarrollo del contenido de mayor calidad de la Web y proporcionar acceso al mismo, es decir, un portal para apoyar a la investigación de alto nivel. Es destacable que la propuesta de establecer este portal sugiera específicamente que puede incluir «elaborados tesauros electrónicos que guíen con precisión a los investigadores a áreas de interés» (9).

También es significativo para el futuro del control del vocabulario la aparición del campo de la gestión del conocimiento. Las actividades de gestión del conocimiento van dirigidas a la organización de los recursos de información de una empresa para hacerlos más rápidamente accesibles. Los grupos de gestión del conocimiento están descubriendo que las herramientas de la biblioteconomía tradicional -la clasificación, la indización, el control del vocabulario- son muy relevantes para el diseño de portales a recursos corporativos internos y a recursos de información potencialmente valiosos accesibles a través de la Web. Muchos de esos grupos emplean graduados de escuelas de biblioteconomía/ciencias de la información y/o personas con preparación en lingüística. Además, las actividades de gestión del conocimiento frecuentemente incluyen la creación de vocabularios controlados del tipo de los tesauros aunque de hecho se les dé otros nombres, como taxonomías. Algunos grupos de gestión del conocimiento están ahora comprometidos en actividades innovadoras de control del vocabulario. Por ejemplo, Microsoft ha desarrollado un tipo de lenguaje de conexión que permite que diferentes grupos de la empresa busquen en todos los recursos corporativos utilizando su propia (es decir, del grupo) terminología preferidas.

(*) *Aunque el registro en tarjetas pueda parecer primitivo en nuestra sociedad electrónica, la mayoría de la gente encuentra más fácil ordenar objetos físicos que trabajar en un entorno completamente virtual.*

REFERENCIAS

Capítulo 7 de: Procesamiento de la información científica. Madrid: Arco/Libros, 2001, pp. 182-193.

(1) LANCASTER, F. W., El control del vocabulario en la recuperación de infamación. Valencia. Universitat de Valencia, 1995.
(2) International Organization for Standardization. Guidelines for the Establishment and Development of Monolingual Thesauri. 2a. ed., Geneva, 1986 (ISO 2788).
(3) LANCASTER, F. W.; KROOKS, D. A., "The evolution of guidelines for thesaurus construction". En: Libri, 1993, 43, 326-342.
(4) SALTON, G.; McGiLL, M. J., Introduction to Modern Information Retrieval. New York: McGraw Hill, 1983.
(5) DOYLE, L. B., "Semantic road maps for literature searchers". En: Journal of the Association for Computing Machinery, 1961, 8, 553-578.
(6) FOWLER, R. H. et al., "Visualizing and browsing WWW semantic content". En: Proceedings of the First Annual Conference on Emerging Technologies and Applications in Communications, 110-113. Los Alamitos: CA, IEEE Computer Society Press, 1996.
(7) ZIZI, M., "Interactive dynamic maps for visualization and retrieval from hypertext systems". En: Information Retrieval and Hypertext (M. Agosti, A. F. Smeaton, eds.), 203-224. Boston: Kluwer, 1996.
(8) WELLS, A. T. et al., The Amazing Internet Challenge. Chicago: American Library Association, 1999.
(9) CAMPBELL, J. D., "The case for creating a scholars portal". En: ARL, 211, August 2000,1-4.
(10) CRANDALL, M., "Microsoft". En: Linkage Inc.´s Best Practices in Knowledge Management and Organizational Learning Handbook, 89-123. Lexington: MA, Linkage Inc. 2000.

ELABORACIÓN DE LOS TESAUROS DE DESCRIPTORES

Miguel Ángel López Alonso
Universidad de Carlos III de Madrid (España)

EVOLUCION HISTÓRICA

Su desarrollo vendrá precedido por diversos estudios lingüísticos sobre lexicología en distintos países europeos. Ya a comienzos del siglo XVIII, en Francia, Girard trata de dar solución a las dudas que presenta el empleo de voces afines en su libro "Synonimes françois" (1741) que orientará en Europa el sucesivo tratamiento de la sinonimia, como afán de producir para cada lengua intelectual -no provincial o artesana- libros que fijen el valor exacto de las palabras de significación semejante. En Alemania, Gottsched publicará en Leipzig su libro "Observaciones sobre el uso y abuso de varios términos de la Lengua Alemana" (1758).

A principios del siglo XIX, en Inglaterra, Crabb escribirá el famoso libro "English Synonymes Explained", que todavía sigue reeditándose con adicciones y puestas al día de diversos autores. En los años sesenta de este siglo, algunos prestigiosos investigadores fundaron serias esperanzas en la automatización del tratamiento de los documentos, debiendo analizar profundamente los conceptos lingüísticos que utilizaban: palabras, frases, resúmenes, descriptores, etc., tanto en la INDIZACION como en la creación, ordenación e impresión de los primeros INDICES (índice KWIC de Peter Luhn, 1957). En este período se multiplicación los estudios lingüísticos sobre Lenguajes de Indización de autores como: Spark Jones y Needham (Inglaterra, 1964); Coyaud (Francia, 1966) que elabora una estructura teórica para el análisis uniforme de los nuevos lenguajes de indización, cuyos constituyentes están tomados de la terminología lingüística (fonemas, semas, etc.); I. Dahlberg (Alemania, 1974) que emprende una búsqueda interdisciplinar de la noción de clasificación, desbordando el campo de la biblioteconomía y pasando por la filosofía, la epistemología, la lingüística, las teorías científicas, etc.; o Hutchins (1975) con una profunda introducción de las "Estructuras Lingüísticas Generales de los Lenguajes de Indización comparados con los Lenguajes Naturales" en sus diversos aspectos: formales, semánticos, pragmáticos, etc.

Será Mortimer Taube (1953) quien dará a conocer el primer sistema de indización con tesauro integrado (Uniterm), todavía usado en grandes sistemas con muy variados documentos y pluralidad de temas (ej.: la Unesco).

Aunque el primero de los tesauros publicados, formalmente constituido, fué el ASTIA, que luego se reconvirtió en el DDC (Defence Documentation Center, 1960); el primero de estos tesauros con vistas a la recuperación automatizada fue desarrollado por la sociedad Dupont de Nemours (Hohn and Rasmussen, 1961), atribuyéndose a Helen Brownson el ser la primera persona en utilizar el término tesauro en este contexto.

Ya previamente en 1956, investigadores de la Unidad de Investigación Lingüística de Cambridge (Inglaterra) adelantaron sus hipótesis sobre la aplicación del concepto de tesauro en la recuperación de la información, mediante la combinación de descriptores (usando el álgebra de Boole) para obtener la información deseada; y las primeras referencias escritas se atribuyen a Luhn (1957), Bernier y Heumann (1957) y Joyce y Needham (1958).

Eugenio Wall, apoyándose en las investigaciones de los anteriores, conceptualizará los principales problemas lingüísticos en la recuperación de la información (sintaxis, semántica, género y sentido) en dos artículos de 1957 y 1959 que determinaron el sentido del tesauro de la Dupont antes citado.

EVOLUCIÓN HISTÓRICA EN ESPAÑA

M. Carrión menciona como primeros libros donde figuran clasificaciones sistemáticas: el Libro de los Epítomes y el Libro de las Proposiciones de Hernando de Colón (1530), y el Dictionarium Historicum de Charles Estienne (1553).
En la época moderna, el primer eco de estas ideas fue el breve libro "Ensayo de los synónimos" de Manuel Dendo y Avila (Madrid, 1757), o el llamado "Examen de la posibilidad de fixar la significación de los sinónimos de la lengua castellana" de José López Huerta (Viena, 1789) que será reeditado en España varias veces y estimulará la aficción por los estudios sinonímicos; seguirán los destacados trabajos de José Joaquín de Mora en su "Colección de Sinónimos de la Lengua Castellana" (Madrid, 1855) y el ya posterior de Richard Ruppert en su libro "Spanische Synonimik" (Heilderberg, 1940).

CONCEPTOS FUNDAMENTALES

Ya desde finales de los años cincuenta, serán tiempos de gran actividad en cuanto al asentamiento de los principios teóricos y de las definiciones conceptuales (tesauro, descriptor, indización, indización coordinada, relevancia, pertinencia, etc.), atribuidas a los investigadores Taube, Howerton, Helen Brownson, Farradane, Jolley, etc.

LÉXICOS

El primer sistema desarrollado para buscar extensos cuerpos de textos legales (en el Centro Jurídico de la Universidad de Pittsburg) utilizaba un tesauro formado por una mera compilación de palabras con significados similares, pareciéndose más al Rogert's Thesaurus (de principios de siglo) que a la estructura de los tesauros actualmente usados en búsquedas documentales.

GLOSARIOS

Estos suelen ser lenguajes controlados, con el control aplicado a la salida; es decir, lenguajes generados automáticamente a la salida que están formados por tablas con nombres y números de identificación, que pueden ser consultados en el lenguaje natural de los usuarios y proporcionarles alternativas a los términos de su mente. Las tablas pueden ser visualizadas en línea y los términos pueden ser seleccionados de ellas, o, también, la tabla completa puede ser incorporada dentro de una estrategia de búsqueda por su número de identificación.
Estas tablas no deben estar restringidas solo a palabras, pudiendo incorporar raíces de términos (prefijos más radicales) para búsquedas booleanas; el vocabulario también puede tener alguna mínima estructura del tipo de referencias cruzadas o tablas interrelacionadas, y muy bien pudieran ser tesauros multilingües para uso en Bancos de Datos Internacionales.

TESAUROS

Concretamente en el campo específico de los Lenguajes Documentales de Estructura Combinatoria: "formados por los términos extraídos del Lenguaje Natural mediante la búsqueda de los llamados unitérminos y sus relaciones paradigmáticas (especialmente la sinonimia)".
Tratando de explicar la evolución sufrida desde los primeros tesauros alfabéticos hasta las últimas modelizaciones lingüísticas, matemáticas, etc., pasando por los más utilizados tesauros temáticos, facetados o mixtos; pero, sin perder de vista su futuro: cada vez más lejano de los convencionales lenguajes controlados jerárquicos (sustituidos progresivamente por los lenguajes naturales) aunque continuando su desarrollo como sublenguajes técnicos generados automáticamente que aumenten su efectividad y disminuyan su alto coste de realización manual.
Partiremos de la definición de tesauro, dada por la norma UNE 50/106 sobre "Directrices para el establecimiento y desarrollo de tesauros monolingües": "Vocabulario de un lenguaje de indización controlado, organizado formalmente, con objeto de hacer explícitas las relaciones a priori entre conceptos de los tipos: más genérico que o más específico que".

RELACIONES ENTRE DESCRIPTORES

En una lista no estructurada de unitérminos, tipo vocabulario o glosario, se dan únicamente relaciones de equivalencia o paradigmáticas entre conceptos individuales, definidas como: "relación entre los términos preferentes y no preferentes cuando se considera, a efectos de indización, que uno o más términos se refieren al mismo concepto".
Esta reciprocidad entre descriptores y sus sinónimos o cuasisinónimos (no descriptores), se expresa mediante las siguientes convenciones:

- USE, que se escribe precediendo al término preferente.
- UP (usado por), que se escribe precediendo al término no preferente.

Cuando la representación de conceptos, mediante todos los posibles términos, precisa no solo las relaciones de equivalencia entre unitérminos sino que "organiza eficazmente los términos (según su significado) en categorías, subcategorías, etc., basándose en grados o niveles de super y subordinación; en los que un término superordenado representa un todo o clase y los términos subordinados corresponden a sus miembros o partes", se dan las llamadas relaciones jerárquicas, entre términos individuales, características de los primeros tesauros o vocabularios controlados.
Esta jerarquía entre términos específicos y genéricos se expresa mediante las siguientes convenciones:

- TG (término general), que se escribe precediendo al término superordenado.
- TE (término específico) se escribe precediendo al término subordinado.

Si añadimos relaciones asociativas o afines (de tipo psicológico) entre los términos no equivalentes ni con relaciones jerárquicas entre sí (sobre la base de que tal conexión podría revelar términos alternativos útiles en la recuperación), estaremos caracterizando los tesauros americanos alfabéticos de finales de los años sesenta, tipo ERIC Descriptors y sus reglas para la preparación de tesauros (Office of Education, 1969).

En el caso de estas asociaciones, es deseable que los términos se encuentren en distintos niveles jerárquicos y que la relación sea recíproca entre términos, expresándose mediante la abreviatura:
- TR (término relacionado) o su equivalente en otros idiomas.
Según esto, todos los términos de indización, preferentes y no preferentes, se organizan en una secuencia alfabética única.
Normalmente, los términos no preferentes sólo se acompañan de reenvíos (ej., USE) a sus equivalentes preferentes, mientras que la información que acompaña a los términos preferentes debe listarse en el siguiente orden:
NA. Notas de aplicación o definiciones términos confusos.
UP. Reenvíos de los términos equivalentes no preferentes.
TG. Referencias a los términos genéricos.
TE. Referencias a los términos específicos.
TR. Referencias a los términos relacionales.
R. Jansen (1974) tratará de demostrar la utilidad de un nuevo tipo de relación que llamará de pertenencia, distinta de la jerárquica, que pone en relación cada término con sus descriptores multitérminos en cuya composición participa.

CONSTRUCCION DE TESAUROS

I. Elección del área temática y razones para su elaboración.
II. Comprobación de la no existencia de otro tesauro similar, tanto en el ámbito nacional como europeo.
Otras de las posibilidades es la adaptación a otro de un nivel conceptual más genérico o macrotesauro (tipo OCDE), adaptándolo a las necesidades de la institución, incorporando términos propios de la documentación específica que se maneja en el sector, excluyendo aquellos términos que no se identifican con la materia de interés.
Van Slype describe el proceso de adaptación de un tesauro de la siguiente forma:
1) Se adquieren los tesauros disponibles sobre la temática,
2) Se prepara una muestra de unos mil documentos para indizar,
3) Se indizan esos documentos en lenguaje natural,
4) Se intentan traducir los descriptores libres de la indización en descriptores controlados, extraídos de cada uno de los tesauros examinados, uno tras otro.
El tesauro que haya proporcionado la mayor cantidad de descriptores, que respondan a las necesidades del sistema documental, será el candidato para la elección.
Por debajo del 95 y más del 80 por ciento de los descriptores necesarios, deberá ser completado por los descriptores que falten para llegar a ser el adecuado.
Por debajo del 80 por ciento deberá construirse un nuevo tesauro para el sistema documental en estudio, utilizando los descriptores como fuentes de terminología del tesauro a preparar.
III. Delimitación del tema principal, sus auxiliares y marginales.
IV. Recopilación de los términos,
- de vocabularios, diccionarios, libros, y otras fuentes secundarias ("a priori"),
- de publicaciones periódicas y otras fuentes primarias ("a posteriori").
V. Normalización y depuración del vocabulario.
VI. Agrupación de los términos y estructuración de las relaciones jerárquicas, asociativas, de equivalencia,
VII. Puntualización de las notas de aplicación.

VIII. Edición.
El tesauro debe ir precedido de una introducción detallada que indique claramente su objetivo, estructura y los campos que cubre, teniendo en cuenta las consideraciones internas referidas al propio centro para el que se desarrolla. También, es necesario indicar brevemente las reglas que han regido su elaboración, así como las fuentes y métodos que se han usado en la selección de los descriptores, y las fechas previstas para su actualización.

TIPOS DE TESAUROS

En cuanto a las materias tratadas y/o su forma de tratarlas, pueden ser:
- generales o especializados,
- monodisciplinares o multidisciplinares.

En cuanto al numero de idiomas en que los descriptores pueden ser consultados, pueden ser:
- monolingües o plurilingües.

Por la cobertura territorial o institucional, pueden ser:
- nacionales o internacionales,
- públicos o privados.

Por el tipo de estructura con que se presentan los descriptores, pueden ser:
- jerárquicos, facetados o mixtos.

En cuanto a la forma de presentar los términos y sus interrelaciones en un tesauro, pueden ser de:

a) Presentación alfabética, con notas de aplicación (SN) e indicadores de interrelaciones terminológicas;
b) Presentación sistemática, apoyada en un índice alfabético (tipo Léxicos);
c) Presentación gráfica, con una sección alfabética; y
d) Presentación mixta, uniendo dos o más de las formas anteriores.

Partiendo de una estructura constructiva facetada y trabajando Barhydt y su grupo de la Universidad Western Reserve (1966) en el desarrollo de un tesauro de términos educacionales que utilizaba el análisis de facetas puro, enviaría a ERIC (Office of Education, 1966) un informe con alrededor de 4500 términos que fue rechazado; aunque posteriormente sería revisado y publicado por el Servicio de Publicaciones de la Universidad Western Reserve.
La primera solución práctica a las dificultades que planteaba un tesauro de facetas puro, la adelantarían Aitchison y otros colaboradores de la English Electric (1969) al presentar el primer tesauro-faceta que adoptaría una solución mixta mucho más eficaz: aunando la clasificación por materias con una subdivisión por facetas que comprendía dos entradas:
- la relación alfabética de los descriptores del tesauro, con las relaciones TE, TG, TR, y el reenvío por un código de tres caracteres a

Es esta área la que facilita la búsqueda inicial y remite al usuario a la parte facetada, mediante un código.
- la clasificación por materias, en la que los términos son reagrupados en facetas o categorías fundamentales.

Es en esta área donde la riqueza documental localiza los descriptores específicos del contexto científico en que se ha elaborado el tesauro.
Este concepto integra en un solo sistema las ventajas de los tesauros alfabéticos junto con una nueva presentación sistemática de los términos, teniendo en cuenta también las relaciones entre las propias jerarquías o categorías; analizando los

términos de un campo temático en clases o conjuntos con una característica común, según los tipos básicos de categorías principales o facetas que representan (entidades, partes, propiedades, procesos, operaciones, agentes, aplicaciones, etc.) y abandonando los campos de interés por materias o disciplinas científicas, usados tradicionalmente.
Para entenderlo mejor, los objetos concretos podrían haber sido subdivididos (por la naturaleza de los conceptos principales) en facetas como por ej., procesos, materiales, útiles, etc.; mientras que los campos temáticos por disciplinas podrían haber sido subdivididos en por ej., Agricultura, Medicina, o Economía.
Dado que las facetas son categorías fundamentales que pueden ser reconocidas en cualquier campo científico y se usan para clasificar distintos conceptos en grupos homogéneos, con las mismas características divisorias; el número de facetas puede variar desde las cinco propuestas por Ranganathan (personalidad, energía, materia, espacio y tiempo) a las propuestas actualmente por el CRG (Grupo de Investigación en Clasificaciones) de Londres: procesos y operaciones, entidades y entidades abstractas, agentes (personas, organizaciones, facilidades, equipos), propiedades, materiales, partes, entidades globales, productos, receptores o pacientes, etc.
La norma francesa AFNOR Z47-100 propone seis facetas fundamentales (objetos, dispositivos, elementos, propiedades, procedimientos y aplicaciones) y algunos ejemplos de facetas anexas que vienen a coincidir con las propuestas recientemente por Alexis Rivier (1991) en su libro "Construcción de los Lenguajes de Indización".
Dentro del tipo de Tesaurofacetas conviene señalar: la "Teoría de la Indización Multimodelo" de D. Soërgel (1985), para quien cualquier tipo de clasificación facetada nos debe conducir al mismo descriptor en un tesauro determinado, según los distintos tipos de indizadores y sus distintas aproximaciones contextuales.

CARACTERÍSTICAS ESTRUCTURALES Y FUNCIONES

La consecuencia de la doble codificación (en la comunicación entre el emisor y el destinatario) es que se aprecia el lenguaje natural como un instrumento no ideal para asegurar el buen funcionamiento de la comunicación documental.
¿Cómo conseguir representar el mismo objeto por medio de una única fórmula en las dos fases de la codificación (la del emisor y la del documentalista), si se utiliza un código tan ambiguo como es el lenguaje natural?.
Recordemos que un lenguaje se llama «biunívoco» cuando además de unívoco, es recíproco:
a) no hay más que un solo término para denominar un mismo objeto, y
b) cada término no designa más que a un solo objeto.
La primera condición es fundamental en documentación ya que de ella depende el éxito de la búsqueda. La segunda es menos importante, pero los términos plurívocos (con varios sentidos) provocan malentendidos, respuestas falsas, o «ruido documental».
Por ejemplo: cualquiera ha tenido un día u otro la experiencia desesperante de una búsqueda infructuosa en la guía telefónica, simplemente, se buscaba como café, bar, restaurante, un establecimiento que estaba ordenado como mesón.

TRATAMIENTO DE LA AMBIGÜEDAD EN EL LENGUAJE NATURAL

Se sabe que el lenguaje natural no es «biunívoco»: esto se debe a sus características internas, cuyos dos aspectos más estudiados son la sinonimia y la perífrasis.

A. La sinonimia:

Se dice que dos palabras son sinónimas cuando tienen el mismo significado. Son sinónimas dos palabras de la misma lengua que tengan la misma denotación y la misma connotación. La denotación establece las relaciones lingüísticas entre una palabra y un objeto (para ser más exacto entre una palabra y la imagen mental de un objeto). Por el contrario, la connotación determina el uso concreto que un emisor puede hacer de una palabra.

La verificación práctica de la sinonimia es, pues, fácil de enunciar: dos sinónimos son sustituibles el uno por el otro, en cualquier contexto y fuera de él. En la práctica, es muy difícil encontrar palabras que reúnan perfectamente estas dos condiciones, salvo en:

- los vocabularios especializados, en los que la connotación tiene un papel poco importante, y en los que, a veces, dos términos son usados indistintamente (cancerólogo o carcinólogo, cefalea o cefalalgia). De lo que resulta que, en:
- el lenguaje natural, la noción de sinonimia es una noción límite y que los sinónimos perfectos no existen.

Para convencerse de que la relación de sinonimia no es transitiva basta con establecer una cadena de sinónimos, por ejemplo: juego, distracción, diversión, recreo, relajación, descanso, ocio, reposo... Tomados de dos en dos, los términos de la cadena son sinónimos, pero la semejanza de significado con el término inicial disminuye a medida que se alejan del mismo. Sin embargo, estas limitaciones no significan que la noción de sinonimia sea banal, ni que la cuestión de la pluralidad de designaciones sea un problema falso.

B) La perífrasis:

La perífrasis es una especie de réplica de la sinonimia en el nivel de la frase, al igual que la sinonimia es una equivalencia de significado entre dos palabras, la perífrasis es una equivalencia de significado entre dos enunciados.

Mi suegro se fue.
El padre de mi cónyuge se marchó.

La perífrasis es un fenómeno complejo que no puede ser ignorado por los especialistas de la información, ya que presenta los mismos problemas que la sinonimia: un contenido idéntico con distintas formulaciones. Pero es más difícil de dominar porque se trata de un fenómeno del "discurso", mientras que la sinonimia es un hecho de "lengua" que se puede recopilar en los diccionarios.

En efecto, la perífrasis no se limita a una acumulación de sinónimos, aunque sean un elemento de la misma, como puede verse en el ejemplo siguiente:

Juan ha dejado de fumar hace dos meses.
Mi hijo no fuma ya desde hace dos meses.

Desde el uno de abril, mi hijo mayor ha renunciado al tabaco.

La sinonimia y la perífrasis tienen importancia en la riqueza de la lengua: permiten variar y matizar la expresión y evitar las repeticiones.

TRATAMIENTO DE LA AMBIGÜEDAD EN LOS LENGUAJES CONTROLADOS

La otra cara de la biunivocidad se llama sencillamente «univocidad» o ausencia de ambigüedad. Un término es unívoco (no ambiguo) cuando se aplica a un único objeto de la realidad, más exactamente a una única representación mental, sea la imagen de un objeto concreto (percepción) o de una noción (concepto).

Por ejemplo, los términos españoles televisión, cuatro, esférico, son unívocos.

Basta con abrir un diccionario de sinónimos para valorar la importancia de la plurivocidad (o, mejor dicho, la «polisemia»). Las dos variantes de la ambigüedad que eliminan los lenguajes controlados son la homonimia y la polisemia.

A) La homonimia:

La homonimia es la similitud formal de palabras diferentes. Si la similitud aparece en la forma oral de la palabra se habla de homofonía; si aparece en la forma escrita, se habla de homografía. Por ejemplo, en español, «vello» y «bello», «ojear» y «hojear», son homófonas pero no homógrafas.

La homografía se distingue de la polisemia en que los términos homógrafos son distintos por su origen y en que el encuentro formal que los ha confundido es simplemente fortuito.

Comencemos por un contra-ejemplo: Una moneda es una moneda.

En la frase las dos cadenas de caracteres «moneda» no son homógrafas sino la simple repetición de una misma palabra (mismo sentido, mismo origen).

Por el contrario, los siguientes ejemplos muestran claramente el aspecto fortuito de la homografía:

La espuma cayó sobre un colchón de espuma.

Ella vino a por una botella de vino.

LA es la quinta nota musical.

B) La polisemia:

Como en el caso de la homografía, la forma de los términos es similar y el sentido es diferente. Pero se trata de una misma palabra que, a partir de un determinado momento se ha enriquecido con un nuevo significado. Generalmente se reconocen los términos polisémicos en que los significados guardan un cierto parentesco (hoja de papel, hoja de árbol - devorar un pollo, devorar un artículo), aunque en algunos casos la relación sea tan lejana que se haya borrado totalmente y sea necesario recurrir al diccionario (pupila del ojo, pupila de la nación).

Hay que indicar que las lenguas generan continuamente nuevos casos de polisemia dadas las posibilidades creativas del lenguaje. Las dos causas más frecuentes de esta proliferación son la metáfora y la metonimia.

C) La metáfora deriva de la comparación. Por poco que una metáfora se incorpore al uso lingüístico, ya hay creado un nuevo caso de polisemia.

Por una tendencia natural a abreviar la expresión, se pasa, sin darnos cuenta, de

Este individuo es voraz como un tiburón a
Este individuo es un tiburón, y de
Esta mujer es insoportable como una cotorra, a
Esta mujer es una cotorra.

D) La metonimia es también un deslizamiento del sentido por reducción de la expresión. Pero, en este caso, la proximidad semántica entre la expresión original y el término resultante es un hecho de la realidad y no-fruto de la imaginación.

Variantes: Se designa el todo por una de las partes, el producto por el productor, el ocupante por el lugar ocupado, el contenido por el continente, la función por un símbolo, el producto de la acción por la acción:

Leer una policíaca.
Una aldea de doce hogares.
Los aplausos del patio de butacas.
Beber una botella.

El diputado ha perdido su escaño.

Utilizar una clasificación.

REFERENCIA

Material bajado de Internet. Extracto del curso Diseño de Sistemas de Indización.

LOS TESAUROS CONCEPTUALES COMO HERRAMIENTA DE PRECISIÓN EN LOS SISTEMAS DE ORGANIZACIÓN CIENTÍFICA

Miguel Ángel López Alonso
Universidad Carlos III de Madrid (España)

INTRODUCCIÓN

Se detecta una tendencia creciente, apoyada en el modelo conceptual, para proporcionar al usuario soluciones de acceso a la recuperación de la información, que mejoren las preguntas iniciales, mediante sesiones de búsqueda en las que este comparta sus propios conocimientos con los existentes en las terminologías a su alcance. En las nuevas aplicaciones de las industrias de la lengua se hace uso de los tesauros conceptuales como elemento de precisión para las búsquedas en lenguaje natural, dentro del vocabulario mixto usado en los sistemas de recuperación de la información.

En una primera fase del procesamiento del lenguaje natural (1), la lingüística automatizada desarrolla diferentes analizadores morfológicos, sintácticos o semánticos, para evitar el análisis de las palabras y de sus relaciones (2).

El análisis por el significante, característico de los analizadores morfosintácticos, obliga a descomponer los descriptores complejos en sus componentes. La utilización de los analizadores semánticos implica el análisis de las palabras para su descomposición en el mayor número de ideas posible (ej.: "retroalimentación", se forma a partir de dos términos simples, alimentación y retro = volver a).

En una segunda fase, la terminología reúne los conceptos obtenidos del análisis contextual de los documentos científicos de las distintas ramas del conocimiento, para la creación de bancos de datos terminológicos.

Finalmente, se utilizan los procedimientos de la inteligencia artificial en la rama del procesamiento del lenguaje natural (3) para facilitar la creación de macrodiccionarios electrónicos o bases de conocimientos terminológicos, que permitan búsquedas automatizados en diferentes bases multilingües, en línea, trabajando cooperativamente a través de la red Intemet.

La normalización del lenguaje puede aplicarse fácilmente al conocimiento codificado de los vocabularios controlados, pero, limitada sólo a pequeños grupos, de usuarios (ej. en la morfología restringida de las nomenclaturas especializadas, en la sintaxis especial de las patentes o dé los contratos legales, etc.). Puesto que muchos textos técnicos se almacenan en soporte digital, se les puede convertir a un formato adecuado para su análisis terminológico con las técnicas desarrolladas por la lingüística automatizada. Si se les analiza y compara con los vocabularios electrónicos, almacenados en los bancos terminológicos, se pueden obtener listados de elementos nuevos, erróneos o necesitados de equivalente en otras lenguas.

En el diseño de los sistemas integrados de gestión de la información, los documentalistas deben fijarse como meta la investigación pormenorizada de la base de conocimientos del sistema, dejando el diseño del motor de inferencia y de la interfase de usuario a los ingenieros del conocimiento. La habilidad del sistema para aplicar la heurística de los distintos profesionales científicos al proceso de

recuperación de la información, proporciona respuestas a las siguientes preguntas de los usuarios:

- ¿Cuándo usar un tesauro?
- ¿Dónde empezar a usarlo?
- ¿Qué relaciones seguir?
- ¿Qué nuevos conceptos seleccionar?
- ¿Cómo incluir los nuevos descriptores en la pregunta?

Esto les permitirá ampliar o restringir la búsqueda, mediante interacciones con el sistema, del tipo:

- Use términos más generales y términos relacionados para un más amplio tratamiento de la búsqueda.

- Muévase más arriba o más abajo en la jerarquía del tesauro para modificar la especialización.
- Use sólo vocabulario controlado para una mayor especificidad en la búsqueda. (4)

Como punto de partida para el diseño de sistemas perfeccionados, se considera imprescindible la fusión de las teorías de la recuperación de la información (IR) y de la organización del conocimiento (KO) que aporta a la primera sus potentes estructuras de conocimientos.

Se han establecido como postulados mínimos para su integración:

1) Que la recuperación de la información, en general, no sea una actividad aislada sino que se integre con la indización previa y esto les permita mantener un alto nivel de correspondencia, y

2) Que la recuperación de la información científica, en particular, se produzca en el marco del trabajo habitual del profesional y de sus habituales sistemas informáticos de tratamiento documental (ej.: archivos documentales, procesadores de texto, correo electrónico, etc.).

LOS VOCABULARIOS CONTROLADOS COMO SUPERACIÓN CONCEPTUAL DE LOS SISTEMAS DE CLASIFICACIÓN

Los sistemas de clasificación de estructura enumerativa aparecieron cuando predominaba la preocupación por clasificar el contenido de los documentos temática y físicamente, mediante su representación sintáctica. Con el posterior crecimiento exponencial de la información, la capacidad de estos vocabularios controlados para representar y recuperar el contenido de los documentos se desborda y obliga a realizar nuevas investigaciones.

Para intermediar en la comunicación entre la ciencia de la documentación y la lingüística documental, aparecen los lenguajes de estructura combinatoria o vocabularios controlados (6), que tratan los documentos prescindiendo de los sistemas de clasificación universales y se especializan por áreas del conocimiento. La clasificación, como ayuda a la catalogación y a la recuperación por materias, se sustituye por la indización mediante descriptores tomados de un tipo de estos vocabularios, los tesauros. Los sistemas integrados de gestión de la información recurren a los descriptores de dichos tesauros específicos, para indizar los documentos por sus títulos o resúmenes y, tras la ponderación de los conceptos más repetidos, proceden a utilizarlos posteriormente como descriptores postcoordinados en la recuperación, mediante combinación con los operadores

booleanos.
Los vocabularios controlados constituidos sobre lenguajes de estructura combinatoria buscan un acercamiento de los conceptos documentales y los lingüísticos que mejore el análisis y la indización de la información almacenada en las bases de datos documentales, para su exacta recuperación en los sistemas anteriores. Aunque los estudios de Markey, Atherton y Newton (7) ya habían señalado que las recuperaciones con lenguaje natural producían mayor exhaustividad y menor precisión que las realizadas con un vocabulario controlado, será Boyce quien afirme:
“Dado que cada documento recuperado mediante una búsqueda, a partir de un término descriptor, es uno de los documentos recuperados por otra búsqueda basada en sus términos componentes en el índice inverso primitivo, será imposible que la primera búsqueda, con vocabulario controlado, recupere más documentos que la segunda con lenguaje libre”, lo que, en consecuencia le induce a afirmar:
”el vocabulario controlado, usado en los grandes sistemas comerciales de gestión de la información, constituye un elemento de precisión en virtud de la propia estructura por defecto de sus registros informáticos, y nunca será un elemento de exhaustividad”. (8)
En esta línea, podemos afirmar que la menor exhaustividad y mayor precisión inherente a las recuperaciones realizadas con vocabularios controlados, dependerán preferentemente del hecho intrínseco que las caracteriza: "ser una reducción en la cantidad de términos utilizados para la búsqueda". Y, en segundo lugar, del tipo de vocabularios controlados utilizados (tesauros, clasificaciones, etc.), del número de términos asignados a cada descriptor, de sus capacidades para evitar homónimos, etc. (9) Con su integración se logra mayor exhaustividad y precisión en las recuperaciones, al eliminar el «ruido documental» derivado de la utilización de los sistemas de clasificación tradicionales.

LA INCORPORACIÓN DE UN TESAURO CONCEPTUAL COMO ELEMENTO DE PRECISIÓN EN LAS RECUPERACIONES CON LENGUAJE NATURAL

Entre los procedimientos propuestos para evitar la falta de precisión del lenguaje natural, se destaca la incorporación de los conceptos obtenidos del análisis de contenido de los documentos en un tesauro conceptual que facilite las recuperaciones desde la interfase sistema - usuario. De esta forma se incrementó la variedad de las ecuaciones de búsqueda del usuario, o al menos le muestra la variedad de las existentes; lo que, de acuerdo con el principio de variedad de requisitos de Ashby (10), incrementa simultáneamente la cantidad de respuestas del sistema consultado.
Los tesauros conceptuales diseñados para ayudar en la enunciación de las preguntas en la fase de recuperación de la información en grandes bases de datos, propuestos por autores como Bates (11), Lancaster (12) o Schnutz-Esser (13), atenúan la indeterminación de las búsquedas en lenguaje natural, especialmente en aquellas bases de datos cuyos documentos con texto completo no han sido previamente indizados con ningún otro tesauro.
Diversos experimentos con ecuaciones de búsqueda, en las que los términos del lenguaje natural de los usuarios se apoyan con términos adicionales extraídos de un tesauro conceptual (14), han llegado a doblar la precisión en el número de

documentos recuperados (15). Otros, que tratan de medir los efectos de la ampliación de las búsquedas mediante este tipo de tesauros, obtienen los mejores resultados cuando los usuarios:

a) Pueden elegir de entre una gran cantidad de términos del campo específico tratado, sean suyos, de la base de datos o incluso de indizaciones previas, y
b) Controlan interactivamente el proceso de navegación por el tesauro, en vez de utilizar procedimientos automáticos (15).

METODOLOGÍA PARA LA COMPILACIÓN AUTOMÁTICA DE TESAUROS CONCEPTUALES

La modelización de los principios teóricos que presiden la estructura de los tesauros ha seguido preferentemente los modelos lingüísticos o matemáticos; sin embargo, las recientes teorías giran alrededor de la noción del motivo o materia de la que tratan los textos, es decir, del concepto semántico o cognitivo.

Maniez propugna (1976) un modelo de tesauro en el que las relaciones no sean lingüísticas: paradigmáticas (pertenecientes a la lengua, fuera de todo contexto), o sintagmáticas (pertenecientes al discurso, integradas en su contexto), sino extrasemánticas o asociativas; de forma que aúnen términos y conceptos reales por su similitud de sentido, en el contexto específico del usuario (17).

Dewéze formaliza (1981) la representación de las relaciones semánticas con la adopción de una teoría semántica extralexical que sitúa a un nivel superior al de los lenguajes naturales, y la perspectiva de construir tesauros multilingües. En esta teoría, un significado es definido como «un conjunto de semas a los que se pueden atribuir posteriormente relaciones lexicales en varios idiomas (18).

La red semántica conceptual facilita:

- el encaminamiento de las hojas hacia una raíz y viceversa, gracias a la función de orientación presentada por las clases de discriminantes o facetas,
- la transición de un árbol a otro atravesando los nudos polijerárquicos o relaciones asociativas,
- la extensión a un árbol completo, la restricción a un subárbol, o incluso la restricción a árboles parciales, o
- la exploración transversal de la red buscando configuraciones de una o varias facetas en una reunión o en una intersección de árboles.

A partir de este concepto de red semántica de Dewéze, Schaüble (1989) propone una nueva estructura de la información, el espacio conceptual. Y construye una teoría de los tesauros conceptuales, modelada como un sistema mediante la lógica matemática del dominio algebraico, que revela una estrecha relación entre los tesauros y el modelo espacial, y en la que las relaciones entre términos son definidas con más precisión que en los tesauros jerárquicos (19).

Durante las dos últimas décadas se han propuesto diversos métodos para la construcción automática de los tesauros conceptuales, pero todos ellos descansan básicamente en análisis estadísticos de la covarianza de las palabras que componen los textos, no alcanzando a utilizar una función realmente apropiada. (20). Aunque no existe una clara demostración de la utilidad de los términos generados mediante dicho análisis, una investigación reciente ha sugerido que la relevancia de los documentos recuperados puede doblarse si el usuario añade a

sus propios términos, los sugeridos por un tesauro autogenerado mediante dicho análisis estadístico (21).
Chen sugiere que para la autogeneración de un tesauro conceptual automatizado, se deben crear las condiciones previas adecuadas para el desarrollo de algoritmos potentes y capaces de producir:

- listas de términos que actúen como filtros para la identificación inmediata de los conceptos más importantes de los documentos,
- indizadores automáticos para la identificación por materias del resto de los conceptos, y
- analizadores semánticos que agrupen los conceptos relacionados, mediante el procesamiento de los diferentes documentos de la base de datos y la obtención de sus índices de covarianza correspondientes (22).

En la misma línea conceptual, Paice propone que tanto las preguntas como los documentos sean representados por «entidades condensadas ideales» o situaciones sacadas del texto, generadas mediante «expansión activada» (23) de los paneles conceptuales:
«Si somos capaces de hacer equivalentes una alusión particular de un texto con un tema conocido a través de dichas entidades, entonces una entidad almacenada podrá ser utilizada como base para la representación de la idea implicada» (24)
Para la construcción de estas entidades, capaces de ser comprendidas, descritas y almacenadas como base para la representación de las ideas implicadas en las preguntas y en los documento (25), se precisa disponer de una base de conocimientos terminológicos suficientemente profusa y exacta, que permita abarcar:

- los términos del tesauro que estén en el fragmento de texto analizado,
- los términos no mencionados en dicho fragmento que aparecen interrelacionados con los anteriores en el tesauro,
- los términos que se interrelacionan con ambos términos anteriores dentro de la red semántica, y
- las propiedades de todos ellos, junto con las relaciones de cualquier tipo que los unan (asociativas, de relevancia, etc.).(26)

Con una exacta medición del grado de solapamiento entre cada dos de estas «situaciones ideales» (27) se puede incrementar la relevancia de los sistemas de recuperación y hacer posible la utilización de un espacio conceptual que refleje relaciones asociativas más estrechas entre los datos e ignore las menos importantes. Las recuperaciones se realizarán colocando un nuevo objeto, correspondiente a una pregunta, dentro de la red neuronal, de manera que pueda localizar sus documentos más próximos conceptualmente. (28)

APLICACIÓN DE LA TECNOLOGÍA HIPERTEXTUAL EN LOS SISTEMAS DE ORGANIZACIÓN DE LA INFORMACIÓN

Mientras la organización de la información contenida en los trabajos impresos está dispuesta secuencialmente (en capítulos y párrafos con una variedad de índices), la información contenida en las bases de datos de texto completo precisa de una estructuración que permita una recuperación no tan rígida, mediante referencias cruzadas entre sus distintos documentos asociados. La organización hipertextual permite moverse a través de una red conceptual de macroestructuras, que se define a partir de las asociaciones de ideas del usuario, derivadas de su nivel de

conocimientos de cada disciplina y de su filosofía personal.
Los sistemas hipertextuales permiten mejorar la organización escasamente estructurada de estas bases de datos, y atenúan la débil relevancia en la recuperación de los documentos científicos, derivada de las características propias de su discurso, del considerable volumen de los textos, del almacenamiento masivo en bases de datos en texto completo, y de la integración de tipos muy diversos de documentos, entre otros factores.
Estos sistemas se caracterizan por:

- Proporcionar una gran potencia conceptual para la gestión del texto de los documentos y de su red de relaciones no estructuradas, y
- Favorecer que la colección documental sea explorada utilizando criterios asociativos definidos por los enlaces hipertextuales entre las diversas partes de sus documentos.

La creación de una ontología hipertextual
Desde que en 1945, Vannevar Bush (asesor científico de Roosevelt) propusiera su sistema de recuperación Memex mediante referencias cruzadas (29), se ha considerado repetidamente la idea de crear sistemas hipertextuales para el procesamiento de las fuentes documentales. Se ha desarrollado dicha idea con el avance de la tecnología informática, siendo Ted Nelson uno de sus más fervientes defensores en el ámbito científico actual, a partir de que en 1981 definiera su sistema Xanadu (30).
La tecnología hipertextual es la mejor herramienta para estructurar enlaces significativos entre los documentos científicos, y para relacionar los descriptores de los tesauros y los códigos clasificatorios. Para mejorar la precisión de las recuperaciones documentales se propone el diseño de ontologías especializadas por áreas del conocimiento que conduzcan hasta las preguntas específicas del área tratada. Deberán ser reutilizadas, en las bases de conocimientos de los agentes expertos de la IA, para la autogeneración de tesauros conceptuales internos que permitan la distinción de los sinónimos, la supresión de los homónimos y la inducción de relaciones asociativas entre los descriptores.
Una ontología para una base de conocimientos de la IA debe abarcar los diferentes tipos de documentos, las descripciones conceptuales, las relaciones entre dichos documentos (citas), y las de éstos con los diferentes problemas científicos; además de índices, descripciones bibliográficas, tesauros, códigos clasificatorios, formalizaciones de validez, información terminológica, etc. Su aplicación debe proporcionar una metavisión de la estructura y de la terminología del dominio que facilite recuperaciones altamente relevantes.

SUPERACIÓN DEL CARÁCTER ESTÁTICO DE LA TECNOLOGÍA HIPERTEXTUAL

Se discute la aplicación de esta organización en los agentes expertos, como apoyo directo en la resolución de problemas del trabajo científico profesional, debido a la complejidad de la tarea intelectual para crear las bases de conocimientos de soporte. Sin embargo, una buena parte de los documentos científicos están especialmente indicados para su organización hipertextual, al estar repletos de referencias cruzadas, expresas o tácitas (ej. las leyes y sus desarrollos reglamentarios, la doctrina jurídica y sus definiciones o comentarios, los casos legales de las diferentes jurisdicciones, etc.) (31)

Dado que las bases de datos documentales no fueron diseñadas para la resolución directa de problemas, la organización de su conocimiento requiere el análisis de cada documento y el resumen de sus notas más destacadas, mediante dos fases consecutivas: primera, la extracción previa de los datos menos relevantes (clasificación de los documentos), y segunda, el posterior examen especializado para estructurar el conocimiento referido a temas específicos.

a. Una de las limitaciones de estos sistemas es el "desbordamiento cognitivo" o desorientación de los usuarios durante su navegación a través de los textos o de las clasificaciones hipertextuales para encontrar documentos relevantes o conceptos apropiados, respectivamente (32). La unión de la información de numerosas fuentes con un único nodo conduce a una sobresaturación de enlaces en dicho nodo. La introducción de muchos de éstos dificulta la navegación, reduce el deseo de los usuarios de explorar su información e incrementa las posibilidades de perderse. Se trata de utilizar sólo enlaces significativos e incluir en cada nodo una idea conceptual completa que sintetice el conocimiento contenido en las fuentes enlazadas.
b. Otra de las limitaciones de estos sistemas se deriva de la necesidad de controlar manualmente los enlaces creados durante el trabajo automático, dado que éstos no contienen información conceptual directa "a priori" sobre el contenido de los documentos relacionados.

Al no poder procesar dinámicamente la información existente en sus fuentes, dichos sistemas recuperan solamente los documentos que tienen enlaces predefinidos con un nodo principal o mantienen relaciones dinámicas con otros. Se trata de proporcionar al usuario herramientas globales de ayuda a la navegación: mapas cognitivos, historia semántica de los nodos visitados, opción de retorno mediante el marcado de nodos, etc. (33)

CONCLUSIONES Y SOLUCIONES PROPUESTAS

- La normalización de la terminología de los documentos es uno de los problemas cruciales para el desarrollo de los sistemas hipertextuales. Se estima que en éstos se precisa de un conocimiento terminológico mínimo de 80.000 términos normalizados. Para ello se propone el desarrollo de aplicaciones hipertextuales que utilicen las tecnologías de la IA para automatizar la compilación de hipertesauros a partir de las bases de conocimientos terminológicos, capaces de abarcar una buena parte del corpus lingüístico de cada una de las lenguas más utilizadas (34).

Entre las soluciones investigadas atenuar el carácter estático de la tecnología hipertextual, se destaca la utilización del tesauro conceptual (35), concebido como una red semántica en la que en cada nodo hay un único concepto semántico que puede contener una serie de descriptores asociados, que también pueden ser identificados en la red de descriptores relacionados según las típicas relaciones de los tesauros: preferenciales, jerárquicas o asociativas. Para evitar la falta de precisión del lenguaje natural, se propone la incorporación de los conceptos obtenidos del análisis de contenido de los documentos científicos en un tesauro conceptual que facilite las recuperaciones desde la interfase sistema usuario del profesional. De esta forma se incremento la variedad de las ecuaciones de búsqueda del usuario, o al menos le muestra la variedad de las existentes; lo que, de acuerdo con el

Principio de Variedad, de Requisitos de Ashby (36), incrementa simultáneamente la cantidad de respuestas del sistema consultado.

- La perfección de los agentes expertos del campo de la IA permitirá mejorar las bases de datos de conocimiento de dichos sistemas, facilitando su viabilidad como herramientas para la organización y recuperación del conocimiento en la actividad científica habitual (37), apoyados en la tecnología hipertextual integrada con las redes semánticas (38). Estas últimas deberán proporcionar un mapa de las variables introducidas en la búsqueda y de las obtenidas en la recuperación, y ser capaces de almacenar información de forma distribuida, aprendiendo de los problemas cotidianos, tras un período de validación que actualice las respuestas (39).

Para su optimización, los sistemas integrados de gestión de la información deberán contar con una interfase hombre-máquina, capaz de relacionar los conceptos científicos a partir de las ontologías preestablecidas. El usuario tendrá acceso a ellos a través de una representación gráfica de la red (tabla de contenidos gráficos con las conexiones visibles mediante enlaces hipertextuales), por la que podrá rastrear los documentos siguiendo su línea de interés (40).

Se deberá priorizar el desarrollo de asociaciones de redes neuronales que de alguna manera conviertan en dinámicos los nodos hipertextuales, permitiendo que todos los conceptos situados en cada uno de ellos puedan ser consultados simultáneamente (no sólo secuencialmente) de acuerdo con las necesidades cognitivas de cada usuario en un momento dado. Para ello será preciso utilizar neurordenadores, con procesamiento “masivamente paralelo”, que trabajen con múltiplos nodos hipertextuales del mismo modo que el crebro humano trabaja con múltiples neuronas (41).

NOTAS Y REFERENCIAS

Artículo publicado en: Revista Interamericana de Bibliotecología, vol. 22, No. 1, enero-junio de 1999, pp. 21-35.

(1) Lo que se ha dado en llamar “Modelización de los Procesos Lingüísticos Naturales”.

(2) Conociendo los complicados mecanismos del razonamiento humano se podrá resolver la ambigüedad léxica u oracional, profundamente difíciles de programar.

(3) Mediante el agente experto, dotado de un «motor de inferencia» que controla cómo y cuándo deberá aplicarse la información recogida (en las dos primeras fases de la modelización), al corpus lingüístico de una lengua viva.

(4) JONES, Susan et al. Interactive Thesaurus Navigation: Intelligence Rules Ok?. Journal of the American Society for Information Science, 1995, 46(1), p. 58-59.

(5) DOMINICH, S. Interaction Information Retrieval. Journal of Documentation, 1994, 50 (3), pp.197-212.

(6) MOREIRO GONZALEZ, J. A. De la Documentación a la Ciencia de la Información: evolución de los conceptos y aplicaciones documentales. Seminario de Humanidades Agustín Millares Carlo, separata Homenaje a Antonio de Bethencourt Massicu, 1995, p. 20.

(7) MARKEY, K., ATHERTON, P. y NEWTON, C. An analysis of controlled vocabulary and free text search statements in online searches. Online Review, 1980, 4, pp. 225-236.
(8) BOYCE, B. R. y McLAIN, 3. P. Entry Point Depth and Online Searching using a controlled vocabulary. Journal of ASIS, 1989, 40 (4), p. 273.
(9) SOËRGEL, D. Indexing and Retrieval Performance: The Logical Evidence. Journal ASIS, 45 (S), 1994, pp. 539-599.
(10) ASHBY, W. R. An Introduction to Cybernetics. London: Methven, 1973, pp. 202-212.
(11) BATES, M. J. Subject Access in Online Catalogs: A Design Model. Joumal of the ASIS, 1986, 37 (6), p. 361.
(12) LANCASTER, F. W. Vocabulary control for information retrieval. Information Resource Press, 1986 (XVII ed), 270 p.
(13) SCHMITZ-ESSER, W. New Approaches in Thesaurus Application. International Classification, 18 (3),1991, pp. 143-147.
(14) CROFT, W. B. y DAS, R. Experiments with query acquisition and use in document retrieval systems. Procecedings of the 13th Conference on Research and Development in Information Retrieval, Brussels, Belgium, 9/1990. KRISTENSEN, J. Expanded End-user's Query statements for free text searching with a search-aid thesaurus. Information Processing and Management, 1993, 29 (6), pp. 733-744.
(15) EKMEKCIOGLU, F. C., ROBERTSON, A. M. Y WILLET, P. Effectiveness of query expansion in ranked-output document retrieval systems. Journal of Information Science, 1992, 18, pp.139-147.
(16) Ibid. Cit. 4.
(17). MANIEZ, J. Los lenguajes documentales y de clasificación. Madrid: Pirámide, 1993, p. 214.
(18) DEWÉZE, A. Réscaux sémantiques: essai de modélisation; aplication á l'indexation et á la recherche documentaire. Lyon: Universidad Claude Bernard. Tesis doctoral, 1981, p. 68.
(19) SCHAÜBLE, P. Information Retrieval Based on Information Structures. Informatik-dissertationen eth Zurich, Verlag der Fachvereine, 1989, 135 p.
(20) La mayoría de las funciones de análisis de covarianza existentes (ej.: coseno, Dice, de Jaccard, etc.) son simétricas por naturaleza, habiendo producido efectos colaterales no deseados y poca exactitud en los resultados.
(21) Ibid. Cit. 15.
(22) CHEN, H., YIN, T. y FYE, D. Automatic thesaurus generation for an Electronic community system. Journal of the ASIS, 1995, 46 (3), p. 179.
(23) Formados por la distribución («mapping») de los términos tomados de un área del conocimiento específica y agrupados en forma de red neuronal.
(24) PAICE, C.D., “A Thesaural Model of Information Retrieval”. Information Processing Management, 1991, 27(5), p. 435.
(25) Existentes separadamente de los textos y de los tesauros terminológicos autogenerados.
(26) GREEN, R. Topical Relevance Relationships. Why topic matching fails?. Journal of the ASIS, 1995, 46 (9), p. 648.
(27) Ej.: incorporando pesos a los vectores que unen los conceptos.
(28) DEERWESTER et al. Indexing by latent semantic analysis. Journal of the ASIS, 1990,41 (6), p. 394.

(29) BUSH, V. As we may think. The Atlantic Monthly, july 1945.
(30) NELSON, T. Replacing the printed word: A complete literary system. Information Processing´ 80, 1980.
(31) BENCH-CAPON, T. An incremental aproach to legal drafting support. Proceedings of JURIX'94, 1994, Lelystad: Kroninklijke Vermande, Netherlands.
(32) ROVIRA, C. Entornos hipertextuales de aprendizaje. Anuario SOCADI, 1997, pp. 121-126.
(33) CODINA, L. El libre digital. Barcelona: Centre d' investigació de la Comunicació, 1996.
(34) Electronic Dictionary Research Project. Japan Key Technology Center. JTEC/WTEC Hyper-Librarian, 1995,
VIVALDI, J. Seminario Procesamiento Lenguaje Natural (SEPLN'95). Universidad de Deusto, Bilbao, 20-23 sep., 1995.
(35) LÓPEZ ALONSO, Miguel A. Un tesauro conceptual para la recuperación de la información jurídica comercial. Revista Española de Documentación Científica, 1998, 21 (2), pp. 164-173. 36.
(36) Ibid, Cit. No. 10.
(37) CORTEZ, E. M. Et al. The hybrid application of an inductive learning method and a neural network for intelligent information retrieval. Information Processing and Management, 1995, 31 (6), p. 790.
(38) BERGER; F. C.; VAN BOMMEL; P. Aumententing a characterization network with semantic information. Information Processing and Management, 1997, 33 (4), p. 453-479.
(39) COHEN; P. R.; KJELDSEN, R. Information retrieval by constrained spreading activation in semantic network. Information Processing and Management, 1987, 23 (4), p. 257.
(40) ALLEN, L. E. Language, law and logic; plain legal drafting for the electronic age. Computer Science and Law, Cambridge U. Press, 1980, pp. 75-100
(41) MOYA, F. de, HERRERO, V. y GUERRERO, V. La aplicación de redes neuronales artificiales a la recuperación de la información. Anuario SOCADI 1998, pp. 147-164.

DISEÑO LÓGICO-CONCEPTUAL DE TESAUROS

Francisco Javier Martínez Méndez
Laura Martínez Méndez
J. Vicente Rodríguez Muñoz
Universidad de Murcia (España)

1. INTRODUCCIÓN

En anteriores trabajos, hemos aportado la idea de la introducción de un Modelo de Datos, el Modelo Entidad Relación en particular, como marco de referencia para la implementación de un Lenguaje Documental de Estructura Combinatoria, concretamente nos referimos a un Tesauro. El concepto de Modelo de Datos se refiere al grupo de herramientas conceptuales utilizadas para la descripción de la realidad de un sistema de información. Este grupo se compone de los datos, sus relaciones, su semántica y sus relaciones; instrumentos que utilizamos para el diseño de una Base de Datos a nivel lógico, dentro de la Arquitectura de Tres Niveles aceptada por la Norma ANSI/SPARC.

Uno de los Modelos de Datos de mayor aceptación y posteriores desarrollos es el Modelo Entidad Relación, introducido por Chen a mediados de los años 70. Este modelo se basa en dos elementos fundamentales:

a) Las Entidades o conjunto de objetos individuales que se distinguen unos de otros por medio de sus atributos y

b) Las Relaciones o asociaciones que se establecen entre las entidades.

Podemos destacar que existe una cierta similitud estructural entre un Tesauro y un Modelo E-R. En un Tesauro, los términos descriptores son distinguibles y además, se establecen entre ellos una serie de relaciones de naturaleza semántica. Por ello, el Modelo de Datos E-R parece muy adecuado para el diseño de un Tesauro, debido a la gran facilidad que nos aporta para la representación de los conjuntos de entidades que participan en un Tesauro y de las distintas relaciones propias de sus términos.

A la hora de la implementación y puesta al marcha de nuestro sistema, tomando como base el Modelo E-R anterior, utilizamos como herramienta el Modelo Relacional. La nueva vista de la realidad que proporciona este modelo, es el marco apropiado para la aplicación de una serie de reglas de inferencia lógica sobre los datos contenidos en el mismo. Así, generamos una Base de Datos Deductiva, que ofrece información adicional a la ya ofrecida explícitamente.

2. EL MODELO ENTIDAD RELACIÓN (MODELO E-R)

Tal como se ha destacado anteriormente, los dos elementos fundamentales de este Modelo de Datos, se encuentran inmersos en su propio nombre: la Entidad o Conjunto de Entidades y la Relación o Conjunto de Relaciones.

Una entidad se distingue de otra por medio de sus atributos, o características de la misma. Por propia definición, no pueden existir dos entidades iguales. El contenido o valor de los atributos se encuentra limitado por un determinado Rango.

Una entidad se puede agrupar con otras del mismo tipo (es decir, que posean los mismos atributos, pero, evidentemente, con contenido diferente). Es decir, una entidad Persona, puede pertenecer al conjunto de entidades Ciudadanos. Una entidad puede pertenecer a varios conjuntos de entidades, o sea, la misma

entidad Persona puede pertenecer al conjunto de entidades Clientes de una determinada empresa.
Al conjunto de atributos que sirve para identificar una entidad de otra, se le conoce como Superclave, y a la superclave mínima (es decir, al mínimo conjunto de atributos válido para efectuar la distinción entre dos entidades), se le denomina Clave Primaria. La Clave Primaria de una entidad, es también la clave primaria del conjunto de entidades del mismo tipo.
Cuando una entidad precise por razones de existencia, de la existencia previa de otra entidad de distinto tipo (es el típico caso de un apunte en una cuenta corriente: no puede existir el apunte si no existe la cuenta), podemos decir que la primera entidad es una entidad dependiente por existencia de la segunda. En este caso, la entidad dependiente se considera que es de naturaleza débil, frente a la otra que se considera de naturaleza fuerte.
Una entidad débil, carece de clave primaria, por lo que para distinguirla de otra se hace necesario recurrir a la entidad fuerte de la cual depende.
Un Modelo de Datos E-R puede trasladarse a un Modelo de Datos Relacional, donde la visión del sistema de información se realiza por medio de tablas (Relaciones). Para ello, se siguen una serie de reglas apropiadas al caso, que explicaremos en el apartado 7.

3. MODELO RELACIONAL

La visión relacional de un determinado sistema de información se corresponde al almacenamiento en forma de tablas (o relaciones), de las distintas tuplas (filas de la relación), que se corresponden a las entidades del modelo E-R. En cada columna de la tabla se depositan los valores de los distintos atributos de las tuplas.
El Modelo Relacional, es con mucho, el más en auge en la actualidad. El aumento considerable de los sistemas gestores de bases de datos relacionales hoy en día, no hace más que afirmar su gran valía como modelo de datos.
Las tuplas se distinguen unas de otras por medio de su Clave Primaria, de igual definición que en el Modelo de Datos E-R. Toda tupla tiene clave primaria, por lo tanto, toda tupla es distinguible. Si en una relación aparece un atributo que es clave primaria en otra relación, se le denomina Clave Ajena.
Es muy importante en este modelo todo lo referente a la integridad y consistencia del mismo. Por ello, se han introducido como norma general dos reglas de integridades o propiedades de tipo semántico que la base de datos debe cumplir:
1. Integridad de Entidad: ningún valor de una clave primaria puede ser nulo.
2. Integridad de Referencia: todo valor de una clave ajena debe ser distinto de nulo y además pertenecer al conjunto de valores de la relación donde dicha clave sea primaria.
Estas dos reglas de integridad se ven complementadas por una serie de restricciones de integridad, que en cada modelo persiguen el objetivo de salvaguardar la consistencia y verificabilidad de los datos. Por ejemplo, no podremos hacerle una nota de préstamo de un libro a un estudiante, si éste no aparece en el listado de los alumnos del centro.

4. BASES DE DATOS DEDUCTIVAS

Una base de datos deductiva es una base de datos en la que podemos derivar información a partir de la que se encuentra almacenada explícitamente. Como

elementos constitutivos de una Base de Datos Deductiva nos encontramos con los Hechos, Reglas de Inferencia y las Restricciones de Integridad.
Los hechos representan la información que se almacena explícitamente; en el diseño e implementación de las reglas de inferencia se toma como base la lógica de primer orden y las restricciones de integridad son de la misma tipología que en el modelo anterior.
La actuación de un conjunto de rutinas lógicas sobre los hechos llega a producir como resultado una información inferida que en un principio no aparece de forma explícita. Es ésta una característica muy a tener en cuenta, ya que así podemos deducir una serie de relaciones existentes entre los términos descriptores que almacenamos en un tesauro soportado por una base de datos relacional pero que no aparecen reflejadas en una primera instancia. Sirva como ejemplo el caso de la siguiente relación, en la que se recogen los datos relativos al parentesco PADRE-HIJO. Hay que destacar que en la misma, no aparece información relativa a la ascendencia en un grado superior (como puede ser el caso del abuelo). A esta relación la denominaremos PADRE.

PADRE	**HIJO**
Antonio	Juan
J. María	Dolores
Juan	Pedro
Pedro	Jesús

Sobre los datos recogidos en la relación anterior, podemos definir las siguientes reglas deductivas:
i)Ascendiente(x,y) <-- Padre(x,y)
ii)Ascendiente(x,y) <-- Padre(x,z) ^ Ascendiente(z,y)

Y por medio de las mismas, queda definida la figura del Ascendiente, de la siguiente manera:

i) todo padre es Ascendiente
ii) una persona X es un Ascendiente de una persona Y si existe un Z tal que X sea padre de Z y Z sea a su vez un Ascendiente de Y.

Aplicando las reglas de inferencia lógica i) y ii), obtendremos otra serie de hechos más amplia, que recogemos en la siguiente relación que vamos a denominar ASCENDIENTE.

ASCENDIENTE	**DESCENDIENTE**
Antonio	Juan
J.María	Dolores
Juan	Pedro
Pedro	Jesús

Antonio	Pedro
Antonio	Jesús
Juan	Jesús

5. EL MODELO DE DATOS E-R MICROTES

A continuación presentamos nuestro Modelo de Datos Microtes, diseñado y adaptado a un Tesauro. Un Modelo de Datos, tal como se ha dicho anteriormente representa una realidad, en este caso el sistema de información es un Tesauro, cuya estructura la detallamos de manera concisa a continuación:

Podemos tomar como definición de Tesauro, la proporcionada por Borko y Bernier: "un Tesauro es una LISTA organizada de términos de un vocabulario especializado elaborada para facilitar la selección de sinónimos y de palabras que sean afines de otra manera".

Aitchison y Gilchrist consideran a los términos de la lista como Términos Indizantes, tomando como base de definición de los mismos la proporcionada por la Norma ISO-2788: "un término indizante (index term), es la representación de un concepto". Puede consistir de más de una palabra, y entonces, se conoce como término compuesto. En un lenguaje controlado un Término Indizante puede ser bien un Término Preferente o bien, un Término No Preferente.

Un Término Preferente es aquél que es utilizado consistentemente en la indexación para representar un concepto dado. Es conocido también como "Descriptor" o "palabra clave" (keyword).

Un término no preferente es el Sinónimo o Cuasi-sinónimo de un término preferente. No es utilizado en la indexación, pero provee de una entrada alternativa desde la que el usuario puede acceder directamente por medio de la instrucción USE al término preferente apropiado. Este tipo de término es también conocido como no descriptor.

Para la clarificación de los Términos Descriptores se hace necesaria, a veces, la utilización de las Notas Explicativas. Por ejemplo:

Bibliografías nacionales SN *Bibliografías de las obras producidas en un país en cualquier lengua que sea y/o en la lengua propia del país.*

De esta breve descripción de los elementos constituyentes de un Tesauro, destacamos a continuación tres conjuntos de entidades:

a) Términos Descriptores.

b) Términos No Descriptores.

c) Notas Explicativas.

Los conjuntos de entidades a) y b) representan a todos los términos Indizantes, que juntos conforman un subconjunto estructurado del lenguaje natural. El conjunto de las Notas Explicativas es de naturaleza débil, pues una Nota depende por existencia del un Término descriptor.

Entre los Términos Descriptores y los No Descriptores se establecen relaciones de equivalencia, que denotaremos con el símbolo USE. Esta relación admite el re-envío o relación en sentido inverso (también conocido normalmente como relación UF). En el Modelo E-R podemos especificar los re-envíos utilizando el concepto de rol (papel que desempeñan las entidades en una relación, según el sentido de la misma). USE asocia un Término Descriptor con uno de sus términos equivalentes.

Entre el conjunto de Términos descriptores y el de las Notas Explicativas se establece la relación SN, que asocia a un Término descriptor una Nota Explicativa. Los atributos correspondientes a estos conjuntos de entidades:

a) Descriptores: Signatura, Término.

b) No Descriptores: Definición (el término en sí)

c) Notas Explicativas: Explicación (la nota explicativa)

Sobre el conjunto de Términos descriptores se establecen una serie de relaciones de naturaleza jerárquica y asociativa. Las mismas se representan en el modelo E-R con la única particularidad de que coinciden los conjuntos de entidades asociados; ya que los términos descriptores pertenecen a un mismo conjunto. Pasemos a continuación a detallar las relaciones que encontramos sobre este conjunto.

NT: o relación jerárquica término amplio. El primero de los términos se considera de un significado más específico y superior que el segundo. Ejemplo: Derecho NT Derecho Civil.

BT: es la relación inversa de la anterior. En este caso, el primero de los términos tiene un significado más amplio que el segundo de los términos. Ejemplo: Bioquímica BT Química.

RT: es la relación asociativa o de afinidad. Son todas las relaciones que no pueden definirse por equivalencia o por jerarquía. Ejemplo: Enseñanza RT Educación.

De esta manera, queda definido nuestro Modelo Microtes. A continuación, pasamos a identificar las claves primarias de los conjuntos de entidades y de las relaciones que participan en el Modelo de Datos Microtes.

Entidades:

Descriptores (Signatura, Término)	CP: {Signatura}
No Descriptores (Definición)	CP: {Definición}
Notas Explicativas (Explicación)	CP: {Signatura, Explicación}

Relaciones:

USE CP: {Signatura, Definición}
SN CP: {Signatura, Explicación}
NT CP: {Signatura TE, Signatura TA}
BT CP: {Signatura TA, Signatura TE}
RT CP: {Signatura, Signatura Tafin}
donde TE: término específico; TA: término amplio; Tafin: término afín

6. DISEÑO RELACIONAL DE MICROTES

Aplicando una serie de normas de frecuente utilización en el entorno relacional de datos, podemos diseñar un Modelo Relacional a partir de un Modelo E-R. Según estas reglas, los conjuntos de entidades se representarán en forma de tablas. Las entidades ocuparán las filas (tuplas), de las tablas y éstas tendrán tantas columnas como atributos tenga la entidad. En el caso de una entidad de naturaleza débil, se incluyen las columnas necesarias para representar los atributos de la clave primaria de la entidad fuerte.

Las relaciones también se representan en forma de tablas. En este caso, las columnas se corresponderán con los atributos de las claves primarias de los conjuntos asociados por medio de la relación.

Del modelo de datos anterior, obtenemos el siguiente sistema relacional:

TABLA DESCRIPTORES

SIGNATURA	TÉRMINO
F04/49	BIOLOGÍA
G09.10	BIOLOGÍA AGRÍCOLA
F12	BIOLOGÍA CELULAR
D96	BIOLOGÍA MARINA
G64.80.10	RADIOISÓTOPO

TABLA NO DESCRIPTORES

DEFINICIÓN
BIOLOGÍA AMBIENTAL
BIOLOGÍA ANIMAL
BIOLOGÍA AGUA DULCE
BIOLOGÍA VEGETAL

TABLA NOTAS EXPLICATIVAS

SIGNATURA	EXPLICACIÓN
Z18.30	Estudio y ...
Z34.30.30	Biblioteca

TABLA USE

SIGNATURA	DEFINICIÓN
F12	CITOLOGÍA ...
F12	CITOQUÍMICA
G64.80.10	RADIOELEMENTOS

A continuación se representará la tabla correspondiente a la relación NT, cuya estructura, aunque, evidentemente, no su contenido; coincide con las tablas de las relaciones BT y RT.

TABLA NT

SIGNATURA TA	SIGNATURA TB
F04/49	G09.10.
F04/49	F05.10
C30/69	F25

F04/49	F25
F52	F52.25
F04/49	F12

7. APLICACIÓN DE REGLAS DE INFERENCIA SOBRE EL MODELO RELACIONAL MICROTES

Ha llegado el momento de plantearnos diversos objetivos para definir las Reglas Deductivas que bien pueden estar orientadas a complementar la información explícita en el M.R o bien, pueden utilizarse para asegurar el mantenimiento de la integridad del sistema de información. A continuación vamos a plantear dos ejemplos de cada uno de estos tipos de aplicación sobre Microtes:

i) En algunos tesauros se utiliza la relación **Top Term** (generalmente representada por TT); un término X es Top Term de otro término Y si existe un determinado término Z tal que X NT Z y Z NT Y. Podemos definirla de la siguiente manera:

TT(x,y) <-- NT(x,z) ^ NT (z,y)

De igual forma que en el ejemplo anterior de los ascendientes y descendientes, podemos generar una serie de reglas con el objetivo de deducir todo el conjunto de relaciones NT que no aparezcan explícitamente en la tabla NT.

NT(x,y) <-- NT(x,y)

NT(x,y) <-- NT(x,z) ^ NT(z,y)

También podemos considerar el caso de la relación asociativa o de afinidad RT, de manera que si un término X cumple la relación RT con otro término Y, cualquier término Z que también cumpla dicha relación con Y también la debe cumplir con Z. En caso de que no se haya considerado algún caso particular de esta relación, aplicando las siguientes reglas deductivas podemos completar la información del sistema.

RT(x,z) <-- RT(x,y) ^ RT(y,z)

T(y,z) <-- RT(x,z) ^ RT(x,y)

De la aplicación de los dos primeros conjuntos de reglas de inferencia lógica, obtenemos como resultado las siguientes tablas:

TABLA TOP TERM (TT)

SIGNATURA TT	SIGNATURA TB
C30/69	G09.30
C30/69	C28.35
C30/69	F25
C30/69	F52.25
Z	Z.08
Z	Z.08.10
Z	Z.08.20

TABLA NARROWER TERM (NT)

SIGNATURA TA	SIGNATURA Tn
Z	Z.08
Z.08	Z.08.10
Z.08	Z08.30
Z	Z08.10
C30/69	C50/53
C50/53	C42/46
C30/69	C42/46

ii) Otra posibilidad de diseño que nos proporcionan las bases de datos deductivas es la de definir restricciones. Aquí podemos orientarlas hacia la preservación del mantenimiento de la integridad del sistema. Por ejemplo, una restricción a considerar es aquella que hace referencia a que si un término X es NT de un término Y, no pueda darse de alta en el sistema una tupla del tipo Y NT X o Y NT X. Esta restricción la podemos representar del siguiente modo:
para todo { NT(x,y) --> no NT(y,x) ^ no NT(y,x) }

8. CONCLUSIONES

De los ejemplos anteriores deducimos que una vez se haya unificado el soporte relacional de un tesauro, de acuerdo a las necesidades de cada compilador de los mismos, podemos centrar nuestros trabajos en el desarrollo de reglas de inferencia adicionales y suplementarias a las indicadas anteriormente con el objetivo de implementar una base de conocimiento.

REFERENCIA

Material bajado de Internet. Texto completo de la comunicación del mismo título presentada a las IV Jornadas Catalanas, 22-23 de enero de 1992. http://www.um.es/~gtiweb/fjmm/disetesa.htm

BIBLIOGRAFÍA

AITCHISON, J; GILCHRIST, A.Thesaurus Construction. A practical manual. Londres, Aslib, 1987.
AMAT NOGUERA, N. Documentación Científica y Nuevas Tecnologías de la Información. Madrid, Pirámide, 1988.
BORKO, H; BERNIER, C. Indexing Concepts and Methods. Nueva York: Academic Press, 1978.
CHEN, P.P.S. "The Entity-Relationship Model-Towards a Unified View of Data". Nueva York. ACM. Trans. on Database Systems 1. Vol 1. No 1. Marzo 1976. pp 9-36.
DATE C.J.Introducción a los sistemas de bases de datos.México, Adison Wesley Iberoamericana, 1986.
MINKER, J. Foundations of Deductive Databases and Logic Programming. Los Altos, California. Morgan Kaufmann Publishers, Inc.
RODRIGUEZ MUÑOZ, J.V.; MARTINEZ MENDEZ F.J.; DIAZ ORTUÑO, P.M. "Los Modelos de Datos como alternativa en la construcción de Tesauros".En: Actas de

las III Jornadas Nacionales de Documentación Automatizada (DOCUMAT-90), Mallorca, 1990.
VAN SLYPE, G. Les langages d'indexation : conception, construction et utilisation dans les systemes documentaires. París, Les Editions d'Organisation, 1987.

CONSIDERACIONES SOBRE LA INDIZACION EN LAS BIBLIOTECAS UNIVERSITARIAS ESPAÑOLAS

José Elías Jiménez Rodríguez
Universidad del País Vasco (España)

LA INDIZACIÓN EN LAS BIBLIOTECAS ESPAÑOLAS

A diferencia del francés o del inglés, idiomas en los que existen grandes repertorios de materia que sirven de referencia, no disponemos en español de un instrumento equiparable a las "Library of Congress Subject Headings" o las "Vedettes-mattière" de la Universidad Laval. Las empresas encaminadas a la redacción de una lista en lengua castellana a lo largo de esta siglo han sido relativamente numerosas (1) pero modestas constituyendo, con seguridad, la lista de encabezamientos de materia del Ministerio de Cultura la iniciativa más exitosa. Pero tanto esta como otras listas existentes se revelaron pronto claramente insuficientes para las necesidades de las bibliotecas universitarias, sobre todo una vez que se pusieron en marcha los procesos de automatización de los catálogos. Por ello, la publicación en 1992 de las materias de la Biblioteca Universitaria de Sevilla encontró una favorable acogida pues por primera vez los bibliotecarios disponían de un instrumento impreso de suficiente envergadura para acometer la tarea de indización en el nivel que requieren las bibliotecas de centros de enseñanza superior. No es ésta, sin embargo, la única lista concebida en una biblioteca universitaria: tanto el CSIC como la Universidad Autónoma de Madrid habían editado con anterioridad sus propios repertorios y posteriormente también la Universidad Complutense ha publicado una nueva edición de sus encabezamientos de materia. De hecho, todas las universidades trabajan día a día en la actualización de sus respectivos índices, que en ocasiones tienen correspondencia impresa, pero en cualquier caso no puede negarse el eco que ha encontrado la lista sevillana como instrumento de referencia para el establecimiento de las autoridades de materia en el mundo universitario.

Partiendo de esta constatación nuestro propósito es trazar un diagnóstico aproximativo del estado de la indización en las bibliotecas universitarias españolas. La inexistencia de una fuente de autoridad castellana al estilo de las Library of Congress suject Headings (en adelante LCSH) en el mundo anglosajón ha dado lugar a una importante dispersión que sólo de forma limitada ha sido paliada por la difusión de los principales listados españoles (especialmente el de la Universidad de Sevilla, pero también el del CSIC y el de la Universidad Complutense de Madrid). Cada universidad ha entendido a su manera la tarea de indización en función de sus necesidades y aunque se emplean en general fuentes comunes para la redacción de los encabezamientos, tanto en castellano como en otros idiomas, los resultados y las soluciones adoptadas difieren notablemente. En definitiva, y como en otros aspectos España arrastra también en éste las consecuencias de la inexistencia de una tradición en el desarrollo de técnicas de representación del conocimiento. Tenemos la sensación de que la asignación de materias es un asunto secundario en las bibliotecas españolas, y quizá existan razones para ello: en una conocida monografía F.W. Lancaster (2) señala diversos criterios para ponderar el balance coste-eficacia en el uso y

desarrollo de vocabularios controlados, y ciertamente no siempre está justificado el esfuerzo que supone una indización exhaustiva. Pero también creemos que la coordinación del trabajo para lograr un vocabulario más consensuado y comprensivo implicaría a medio plazo un importante ahorro de energía intelectual para los indizadores así como un incremento en el éxito de las búsquedas por materias de nuestros usuarios, cuyos decepcionantes resultados son de momento la razón última de la decadencia de la búsqueda por materias y consiguientemente de la indización en las bibliotecas.

Así pues la indización como proceso técnico (3) adquiere una relevancia especial en el momento en que la automatización de los catálogos permite una recuperación mucho más versátil y efectiva. Es entonces cuando las bibliotecas universitarias se proponen una seria actualización de sus autoridades de materia que van incorporándose a los correspondientes índices consultables en los OPACs. La mencionada falta de un gran repertorio en castellano implica inicialmente la necesidad de recurrir a fuentes de autoridad en idiomas extranjeros, destacando la utilización del repertorio de la Université Laval y los LCSH. La publicación relativamente temprana del repertorio de Sevilla lo convierte también, como ya se ha dicho, en una de las fuentes de referencia básicas, junto, en menor medida, a la lista de encabezamientos de la base de datos CIRBIC del CSIC, los encabezamientos de la Universidad Complutense y las autoridades de la Biblioteca Nacional, los otros tres principales repertorios españoles. Semejante dispersión en el uso de fuentes está provocando una importante disonancia terminológica así como una notoria diversidad en políticas de indización. Sus raíces, alcance y consecuencias son el objeto del presente estudio.

LA SICOLOGÍA EN LAS LISTAS DE MATERIAS: COBERTURA CONCEPTUAL Y PROXIMIDAD LÉXICA

Antes de continuar conviene hacer notar que nuestro objetivo no es valorar la calidad de la indización estudiando su pertinencia. sino juzgar las posibilidades que ofrecen los repertorios (o mejor aún, los índices de materia) como vocabularios controlados, tanto desde la perspectiva del usuario, que precisa de un léxico que responda eficazmente a sus interrogaciones de búsqueda, como desde la perspectiva del catalogador, cuya labor está supeditada a las características del vocabulario con el que indiza los documentos. En ambos sentidos cabe considerar la incidencia que en las distintas listas tienen aspectos como el grado de precoordinación (o, en su caso, de "tesaurización"), la capacidad orientadora de las referencias de equivalencia semántica, la determinación del elemento inicial en los conceptos compuestos, el uso de singulares o plurales y la adecuación de las traducciones de términos tomados de repertorios extranjeros al principio de uso en lengua castellana. No se trata, en cualquier caso, de determinar qué repertorio es mejor sino de, por medio de un estudio comparativo, ofrecer pistas sobre las tendencias de la indización (o con más precisión, la construcción de vocabularios controlados) en el mundo universitario.

Para ello hemos analizado la cobertura ofrecida por distintos repertorios españoles en aquel campo del conocimiento en el que con mayor énfasis hemos trabajado durante estos últimos años, la sicología. Somos conscientes de que limitar el estudio a un área concreta no permite establecer conclusiones definitivas pues la situación puede variar en otros campos científicos aunque bien es cierto que las tendencias detectables aquí poco diferirán si analizamos otra disciplina. Por tanto,

sin pretender que nuestras conclusiones tengan valor terminante, creemos que pueden dar noticia del estado actual de la indización. Este análisis se ha desarrollado en dos fases: en una primera, se compararon exhaustivamente los repertorios de las universidades del País Vasco (UPV), Sevilla (USE) y Complutense de Madrid (UCM), la primera por tratarse de aquélla en la que venimos trabajando, y las otras dos por ser responsables de dos de los repertorios españoles más empleados por otras universidades. Para ello se partió como muestra de una selección de 1.045 conceptos tomados del tesauro ISOC de sicología, estudiándose la cobertura conceptual que ofrecían los tres repertorios así como su proximidad terminológica. En una segunda fase, del total se extrajeron 46 conceptos considerados como especialmente significativos y útiles para un estudio comparativo, analizándose su representación, además de en los tres repertorios mencionados, en los de las siguientes bibliotecas universitarias: Granada (UGR), Zaragoza (UZA), Autónoma de Madrid (UAM) y CSIC (que en rigor no es una biblioteca universitaria, aunque su catálogo se asemeja notablemente al de éstas). Debe señalarse que las consultas han sido realizadas a través de los índices de materias de los respectivos OPACs, cotejándose con la versión impresa siempre que esto ha sido posible. Pensamos que esta vía ofrece una perspectiva más parecida a la del usuario y también a la del catalogador, así como un nivel de actualización mucho más alto, aunque siempre han de analizarse con suma cautela los índices en línea, pues presentan mayor número de errores tipográficos y entradas falsas que sus versiones impresas (4).

a) La cobertura conceptual

Para la comparación de la representación de conceptos de sicología en los índices de materias de las universidades del País Vasco, Sevilla y Complutense se seleccionaron 1.045 descriptores del tesauro ISOC de sicología, revisándose la constancia de los conceptos representados por los descriptores en los tres índices de materias. Se tuvo en cuenta no sólo la existencia del término usado como descriptor sino también la posibilidad de que el concepto apareciera en los repertorios representado por un no-descriptor semánticamente equivalente En definitiva se trataba de determinar la cobertura conceptual para después examinar el grado de proximidad terminológica que mantienen entre si los tres repertorios (en adelante LEM).

En la selección de los descriptores de la muestra se optó por tomar, de los 2.788 existentes en el tesauro, aquellos que pertenecían al campo de la sicología o áreas estrictamente afines (psiquiatría, etología) descartándose los que en rigor deben considerarse terminología de otras disciplinas (sociología, pedagogía, lingüística, trabajo social, etc.). Finalmente, como se ha dicho, fueron seleccionados 1.045 conceptos que permitieron inferir los siguientes resultados:

COBERTURA CONCEPTUAL

	Número de conceptos	Porcentaje de cobertura conceptual
UPV	598	57,22
USE	534	51,1
UCM	547	52,3

Como puede observarse, la cobertura es similar en las tres LEM, rondando el 50% y sin perder de vista que habría sido sensiblemente superior de considerar el conjunto de descriptores del tesauro ISOC, que incluye numerosos términos de rango más genérico pertenecientes a otras disciplinas. Además debe tenerse en cuenta la existencia en los repertorios de numerosos encabezamientos sin equivalente en el tesauro, especialmente aquellos consistentes en una frase de dos o más sustantivos vinculados mediante algún nexo gramatical, que en un lenguaje poscoordinado como es el tesauro encuentran mejor traducción en alguna combinación de términos (por ejemplo, "ansiedad de separación en el niño", "conducta, trastornos de la, en el niño). En otros casos en los que el tesauro si presenta descriptores sintagmáticos ("miedo a hablar") puede constatarse la existencia en las LEM de encabezamientos similares sin correlato ("miedo a la muerte") No obstante, la abundancia de términos específicos es obviamente mayor en el tesauro, siendo relativamente frecuente que un término muy específico que en el tesauro es descriptor, en las listas de materia sea un término rechazado que remite a otro más genérico usado en su lugar (por ejemplo, "distimia" remite a "trastornos afectivos" en la LEM de la UPV). En este caso hemos entendido que la cobertura conceptual estaba garantizada, si bien condicionando una indización menos precisa.

No conocemos ningún estudio semejante que permita comparar el grado de especificidad de nuestros repertorios, pero debe considerarse que los índices de materias crecen día a día y por tanto aumenta su comprensividad. En cualquier caso da la impresión de que, al menos en el campo de la sicología, la cobertura es bastante satisfactoria para las expectativas de búsqueda de los usuarios y quizá algo menor para las necesidades de una indización que se quisiera precisa, consistente, exhaustiva y específica, principales factores que según, Aluri, Kemp y Boll afectan a la calidad de la indización (5) . Pero no es lo mismo indizar artículos de revista que libros, y sin duda las búsquedas que, por la especificidad del término, no son satisfechas en el índice de materias, suelen encontrar respuesta en los índices de títulos cuando el usuario mantiene la precaución de formular la interrogación en los idiomas en los que previsiblemente existan títulos que contengan el término de búsqueda (por ejemplo, ocurre con el término "alexia" en los ficheros de la UCM).

b) Proximidad terminológica

Por otra parte, Se analizó la proximidad terminológica que, para la representación de los conceptos, mantienen entre sí los tres repertorios tomados en consideración. Para ello se tuvieron en cuenta, obviamente, aquellos conceptos comunes que compartían una y otra LEM en cada comparación emparejada, deduciéndose en cada caso el porcentaje de conceptos representados por el mismo término del total de los compartidos.

PROXIMIDAD TERMINOLÓGICA

	Conceptos comunes	Términos idénticos	Proximidad terminológica (%)
UPV/USE	496	438	88,3
UPV/UCM	457	291	64,98
USE/UCM	434	287	66,12

Nuestro examen revela una acusada afinidad entre los repertorios de Sevilla y de la Universidad del País Vasco. Es lógico si tenemos en cuenta que la UPV tiene en el repertorio de Sevilla su primera fuente de autoridad en castellano; pero además, ambas universidades utilizan la lista de la Universidad Laval como fuente principal, con las consecuencias de orden léxico que ello comporta. La Universidad Complutense, por su parte, emplea como fuente básica los LCSH, siendo la lista del CSIC su primer referente en lengua castellana. Esta discrepancia no implica diferencias esenciales en la filosofía global de la indización, en aspectos como la precoordinación, el uso de especificadores u otras cuestiones que ya se tratarán más abajo. Sin embargo sí da lugar a diferencias léxicas importantes (los porcentajes lo expresan con claridad) de las que podemos dar algunos ejemplos indicativos: "animales-sicología" (UPV/USE) vs. "sicología animal" (UCM); "relaciones humanas" (UPV/USE) vs. "relaciones interpersonales" (UCM); "sicología del trabajo" (UPV/USE) vs. "sicología industrial" (UCM). Los ejemplos son numerosos y aunque no siempre, con frecuencia puede observarse que la adopción de un término u otro está directamente relacionada con la fuente extranjera que se emplea (Laval en el caso de la UPV y USE, LCSH en el de la UCM). Volveremos sobre esta cuestión (y sobre el asunto de la proximidad léxica que mantienen los repertorios) al tratar los resultados del examen global de las siete LEM que hemos analizado en nuestro estudio.

POLÍTICAS DE INDIZACIÓN EN SICOLOGÍA: LA CREACIÓN DE VOCABULARIOS CONTROLADOS

Decíamos más arriba que en una segunda fase se seleccionaron 46 conceptos considerados especialmente significativos (es decir, presentes en los tres repertorios iniciales y susceptibles en su mayoría de ser expresados de formas diversas) para estudiar su representación en otras cuatro LEM y así obtener una panorámica más amplia de las tendencias del control del vocabulario psicológico en las bibliotecas universitarias. La reducción de la muestra, de 1.045 a 46 conceptos, hace mucho menos fiable el estudio de la proximidad terminológica entre los distintos repertorios aunque se mantienen, de forma aproximada, las proporciones que habíamos señalado para los repertorios de UPV, USE y UCM. Atribuir la mayor-menor afinidad a la coincidencia de fuentes de autoridad no es asunto fácil: las siete universidades estudiadas utilizan distintas fuentes que a veces coinciden. La biblioteca de la Universidad de Granada (UGR) emplea, a modo de descriptores, los encabezamientos de materia de la Organización de los Estados Americanos, y su influjo léxico, netamente hispanoamericano, se deja notar en términos como "psicoterapia marital" (psicoterapia o terapia de pareja o conyugal en todos los demás repertorios) o "pruebas psicológicas" (traducción infrecuente del anglicismo tests). La biblioteca de la Universidad de Zaragoza (UZA) confía básicamente en la lista de

la red de bibliotecas del CSIC y, de forma secundaria, en los LCSH. La Universidad Autónoma de Madrid (UAM), que en 1990 publicó su propio repertorio, utiliza el de la Universidad Laval así como los LCSH y los encabezamientos del CSIC. En cuanto a la red de bibliotecas del CSIC, que en 1995 publicó una nueva edición en papel de su lista de autoridades de materia, emplea como fuente principal los LCSH, y secundariamente las materias de Laval así como otros listados españoles y extranjeros.

El panorama es pues diverso y consiguientemente da lugar a una significativa **disparidad léxica**. Sólo seis de los 46 conceptos presentan idéntica representación en los siete repertorios y se trata en casi todos los casos de conceptos simples sin sinónimo usual en castellano ("sícolingüística", "etnosicología") o conceptos compuestos cuya formulación habitual goza de gran tradición ("sicología evolutiva", "sicología social"). Otros dos términos también eran idénticos, aunque no estaban presentes en todos los repertorios. Lo normal es que en todos los términos existan diferentes fórmulas, tratándose a veces de pequeñas variaciones, como el uso o no de un cualificador ("estrés" o "estrés (sicología)") o de un artículo ("psicoterapia de pareja" o "psicoterapia de la pareja") y en otras ocasiones de diferencias importantes, como la presentación del encabezamiento con un subencabezamiento precoordinado ("lenguaje-trastornos") o de forma sintagmática invertida ("lenguaje, trastornos del"). Sin que deban tomarse en gran consideración las cifras, dada la escasa significación estadística de la muestra, señalaremos que los índices de equivalencia terminológica han rondado en casi todos los casos el 50-60 %, alcanzando valores más altos en la correspondencia UPV/USE (80 %), UAM/CSIC (69,4 %) y UZA/CSIC (66,6 %). Quizás pueda destacarse que tanto UAM como UZA cuentan con el repertorio del CSIC como fuente de referencia, mientras que la UPV, como ya se ha señalado, emplea el de la Universidad de Sevilla.

No hemos observado ningún rasgo de los conceptos que incide en la disparidad terminológica: hay conceptos compuestos que coinciden en todos los repertorios y conceptos simples expresados de formas diversas. Quizá pueda apuntarse una cierta tendencia a la uniformidad en conceptos muy específicos sin sinónimos usuales ("discalculia", "afasia") o como se ha señalado, en conceptos más genéricos con gran tradición de uso ("sicología evolutiva", "psicosis") y conceptos de uso coloquial ("memoria", "atención"). Por lo demás las disonancias léxicas no parecen atribuibles a otra razón más que a la fuente que en su momento se adoptó para cada concepto en cada universidad. En unos repertorios encontramos términos claramente derivados de tos LCSH cuyos equivalentes en otras LEM parecen adaptados de la lista de Laval: "sicología industrial", adaptación del término "psychology, industrial" frente a "sicología del trabajo", traducción de "psychologie du travail", equivalente en Laval del término americano.

La traducción de términos de LEM en lenguas extranjeras implica un esfuerzo de adaptación que no siempre se ha invertido. Así, existen términos como "sicología correccional", adaptación literal del inglés o del francés ("correctional psyclogy"/"psychologie correctionnelle") cuyo equivalente usual en castellano sería "sicología penitenciaria". No obstante, en general se respeta el principio de

adaptación al uso del término en castellano siendo extrañas las traducciones improcedentes.
Junto a la falta de uniformidad léxica los repertorios españoles presentan otros rasgos, en muchos casos heredados de aquellos que les sirven de fuente de referencia, merecedores de un examen detenido. En primer lugar cabe destacar un cierto descuido de **las referencias de equivalencia semántica**. Sabido es que no siempre se cumple el principio de uso en la elección del término aceptado como encabezamiento, pero este problema puede soslayarse con un aparato de referencias conveniente, que se adecue al lenguaje que previsiblemente emplearán los usuarios. Entre los conceptos por nosotros considerados hemos observado la inexistencia en prácticamente todos los repertorios de la expresión "evaluación psicológica", cuyo uso es muy frecuente entre la comunidad universitaria (basta con acercarse a cualquier índice de títulos para comprobarlo). Pues bien, si un usuario desea localizar en nuestros OPACs documentos sobre tal materia no encontrará en ningún índice (salvo en el de UZA) referencia alguna que le remita al término que en su lugar se emplee para expresar el concepto. Esta situación afectará también a la calidad de la indización, por cuanto sólo un catalogador mínimamente familiarizado con esta carencia podrá asignar con consistencia un encabezamiento sustitutivo. El problema de le falta de referencias de equivalencia es especialmente notorio en los encabezamientos con subdivisión. Una entrada del tipo "lenguaje-trastornos" puede ser difícilmente localizable para muchos usuarios insuficientemente avezados en la navegación por los OPACs. Entre los encabezamientos de la muestra se han aplicado criterios muy diferentes, incluso dentro de una misma LEM, de forma que en ocasiones encontramos la fórmula en lenguaje usual remitiendo a la fórmula adoptada con subencabezamiento ("trastornos de la personalidad", véase "personalidad-trastomos") mientras en otras ocasiones no ocurre así ("lenguaje-trastomos", que al igual que el encabezamiento anterior está en la LEM de UCM, carece de referencias de equivalencia).
Otro aspecto observado es el desigual **uso de la precoordinación**. El desarrollo de los catálogos automatizados ha agudizado la crisis de los encabezamientos de materia tradicionales, dando pie al debate sobre su ineficacia y obsolescencia y dejando así las puertas abiertas a una progresiva "tesaurización" de las LEM. Entre los repertorios estudiados sólo uno, el de la Universidad de Granada, emplea descriptores en sentido estricto. Los demás continúan apegados a la tradición de las materias precoordinadas que tanta contestación ha encontrado últimamente. Lo cierto es que por lo que respecta a la sicología, los efectos negativos de la precoordinación se dejan notar bastante menos que en áreas como la historia o el derecho, ya que las subdivisiones cronológicas y geográficas tienen poco uso. Así que el desarrollo de un vocabulario con un grado considerable de especificidad hace que en muchos aspectos las LEM parezcan tesauros si bien aún persiste, especialmente en lo que se refiere a la expresión de ciertas materias, la tendencia a usar el lenguaje precoordinado: por ejemplo, casi todos los repertorios emplean la subdivisión "trastornos" a continuación de un encabezamiento principal y cuando no es así se formula la expresión invirtiendo el orden de las palabras ("personalidad, trastornos de la", tal y como aparece en UPV y USE).
Al hilo de este último ejemplo señalaremos que la **inversión del orden natural de las palabras** es un claro residuo de le catalogación manual que, especialmente en

los sistemas que permiten búsquedas permutadas en los índices de materias (como es el caso de DOBIS-LIBIS), carece de sentido, pues aunque ciertamente posibilite insertar el encabezamiento junto a aquellos que tienen la misma raíz que el término focal, raro será que el usuario acceda directamente por le forma autorizada beneficiándose así de la supuesta ventaja. De hecho, ciñéndonos e la LEM de la UPV, en algunos casos se ejecuta la inversión ("lenguaje, trastornos del") mientras que en otros se respeta el orden natural ("sistema nervioso"). En general, y considerando el conjunto de los repertorios, no hay una tendencia fija: sí suelen modificarse los términos que en su expresión coloquial comienzan por palabras del tipo "sistema", "teoría", pero en otros términos difieren las soluciones adoptadas por las distintas LEM ("toma de decisiones" en el CSIC, "decisiones, toma de" en UZA o UAM). Es llamativo el caso de los conceptos cuya expresión usual comienza por la palabra "trastornos", que tanto en la UPV, como en USE dan lugar a encabezamientos formulados en orden Invertido ("personalidad, trastornos de la", por citar otro ejemplo) mientras el resto de las universidades emplean el encabezamiento con subdivisión ("personalidad-trastornos") excepto la Universidad de Granada, que como se ha dicho usa descriptores ("trastornos de la personalidad").

La **estructura de las LEM** se va aproximando **a** la de los tesauros (6), especialmente mediante el desarrollo de relaciones semánticas entre las unidades léxicas. No nos hemos detenido en este aspecto, pues nuestro interés era estrictamente las fórmulas de representación de los conceptos así como las semejanzas y diferencias léxicas entre las distintas LEM, pero no la estructura interna de éstas. No obstante opinamos que en general se presta menos atención de la que merece al desarrollo de relaciones semánticas, especialmente a las relaciones jerárquicas y asociativas, aunque también, como ya se ha dicho, a las de equivalencia (y esto tiene consecuencias más graves).

Por último, entre otras cuestiones, puede señalarse cierta indecisión en el uso de plurales o singulares ("actitud"/"actitudes", "arquetipo"/"arquetipos") así como el empleo abundante de cualificadores, que en sicología se hacen especialmente necesarios al ser frecuentes los términos que, con diferente significado, provienen de otras disciplinas: "cooperación (sicología)", "autonomía (sicología)". Hemos observado que en estos casos el grado de proximidad terminológica suele ser muy alto, coincidiendo casi todos los repertorios.

BALANCE Y PERSPECTIVAS

Recapitulando, puede afirmarse que los repertorios españoles vienen siendo redactados con las servidumbres que implica depender de versiones previas en otros idiomas. No queremos dar a entender que la indización en las universidades españolas sea de baja calidad. Al contrario, la cobertura conceptual, ya se he señalado, es satisfactoria y la elección de términos, salvo excepciones, es correcta y permite con más o menos rodeos llegar al usuario hasta el concepto que busca. Sin embargo creemos que la disparidad de los vocabularios controlados que mantienen las universidades es un síntoma de la escasa atención que se presta a la catalogación por materias. De la misma forma que se han establecido redes cooperativas con ánimo de trabajar coordinadamente en aspectos como el préstamo interbibliotecario o la catalogación, somos partidarios de un esfuerzo colectivo para el establecimiento de un gran vocabulario controlado en castellano (7).

Uno de los tópicos de la biblioteconomía actual es la decadencia de la búsqueda y, consecuentemente, de la indización o catalogación por materias. Se ha destacado que casi la mitad de las búsquedas por materias producen resultados nulos, mientras buena parte del resto ofrecen un número excesivo de registros recuperados (8). Es paradójico que, conforme la disponibilidad de recursos informativos va creciendo de une manera exponencial, los resultados obtenidos por los usuarios en sus búsquedas siguen siendo alarmantemente bajos En buena medida la razón de esta situación reside en lo que en inglés ha sido referido como "user's information iliteracy", es decir la incompetencia de los usuarios para la recuperación de información, especialmente cuando el criterio de la búsqueda es temático (9). Pero junto a ello, y especialmente en los OPACs. Las imperfecciones de los vocabularios controlados juegan un importante papel en el fracaso de las búsquedas por materia. Ciertamente, los usuarios están por norma poco familiarizados con el uso de los encabezamientos de materia, pero el lenguaje de éstos es con frecuencia poco asequible e inadecuado para un uso fructífero En las páginas anteriores se ha dado más de un ejemplo en el que la recuperación documental era verdaderamente dificultosa con los encabezamientos disponibles en nuestras LEM. Pero el gran interrogante es si realmente merece la pena el esfuerzo que requeriría una indización de mayor calidad. La cuestión tiene aspecto de círculo vicioso, pues ¿qué es anterior, la insatisfacción de los usuarios o la inadecuación de los lenguajes de indización? Si mantenemos que no es rentable invertir en indización porque la consulta por materias no seduce a los clientes, también puede sostenerse que éstos no usan los índices de materias porque les ofrecen resultados pobres.

El estudio comparado de los repertorios españoles nos ha permitido constatar, como hemos ido señalando, diferencias en la elección de términos, en el uso de subdivisiones, en las formulaciones sintácticas y en otros aspectos que van determinando listas de materias con vocabularios distantes y deudoras de fuentes muy diversas. ¿No seria deseable una mayor uniformidad? En un momento en el que comienza e clamarse por la necesidad de vocabularios controlados para la gestión de los inmensos e incontrolados recursos disponibles en Internet (10) sorprende que no dispongamos aun en castellano (que no lo olvidemos, es un idioma con cientos de millones de hablantes) de un instrumento léxico estandarizado para la indización de los, en otros sentidos, tan bien controlados recursos de las bibliotecas. Dando por sentado que la lista del Ministerio de Cultura es insuficiente, no cabe pensar sino en las bibliotecas universitarias y en la Biblioteca Nacional como únicas instituciones capaces de canalizar el esfuerzo necesario para su creación. Es mucho el trabajo realizado hasta el momento como para no sacarle provecho: piénsese que las bibliotecas universitarias disponen de LEM considerablemente comprensivas y estructuradas.

Evidentemente, una mayor uniformidad de los repertorios no garantizaría una catalogación más uniforme, en la medida que la asignación de encabezamientos de materia es responsabilidad de distintos indizadores que en cada biblioteca pueden usar encabezamientos diferentes en función de sus necesidades. Pero mucho más allá de la uniformización de la catalogación, incluso de la unidad terminológica que podría beneficiar a las redes cooperativas (en asuntos como la captación de registros o la consulta de los catálogos colectivos) la estandarización del vocabulario facilitaría la labor de los indizadores y mejoraría los resultados de las búsquedas. Tenemos la convicción de que la acumulación de las experiencias

de los distintos actores susceptibles de intervenir en semejante empresa provocaría efectos sinérgicos mejorando claramente la calidad de nuestros índices de materias.

En fin, no compartimos la idea de que la búsqueda por materias haya entrado en una crisis irresoluble. El desarrollo que están adquiriendo los recursos informativos disponibles en Internet hacen cada día más necesario algún tipo de instrumento de control temático, más allá de la recuperación mediante indización automática que permiten los buscadores. Por lo que respecta a las bibliotecas la búsqueda por materias va a seguir concitando el interés de los usuarios siempre que sepamos dar respuesta adecuada a sus necesidades. Ello hará necesario seguir disponiendo de instrumentos eficaces de control del vocabulario, con forma bien de tesauro o bien de lista de encabezamientos de materia adaptada a las características de los nuevos sistemas de Información.

REFERENCIAS

Artículo bajado de Internet.

(1) BENEDITO, Pilar: "Clasificación e indización en las bibliotecas españolas", *Boletín de la ANABAD,* XLIV (1994), $n^0$1, PP. 69-8O

(2) LANCASTER. Fredwick W.: *El control del vocabulario en la recuperación de información,* Valencia: Universidad, 1995, p.251y ss.

(3) Preferimos hablar de indización antes que de catalogación por materias, entendiendo que aquél término es más específico que éste. Véase ALURI, Rao,

(4) KEMP; D. Alasdair; BOLL, John J.: *Subject analysis in online catalogs,* Englewood, Colorado: Libraries Unlimited, 1991, p. 55

Aquí se hablará indistintamente de "índices" o "repertorios", si bien en rigor éste es una reproducción estructurada de aquél

(5) ALURI, Rao; KEMP, D. Alasdair; BOLL, John J., op. cit., pp. 56 y ss.

(6) IZQUIERDO, José Mª; MORENO, Luis Miguel: "Listas de encabezamientos de materia y thesauri en perspectiva comparada", *Documentación de las Ciencias de la Información,* 17(1994), pp. 287-310

(7) RODRÍGUEZ RICARD, Teresita; FRANQUI, Delsi Trejo: "Estudio comparativo de tres listas de encabezamientos de materia en español", *Revista Española de Documentación Científica,* 12, 4 (1989), pp. 422-440

(8) TAYLOR, Arlene G.; "On the subject of subjects", *Joumal of Academic Librarianship,* 21, 6 (1995), pp. 484-491

(9) ESPELT, Constança: "Improving subject retrieval: user-friendly interfaces and effectiveness", *BID: Biblioteconomia i documentació,* 1(1998): http://www.ub.es/div5/biblio/bid/espel98.htm#inici

(10) TAYLOR, Arlene G., op. cit., p.486

PROCESO DOCUMENTAL, DEL ANÁLISIS A LA RECUPERACIÓN: INDIZACIÓN, RESUMEN Y LENGUAJES DOCUMENTALES

Juan Marcos
Universidad Complutense de Madrid (España)

LA INDIZACIÓN

Introducción

La indización consiste en extraer los conceptos representativos del contenido de un documento con la ayuda de un lenguaje documental o lenguaje controlado. Se pueden emplear materias, palabras claves o descriptores.
La indización no se limita sólo a detectar los vocablos presentes en el documento, sino también su traducción e interpretación para pasar del lenguaje natural al lenguaje documental.
El indizador, cuando ya tiene el documento original o su expresión condensada, retiene unas cuantas nociones que representan su contenido con la máxima fidelidad.
Una correcta indización no lleva a descubrir lo que dice el documento indizado, pero sí desvela sobre qué trata y esto en el mundo de la documentación es suficiente muchas veces.
Los teóricos de la documentación consideran que la indización se realiza, casi siempre, después de la condensación, ya que si el resumen es correcto y lo ha realizado un experto, resulta más fácil y más operativa la operación de indizar y ésta la puede llevar a cabo un documentalista sin necesidad de ser un especialista en el tema.

Concepto

Coll Vinent y Bernal Cruz: "aquella operación documental que consiste en extraer de un documento original o de su resumen unos vocablos especialmente expresivos y con enorme carga informativa -palabras claves-, muy indicativos del contenido esencial del documento indizado. Es el acto de retener una o más nociones que representan el contenido de un documento y adecuarlas al lenguaje natural o documental previamente escogido por el analista".
Unesco: "describir y caracterizar un documento con la ayuda de representaciones de los conceptos contenidos en dicho documento".
Guinchat y Menou: "la operación que permite elegir los términos más apropiados para representar el contenido de un documento".
Consideran que es la actividad más importante de todo el proceso documental.
Nuria Amat: "consiste en retener una o más nociones que representan el contenido de un documento o los conceptos de una búsqueda bibliográfica. Para ello asignaremos términos a un documento con el objeto de representarlo temáticamente y para facilitar la formulación de búsquedas bibliográficas".
Van Slype diferencia entre clasificación e indización.
- Considera la Clasificación como la expresión más general del contenido (destaca el tema general) y se lleva a cabo a través de lenguajes de estructura jerárquica.
- La Indización analiza el documento y los conceptos que lo constituyen, asigna descriptores, generalmente extraídos de un Tesauro, para describir el contenido conceptual del documento.

Indización manual y automática

1. Humana, inteligente, manual

Exige un gran esfuerzo de síntesis para extraer las palabras claves más significativas del contenido de un texto. Se trataría de aislar los conceptos más representativos, aquellos que pudieran interesar en el tiempo y en el espacio al usuario.

2. Automática

Trabaja con unitérminos, de modo que va leyendo una a una todas las palabras que componen el texto, a excepción de los contenidos en un tesauro negativo de palabras vacías. El ordenador las rechaza. Se conocen con el nombre de Stop word list.

Ventajas e inconvenientes

1. La humana parece más eficaz en cuanto a captación y desestructuración de contenidos, pero presenta como inconveniente la velocidad, ya que es más lenta.
2. La automática: Es más o menos eficaz en función de los programas. No lee el texto de una forma convencional, sino que lo hace de forma secuencial.

Metodología para efectuar la indización

No siempre se aplica de la misma forma la metodología. Depende, la mayor parte de las veces, de los criterios que se han establecido en el propio centro de documentación. En todo caso, es conveniente seguir unas normas generales.

Consideraciones previas a la indización (sirven también para el resumen)

1. Conocer a fondo el documento original o el resumen del mismo en su caso.
2. Penetrar en la idea central de dicho documento.

Identificar las palabras claves que contienen más información y que ofrecen datos sobre la idea central.

3. Considerar la pertinencia de las palabras claves seleccionadas en orden a la búsqueda ulterior del documento indizado.
4. Escoger los descriptores del tesauro que más se adecuen en significación a las palabras claves.
5. Reunir un número suficiente de palabras claves para que con ellas se logre el contenido temático del documento original que se quiere indizar.

Criterios a tener en cuenta

1. Exhaustividad: Todos los conceptos básicos, nombres propios importantes, lugares geográficos significativos, han de estar representados en la indización.
2. Concreción: Han de evitarse palabras demasiado genéricas, expresiones vagas o ambiguas o cualquier tipo de generalización.
3. Pertinencia: Si no puede retenerse todo, ni lo más importante, hay que ser exigentes en la selección de los vocablos más expresivos y mas significativos.
4. Uniformidad: Es el más difícil y el que exige una actitud mas positiva por parte de quien indiza, sobre todo cuando no existen palabras similares a la escogida por el documentalista.
5. De autor: El que realiza la indización ha de tener sus criterios, pero debe aceptar primero los que impone el centro para el que trabaja. No tiene porque coincidir ni con el del documentalista, ni con el del usuario que solicita la información.

6. Interés del usuario: Esta será siempre la principal misión de la indización si se quiere ofrecer un servicio útil.
7. Estadístico: Frecuencia del uso de un término en el texto, es decir, el número de veces que aparece.
8. Especificidad: Se utiliza en el lenguaje controlado, y expresa la exactitud con la cual unos determinados términos representan a ideas concretas.
9. Precisión: Aquello que mide la habilidad o la aptitud de un sistema de información para encontrar, en respuesta a una pregunta formulada correctamente, unos resultados precisos.

Etapas de la indización (sirven también para el resumen)
1. Familiarización: Llevará ventaja a la ahora de indizar aquella persona que tenga conocimientos del documento que va a analizar, aunque sólo conozca sus líneas generales. A un indizador hay que pedirle, como mínimo, que conozca las grandes líneas del documento sobre el que va a trabajar.
2. Determinación de su tema principal: Una vez que conoce las líneas generales del documento, está ya en condiciones de realizar el análisis en virtud del cual extrae los términos más significativos del contenido. Las grandes agencias y empresas de servicios que se dedican a indizar dan órdenes muy precisas en este sentido, aunque luego la intuición del indizador es casi más importante que estas medidas.
3. Verificación de la pertinencia de los términos elegidos: Para evitar luego tener que desechar algunos términos que venían encuadrados en otros conceptos.
Tener presente el tipo de usuario que se interesa por el documento: Siempre se ha de trabajar pensando en el servicio que vamos a ofrecer y en las necesidades del usuario.
4. Traducción o conversión del lenguaje natural a los términos correspondientes del lenguaje documental: Esta tarea puede resultar fácil si se dispone de un tesauro; en caso contrario, se debe realizar la misma operación, pero valiéndose de palabras claves decididas por el mismo indizador.
5. Presentación de descriptores: Los descriptores seleccionados en la operación de indización podemos presentarlos con enlaces sintácticos entre ellos. Los más habituales son:
- Yuxtaposición: se separan los descriptores con un signo de puntuación llamado separador: ; /.
- Ponderación: al indizar se distinguen los descriptores principales de los secundarios de manera que representen el contenido esencial y el complementario.

Sistemas de indización
Un sistema de indización: es el conjunto de procedimientos prescritos para organizar los contenidos de los registros de información, a fin de conseguir su recuperación y difusión.
1. Encabezamientos de materia: Fueron las bibliotecas, a través de clasificaciones enciclopédicas y facetadas, quienes emplearon el concepto de indización por materias. Esta indización consiste en la correlación sucesiva de diferentes encabezamientos que expresan el tema o temas de un documento.
2. Unitérminos o palabras claves: Se corresponde con el lenguaje natural o libre. Se presenta en fichas de tamaño normalizado, en las que se sitúan los números de

registros de los documentos que contienen cada unitérmino. Fue Mortimer Taube quien expuso este método por primera vez en 1955. Cuando los documentos entran en el centro de documentación se les otorga un número y ese número se anota en las fichas correspondientes de los unitérminos que reciben el contenido de ese documento. Hoy en día apenas se utiliza en los centros de documentación.
3. Descriptores: Simples listas de ellos, o tesauros con relación semántica entre sus términos. Fue Calvin Mooers quien utilizó por primera vez el término descriptor. Este lenguaje surgió en el *Defence Documentation Center* (DDC) de EE.UU.

Descriptor
Es un término normalizado o controlado que expresa el contenido significativo del documento.

Tipos de descriptor
- En función de su composición
1. Simples o unitérminos: formados por un solo término.
2. Compuestos o sintagmáticos: aquellos que para representar un concepto requieren de varias palabras. Se utilizan para evitar la ambigüedad.
Los descriptores sintagmáticos se construyen de dos formas:
- Unión morfológica (sintáctica): donde dos o más términos se unen utilizando preposiciones o artículos. Por ej.: permiso de caza. Es habitual en los tesauros. Cuando llevamos a cabo esta fusión de conceptos es porque vamos a unir uno o más términos de tal manera que la estructura significante no varíe sustancialmente.
- Unión semántica (lexicológica): según la cual dos o más términos se funden en uno por significación recíproca de contenido. Se lleva a cabo por la fusión de conceptos, en la que cambia la forma parcial o totalmente. El resultado será parcial o irreconocible para los descriptores de los que se parte. Ej.: Información (unitérmino general)+distancia (unitérmino vago)= teleinformación. Otro ejemplo: Telecomunicación+informática=telemática.
- En función de su carga informativa
1. Primarios: se trata de un unitérmino o conjunto de términos que representan un concepto de manera unívoca. Son significativos y relevantes. Ej. tierra, fuego, agua, etc.
2. Secundarios: necesitan ir acompañados de otros descriptores para darle apoyo a otros términos. Tienen una gran fuerza dentro del lenguaje.
3. Infraconceptos: elementos exentos de significación pero que acompañados de un descriptor primario o secundario pueden modificarlo. Apenas se utilizan. Son los prefijos, sufijos: extra, super, mini, etc.
- En función de su cobertura
1. Onomásticos: llamados también personales o corporativos, representan un nombre de persona o institución. Son descriptores fáciles de encontrar en el texto. Ej.: Juan Carlos I, Rey.
2. Geográficos o territoriales: abarcan todo tipo de conceptos vinculados con lugares, países, mares, montañas, etc.
3. Temáticos o de materias: son los más importantes y más difíciles de buscar. Pueden representar cualquier contenido de una disciplina.

4. Cronológicos o temporales: representan fechas, períodos, pero por su denominación. Ej. Renacimiento, Edad Media.

¿Cómo hacer más eficaz la indización?

Han de considerarse estos principios:

1. Sinónimos: Aquellos vocablos de parecido significado, pero cuya fusión o combinación puede perjudicar la búsqueda que se realice a través de ellos. En un lenguaje técnico puede no ser válido lo que sí lo es en el lenguaje natural.
2. Expresiones compuestas: El significado de una expresión compuesta puede variar según sea el orden en que se colocan los distintos vocablos que la componen. Ej. No es lo mismo a la hora de indizar decir convenio franco-español, que convenio hispano-francés. Esta es una cuestión sintáctica que al igual que las lingüísticas y de orden semántico, introducen elementos decisivos para una indización acertada o desacertada.
3. Homógrafros: Son aquellas palabras que se deletrean del mismo modo, pero cuyo significado es distinto. Suenan fonéticamente igual, pero no tienen nada en común.
4. Más de un vocablo para que tengan sentido: En este caso, la indización ha de ser realizada de modo que cada uno de los vocablos encabece una palabra distinta.

Normas básicas de presentación de descriptores

1. Han de tener forma sustantiva si se trata de unitérminos.
El descriptor sintagmático al menos será en la base un sustantivo.
2. Unificar género y número según la mayoría de términos utilizados. UNESCO, como norma, recomienda que se utilice el masculino singular.
3. Utilizar la forma desarrollada en general. Son acrónimos: radar, láser, etc. Se deben utilizar de esta forma si así son más conocidos. También si se conoce en cualquier idioma se puede usar en la forma abreviada. Ej. UNESCO.
4. Utilizar la secuencia lineal normal. Ej.: Ley de Mendel y no Mendel, ley de.
5. Seleccionar el más comúnmente aceptado o más extendido entre la comunidad científica, o entre los usuarios de la base de datos. Hay que utilizar siempre el término más sencillo, más habitual, el que mejor conoce la gente.

EL RESUMEN

Introducción

El volumen creciente de documentos académicos, científicos, técnicos y otros documentos informativos e instructivos hace que sea cada vez más importante, tanto para los lectores del documento primario como para los usuarios de los servicios secundarios, que el contenido básico del documento sea identificado de la manera más rápida posible.

Esta identificación rápida se facilita si el autor del documento primario (ayudado por los editores) lo encabeza con un título significativo y un resumen elaborado. Las directrices básicas sirven para la preparación de resúmenes tanto por los autores como por otras personas y se incluyen reglas específicas para su presentación en publicaciones y servicios secundarios.

Concepto

Resumir es una operación que permite disminuir considerablemente el volumen de la información primaria y destacar los aspectos que tienen especial interés para el usuario.

Un resumen documental debe ser la representación condensada del contenido de un documento.

En esta operación se usa el lenguaje natural, si bien éste sufrirá modificaciones.

Un resumen no es una simple enumeración de ideas esenciales, realmente se trata de la representación o reconstrucción de ese texto condensado.

Para ello, es preciso conocer el texto en profundidad (leerlo todo, al completo) y luego alejarse del texto a la hora de resumirlo (no tenerlo delante)

El resumen es una representación sintética del contenido de un documento. Pero, hay que distinguir entre:

- Una anotación es un comentario o explicación breve acerca de un documento o de su contenido; o también una descripción muy breve del contenido, a menudo añadida como una nota a continuación de la referencia bibliográfica del documento.
- Un extracto es una o más partes del documento seleccionadas para representar el todo.
- Un resumen de conclusiones, si se necesita, es una exposición breve (generalmente colocado al final del documento), de sus principales hallazgos y conclusiones, que intenta completar la orientación del lector que ha estudiado el texto precedente.

Metodología para efectuar el resumen

Elementos que intervienen

1. Sujeto: ¿Quién hace el resumen?
• El documentalista-analista.
• El Autor.
• En todo caso, ambos deben esta familiarizados con la temática del documento a resumir.
2. Objeto: ¿Qué se puede resumir?
• Todo es factible de ser sometido a resumen.
• Sin embargo, los resúmenes se aplican sobre todo a las revistas científicas, los libros, los informes técnicos, las tesis, las patentes y las conferencias.
3. Producto final: Se entiende como el resultado del contenido del documento original.
4. Destinatario: científicos, bibliotecas, educación, profesionales (médicos, abogados, periodistas...).

Pasos a seguir a la hora de hacer el resumen

- Lectura

La persona que va a resumir debe conocer el texto a resumir y para ello empleará su propio estilo. Hay que hacer dos lecturas previas y una posterior del texto para analizar en profundidad el texto original.

* Primera lectura. (Llamada también lectura recuperadora)

Es realizada por el documentalista de una forma rápida, con el fin de localizar en el texto aquellas partes que contengan información importante para el resumen sobre objetivos, alcance, métodos, resultados, conclusiones o recomendaciones.
Esta lectura se efectúa de una sola vez, sin pararse, con un número concreto de retrocesos y sin fijaciones.
Mientras leemos tenemos que ir anotando mentalmente o en el margen de papel qué partes del material, o del texto, contienen información sobre estos aspectos:
• Información clave o básica
• Objetivos, alcance y métodos
• Resultados y conclusiones o recomendaciones.
* Segunda lectura. (Llamada también lectura creativa)
Es aquella en la que el analista vuelve a leer el material identificado durante la primera lectura para seleccionar, extraer, organizar y escribir la información pertinente para el resumen.
En este punto ya hay que tener en cuenta las normas de estilo, pues tras la lectura se procede a la redacción, cuidando especialmente la frase anotativa.
• La frase anotativa es la primera que se hace.
• Debe, en lo posible, contener la idea esencial del texto original si no se encuentra expresada en el título del documento.
• Esta frase será, aconsejablemente, de unas tres líneas y a lo largo de todo el texto es recomendable redactar párrafos que a ser posible no excedan de las seis líneas.
* Tercera lectura. (Llamada también crítica)
Es aquella en la que el analista lee el resumen escrito cualitativamente con el objeto de corregirlo, atendiendo a su unidad y concisión y para cerciorarse de que siguió las normas de estilo.

Análisis
Hay que distinguir dos tipos de análisis:
El formal con el que se elabora la referencia del resumen.
El temático, empleado en la reducción del contenido documental.
Lo primero que habrá que hacer es desmenuzar todo el contenido de la información, lo que nos permitirá eliminar aquella información secundaria o poco importante.
Para realizar un análisis temático -el más usado- habrá que tener en cuenta estos puntos:
1. Objetivos y alcance: el analista debe descubrir el propósito del autor a la hora de escribir el documento para comprender mejor el alcance del análisis.
2. Metodología: se apuntarán las técnicas y métodos empleados por el autor del documento, el equipo y el material empleado.
3. Resultados: serán expuestos de forma clara y recogerán los descubrimientos de manera concisa aunque informativa.
4. Conclusiones: describe las implicaciones de los resultados y, especialmente, cómo éstos se relacionan con el propósito de la investigación. Pueden ser recomendaciones, sugerencias, aplicaciones, nuevas relaciones, hipótesis, etc. Es la esencia de la investigación y por consiguiente de gran valor e interés para los investigadores.

Para realizar un análisis formal hay que tener en cuenta la confección de la referencia del resumen, es decir, aquellos elementos simples y convencionales que posibilitan la identificación precisa y forma del documento original.

Otras consideraciones a tener en cuenta

1. Evitar expresiones tales como: según dice el autor, parece que, el artículo trata de...
2. Excluir, a menos que sea imprescindible, la utilización de gráficos, diagramas, esquemas, etc.
3. Evitar redundancia en las expresiones y añadidos que generen frases fuera de contexto
4. No conviene traducir el título.
5. El contenido se conocerá sin recurrir al original.
6. No citar, a menos que sea imprescindible, los mismos ejemplos que cita el autor.
7. Evitar neologismos o palabras de otros idiomas.
8. Se deben utilizar los verbos en forma activa siempre que sea posible; esto contribuye a una redacción clara, breve y rotunda.

La confección de un resumen depende de sus objetivos, de sus contenidos temáticos y de las directrices del organismo para el que se elabora.
El lenguaje adoptado será correcto; legible (breve y simple); la redacción clara –evitando ambigüedades, palabras mal empleadas o términos no específicos-; además de apropiada (en estilo y tono) y concisa.
En la elaboración del texto final habrá que tomar en consideración las necesidades de los lectores y reflejar la estructura del documento original, manteniendo una postura objetiva.

Principales características del resumen

Una de las formas de caracterizar los resúmenes es por su extensión; sin embargo, según Lancaster, no hay ninguna razón para que todos los resúmenes tengan aproximadamente el mismo tamaño, ya que los factores que lo determinan varían en función de:

1. La extensión del ítem que se está resumiendo
2. La complejidad del contenido temático
3. La diversidad del contenido temático.
4. La importancia del ítem para la institución que elabora el resumen
5. La accesibilidad del contenido temático
6. El coste
7. La finalidad

No siempre resulta más caro elaborar un resumen más extenso.

Clasificación de los resúmenes (Fondin)

1. Telegráfico: está formado por una frase constituida por la unión de palabras claves.
2. Indicativo: Es una breve y exacta representación del contenido de un documento.

A diferencia del informativo indica de forma superficial los temas abordados. Describe la información del documento relativa a objetivos y métodos.

• La extensión oscila entre las 50-100 palabras.

• La función principal es alertar al usuario, anunciándole la existencia del documento y ofreciéndole la información suficiente para que decida si merece la pena iniciar la lectura.
• En ningún caso un resumen indicativo sustituirá la lectura del texto, a diferencia del informativo que si puede hacerlo.
3. Informativo: se diferencia del indicativo en que añaden a los objetivos y métodos aspectos relativos a resultados y conclusiones del autor del texto.
• Tiene una extensión entre las 100-250 palabras.
• Representa todos los aspectos significativos y relevantes del documento primario, mediante una relación lógica y lineal de los temas tratados.
• También se conoce como analítico porque está elaborado por el autor de la obra, de ahí que se le achaca falta de objetividad, porque el criterio del autor a veces no es el que más importa al usuario, ni el más pertinente para el proceso documental.
4. Competio: versión abreviada del texto del documento. Tiene una forma cuidada y literaria. De un 20 a un 50% del documento original.
5. Reseña: se trata de un resumen poco objetivo en el que el analista aporta comentarios críticos, con respecto a las ideas expuestas por el autor. García Gutiérrez, destaca uno más:
6. Reseña sintética: es el aglutinamiento de resúmenes correspondientes a varios temas afines o a un tema de desarrollo cronológico realizado habitualmente en la depuración del archivo.

¿Cómo se puede evaluar un resumen?
Borko y Bernier aportan estas ideas:
1. Una consideración global de calidad definida por evaluadores humanos
2. La medida en que se respetan en los resúmenes las normas de estandarización
3. La inclusión de información importante y la exclusión de información sin importancia.
4. Ausencia de errores
5. Coherencia de estilo y legibilidad
6. Previsibilidad de la relevancia
7. Capacidad de servir como sustituto del original en el caso de los resúmenes informativos
8. Adecuación como fuentes de términos de indización.
Payne, por su parte, utiliza tres conceptos:
1. Coherencia (similitud en la cantidad de información presentada entre distintos analistas)
2. La cantidad de reducción de texto obtenido
3. Utilidad (comparada y averiguada a través de la confrontación de las opiniones de especialistas en la materia).

Utilidades de un resumen
1. Sirve de anticipo del documento original, permitiendo a los usuarios decidir sobre la conveniencia o no de consultar el texto original.
2. Actúa a veces como sustituto del documento original, siempre y cuando la información que aporte sea satisfactoria para el receptor.
3. Actualiza los conocimientos del especialista sobre los desarrollos habidos en su campo teórico, ahorrándole tiempo y esfuerzo.

4. Contribuye a la superación de las barreras del lenguaje.
5. Ayuda en las tareas de búsqueda retrospectiva y recuperación de información, cumpliendo un papel importante en la estructura de los sistemas automatizados, ya que muchas bases de datos incluyen, junto a las referencias bibliográficas, resúmenes que permiten la localización y selección del texto completo o documento original.
6. Facilita la indización, ya que concreta la materia indizable y elimina los problemas del lenguaje, por eso, en muchos Centros de Documentación e Información se han empleado los resúmenes como base para la confección de índices, al contener, por regla general, la macroestructura del texto y la información esencial de su contenido.

LENGUAJES DOCUMENTALES

Introducción

Los lenguajes documentales aparecen como una necesidad para estructurar el pensamiento, agrupando y asociando cada documento a una lista clasificatoria, o bien mostrando el contenido, sobre todo a través de resúmenes o palabras claves.
En el comienzo, los lenguajes documentales se presentaron como sistemas clasificatorios. Algunos teóricos de la documentación consideran que el primer sistema clasificatorio fue el realizado por Brunet en 1804, aunque terminó por imponerse la Clasificación Decimal de Dewey. Melvin Dewey, en 1876, clasificó y dividió el pensamiento en varias clases y a cada una de estas las volvió a dividir y subdividir sucesivamente, dando origen de esta manera a una Clasificación Decimal.
El Instituto Internacional de Bibliografía adoptó este sistema, publicando en 1905 la primera edición internacional. Con posterioridad, la federación Internacional de Documentación realizó varias modificaciones hasta que se constituyó la Clasificación Decimal Universal.

Concepto de lenguaje documental

El lenguaje documental es un sistema convencional que utiliza una unidad de información para describir el contenido de los documentos, con miras a su almacenamiento y recuperación. Por regla general, un documento trata de más de una noción, más de un contenido.
El lenguaje documental (AMAT): "Es un conjunto de términos o procedimientos sintácticos (frases nominales) y convencionales utilizados para representar el contenido de un documento, con el fin de permitir su recuperación. Se le denomina también lenguaje de indización".
Es, por tanto, un lenguaje artificial para diferenciarlo del lenguaje natural, aunque en algunos casos se empleen los mismos términos.
Lenguaje documental, según Inocencia Soria, de la Unidad Coordinadora de Bibliotecas del CSIC: "Es el sistema convencional creado para poder expresar el contenido de los documentos sin los impedimentos del lenguaje natural. Simplifica el lenguaje común utilizando sólo una pequeña parte del léxico, algunas formas y poco o casi nada de gramática. Los lenguajes documentales pueden consistir

simplemente en una lista de palabras admitidas aunque lo más frecuente es que consten de un sistema estructurado que relacione sus distintos términos".
Cuando se incluyen relaciones sintácticas permiten recuperar la información de términos relacionados. Si son relaciones jerárquicas como las que se encuentran en las clasificaciones enciclopédicas, alfabéticos de materias y tesauros, permiten respetar el nivel específico del texto, es decir, contemplar los términos en su profundidad.
Para Coll-Vinent y Bernal Cruz, lenguaje natural es aquel en el que están escritos todos los documentos primarios sometidos al análisis documental.
Es también el lenguaje hablado y el que se emplea en todas las informaciones originales, sea cual sea su soporte. Lenguaje artificial o convencional: es un lenguaje estructurado con un propósito particular y con unas características que le son propias. Para ellos, el lenguaje documental es el lenguaje convencional que apunta a la descripción del contenido de un documento primario en orden a su almacenamiento y ulterior recuperación. Es, además, un lenguaje que produce la transformación de un texto original en otro distinto mucho más breve, en el que queda fielmente representado.
En resumen: el lenguaje documental actúa como vehículo de comunicación entre el contenido del documento y el usuario y con él se pretende reducir y, a ser posible, evitar la multiplicidad de sentidos que tiene el lenguaje natural.

Diferencias entre lenguajes documentales y lenguajes naturales
Si el lenguaje natural se usa para la comunicación inmediata, esta que realizamos ahora mismo; el lenguaje documental se emplea para conseguir una comunicación, que es primordialmente un medio, un código unívoco y estereotipado, controlado y no libre, normalizado y no arbitrario.
En el lenguaje natural coexisten diferentes significados para un solo significante o diversos significantes sinónimos. Por el contrario, el lenguaje documental ejerce un control léxico que impide la utilización de distintos significantes libres para un mismo significado con el objeto de sobrevivir dentro de un código normalizado.

Funciones del lenguaje documental
El lenguaje documental sirve como elemento aglutinador de todos los trabajos que se lleven a cabo en el proceso documental. Así, puede considerarse la expresión lenguaje natural como sinónimo de discurso común, el lenguaje normalmente usado en la escritura y la conversación. Pero, en el contexto de la recuperación de la información, la expresión usualmente se refiere a las palabras que ocurren en los textos impresos y "texto libre" hay que considerarlo como sinónimo.
El lenguaje documental ha de precisar cada uno de los diferentes encuadres que presenta el documento para facilitar la recuperación. Se trata de un lenguaje que recoge los elementos más importantes cuando se está elaborando el análisis y los aplica para hacer efectiva la recuperación.
Sin embargo, existen una serie de problemas:
- Las palabras contienen varios significados.
- Las traducciones de otros idiomas.
- Los nuevos términos que se introducen.
- Las especificaciones técnicas resultan para la mayoría de los usuarios incomprensibles.
- La forma en que se presentan los nuevos documentos, avalados por unas

tecnologías que no disponen de aspectos comunes adaptados internacionalmente
• El desacuerdo entre los documentalistas a la hora de aplicar los diferentes lenguajes.
Para Blanca Gil, otra de las funciones pasa por una normalización y una inducción: "El lenguaje documental tiene capacidad para representar los mensajes contenidos en los documentos, lo que permite cumplir dos objetivos fundamentales: el de normalización y el de inducción, estando encaminadas a éste último todas las demás funciones que desempeña a lo largo del proceso documental".
El objetivo básico de un lenguaje documental es suministrar los conceptos que aporta cada palabra, una vez efectuado el análisis. Se trata de reducir los términos que aporta el lenguaje documental, como precisa Blanca Gil: "El lenguaje documental reduce considerablemente el volumen de términos del lenguaje natural no tomando en consideración más que los sustantivos o los sintagmas nominales".

Tipología de lenguajes documentales
Hay diversas formas de tipificar los lenguajes documentales; sin embargo, las más usadas son:

- Dependiendo del control ejercido sobre el vocabulario
1. Lenguajes controlados (Clasificaciones, tesauros, etc.): aquellos que han establecido una lista de descriptores antes de proceder al análisis documental. Esta es cerrada y nominativa. Define todos los términos y únicamente aquellos que se pueden utilizar para presentar el contenido de un documento. La búsqueda y recuperación es más rápida y eficaz.
2. Lenguajes libres (Listas de descriptores libres). Al contrario, cuando se trabaja con un vocabulario o lenguaje libre no se conocen a priori listas de términos autorizados. Basta con extraer de los documentos los términos más apropiados. La búsqueda es más lenta y menos eficaz.
De todas formas, hay que señalar que ningún lenguaje es completamente puro, ninguno es pre o poscoordinado, libre o controlado.

- De acuerdo con la coordinación de términos
1. Lenguajes precoordinados (clasificaciones, listas de encabezamiento de materias): aquellos que coordinan los diferentes conceptos que forman un tema, antes de memorizar los documentos. Son lenguajes utilizados en bibliotecas: sistemas de clasificación y listas tradicionales de alfabéticos de materias. Permiten pocos términos de indización por documento
2. Lenguajes poscoordinados (Listas de descriptores libres, listas de palabras claves, tesauros): los que permiten yuxtaponer los conceptos en el momento del análisis, de manera que se pueda coordinar después del almacenamiento. Hay que recordar siempre que en los lenguajes poscoordinados es necesaria la utilización de ficheros suplementarios, llamados también ficheros inversos.
En resumen:
La precoordinación permite pocos términos de indización por documento, pero proporciona en una sola búsqueda los elementos esenciales de la información.
La poscoordinación permite utilizar un gran número de vías de acceso a los documentos, pero teniendo como intermediario ficheros especiales que necesitan

una búsqueda en dos tiempos; primero la identificación de los documentos pertinentes y después su localización.

- De acuerdo con su estructura

a) Lenguajes de estructura jerárquica o clasificatoria (clasificaciones jerárquicas) aquellos que siguen un orden lógico que agrupa y aproxima los conceptos más sencillos o específicos dentro de los conceptos más generales. Este tipo de lenguaje se puede emplear para localizar un documento, **pero no para indizar con profundidad**. Se establece una clasificación sistemática lineal, en la cual los conceptos se encuentran ordenados siguiendo una jerarquía natural, definida por el estado de los conocimientos en el momento en que ha sido elaborada. Cada concepto de estas estructuras jerárquicas se halla representado por un símbolo numérico, alfabético o alfanumérico que indica la situación de cada materia.

Los lenguajes de estructura jerárquica se dividen en:

1. Clasificaciones enciclopédicas:

• Permiten la organización de documentos que tratan sobre cualquier materia: son de ámbito universal y multidisciplinario.

• Presentan dos inconvenientes: el objetivo de su universalidad limita la descripción de un documento especializado y su rigidez dificulta una puesta al día ágil y rápida.

- Clasificación de Dewey (1876): divide el conjunto de los conocimientos en 9 clases principales, designadas en números arábigos del 1 al 9, reservando el 0 para las generalidades. Cada clase se subdivide sucesivamente en 10 subclases y así sucesivamente, con números que se dividen en grupos de tres por medio de puntos para hacer más fácil la lectura.

- Clasificación expansiva de Cutter (1891): se compone de 7 tablas o esquemas, cada una de los cuales incluye la totalidad de los conocimientos, pero con una complejidad progresiva.

- Clasificación de la Library of Congress (1897): cuenta con 21 clases principales, tomadas del sistema de Cutter, que designa con otras tantas letras mayúsculas, dejando las restantes para futuras ampliaciones.

- Clasificación bibliográfica de Bliss (1935): está formada por 4 esquemas generales que son filosofía, ciencias zoológicas, físico y social que se dividen en un total de 26 clases principales. Es muy similar a las facetadas.

- Clasificación Decimal Universal (1905): es una ampliación de la clasificación decimal de Melvil Dewey (se basa en su quinta edición). Es numérica hasta cierto punto, precoordinada, universal, multidimensional y arborescente. Está agrupada en 10 clases, reducidas a 9 por la fusión de lingüística, filología y literatura, dejando la clase 4 vacía. Cada una de ellas se subdivide: 0 generalidades. 1 Filosofía. 2 Religión. Teología. 3. Ciencias Sociales. Estadística. Política. Economía. Derecho. Administración. Asistencia social. Seguros. Educación. Etnología. 4 sin ocupar. 5 Ciencias puras. Ciencias exactas y naturales. 6 Ciencias aplicadas. Medicina. Técnica. 7 Arte. Artes industriales. Fotografía. Música. Juegos. Deportes. 8 Lingüística. Filología. Literatura. Crítica literaria. 9 Arqueología. Prehistoria. Geografía. Biografía. Genealogía. Historia. La tarea de actualizar la CDU está encomendada a la FID (Federación Internacional de Documentación), si bien como señala Inocencia Soria: "En 1992 se constituyó el consorcio CDU, que asumió las responsabilidades que antes tuviera la FID sobre su edición, actualización, versiones, etc. Este consorcio, cuyos socios fundadores

son Bélgica, España, Holanda, Japón, Reino Unido y la propia FID, se comprometió a organizar y mantener la CDU y sus esfuerzos ya han dado algunos importantes frutos: se ha creado un fichero informático con más de 60.000 entradas que está sirviendo de base para facilitar su manejo y actualización".

2. Clasificaciones especializadas: Son instrumentos de indización que abarcan disciplinas o campos especializados (medicina, derecho, economía, etc.) que no quedarían profunda y ampliamente representados en una estructura jerárquica de ámbito universal y multidisciplinario. Por ejemplo, la mayoría de las bases de datos disponen de este tipo de clasificaciones.

3. Clasificaciones de facetas: Son de origen enciclopédico, pero su organización permite construir áreas concretas de los conocimientos, ya que faceta es cada uno de los aspectos o puntos de vista que pueden incluirse en un área concreta. Se basan en:

- Clasificación colonada de Ranganathan (1933): Se llama colonada porque utiliza el colon como único signo de síntesis. No es, por tanto, una división lineal y jerárquica como el resto sino la aplicación que tienen algunas materias para descomponerlas. Son muchas su características, pero en el campo de la biblioteconomía se reducen a cinco: personalidad, materia, energía, espacio y tiempo. Sin embargo, tiene subdivisiones de lengua, geografía, cronología, etc.

b) Lenguajes de estructura asociativa (Léxicos documentales, tesauros): aquellos que se organizan por orden alfabético en términos que expresan los conceptos retenidos durante la indización. Los términos o descriptores se combinan libremente entre sí sin quedar sujetos a una posición determinada del lenguaje, según las necesidades de los documentos. Esta estructura proporciona un acceso más inmediato a la información, pero tiende a dispersar los términos parecidos.

Los lenguajes de estructura asociativa se dividen en:

1. Alfabético de materia: Se organizan alfabetizando encabezamientos de palabras o grupos de palabras que expresan conceptos. Tienen como requisito la uniformidad de los términos empleados, estableciendo un juego de referencias cuando sea necesario para relacionar, completar o no marginar temas. Utilizan un lenguaje precoordinado y un vocabulario controlado.
2. Uniterms: Se caracteriza por corresponder a una estructura asociativa alfabética, pero cada término (uniterm) representa una palabra-clave sin determinar el nivel que hay de asociación.
3. Descriptores: Un término o grupo de términos que representan un concepto preciso.
4. Índices permutados: se permuta en forma circular todas las palabras del texto o del título para distinguir la palabra que se utiliza como descriptor.
   - Índices KWIC (Key word in context): se recoge cada palabra significativa del texto o título en una lista alfabética, de tal manera que están las más significativas. Y de ellas se elige la que aparece siempre en el mismo lugar: el centro.
   - Índices KWOC (Key word out of context): Se buscan de la misma forma, pero se obtendrán aquellas palabras que sobresalgan fuera del título.
   - Otros índices: Cruzados, Acumulativos e Índices de citas.

REFERENCIA

Material bajado de Internet. Extracto del curso Fundamentos de Información y Documentación.
http://www.ucm.es/info/multidoc/prof/periodismo/curso2004_tem_periodismo.htm

BIBLIOGRAFÍA

Amat Noguera, Nuria: Documentación científica y nuevas tecnologías de la Información. Madrid: Pirámide, 1989. Véase especialmente el capítulo 5: Lenguajes documentales y Thesaurus, p. 189-234, de donde se han obtenido las principales aportaciones de este tema. Véase además:
Chaumier, J.: Análisis y lenguajes documentales: el tratamiento lingüistico de la información. Barcelona: Mitre, 1986.
Currás, Emilia: Thesaurus, lenguajes terminológicos. Madrid: Paraninfo, 1991.
Díez Carrera, C.: Técnicas y régimen de uso de la CDU. Gijón: Trea, 1999.
García Gutiérrez, Antonio: Lingüística documental: aplicación a la documentación de la comunicación social. Barcelona: Mitre, 1984.
Gil Urdiciain, Blanca: Manual de lenguajes documentales. Madrid: Noesis, 1996.
Gleyze, A. Pour une méthode d'indexation alphabétique de matières. Villeurbanne: E.N.S.B., 1983.
Lancaster, F.W.: El control del vocabulario en la recuperación de información. Valencia: Universitat de Valencia, 1995.
López-Huertas Pérez, Mª. J.: Lenguajes documentales: aproximación a la evolución histórica de un concepto. En: Boletín de la ANABAD, XLI (1991), núm. 1, enero-marzo, p. 61-70.
López-Huertas Pérez, Mª. J.: Lenguajes documentales: terminología para un concepto. En: Boletín de la ANABAD, XLI (1991), núm. 2, abril-junio, p.171-188.
López Yepes, José (Comp.). Manual de información y documentación. Madrid: Pirámide, 1996.
Maltese, D.: Elementi di indicizzazione per sogetto. Milan: Bibliografica, 1982.
Maniez, J.: Los lenguajes documentales y de clasificación: concepción, construcción y utilización en los sistemas documentales. Madrid: Fundación Germán Sánchez Ruipérez, 1992.
Pinto, María. Manual de clasificación documental. Madrid: Síntesis, 1997.
Soria González, Inocencia. La organización de la información, los lenguajes documentales y la normalización. Consejo Superior de Investigaciones Científicas.
Van Slype, G.: Los lenguajes de indización. Madrid: Fundación Germán Sánchez Ruipérez, 1991.

LA PRODUCCIÓN DE RESÚMENES CIENTÍFICOS

María Pinto Molina
Universidad de Granada (España)

1. INTRODUCCIÓN

Como hemos apuntado en el capitulo anterior, los procesos de resumir son actividades cognitivas que se llevan a cabo sobre objetos lingüísticos mediante el empleo de herramientas lógicas con un objetivo documental. El resultado es el resumen documental, un breve texto representativo, intencional, no unívoco y con vocación de sinónimo, que debe responder a determinadas características derivadas de su equivalencia funcional con el texto origen, como son su legibilidad, exhaustividad, precisión y entropía. Por encima de todas estas propiedades, podemos asegurar que un resumen no sirve para nada si no es comprensible (1), y esa comprensibilidad depende directamente de su coherencia como unidad textual: se trata de producir ante todo y sobre todo un texto coherente. Sin lugar a dudas las tareas resumidoras se encuentran en la cúspide de las actividades documentales, dada la categoría textual de los productos referenciales que de ellas derivan. El hecho de que a los resúmenes se les exija un cúmulo de propiedades recuperativas, indicativas, informativas, supletorias y orientativas obliga a sus productores a un esfuerzo intelectual máximo en pos de ese amplio y a veces impreciso abanico de objetivos. Esta riqueza teleológica provoca la multiplicidad de sus resultados, teniendo en cuenta que se pueden generar varios resúmenes para un mismo documento en función de las variables puestas en juego durante el proceso resumidor.

Considerando que estas tareas se apoyan en el triángulo documento base (texto o multimedia), situación (contexto operativo) y agente (resumidor) comprenderemos fácilmente las divergencias que pueden surgir a la hora de operar sobre un documento común. Pero si nos restringimos al ámbito dé los documentos científicos también deducimos de un modo espectacular esas probables diferencias entre sus distintos resúmenes, al encontrarnos en un territorio cognitivo muy estructurado retórica y extremadamente preciso en lo que a información implícita y manifiesta se refiere.

2. EL ENTORNO DE LA PRODUCCIÓN DE RESÚMENES

El texto, como materia prima a representar, y campo de operaciones de la compleja actividad resumidora, es un producto muy elaborado cuyo tratamiento analítico resulta complicado. Como es fácil intuir, son grandes las posibilidades de acometer de un modo reglado el resumen de un texto científico, en un entorno científico y para una clientela científica, sobre todo si se le compara con el otro extremo de la producción textual, el texto literario. A diferencia de los textos científico/técnicos con un discurso de intención objetiva, analítica o descriptiva, que dejan poco espacio a la personalidad de su creador, el texto literario es por el contrario un texto de autor, que se caracteriza por sublimar la función expresiva mediante el empleo de un lenguaje complejo significativamente, y por consiguiente polisémico y ambiguo. El texto literario pertenece menos a un contexto racional (objetivo) y más a un entorno sensible (subjetivo), con un alto componente emocional o afectivo que se manifiesta de un modo individual, personal y hasta a veces intransferible (2).

Dado el creciente protagonismo de los contenidos multimedia en todos los ámbitos de nuestra sociedad, debemos detener nuestra atención en estas nuevas formas documentales caracterizadas por la convivencia de los mensajes textuales, sonoros y visuales. Aunque su representación resumida tendrá un carácter textual, la metodología a emplear en el procesamiento de unos documentos caracterizados por su juventud y su mayor complejidad, no es tan precisa como pueda ser su equivalente en los entornos exclusivamente textuales. El resumidor tiene ahora que efectuar un salto, no ya entre dos niveles de descripción textual como sucede con los documentos bibliográficos, sino entre un plano de organización audiovisual y otro plano textual. El preceptivo proceso de interpretación es, en el caso de los documentos multimedia, mucho más arriesgado y el resumidor tendrá que recurrir a todas sus capacidades perceptivas, cognitivas, interpretativas y descriptivas, pues en este nuevo contexto los mensajes son cada vez más un proceso abierto y modificable, la pantalla se transforma en un elemento del mensaje y los contenidos se atomizan en un mosaico de elementos cuyo sentido es reconstruido a su manera por el usuario.

2.1 Importancia del dominio discursivo

Puesto que no se puede suministrar una descripción (representación) completa de cualquier fenómeno desde todos los puntos de vista posibles, el resumidor debe recurrir a referencias situacionales para llevar a cabo su actividad práctica (3). El contexto adquiere un papel vital en todos aquellos estadios en los que interviene el razonamiento para la adquisición de información, y la comprensión se reconoce como un proceso activo de contextualización del mundo. Consiguientemente, las investigaciones se están desplazando hacia un cuadro más contextualizado del resumidor que juega un papel crucial a la hora de ampliar el significado de los textos leyéndolos dentro del contexto de su propia vida. De las múltiples situaciones que se pueden presentar en el entorno documental en que nos movemos, el resumidor se verá afectado por los contextos individual, social, de dominio, y documental. Y si el contexto individual deriva de sus propios esquemas personales, tendrá también que considerar el contexto social, tanto a la entrada como a la salida de los procesos, porque éstos se llevan a cabo en determinados entornos sociales, (científico/técnico, literario, económico, industrial, empresarial, deportivo, ...) y cada uno de ellos tiene unas características especificas que condicionan sus discursos. Pero si pensamos que los dominios discursivos y las comunidades de conocimiento son las unidades apropiadas de estudio, el contexto determinado por dichos dominios adquiere un protagonismo fundamental en las operaciones resumidoras. Finalmente el entorno de la tarea, o entorno documental, plantea un contexto documental determinante del tipo de resumen y consiguientemente los procesos resumidores. Establecidos los contextos, el camino hacia el resumen será mucho más fluido y preciso.

2.2 Agentes

La segunda variable es el resumidor, y sus capacidades de almacenamiento y procesamiento a través de la memoria. Si la memoria a corto plazo tiene un período de actuación limitado y se utiliza para acumular informaciones de

estructura superficial, mayores posibilidades conservadoras y procesadoras tiene la memoria a largo plazo, cuyo objetivo prioritario son las estructuras semánticas. Pero sólo somos capaces de retener en la memoria la enorme cantidad de información que necesitamos si ésta se encuentra eficazmente estructurada pues el almacenamiento se hace de una manera estructurada y frecuentemente jerárquica, constituyendo lo que se conoce como conocimiento previo del individuo. Esta noción se vincula a la teoría de los «esquemas», especie de cuadro de referencia de entidades lingüístico/conceptuales que condiciona la comprensión y construcción de nuevos conocimientos, aunque esta comprensión depende sobre todo del poder del hombre corno pensador o escrutador de la información textual.

Reconociendo que los sistemas de producción y comunicación electrónica de información no requieren la producción inteligente de resúmenes, se impone la automatización de tales procesos, un objetivo prioritario en el que andan empeñados investigadores y empresas, tratando de potenciar las posibilidades del ordenador al servicio de los procesos de reducción informativa. Asumiendo la incapacidad de los actuales programas informáticos para producir resúmenes corno unidades textuales autónomas y coherentes, hemos de admitir su capacidad para generar otras formas documentales que, aunque de inferior categoría informativa, son muy útiles porque se adecuan a unos nuevos entornos menos cerrados a la unidad textual y más abiertos a las relaciones hipertextuales. Nos referimos a los diferentes tipos de microtextos representativos generados en el entorno digital, entre los que se encuentran los extractos y también los sumarios. El creciente protagonismo de estos sucedáneos del resumen se basa en su origen automático y su consiguiente inserción instantánea en las nuevas redes teleinformáticas.

2.3 Objetivos documentales y tipología del resumen

Sobre las distintas variables implicadas en los procesos de producción de resúmenes flotan los tres objetivos documentales ya clásicos en la trayectoria del resumen como figura documental de referencia: indicar, informar y sustituir. Si se trata de indicar basta con generar un documento que suministre indicios sobre el documento que representa. Cuando su objetivo es informar el resumen se equipara a cualquier otro documento primario, siendo ésta quizás su más noble cualidad documental.

También puede darse el caso que la información anticipada sea suficiente para el usuario, ejerciendo el resumen la función suplente del original. En los actuales entornos ciberdocumentales uno puede fácilmente perder la brújula y necesitar instrumentos orientadores adecuados a las nuevas posibilidades viajeras. Los resúmenes responden a ese cuarto objetivo (orientar) en una doble dirección: como etiquetas metadatos de los documentos electrónicos, y como unidades independientes en una red de resúmenes hiperenlazada, con enlaces inter e intradocumentales. El resumen, como cualquier otro texto, depende de unas condiciones de producción que determinan el tipo resultante. El entorno bibliográfico nos remite a la capacidad informativa de los resúmenes, para distinguir los resúmenes indicativos, más elementales y esquemáticos, en la frontera con el índice y con una función de alerta, y los resúmenes informativos, que son los auténticos representantes del resumen como figura documental (un resumen informativo no se distingue de cualquier otra fuente primaria de

información) y responden a las funciones sustitutivo y recuperativa: es el resumen en «texto libre». El entorno electrónico se basa en el grado de estructuración en los resúmenes, distinguiéndose su estructuración interna, según la cual los resúmenes se integran en el documento origen mediante etiquetas o utilizando determinadas estructuras retóricas: es el resumen «aplantillado». La estructuración externa nos remite a resúmenes integrados en una colección documental, para lo cual se les requiere determinadas cualidades hipertextuales: es el resumen «hipertextual».
El entorno digital reticular suele suministrar documentos «formales» según el molde tradicional y también un amplio rango de documentos «digitales» (paginas Web personales, listas de enlaces a otros recursos, comentarios, ...). La función resumidora para estos documentos más novedosos puede satisfacerse con los recientemente desarrollados «microtextos», que son el equivalente funcional del viejo resumen puesto que también se basan en el documento fuente, si bien varían sensiblemente los procedimientos empleados y el producto resultante. Por el momento se han estudiado poco las características físicas, intelectuales y operativas de estas nuevas formas de sustituto documental y sus relaciones con los clásicos resúmenes de las bases de datos.

3. PROCEDIMIENTOS DE ELABORACIÓN DE RESÚMENES

En el ejercicio de la actividad resumidora proponemos un modelo integrador (4) basado en tres etapas: Selección, Interpretación y Producción.

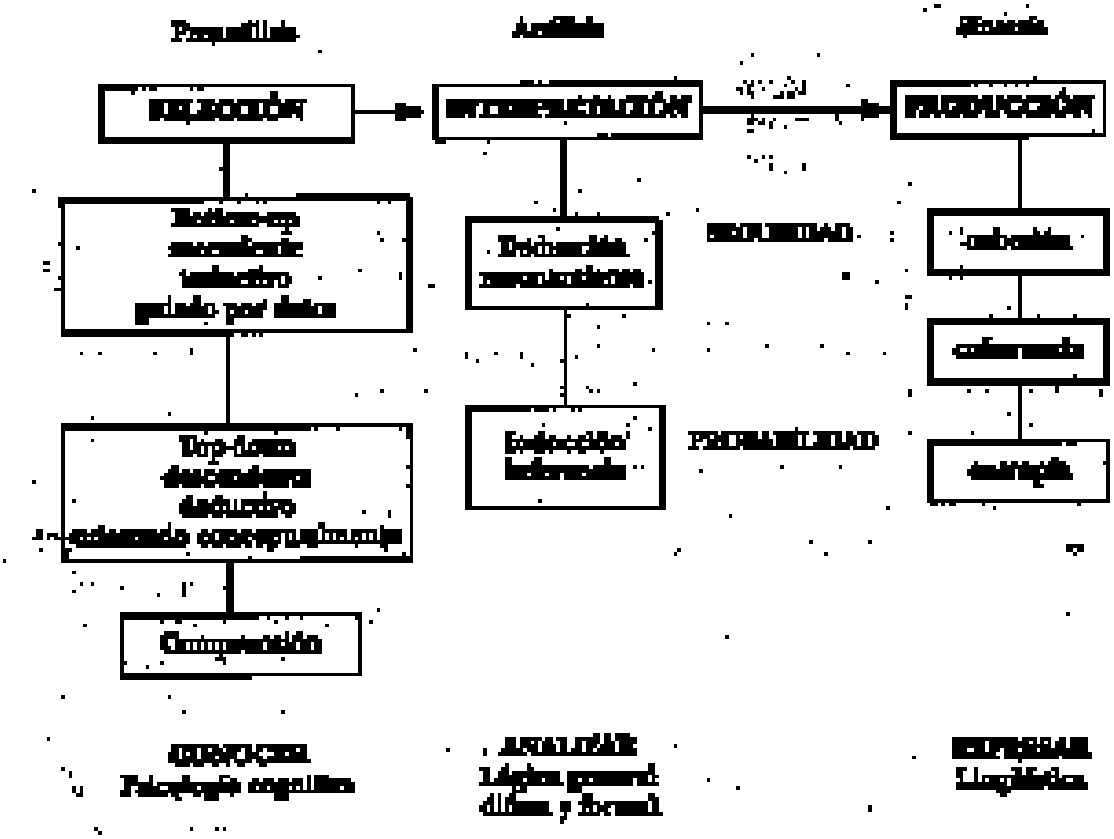


Fig. 1 Procedimientos para la elaboración del resumen.

3.1 Estrategias y técnicas para resumir

Selección de información

La lectura, único modo de percibir la información textual, está condicionada por el papel de las frecuencias de visión, van construyendo poco a poco en la mente los cuadros de comprensión; la acción de la memoria, que puede movilizar sus «cuadros» más profundos relacionando lo desconocido y lo conocido; y la intervención de la razón en sus actividades complementarias de análisis y síntesis. Este medio perceptor pretende lograr un estado óptimo de comprensión textual basado en los principios de segmentación, ya que todo individuo es capaz de segmentar señales del flujo continuo de la lengua; categorización, proceso que se refiere a formas de palabras y a sus categorías sintácticas correspondientes, en consonancia con la ver tiente paradigmático de la lengua; combinación, porque los fonemas y los morfemas se yuxtaponen de acuerdo con las estructuras sintagmáticas, o combinatorias; e interpretación, puesto que a palabras y oraciones se les asigna un determinado significado convencionalmente establecido.

La comprensión de la información se basa sobre todo en la interpretación, es decir, en la adjudicación de significados a señales (perceptibles), y esto tan sólo es posible como consecuencia de operaciones mentales previas de segmentación, categorización y combinación de lo percibido. Desde una perspectiva pragmática el resumidor debe hacer una primera lectura rápida del documento original para centrar la atención en sus características fundamentales (forma, clase, estructura de la información, ...), teniendo en cuenta la distinción entre dos grandes categorías temáticas: la que aglutina los temas principales, o explícitos, relacionados directamente con el contenido exclusivo del trabajo; y la que agrupa a los secundarios, o implícitos, que son tratados paralelamente por necesidades expositivas. Será necesaria una segunda lectura, cuidadosa y activa, centrada en los distintos epígrafes del documento y en sus secciones claves (objetivos, metodología, resultados y conclusiones), pues por regla general contienen la esencia conceptual del documento. Es la lectura «recuperativa», porque pretende identificar sólo los pasajes que contengan información merecedora de ser incluida en el resumen. En el caso de los documentos científicos, es importante considerar los primeros párrafos (focalización de la atención por parte del autor) y también los últimos, dado el alto grado de epitomación informativa que podemos encontrar en ellos. La práctica de la lectura se ve condicionada por múltiples factores, como los objetivos resumidores, derivados del entorno operativo/documental, y el tipo de documento pues no es igual, ni tan siquiera parecido, enfrentarse a sendos textos científicos y literarios.

Van Dijk (5) utiliza la estrategia de las macrorreglas, como instrumento que posibilita la unión entre las estructuras superficial y profunda. La macrorregla omitir opera negativamente eliminando lo innecesario o irrelevante y la macrorregla seleccionar opera positivamente extrayendo lo necesario y relevante (6). Ambas están relacionadas esencialmente con la identificación de las proposiciones «importantes», y se pueden equiparar con lo que se denomina contracción, un método frecuentemente empleado en el aprendizaje de las lenguas y en particular de las técnicas de expresión, que consiste en eliminar la redundancia. Teniendo en cuenta que cualquier texto puede en principio reducirse a la mitad sin

menoscabo de su información significativa, la contracción llevada más allá de cierto límite finaliza con la evaporación de su sentido.

Interpretación de información

El análisis de los documentos escritos pretende descomponer, o desmontar, su estructura discursiva como único modo de comprender su funcionamiento interno. El primer paso es la segmentación, o descomposición provisional del texto en magnitudes más manejables, mediante la división en segmentos, una de cuyas formas consiste en recopilar el texto con margen flotante a la derecha manteniendo en cada línea los conjuntos cuya cohesión interna es suficientemente fuerte (reescritura segmentada). Así se pueden identificar las partes significativas o unidades de significación (palabra y frase). Conviene distinguir las palabras útiles, que expresan relaciones sintácticas u operatorias y pueden ser omitidas por redundantes; y las nocionales o informativas, con tres subcategorías: generales, que pertenecen al vocabulario base de la lengua; circunstanciales, que corresponden a matices del autor, y específicas del lenguaje particular de un determinado dominio. El significado de una frase depende de su estructura sintáctica y de los significados particulares de sus elementos, pudiendo distinguirse las frases estructurales, que no pueden ser ni sustituidas ni suprimidas sin destruir el texto y mutilar el sentido; las frases circunstanciales o permutables, cuya supresión no altera la estructura profunda del texto, y las frases estilísticas, formadas por configuraciones léxicas de las cuales algunas pueden ser figuras retóricas. La frase estructural es una moneda mucho más fácilmente cambiable que la palabra y particularmente mucho más fácil de traducir de una lengua a otra. Sin embargo la unidad de base de la representación, la más pequeña unidad de traducción de conocimientos, es la proposición, o conjunto de conceptos relacionados que se reagrupan para formar episodios o contextos (7).

La disposición o estructuración de la información textual conduce al concepto de estructura retórica, discursiva o esquemática (superestructura), especie de esquema al que el texto se adapta con independencia de su contenido, y que supone el camino más intuitivo y fiable para el establecimiento de una clasificación textual. El abordaje de estas estructuras organizativas, retóricas, discursivas, esquemáticas, se encuentra directamente vinculado al concepto de género textual, y resulta sumamente interesante de cara al estudio de los procesos de resumir, pues el sentido común nos hace ver

que una adecuada correlación entre las estructuras retóricas del texto original y su resumen correspondiente garantizaría el transvase de «la lógica» del discurso, facilitando los procesos resumidores y la mayor representatividad del producto resultante. En realidad, cada género textual provoca la necesidad de unas técnicas especificas que favorezcan su procesamiento y respeten su personalidad. El planteamiento de estrategias características, convenciones que gobernarían los procesos de resumir, es ya una prometedora senda investigadora. Desde un punto de vista semántico pragmático, el resumidor debe no solo generalizar, sustituyendo una serie de conceptos por el sobreconcepto compartido sino también construir o integrar, reemplazando la información por otra nueva, de acuerdo con el principio de implicación semántica. Se produce de este modo lo que normalmente denominamos abstracción.

Producción del resumen

Los procesos de síntesis (producción) son parcialmente reproductivos y parcialmente constructivos (8), suponiendo el «más difícil todavía» del periplo resumidor, pues si hasta aquí las actividades analíticas se han podido someter a unas técnicas más o menos afortunadas, es prácticamente imposible establecer unos mecanismos sintetizadores que sean válidos para todo tipo de documentos y de resumidores. Nos encontramos ante la tarea suprema, exclusiva del resumidor, donde éste deberá poner en juego sus cualidades, habilidades, conocimientos e intereses documentales para representar el documento original a escala reducida. Esta última y definitiva etapa sintetizadora se corresponde con la transformación discursiva de la estructura cognitiva obtenida en el proceso analítico y equivale a una expansión exclusivamente lingüística que se realiza a través de dos mecanismos fundamentales: reformulación, mediante la cual se incorpora una macroestructura a un grupo de ellas ya existente; y asimilación, cuando determinadas macroestructuras se unen a través de una operación de síntesis. De este modo, y tras la aplicación reiterada de estos mecanismos, obtenemos el producto final o resumen, documento secundario autónomo, texto breve y completo gramaticalmente que recoge el contenido esencial del documento original cuyo mensaje tiene significación e importancia por si mismo sin necesidad de recurrir al documento original.

La elaboración de los resúmenes estará presidida por los criterios de fidelidad al original que deberá ser respetado en su contenido, sin omisión de partes sustanciales, evitándose cualquier apreciación personal; precisión, con el empleo de términos justos, eludiendo la redundancia y la repetición; claridad expositiva, utilizando la terminología apropiada; y entropía, dando a la frase la plenitud de sentido con el mínimo de palabras. Lo difícil es compatibilizar estos criterios pues algunos son antitéticos, aunque en todo caso la base para valorar la utilidad del resumen como forma de representación documental debe ser más la adecuación que la corrección ya que una representación puede que no sea correcta ni tampoco incorrecta, pero sí adecuada en el contexto de un dominio determinado o para una tarea particular (9). En resumidas cuentas, se debe velar no ya por la eficiencia sino sobre todo por la eficacia de los procesos y productos de la operación de resumir.

Procedimientos subsidiarios

La idea de conseguir acceso al contenido «hojeando» un espacio informativo, y navegando» a través de ese espacio hasta conseguir la información más útil e interesante, resulta tan atractiva y plausible que hoy día es difícil ponderar el presente y el futuro del acceso cognitivo a la información sin considerar los conceptos de visualización y navegación. El énfasis en el hojeo (browsing) en lugar de la búsqueda tradicional, implícito en el enfoque visualizador/navegador, reduce la dependencia de los sistemas de información de la inmensamente defectuosa tecnología de búsqueda Booleana (10). Pero la visualización de un espacio informativo es sólo posible si existen maneras de estructurarlo. Por eso un área crucial de investigación es la interacción entre las facilidades conceptuales y espaciales y sus respectivas estructuras cognitivas, ya que esta interacción es central al desarrollo de mapas espaciales exhaustivos que correspondan a las estructuras conceptuales de los espacios informativos. La existencia de distintas habilidades cognitivas para los procesos conceptuales y espaciales sugiere que los usuarios pueden tener diferentes niveles de éxito a la hora de utilizar la

visualización y la navegación para encontrar los temas que sean de su interés, aunque los resultados experimentales sugieren que el procesamiento espacial puede reemplazar al procesamiento conceptual sin afectar a la naturaleza y cantidad de información manejada. El término mapa conceptual engloba una amplia gama de representaciones diagramáticas del conocimiento mediante diferentes lenguajes visuales como alternativa complementaria al lenguaje natural en la función de comunicar conocimiento.

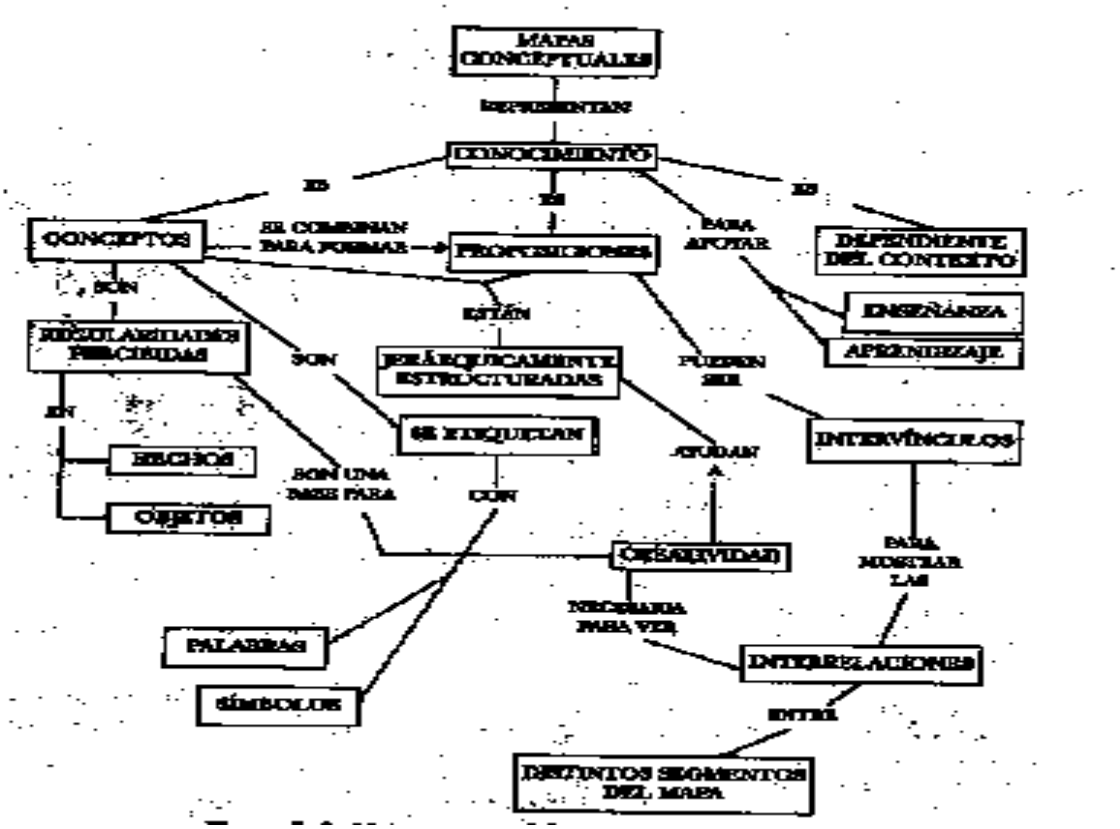


Fig. 2 Mapa conceptual de un Mapa conceptual.
Fuente: Novak, Joseph D., Conocimiento y aprendizaje: los mapas conceptuales como herramientas facilitadoras para escuelas y empresas. Madrid: Alianza Editorial, 1998.

3. 2 Métodos automáticos

Los métodos automáticos de resumir que de un modo global se han multiplicado en las últimas décadas ponen de manifiesto una imperiosa necesidad, y tienen al menos dos puntos en común. En primer lugar, todos ellos pretenden actuar a imagen y semejanza del hombre, tratando de conocer sus métodos espontáneos de percepción, interpretación y producción, y en definitiva intentando copiar sus modos de proceder: el resumidor humano es un modelo al que se trata de imitar con más o menos fortuna. Por otra parte, ninguna de estas propuestas de automatización ha logrado dar una respuesta plenamente satisfactoria al problema planteado. Nos centraremos sucesivamente y en grado creciente de dificultad en las técnicas léxico/sintácticas de selección, en las actividades lógico/semánticas de interpretación y finalmente en las tareas pragmático/documentales de producción.

La investigación sobre la producción automática de resúmenes se ha basado tradicionalmente en métodos estadísticos y en técnicas de extracción ayudadas por diferentes clases de análisis lingüístico de tipo superficial. En sus comienzos se situó en ese primer estadio léxico/sintáctico de selección, iniciado por Luhn (1958) al plantear la posibilidad de producir resúmenes automáticamente seleccionando del texto fuente frases con agrupaciones de «palabras significativas». El método era muy sencillo: cada agrupación recibiría una puntuación en consonancia con su número de palabras significativas a cada frase se le asignarían los puntos de su agrupación mejor valorada, y las frases cuya puntuación excediera un umbral preestablecido se extraerían para su inclusión en el resumen. La primera dificultad estribaba en como reconocer esas palabras significativas, aunque también se comprobó que la agrupación de palabras clave no era la única pista para el significado de las frases. Comparados con los resúmenes, los extractos derivados de estos procedimientos de extracción automática son de una categoría referencial inferior, pero ello no nos impide reconocer su eficacia en determinados contextos. Ya en esta etapa preliminar surgen las dificultades derivadas de los problemas lingüísticos de sinonimia, polisemia, anáfora, ...

Si tenemos en cuenta que los párrafos primero (segundo) y último (penúltimo) de un trabajo científico parecen disponer de elementos informativos suficientemente representativos, en respuesta a los fenómenos de localización de la atención característico del arranque textual y de epitomación informativa propio de su remate, comprenderemos que un extracto producido a partir de esos párrafos extremos será razonablemente representativo y podrá actuar como anuncio de la relevancia del documento origen. Este es un planteamiento extractor no exento de atractivo dada la simplicidad y economía de medios con que se concibe.

Fig. 3 Ejemplo de auto-resumen de Luhn, 1958.

SOURCE
The Scientific Anerica, Vol. 196, No. 2, pp. 68-94, February 1958
TITLE
Mesengers of the Nervous System
AUTHOR
Amodeo S. Marazzi
EDITOR´S SUB-HEADING
The internar communication of the body is mediated by chemicals as well as by nerve impulses. Study of their interaction has developed important leads to the undersanding and therapy of mental lllness.
AUTO-ABSTRACT
lt seems reasonable to credit the sinle-celled organism also with a system of chemical communication by diffusion of stimulating substances through the cell, and these correspond to the chemical messengers (eg., hormones) that carry stimuli from cell to cell in the more complex oganisms. (7.0)
Finally, in the vertebrate animals there are special glands (e.g., the -adrenals) for producing chemical messegers, and the nervous and chemical communication systems are intertwined: for instance, release of adrenalin by the adrenal gland is subject to control both by nerve impulses and by chemicals brought to the gland by the blood. (6.4)

The experiments clearly demonstrated that acetylcholine (and related substances) and adrenalin (and its relatives) exert opposing actions which maintain a balanced regulation of the tansmissions of nerve impulses. (6.3)
It is reasonable to suppose that the tranquilizing dugs counteract thel inhibitory effect of excessive adrenalin or serotonin or some related inhibitor in the human nervous system. (7.3)

Aunque no es fácil lograr una herramienta óptima que garantice la calidad de un buen extracto, destacamos algunos procedimientos relativamente útiles para determinar el contenido de un documento mediante la extracción de frases significativas. Una de las experiencias de mayor resonancia ha sido el sistema SMART, diseñado por Salton (11) como herramienta para la recuperación de información. Basado en el modelo de espacio vectorial, permite que tanto documentos como consultas de los usuarios sean representados por vectores o conjuntos de términos, bien en forma de palabras o de frases extraídas de los documentos una vez eliminadas las palabras vacías y los sufijos. Es importante establecer un sistema de ponderación ya que no todos los términos tienen el mismo valor en la representación del contenido. Se podría conocer el grado de similitud vectorial global entre los textos si representamos documentos y consultas mediante vectores de los términos ponderados y si verificamos que ciertas estructuras del texto (frases, segmentos, ...) se dan localmente en contextos similares. De esta forma se podría resolver ciertas ambigüedades lingüísticas. La ventaja del sistema SMART es que permite la descomposición y estructuración de los documentos, pues no solo usa textos completos sino también segmentos de longitud variable (secciones, párrafos, grupos de frases o frases sueltas). De esta manera se podrá establecer relaciones entre textos y partes de textos y generar mapas relacionales que muestren las similitudes de los textos que han superado un determinado valor. En teoría, partiendo del análisis del mapa y del grado de concentración/densidad de los nodos de información, podría ser fácil definir de manera automática el contenido básico del texto explorando selectivamente determinadas partes del mismo.
Aunque con los métodos de extracción textual no se logra producir resúmenes de calidad, si que se puede elaborar extractos que identifiquen colecciones de segmentos de textos en áreas temáticas de interés, siempre y cuando dispongamos de mapas conceptuales relacionales homogéneos. Cuando el mapa está desconectado, el proceso de análisis textual producirá extractos parciales, que se podrán mejorar si cogemos el párrafo inicial del documento principal, seguido del párrafo más adecuado a cada tema.
El sistema ANES (Sistema de extracción automática de noticias) (12) fue diseñado para realizar resúmenes sobre noticias periodísticas y se basa en la combinatoria de métodos estadísticos/heurísticos sobre las palabras de los documentos para determinar las frases más representativas. Mediante el análisis estadístico de un corpus documental se generan las frecuencias de los términos del documento y esta información se utiliza para calcular el peso de cada término y para determinar el identificador de cada palabra. Para la elaboración del resumen, ANES selecciona frases utilizando la lista de identificadores de palabras anteriormente generada. Los pesos de las palabras se calculan empleando las frecuencias de los términos en el documento siguiendo la fórmula tf* idf= frecuencia total del término/frecuencia del término en el documento, de manera que la importancia de

un término es proporcional a la frecuencia de ocurrencia en el documento e inversamente proporcional al número de documentos en que aparece. De esta forma los términos con frecuencias menores en el corpus tienen mayores pesos. El sistema selecciona las frases que contienen las ideas principales del documento mediante la suma de los pesos de los términos que contiene cada frase seleccionando aquellas con pesos mayores.

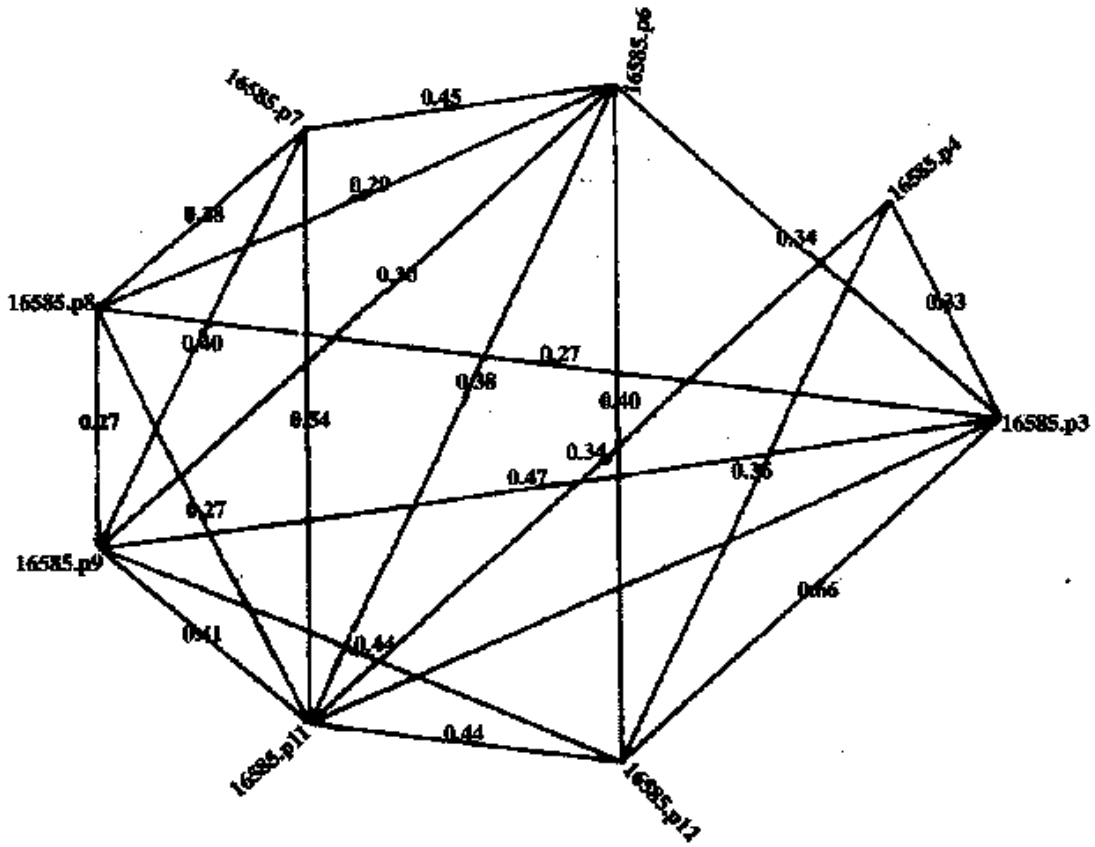


Fig. 4 Ejemplo de mapa relacional textual de Salton.

A pesar de no haberse resuelto en su totalidad la problemática extractora, se atreven algunos expertos a plantear un nuevo frente en la segunda etapa lógico/semántica de interpretación: mediante técnicas de inteligencia artificial se analiza el texto y se construye una representación semántica de su significado utilizando un conjunto de marcos adecuados al dominio de aplicación. Tras este análisis y representación se emplean plantillas de salida para generar el correspondiente sumario. En esta línea de procesamiento conceptual Niggemeyer plantea su modelo natural para sumar textos, basado en un sistema mixto resultante de la experiencia acumulada por resumidores profesionales en sus procesos de trabajo y de la aplicación de la metodología KADS (sistema experto de ingeniería del conocimiento) que emplea una arquitectura de pizarra para la representación cognitivo- técnica de la información. En base a conocidos postulados (Fidel, Creemins, Lancaster, Pinto, ...) y analizando detenidamente los pasos básicos de resumidores expertos, plantea la simulación de un modelo de resumen automático que trabaje emulando la forma de operar del resumidor

humano. El sistema, denominado SIM-SUM (13), es compatible tanto para aplicaciones Macintosh como Windows y describe el conjunto de herramientas automáticas empleadas en la elaboración de sumarios de documentos en dominios restringidos. Se apoya en la arquitectura cognitiva de pizarra que permite la interacción compleja de varios módulos simples, y en la aplicación de la teoría de la estructura retórica muy relacionada con el contenido de los documentos, al aportar pistas estructurales sobre el esquema conceptual y organizador de las distintas unidades que conforman un texto, asignando a las mismas una secuencias. El programa SIM- SUM consta de una estructura global de datos, varias fuentes de conocimiento y un componente de control. Pese a sus prestaciones, las aplicaciones reales siguen estando limitadas a dominios restringidos pues la base de conocimiento requerida para el correcto funcionamiento de estos sistemas debe ser necesariamente grande, compleja y específica del dominio de aplicación, por lo que parece haber pocas perspectivas cuando tratamos de ampliar el ámbito de actuación (15).
Estancados en esa segunda fase sumista, y con problemas no resueltos en la etapa extractora, los sistemas actuales no están en condiciones de acometer una nueva escalada a la cúspide de la síntesis/producción de resúmenes, aunque la investigación continúa.

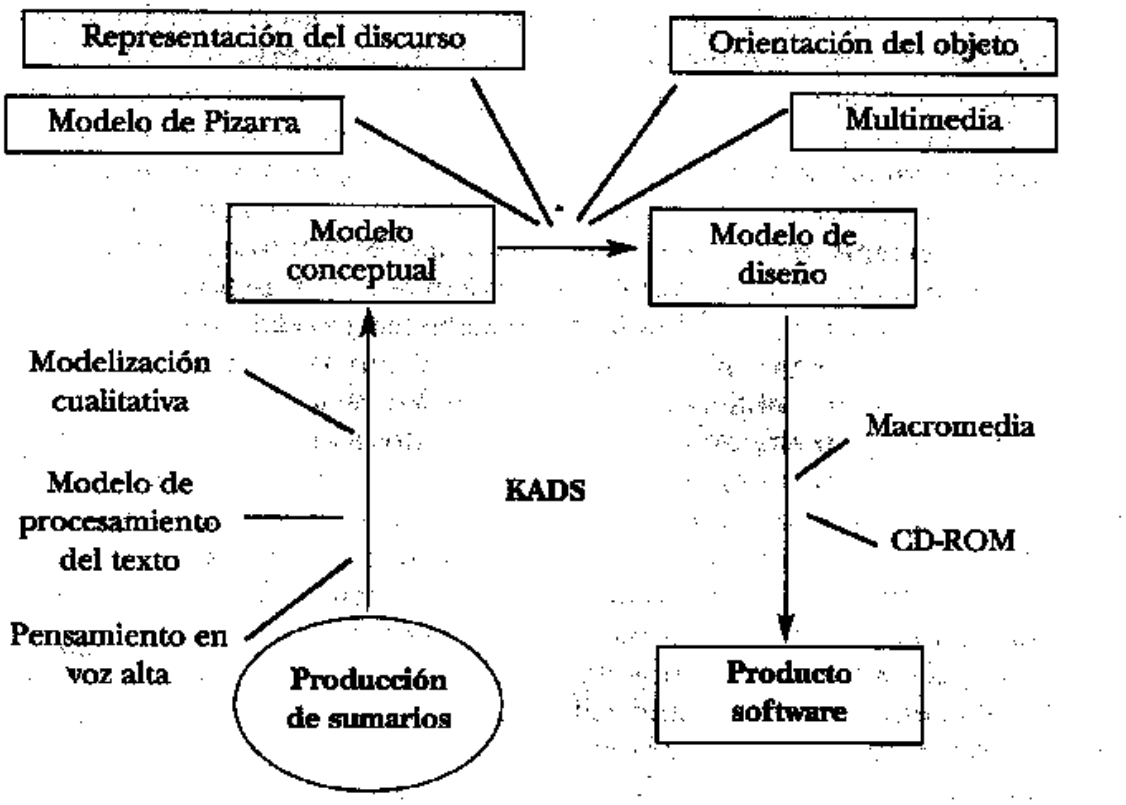


Fig. 5 Arquitectura del sistema SIM-SUM.
Fuente: Endres-Niggemeyer, B., Summarizing Information. Berlin, Springer-Verlang, 1998.

Un objetivo sensato a corto plazo en el que andan empeñadas algunas investigaciones es conseguir un sistema híbrido en el que ciertas tareas se realizaran por el resumidor humano y otras por el software (16).

4. LA RETÓRICA DEL RESUMEN CIENTÍFICO

Sabiendo que los humanos tenemos esquemas para un amplio rango de situaciones, la lingüística del discurso sugiere que tales esquemas también existen para los tipos de texto que participan en la comunicación compartida entre una determinada comunidad de usuarios (17). Los autores de determinados tipos de texto están condicionados por los esquemas de esa tipología textual y cuando escriben consideran no sólo el contenido específico a comunicar sino también la estructura usual para este tipo de texto. En su modo más abstracto y universal, la estructura retórica de los documentos científicos responde a la secuencia OMRC (objetivos, metodología, resultados y conclusiones).

Resúmenes estructurados

El concepto estructura se emplea con cierto grado de ambigüedad para referirnos al modo de relacionar e integrar un determinado conjunto de elementos. Desde nuestra perspectiva resumidora, y teniendo en cuenta la revolución informativo/documental en que nos hallamos inmersos, resulta obligado distinguir en los resúmenes, al igual que sucede con los textos en general, al menos dos territorios estructurales: el de sus estructuras estáticas, internas, o textuales, identificable con su configuraciones esquemático/retórica y conceptual; y el de su estructura externa, dinámica, o hipertextual, estrechamente vinculada a la más reciente noción de hipertexto. Nada tienen que ver, a priori, ambos modos de estructuración vinculados a dos conceptos plenamente distintos y que, lejos de ser incompatibles, se complementan mutuamente. Por eso cuando afirmamos que un resumen está estructurado, podemos referirnos tanto a su estructuración interna como a la integración de tal resumen en los modernos sistemas teleinformáticos, e incluso a ambas cuestiones simultáneamente. Lo que sí resulta evidente es que la estructuración de los textos, y por ende de los resúmenes, facilita su manipulación y favorece notablemente sus potencialidades informativas.

Desde el prisma textual el resumen científico es estructurado, pues debe responder a determinadas estructuras lógico/esquemáticas y conceptuales. Si somos rigurosos con el lenguaje comprenderemos que hablar de resúmenes estructurados es como aludir a «humareda de humo», lo que provoca confusión e incluso incertidumbre. Dando por hecho que todo resumen es estructurado, su estructuración puede ser retórica, cognitiva o la combinación de ambas. Desde un punto de vista retórico los resúmenes científicos se consideran estructurados cuando poseen una estructura que viene determinada desde una plantilla previamente establecida. En este caso el resumen es una sucesión de párrafos cuya coherencia está garantizada por la propia cohesión conceptual del dominio a que pertenece. Este tipo de resumen «aplantillado», cuyo nivel de textualidad (cohesión y coherencia) es o puede ser sensiblemente inferior al del resumen en «texto libre», tiene un alto grado de aceptación en los entornos automáticos, y se reconoce vulgarmente como resumen estructurado. Pero en el caso del documento científico, y por tanto de su resumen, se confunden a veces sus estructuras esquemática y conceptual o funcional. Si bien un análisis minucioso de la estructura funcional del resumen científico ha llevado a descubrir hasta ocho

tipos de funciones distintos (antecedentes, tema, método, resultados, ejemplos, aplicaciones, comparaciones y discusión) (18), lo clásico es la secuencia Objetivos, Metodología, Resultados y Conclusiones (OMRC) que preside el organigrama funcional/conceptual del texto científico. Y en este esquema conceptual se inspiran la inmensa mayoría de plantillas existentes para estructurar resúmenes en ciencias naturales, aunque parece que no será difícil adaptar este sistema al entorno de las ciencias sociales variando ligeramente los encabezamientos: Antecedentes, Objetivos, Metodología, Resultados, Conclusiones y Comentarios (19). En cualquier caso, los resúmenes estructurados, a pesar de la perdida de textualidad que conllevan, son un valor en alza en virtud de las facilidades que proporcionan a la hora de producirlos y de recuperarlos.

En un entorno hipertextual podemos afirmar que un resumen científico está estructurado cuando se integra en un entramado de relaciones interactivas internas (entre partes del mismo) y externas (con otros resúmenes dentro de una colección) que lo transforman en un hipertexto al que denominamos hiperesumen. En todo caso debemos reconocer la multiplicidad estructural del resumen que nos obliga a matizar cada una de sus estructuras (funcional, formal o hiperestructura) pues la diferencia es verdaderamente significativa.

Fig. 6 Resumen estructurado.

TITLE
Assessing the effectiveness of locally delyvered chlorhexidine in treatment of periodontitis.
AUTHOR (S)
Killoy-WJ
SPIRCE (BIBLIOGRAPHIC CITATION)
J-Am-Dent-Assoc. 1999 Apr.; 130(4): 567-70
ABSTRACT BACKGROUND
Several multicenter random clinical trials have studied a secondgereration easy-to-use chorhexidine local delivery systern to assess its effectiveness as an adjunct to oscaling and root planing, oor SRP. METHODS: The author reviews the pharmacodinetics of the loocal delivery system and two of the multicenter randomized clinical trials. One study evaluated 118 patients using splitarch design and the other stydy 447 patients using parallel design. All patients underwent SRP. Test sites, which had pocket depths of 5 millimeters of large, received a cholorhexidine chip (in both studies) or a placebo chip (the parallel stydy only). Test sites that remained 5 mm or larger were the chlorgexidine chip was used in conjunction with SRP than when SPR was used alone. (1. 16 mm vs. 0.7 mm, P<.0001, in the split-arch-desing study and 0.95 mm vs. 0.65 mm, P=.00001, in the parallel-desingn study). CONCLUSIONS: Use of the chlorhexidine chips has significantly iinproved rhe clinical parameters or periodontitis when used as an adjunct to S. CLINICAL IMPLICATIONS: When used with SRP, the chiorbexidine chip offers the elinical a new method of achiccing and maintaining perio- dontal stability.

5. PERSPECTIVAS DEL RESUMEN EN LA ERA DIGITAL

Admitiendo lo inadecuado que resulta el resumen tradicional, inteligente (texto libre), en los nuevos entornos digitales, justo es reconocer el rendimiento de los programas informáticos en el procesamiento de los textos como objetos físicos. Los distintos modelos selectivos basados en la estadística textual generan

distintos tipos de extractos, simples o estructurados, que satisfacen algunas de las propiedades exigibles al resumen tradicional, pues mediante estos sistemas se pueden lograr algunas buenas cualidades como brevedad, legibilidad y precisión. Si bien estas características garantizan un cierto grado de eficacia en el almacenamiento y recuperación documental, el caballo de batalla sigue residiendo en otras propiedades mucho más alusivas, como son la equivalencia cognitiva, la cohesión y la exhaustividad de los resúmenes. Se investigan mediante programas «inteligentes» basados en los aspectos sociocognitivos de los documentos, modelos interpretativos y productivos que pudieran garantizar estos atributos sin necesidad de recurrir, como sucede en la actualidad, a importantes restricciones retóricas o dominicales. En este sentido el resumen informativo (no estructurado, o en texto libre) es la forma de representación más problemática y esquiva, pero también la más poderosa porque puede capturar la estructura argumentativa de los documentos así como las palabras clave que proporcionan un cuadro global de su contenido (20). Asumiendo que ningún sistema es lo suficientemente eficaz si no admite la posibilidad de su renovación, una metodología para resumir debe estar abierta a previsibles (y no previsibles) desarrollos, uno de los cuales es el manejo de los resúmenes en texto libre, aunque no tengamos claro qué tipo y tamaño es el que resulta más eficaz.

De todos modos, probablemente pase mucho tiempo antes de que las máquinas sean lo suficientemente inteligentes para sustituir completamente a los humanos en estas importantes tareas (21).

En un determinado espacio informativo, la aplicación simultánea de las distintas técnicas y modos de resumir sobre un mismo documento generará distintos resúmenes y otras formas documentales afines. El reto para cualquier sistema de intermediación documental será potenciar la posterior interpretación y reconstrucción de esta múltiple documentación referencias en función de las necesidades y el espacio mental del usuario. No olvidemos que sin consultar a los usuarios reales y estudiar sus opiniones y comportamiento, la investigación sobre el resumen corre el riesgo de transformarse en una actividad parroquiana e inútil (22), pues sólo ellos pueden decir que necesitan estos microtextos en sus diferentes contextos operativos. En cualquier caso, ahora mas que nunca los documentos científicos precisan referentes representativos que faciliten su búsqueda y recuperación en un entorno ilimitado, inteligente, dinámico e interactivo. Precisamente este contexto permite otras formas de representación documental, los microtextos interactivos, cuyos prometedores niveles de eficacia están todavía por descubrir.

REFERENCIAS

Capítulo 5 de: Procesamiento de la información científica. Madrid: Arco/Libros, 2001, pp. 103-142.

(1) ENDRES-NIGGEMEYER, B., Summaiizing information. Berlin, Springer-Veriag, 1998.

(2) PINTO, M., “Competencias del Traductor de Textos Literarios desde la Perspectiva Documental”. En: Terminologie et Traduction. 1999, 3, 99-111.

(3) FOSTER, J., “On the Interpretative Authority of Information Systems”. En: WILSON and ALLEN (eds.), Exploring the Context of Information Behaviour London, Taylor Graham, 1999, 506-518.

(4) PINTO, M., “Documentary Abstracting: Toward a Methodological Model”. En: Joumal of the American Society for Information Science. 1995, 46, 3, 225-234.
(5) VAN DIJK, T., La ciencia del texto. Barcelona: Paidos, 1978.
(6) HUTCHINS, W. J., “Information Retrieval and Text Analysis”. En: T. A. VAN DIJK, (ed.), Discourse and Communication. New A pproaches to the Analysis of Mass Media Discourse and Communication. New York, Walter de Gruyter, 1985,106-125I.
(7) MONDAY, l., Les Processus Cognitifs et la Redaction de Resumes. Documentation et Biblioteques, Av-Jun, 1996, 55-63.
(8) KINTSCH, W.; Van DijK, T. A,. “Toward a Model of Text Comprehension and Production”. En: Psychological Review. 1978, 85, 5, 363-394.
(9) JACOB, E.; SHAW, D., “Sociocognitive Perspectives on Representation”. En: Annual Review of Infmation Science and Technology (ARIST), 1998, 33, 131-185.
(10) ALLEN, B. L., “Visualization and Cognitive Abilities». En: ATHERTON and JOHNSON (e,ds.), Visualizing Subject Acces for 21st Century Information Resources. Ilinois University, 1998.
(11) SALTON, G.; ALLAN, J.; BUCKLEY, C.; SINGHAL, A., “Automaüc Analysis Theme Generation, and Summarization of machine-readable Texts”. En: SPARCK JONES, K; WILLET, P., Reading in Information Rdiieval. San Francisco: Morgan Kauffman, 1997.
(12) BRANDOW, R.; MITZE, Y,; RAU, L., “Automatic condensation of electronic publication by sentence selection”. En: Information Processing and Management, 1995, 31, 5, 675-685.
(13) ENDRES-NIGGEMEYER, B., Summarizinginformation. Berlin: Springer-Verlag, 1998. (14) MANN, W. C.; THOMPSON, S.A., “Rhetorical Structure Theory: A Theory of Text Organization”. Technical Report ISl/RS-87-190,June 1990.
(15) PAICE, C.; JONES, P., “The Identification of important concepts in highly structured technical papers”. ACM-SIGIR´93, Pittsburgh, 1993,69-78.
(16) CRAVEN, T. C., “Abstracts produced using computer assistance». En: Journal of the American Society for Information, 2000, 51, 8, 745-756.
(17) LIDDY, E., “The discourse-level structure of empirical abstracts: an exploratory study». Information Processing and Management, 1991, 27, 1, 55-81.
(18) MAEDA, T., “An approach toward functional text structure analysis of scientific and technical documents”. En: Information Processing and Management, 1981, 17, 6, 329-339.
(19) HARTLEY, J.; SYDES, M., Structured abstracts in the social sciences. presentation, readability and recall. British Library R&D Report 6211, 1995.
(20) TiBBO, H. R., “Abstracting across the disciplines: a content analysis of abstracts from the natural sciences, the social sciences and the humanities with implications for abstracting standards and online information retrieval”. En: Library and Information Science Research, 1992, 14, 31-56.
(21) LANCASTER, F. W., Indexing ans Abstracting in Theory and Prractice. 2a. ed. Urbana- Champaign: University of Illinois, Gradu ate School of Library and Information Science, 1998.
(22) WHEATLY, A.; ARMSTRONG, C. J., A Survey of the Content and Characteristics of ELectronic Abstracts. London: Library Information Technology Centre, 1997.

NC ISO 5963:
MÉTODOS PARA EL ANÁLISIS DE DOCUMENTOS, DETERMINACIÓN DE SU CONTENIDO Y SELECCIÓN DE LOS TÉRMINOS DE INDIZACIÓN

1 ALCANCE

1.1 Esta norma recomienda procedimientos para el análisis de documentos, determinación de su contenido y selección de los términos de indización. Se imita a las primeras etapas de la indización y es independiente de la práctica del sistema de indización, ya sea precoordinado o poscoordinado. Describe métodos generales de análisis de documentos que deben aplicarse en cualquier situación. Sin embargo, estos métodos están destinados especialmente a los sistemas de indización en los que las materias de los documentos se expresan de forma abreviada con ayuda de los términos de un lenguaje de indización controlado. En este contexto, el lenguaje controlado está constituido generalmente por un subconjunto de términos extraídos del lenguaje natural y estructurado, por ejemplo, mediante un tesauro. Estos métodos se pueden aplicar, normalmente a sistemas en los que, con fines de recuperación, los conceptos se representan por símbolos seleccionados del esquema de clasificación.

1.2 Las técnicas descritas en esta norma se pueden utilizar por cualquier organismo en el que se empleen indizadores humanos para analizar temáticamente el documento y expresar el contenido en forma de términos de indización. No se aplican en instituciones que utilizan técnicas de indización automáticas en las que los términos existentes en los textos se organizan en conjuntos o clases según criterios que pueden aplicarse mediante un ordenador, por ejemplo, por frecuencia de aparición y/o adyacencia en el texto, aunque la finalidad de estos sistemas sea la misma.

1.3 Esta norma debe, en primer lugar, servir de guía a los indizadores para las etapas de análisis de los documentos e identificación de los conceptos. Puede aplicarse también a la búsqueda documental para transformar las peticiones de los usuarios en términos de indización controlados. También puede servir como guía para la elaboración de resúmenes analíticos, teniendo presente, sin embargo, que estas tareas, aunque análogas, no son idénticas.

1.4 Esta norma se destina a promover la utilización de una práctica normalizada

a) en una institución o una red de centros o instituciones;
b) en diferentes servicios de indización, en especial en aquellos que intercambian registros bibliográficos.

2 TÉRMINOS Y DEFINICIONES

Para los fines de esta norma, se aplican las definiciones siguientes:

2.1 Documento

Cualquier fuente de información, impresa o no, que se pueda catalogar o indizar.

NOTA: Esta definición se refiere no sólo a los materiales escritos e impresos en papel o microforma (por ejemplo, libros, revistas especializadas, diagramas, mapas, etc.), sino también a medios no impresos (por ejemplo, registros legibles por ordenador, películas y grabaciones sonoras) ya objetos de colección.

2.2 Noción o concepto

Una unidad de pensamiento. El contenido semántico de un concepto puede ser ex-presado por una combinación de otros diferentes que pueden variar de un idioma a otro.

2.3 Materia

Cualquier concepto o combinación de conceptos que representa el tema de un documento.

2.4 Término de indización

La representación de un concepto en forma de:

- un término derivado del lenguaje natural, preferiblemente un sustantivo simple o compuesto.
- un código de clasificación.

NOTA: Un término de indización puede constar de más de una palabra. En un lenguaje de indización controlado, un término se designa como descriptor o no-descriptor.

2.5 Descriptor

Término usado siempre, en la indización, para representar un concepto dado, conocido también como "término preferente".

2.6 No-descriptor

Sinónimo o cuasi sinónimo de un descriptor. Los no-descriptores no se asignan a los documentos pero pueden servir como puntos de entrada en un índice, dirigiendo al usuario mediante una instrucción (por ejemplo: ver o véase) al descriptor. Se llaman, también "términos no preferentes".

2.7 Índice

Lista alfabética y sistemática de materias que señala el lugar en que se encuentra cada materia en un documento o en una colección de documentos'.

2.8 Indización

Acción de describir o identificar un documento en relación con su contenido.

3 PROCESO DE INDIZACIÓN

3.1 La indización no concierne a la descripción de un documento como entidad física (por ejemplo, no indica la forma, editor, fecha, etc.), aunque estos factores pueden estar incluidos en un índice de materias si esta información puede permitir a un usuario determinar, de forma más precisa, si un documento dado es relevante para su necesidad de información.

3.2 Durante la indización los conceptos se extraen del documento mediante un proceso de análisis intelectual y después se transforman en términos de indización. Tanto el análisis como la transcripción deben realizarse con ayuda de herramientas de indización, como tesauros y sistemas de clasificación.

3.3 La indización consiste esencialmente en tres etapas, que tienden a solaparse en la práctica:

a) examen del documento y determinación de su contenido;
b) identificación y selección de los conceptos principales del contenido;
c) selección de los términos de indización.

Cada una de estas etapas, junto con un capítulo sobre el control de calidad, son tratadas en los capítulos 4a 7.

4 EXAMEN DEL DOCUMENTO

4.1 La precisión con que se puede examinar un documento depende en gran manera de su forma física. Se pueden distinguir dos casos diferentes: documentos impresos y documentos no impresos.

4.2 Los documentos impresos constituyen el material habitual de las bibliotecas y centros de documentación cuyo fondo consiste principalmente en libros, revistas, informes, actas de congresos, etc. De forma ideal la compresión completa de estos documentos requiere su lectura detallada. Sin embargo, una lectura completa es a menudo impracticable y no siempre necesaria, pero el indizador debe asegurarse de que no se ha descuidado ninguna información útil. Las partes importantes del texto deben examinarse cuidadosamente y se debe prestar especial atención a las siguientes:

a) título;
b) resumen, si lo tiene;
c) sumario o tabla de contenido;
d) introducción, párrafos iniciales de los distintos capítulos o apartados y conclusiones;
e) ilustraciones, diagramas, tablas y su leyenda o explicación;
f) palabras o frases que están destacadas mediante una tipografía diferente o subrayadas.

Todos estos elementos deben examinarse cuidadosamente por el indizador durante el estudio del documento. No se recomienda la indización a partir del título solamente, y el resumen si existe, no se debe considerar como un sustituto del examen del texto. Hay títulos que pueden inducir a errores; existen resúmenes que son insuficientes y ni los unos ni los otros constituyen una fuente segura del tipo de información que necesita el indizador.

4.3 Los documentos no impresos, tales como medios audiovisuales, visuales y sonoros, requieren procedimientos diferentes. En la practica no siempre es posible examinarlos en su totalidad. La indización hay que realizarla en estos casos a partir del título y de la sinopsis o reseña. Sin embargo si estos son inadecuados o insuficientes, el indizador debe visualizar o escuchar el documento.

5 IDENTIFICACIÓN DE LOS CONCEPTOS

5.1 Después de examinar el documento el indizador debe identificar las nociones que son elementos esenciales de la descripción del contenido. Las instituciones que patrocinan la realización del índice deben establecer los factores que se consideran importantes en el campo temático cubierto por el índice.

Algunas cuestiones que ilustran los ejemplos de criterios a retener son:

a) ¿Trata el documento de algún objeto sometido a una acción?
b) ¿Contiene algún concepto activo? (por ejemplo, una acción, un procedimiento, etc.)
c) ¿Se ve afectado el objeto por la acción identificada?
d) ¿Trata del agente causante de la acción?
e) ¿Se describen los medios para llevar a cabo la acción? (por ejemplo, instrumentos, técnicas o métodos especiales)
f) ¿Existen factores considerados en un medio o lugar particular?
g) ¿Se identifican variables dependientes o independientes?
h) ¿Se trata el tema desde un punto de vista particular no asociado normalmente a ese campo? (por ejemplo, estudio de la religión desde un punto de vista sociológico).

Estos son ejemplos de criterios susceptibles de aplicación en muchos campos; en disciplinas particulares puede ser necesario formular otras cuestiones.

5.2 El indizador no tiene necesariamente que utilizar como términos de indización todos los conceptos identificados durante el examen del documento. La selección o el rechazo de conceptos depende de la finalidad con que se van a utilizar los términos de indización, que pueden variar desde la producción de Indices alfabéticos impresos hasta la creación de una base de datos bibliográfica informatizada. La identificación de conceptos puede también estar influida, como se indicó anteriormente, por el documento a indizar. Por ejemplo la indización de libros puede diferir de la de artículos de revistas.

Las dos características de un índice más afectadas por la selección de los términos de indización son la exhaustividad y la especificidad.

5.3 La exhaustividad esta relacionada con el número de conceptos que se tienen en cuenta, y que caracterizan el contenido íntegro de un documento.

5.3.1 Un indizador que sigue los procedimientos indicados antes, debe poder identificar todos los conceptos de un documento, que tienen valor potencial para los usuarios de un sistema de información. En algunos casos, en un mismo documento, se presentan independientemente dos o más temas dentro del campo cubierto por la indización. En ese caso los temas deben tratarse de forma separada y, si es necesario, por diferentes especialistas.

5.3.2 La cobertura de la indización no debe interpretarse de una forma demasiado estricta. Hay que tener en cuenta qué términos de indización creados inicialmente para un grupo de usuarios (por ejemplo, científicos y técnicos) pueden utilizarse por otros grupos (por ejemplo, economistas). Se aconseja que los indizadcres de literatura científica y técnica tengan presentes Otros aspectos del tema, en particular, los sociales y económicos.

5.3.3 El principal criterio de selección de conceptos debe ser su valor potencial como elemento de expresión del tema del documento para su recuperación.

En la selección de conceptos, el indizador debe tener en mente las preguntas que se pueden hacer al sistema de información, en la medida en que dichas preguntas se pueden conocer. En efecto, este criterio constituye la principal función de la indización. Dentro de este contexto el indizador debe:

a) elegir las nociones mas apropiadas para un grupo de usuarios dado, sin perder de vista el objetivo de la indización;
b) modificar, si es necesario, tanto las herramientas como el procedimiento de indización, como resultado de las preguntas hechas al sistema. Dichas modificaciones no deben producir distorsión de la estructura o de la lógica del lenguaje de indización.

5.3.4 El número de términos o descriptores que se pueden asignar a un documento no debe limitarse de forma arbitraria. Debe determinarse enteramente por la cantidad de información contenida en el documento en relación con las necesidades supuestas de los usuarios a que va destinado el índice. Imponer un límite arbitrario puede conducir a una pérdida de objetividad en la indización y a una deformación de la información que se podrá utilizar en la recuperación.

Si por imposiciones establecidas es necesario limitar el número de términos, la selección de conceptos debe ser guiada por el juicio del indizador sobre el papel de cada término para expresar el contenido total del documento.

5.4 La especificidad está relacionada con la exactitud con que un concepto particular que aparece en un documento está representada por un término de

indización. Se produce una pérdida de especificidad cuando un concepto particular está representado por un término que tiene un significado más general.
Las nociones deben identificarse de la forma más específica posible. Sin embargo, pueden preferirse nociones más generales en los casos siguientes:

a) cuando el indizador considere que un exceso de especificidad puede actuar de forma negativa sobre el sistema de indización. (Por ejemplo, puede decidir que un modelo muy específico de una máquina se represente por el término más genérico de ese tipo de máquinas, en especial cuando esas nociones aparecen sólo en áreas muy restringidas del campo temático cubierto por el índice);
b) cuando se trate de una idea no completamente desarrollada, o de la que se hace sólo una alusión por el autor, estará justificada la indización a un nivel más general.

6 SELECCIÓN DE LOS TÉRMINOS DE INDIZACIÓN

6.1 Cuando los conceptos se traducen en términos de indización, el indizador debe observar las reglas siguientes:
a) Los conceptos ya presentes en el lenguaje de indización deben retenerse como descriptores.
b) Los términos que representan nuevos conceptos deben comprobarse, en cuanto a su exactitud y su aceptación, con ayuda de obras de referencia tales como:

- diccionarios y enciclopedias, de autoridad reconocida en la materia en cuestión;
- tesauros;
- clasificaciones temáticas.

Se puede también consultar a especialistas en la materia, prefiriéndose aquellos que tienen conocimientos de indización y documentación.
6.2 El indizador debe estar familiarizado con estas obras y con las limitaciones que presentan, por ejemplo una lista de encabezamientos de materia o un esquema de clasificación puede no permitir la representación exacta de un concepto encontrado en un documento. Si los conceptos están representados por códigos de clasificación, necesita saber que estos códigos designan generalmente un contexto más amplio o más restringido, que puede no ser completamente apropiado para el documento estudiado.
6.3 Si un lenguaje de indización incorpora un tesauro, el número de términos asignados al documento y la multiplicidad de entradas, pueden reducirse sin perdidas, ya que los términos generales y otras relaciones pueden establecerse con el propio tesauro.
Cuando se utiliza un tesauro debe seleccionarse el término más específico existente para representar un concepto dado.

6.4 Algunos sistemas de indización utilizan indicadores de función de enlace, de ponderación, etc. El indizador debe estar familiarizado con todas las reglas asociadas con el uso de estos mecanismos.
6.5 En la práctica el indizador encontrará con frecuencia conceptos que no existen en ningún tesauro o esquema de clasificación. Según el sistema utilizado dichos conceptos deberán tratarse de diferentes formas, por ejemplo:
a) expresarlos por términos o descriptores y añadirlos inmediatamente al lenguaje de indización.

b) representarlos temporalmente por términos más generales, y proponerlos como candidatos para una adición posterior.

7 CONTROL DE CALIDAD DE LA INDIZACIÓN

7.1 La calidad y la coherencia de la indización dependen de factores tales como:

a) la competencia del indizador;

b) la calidad de los instrumentos de indización.

En una situación ideal, los términos de indización asignados a un documento y el nivel de exhaustividad conseguido son idénticos con cualquier indizador. Para un mismo sistema de indización estos factores deben mantenerse relativamente estables en el tiempo. La coherencia es un factor importante en el comportamiento de un sistema de indización, en especial cuando la información se va a intercambiar entre diferentes centros de una red documental.

7.2 La imparcialidad total del indizador es un factor necesario para conseguir la consistencia de la indización. Un juicio subjetivo en la identificación de los conceptos y en la elección de los términos de indización, afectarán inevitablemente al comportamiento del sistema de indización. La consistencia es más difícil de conseguir con un equipo de indización formado p~r muchos miembros o cuando la indización se lleva a cabo por equipos de indizadores que trabajan en lugares diferentes, por ejemplo en un sistema descentralizado. En estas situaciones, se recomienda una etapa de comprobación centralizada con devolución a los indizadores.

7.3 El indizador debe tener un buen conocimiento del campo de que tratan los documentos a indizar. Debe comprender los términos que se encuentran en los documentos y las reglas y procedimientos del lenguaje de indización específico. Los centros que manejan documentos en lenguas extranjeras deben disponer de especialistas en esas lenguas.

7.4 La calidad de la indización se podrá conseguir de manera más efectiva silos indizadores tienen contacto directo con los usuarios. Estos podrían, por ejemplo, determinar si ciertos términos o descriptores son susceptibles de producir combinaciones falsas, dando lugar a salidas no pertinentes.

7.5 La calidad de la indización depende también de la posibilidad de poner al día el lenguaje de indización. Es esencial que el sistema permita la introducción de nuevos términos en el lenguaje o cambios en la terminología que respondan a nuevas necesidades de los usuarios.

7.6 Cuando sea posible, debe comprobarse la calidad de la indización, analizando los resultados de la recuperación de documentos, por ejemplo, calculando los porcentajes de exhaustividad y de precisión.

NC ISO 214:

RESÚMENES PARA PUBLICACIONES Y DOCUMENTACIÓN

1 ALCANCE

Esta Norma internacional presenta lineamientos para la preparación y presentación de resúmenes de documentos. Se hace énfasis en los resúmenes preparados por los autores de documentos primarios, y en su publicación, debido a que tales resúmenes pueden ser de ayuda a los lectores de estos documentos y, al propio tiempo, se puede reproducir en publicaciones y servicios secundarios sin cambio o con cambios mínimos. La guía básica es útil también para la preparación de resúmenes por personas que no sean autores y facilita la presentación de dichos resúmenes en publicaciones y servicios secundarios.

2 DEFINICIONES

En esta Norma Internacional, el término resumen significa una representación abreviada y precisa (exacta) del contenido de un documento, sin añadirle interpretaciones o crítica y sin distinción en cuanto a quien escribió el resumen.

Un resumen debe ser tan informativo como lo permita el tipo y el estilo del documento, es decir, debe presentar, de acuerdo con las posibilidades, toda la información cuantitativa y cualitativa contenida en el documento. Los resúmenes informativos son especialmente apropiados para textos que describen trabajo experimental y documentos dedicados a un solo tema. Sin embargo, se pueden preparar resúmenes indicativos de algunos documentos de gran extensión como trabajos de reseña o monografías completas. Un resumen indicativo es una guía descriptiva del documento que cubre las materias principales y de la forma que se tratan los hechos y datos. Un resumen informativo indicativo combinado se debe preparar frecuentemente cuando las limitaciones sobre la extensión del resumen o por el tipo y estilo del documento y se haga necesario ajustar la exposición a los elementos primarios del documento y relegar otros aspectos a exposiciones indicativas. Ver los ejemplos del 1 al 3.

No debe confundirse los resúmenes con términos relacionados pero diferentes: anotación, extracto y sumario. Una anotación es un breve comentario o explicación acera de un documento o su contenido o hasta una breve descripción, que se añade usualmente como nota después de las citas bibliográficas del documento. Un extracto es una o más secciones o partes de un documento, seleccionadas para representar el total del mismo. Un sumario, si se necesita, es una pequeña reiteración, dentro del propio documento (usualmente al final), de sus hallazgos y conclusiones más sobresalientes y se utiliza para completar la orientación al lector que ha estudiado el texto precedente. (Debido a que otras partes del documento, por ejemplo objetivos, metodologías no se condensan en este tipo de sumario, el término no debe emplearse como sinónimo de resumen y el resumen tal como se define anteriormente no debe ser llamado sumario y si se usa el sumario, este no debe duplicar ni incluir todo el alcance de un resumen).

3 OBJETIVOS Y USO DE LOS RESÚMENES

3.1 Determinación de la relevancia

Un resumen bien preparado permite a los lectores identificar el contenido básico de un documento de manera rápida y precisa, determinar su relevancia respecto a sus intereses en particular y así, decidir si necesitan leer el documento completo.

3.2 Evitar la lectura del texto completo del documento de interés colateral

Los lectores con interés colateral en el documento obtienen suficiente información en los resúmenes y no necesitan leer el documento completo.

3.3 Utilidad en la búsqueda automatizada de textos completo

Los resúmenes son de gran utilidad para la recuperación y la alerta informativa automatizada.

3.4 Utilización en documentos primarios específicos

Las recomendaciones siguientes son para autores y editores de documentos y publicaciones específicas, tales como revistas, informes y tesis, monografías y proceedings y patentes.

3.4.1 Revistas

Se debe incluir un resumen en cada artículo de revista, ensayo o discusión. Las notas, comunicaciones breves, editoriales, y "cartas al editor", que tengan un contenido sustancial técnico o académico, deben llevar también un resumen breve.

3.4.2 Informes y tesis

Se debe incluir un resumen en cada informe, folleto o tesis publicado por separado.

3.4.3 Monografías y proceedings

Un solo resumen puede ser suficiente para un libro o monografía que trate un tema homogéneo, sin embargo, se necesita también un resumen por separado para cada capítulo si el libro cubre diferentes tópicos o si es una colección de trabajos realizados por distintos autores.(por ejemplo, los proceedings de un simposio o de una reunión), Véase ejemplo 4.

3.4.4 Patentes

Cada patente o solicitud debe estar acompañada de un resumen, conforme a las reglas del país que expide la aceptación o de la organización internacional.

3.5 Utilización en publicaciones y servicios secundarios

Las publicaciones y servicios secundarios pueden a menudo hacer uso literal de los resúmenes que aparecen en los documentos primarios, si los mismos han sido preparados cuidadosamente y no están sujetos a las restricciones del derecho de autor.Tales resúmenes de autor pueden también brindar base adecuada para un servicio secundario que oriente sus resúmenes hacia un grupo de usuarios diferentes de aquellos hacia los que apunte la bibliografía. Un resumen totalmente nuevo sólo se escribirá, usualmente, cuando aspectos breves o subordinados del documento correspondan al área cubierta por la publicación secundaria.

3.6 Utilización en fichas de documentación

Las fichas de documentación pueden ser preparadas convenientemente o separadas de las hojas de resúmenes de revistas y proceedings, que incluyen y presentan adecuadamente tales páginas de resumen. Además cuando las fichas de documentación acompañan documentos tales como los informes, estas fichas deben llevar los resúmenes que esos documentos contienen.

4. Tratamiento del contenido del documento

En numerosas disciplinas los lectores se han acostumbrado a un resumen que establece el objetivo, la metodología, los resultados y las conclusiones que se presentan en el documento original. La mayoría de los documentos que describen trabajos experimentales pueden ser analizados de acuerdo con estos elementos, pero su secuencia óptima puede depender de aquellos para los cuales el resumen se ha confeccionado. Los lectores interesados en aplicar nuevos conocimientos obtendrán una información más rápida a partir de un resumen con una disposición orientada hacia los resultados, en el que estos y las conclusiones más importantes aparezcan al principio, seguidos por los detalles que los apoya, otros resultados y la metodología. Véase parte Ay B del ejemplo 5.

Las reglas siguientes son óptimas para resúmenes informativos. Los redactores de resúmenes informativo-indicativo e indicativo solamente, deberán seguirlas en la medida en que les resulten prácticas.

4.1 Objetivo

Establezca los objetivos principales y el alcance del estudio o las razones por la cual el documento fue escrito, a no ser que se encuentren implícitas claramente en el título del documento, o se puedan inferir del resto del resumen. Refiérase a literatura anterior sólo si es parte esencial del objetivo.

4.2 Metodología

Describa las técnicas o propuestas sólo hasta el grado necesario para que sean comprendidas. Sin embargo, identifique las nuevas técnicas claramente y describa el principio metodológico básico, el límite de la operación y la precisión o exactitud obtenida. En documentos referidos a trabajos no experimentales, describa las fuentes de los datos y su manipulación.

4.3 Resultados y conclusiones

Se deben presentar claramente. Se deben resumir juntos para evitar redundancia, pero se deben distinguir las conjeturas de los hechos.

4.3.1 Resultados

Describa los hallazgos tan concisa e informativamente como sea posible. Los resultados obtenidos pueden ser experimentales o teóricos, los datos recopilados, las relaciones y correlaciones notables, los efectos observados, etc. Aclare si los valores numéricos están sin clasificar o son derivados y si son la resultante de una observación única o mediciones repetidas. Cuando los hallazgos sean tan numerosos que no se puedan incluir todos, de prioridad a lo siguiente: eventos nuevos y verificados, hallazgos de un valor a largo plazo, descubrimientos significativos, hallazgos que contradigan teorías previas o descubrimientos que el autor conozca que son relevantes para un problema práctico. Se deben indicar los límites de precisión y fiabilidad, así como los intervalos de validez.

4.3.2 Conclusiones

Describa las implicaciones de los resultados y especialmente como estos están referidos al objetivo de la investigación o a la preparación del documento. Las conclusiones se pueden asociar con recomendaciones, evaluaciones, aplicaciones, sugerencias, nuevas relaciones e hipótesis aceptadas o rechazadas.

4.4 Información colateral

Incluya hallazgos o información incidental referida al objetivo principal del documento, pero con valor fuera de la materia fundamental (por ejemplo, modificaciones de métodos, nuevos compuestos, constantes físicas determinadas recientemente y documentos de fuentes de datos recién descubiertos). Infórmelo

claramente, pero de modo tal que no desvíen la atención sobre el tema principal. No exagere su importancia relativa en el documento resumido.

5 PRESENTACIÓN Y ESTILO

5.1 Ubicación del resumen

Coloque el resumen (al menos uno en el idioma original del documento) al principio de cada documento, siempre que sea posible.

En una revista, el resumen debe aparecer en un lugar bien visible en la primera página de cada artículo u otro documento resumido, preferentemente entre su título y la información sobre el autor y el texto. Se recomienda incluirlo en una "hoja de resumen" preparada.

En un informe publicado separadamente, coloque el resumen en la portada (si es posible), en la "página de informe de documentación" (si la hay) o en la página de la derecha que precede a la tabla de contenido.

En un libro, monografía o tesis, coloque el resumen en el reverso de la portada o en la página de la derecha a continuación de ésta. Ubique resúmenes separados de los capítulos en la primera página de los mismos o en la página que los antecede.

5.2 Información bibliográfica

En publicaciones primarias, incluya una información bibliográfica del documento en la misma página del resumen y en una posición adecuada, por ejemplo en el encabezamiento o al final. En publicaciones secundarias o cuando el resumen se publica separado del documento, coloque la información bibliográfica antes o después del resumen. En el ejemplo 6 se dan 3 variantes de este caso.

5.3 Fichas de documentación

Es muy conveniente la presentación del resumen y sus referencias bibliográficas en un formato adaptable a las fichas de documentación. Es preferible el empleo de cartulina para las páginas de resumen, como para las fichas de documentación que acompañen al documento, pero si el resumen está impreso en el mismo papel que la publicación, debe estarlo por una sola cara, de forma que pueda recortar y pegar en fichas en blanco. Las dimensiones de las partes impresas no deben exceder los 64mm X 95mm.Para permitir la utilización de fichas de 74mm X 105mm o fichas de 75mm X 125mm. (formato internacional de las fichas catalográficas para bibliotecas).

5.4 Exhaustividad, precisión y extensión

El lector debe ser capaz de leer el resumen sin referirse al documento original, por tanto haga el resumen independiente. Retenga la información básica y el tono del documento original. Sea tan conciso como pueda, pero llene los requerimientos en cuanto a contenido, no sea críptico (misterioso, enigmático) u obscuro). Diga los antecedentes someramente. No incluya afirmaciones ni aseveraciones no contenidas en el documento original.

Para la mayor parte de los trabajos y parte de monografías, un resumen de menos de 250 palabras es adecuado. Las notas y comunicaciones cortas serán menores de 100 palabras. Los editoriales y cartas al editor sólo requerirán una línea. En documentos largos, como informes y tesis, el resumen debe contar con menos de 500 palabras y ser suficientemente corto como para caber en una página. Los contenidos de un documento son a menudo más significativos que su longitud y esto se debe tener en cuenta al determinar la extensión del resumen.

5.5 Estilo

Comience el resumen con una oración temática que sea exposición de la materia central del documento, a no ser que esta ya haya sido bien establecida en el título del documento que precederá al resumen. Los resúmenes especialmente escritos y modificados para uso secundario, mencionan el tipo de documento al principio del resumen cuando esto no es evidente en el título o editor del documento, o no queda claro en el contenido del documento. Explique el tratamiento que el autor da a la materia o la naturaleza del documento, por ejemplo, tratamiento teórico, historia, informe sobre el estado de la técnica, revisión histórica, informe de investigación original, encuesta de literatura, etc.

5.5.1 Párrafos, oraciones completas

Escriba un resumen con un solo párrafo, pero en los resúmenes largos utilice más de un párrafo. Escriba el resumen en oraciones completas, especialmente en los resúmenes informativos, utilice palabras de transcripción y frases a los efectos de la coherencia. El texto del resumen puede ser seguido por un grupo de palabras clave para la indización (separadas por una puntuación determinada) o sustituidas por estas cuando se desee emplear un resumen indicativo.

5.5.2 Utilización de verbos en voz activa y de pronombres personales

Siempre que sea posible utilice verbos en voz pasiva, ello contribuye a un escritura clara, corta y enfática. Sin embargo, la voz pasiva puede ser utilizada en expresiones indicativas e informativas en las cuales se debe enfatizar el receptor de la acción.

Ejemplo:

Diga: "gasolinas endulzadas en presencia del aire por bauxitas contentivas de hierro"

No: "las gasolinas son endulzadas por bauxita contentivas de hierro en presencia del aire"

Pero: "Fueron medidas coeficientes de absorción relativa de éter, agua y acetileno".

Utilice la tercera persona a no ser que la primera persona evite una construcción oscura de las oraciones y conduzca a mayor claridad.

5.5.3 Terminología

Utilice palabras significativas del texto que puedan ayudar a la búsqueda automatizada. Evite términos extraños, acrónimos, abreviaturas o símbolos o defínalos la primera vez que aparezcan en el resumen. Utilice unidades, símbolos y terminología ISO, siempre que sea posible, en su defecto emplee normas nacionales.

5.5.4 Material no textual

Incluya tablas coartas, ecuaciones, fórmulas estructurales y diagramas sólo Cuando sea necesario para mayor brevedad y claridad cuando no exista otra alternativa.

ANEXO

(informativo)

EJEMPLOS DE RESÚMENES

EJEMPLO I - Resúmenes informativos típicos

Alteración de grasas usadas en fritura. Correlación entre índices analíticos *y métodos de evaluación directa de compuestos de degradación*

En este trabajo se estudian las posibilidades de utilización de índices metílicos simples y rápidos para medir la alteración producida en las grasas de fritura1 en comparación con métodos cromatográficos que evalúan directamente los compuestos nuevos originados en el proceso.
Se han analizado 140 muestras divididas previamente en 3 grupos, según su origen (aceites termoxidados, grasas procedentes de freidoras industriales y aceites procedentes de freidoras domésticas). A partir de los resultados obtenidos de acidez libre, punto de humo, prueba colorimétrica de Perevalov, glicéridos polares, esteres metílicos polares y dímeros no polares, se han calculado las correlaciones entre las distintas determinaciones para cada grupo de muestras y para todas las grasas de fritura.
Los resultados demuestran que los índices elegidos son aplicables cuando existen valores bien definidos para los parámetros implicados en el proceso de fritura, mientras que es necesario utilizar un método cromatográfico para la evaluación de muestras de historia desconocida.

Anodización del aluminio. Adición de acetato *sódico en el baño de sellado*

Se estudia la respuesta de los ensayos de control de calidad del sellado de recubrimientos anódicos sellados en agua con adición de acetato sódico. Los ensayos de control son el de la medida de la admitancia del recubrimiento, la gota de colorante, inercia a la disolución química en medio fosfocrómico y, como ensayo adicional, se determina la relación de sellado. La respuesta de estos ensayos se estudia empleando distintos tiempos de sellado y se compara con los resultados obtenidos cuando éste se lleva a cabo en agua destilada sin adiciones y en las mismas condiciones de trabajo. Las conclusiones muestran que los cuatro ensayos de control suministran una respuesta más favorable cuando se emplea la adición de acetato sódico. Los resultados de una serie de ensayos de corrosión atmosférica llevados a cabo en una estación urbana-industrial durante doce años, ponen de manifiesto una degradación del aspecto superficial y pequeñas picaduras cuya evaluación diferencial se verificará posteriormente.

EJEMPLO 2 - Resúmenes informativo - indicativos típicos

Lenguaje y esquizofrenia

Se revisan los trastornos del lenguaje en la Esquizofrenia, señalando los problemas derivados de los criterios diagnósticos específicos, de los efectos de la institucionalización y de las formas aguda versus crónica, A pesar de los diversos problemas metodológicos, el estudio de las anomalías del lenguaje en la Esquizofrenia, en los últimos veinte años, ha revelado en este síndrome un defecto en la organización formal del lenguaje en los niveles semántico y léxico del discurso, un déficit en la comprensión y una correspondencia entre los trastornos del pensamiento, anomalías estadísticas en la frecuencia relativa de determinados elementos lingüísticos y anomalías en la motilidad. Se establecen diferencias en relación con los trastornos afásicos y con otros trastornos

psiquiátricos Se estudia, finalmente, la desviación de la comunicación familiar como factor de riesgo para el comienzo de la Esquizofrenia.

Determinación de los parámetros energéticos en un proceso de molienda de escombros

Se describe un método de laboratorio para el estudio de los parámetros característicos de los procesos de molienda. El método se aplica, en este caso, al estudio de la molienda de minerales lateríticos procedentes de los yacimientos existentes en la región occidental de la República de Cuba. El estudio se realiza a partir de experiencias en un molino estándar y la aplicación de la Tercera Teoría de Bond, lo que permite obtener los parámetros energéticos del proceso (Work Index y Potencia Consumida) que, posteriormente, se utilizan en el dimensionado de lo molinos industriales. El método descrito consiste, básicamente en estudiar la influencia del tiempo de residencia y de masa de cuerpos moledores en la generaci6n de un producto molido de características previamente establecidas y en optimizarlos a partir de los resultados experimentales y de las correlaciones establecidas según la Tercera Teoría de Bond.

Determinación colorimétrica de lantánidos totales en aceros inoxidables

Se presenta un método sensible sin separaciones para la determinaci6n colorimétrica de lantánidos total es. El complejo rosa-carmfn con Arsenazo III cumple la ley de Beer entre 1 ug/50 ml y 40 p9150 ml. El pH de formación del complejo es de alrededor de 1 y la longitud de onda de mjxima absorbencia es de 650 nm en solución acuosa que contenga 60%-70% de etanol. Se discuten los efectos de los diferentes metales que normalmente forman la matriz de acero inoxidable. El método posee buena selectividad y puede aplicarse a la determinación espectrofotométrica di recta de lantánidos en aceros inoxidables.

EJEMPLO 3- Resumen indicativo típico

Determinación de antimonjo en aceros par espectrometría de emisión óptica y fluorescencia de rayos X (dispersión de energías)

Debido al creciente interés por conocer la concentración de antimonio en aceros, dada su influencia en las propiedades mecánicas de los mismos, se ha puesto a punto un método para la determinación de dicho elemento en aceros, en una amplia gama de concentraciones, mediante la técnica de fluorescencia de rayos X (dispersión de energías) y contraste de la misma con la espectrometría de emisión óptica).

EJEMPLO 4 - Resumen de monografías y capítulos *Biología de Artemia*

Biología de Artemia

La importancia que ha adquirido Artemia en el cultivo de peces y crustáceos marinas, como presa viva o alimento primordial de sus formas larvarias, ha contribuido a que en general el conocimiento de Artemia se centrara especial mente en su forma naupliar. Este trabajo intenta dar a conocer con más amplitud una serie de fenómenos biológicos que se producen antes y después de la aparición del nauplio, ofreciendo con ello una visión más amplia de las posibilidades que ofrece, tanto en el ámbito de la acuicultura, como en el de las ciencias básicas y experimentales que requieran un instrumenta de trabajo manejable y útil.
Tras una consideración taxonómica aclaratoria sobre la conveniencia de emplear exclusivamente la denominación genérica de Artemia, se hace una sencilla descripción morfológica que facilitará el análisis diversificador de las distintas cepas o razas autóctonas de nuestra área geográfica, al tiempo que las diferencias de la cepa original de la bahía de San Francisco, California (USA) y la forma de Artemia mis ampliamente conocida, hasta ahora, en el mundo. La posibilidad de estudiar un buen numero de cepas americanas procedentes de distintas localidades geográficas de ambos continentes, permite también llevar a cabo unas reflexiones ligadas a su distribución geográfica o latitudinal.
A continuación se hace una breve descripción del género a través de los diferentes sistemas biológicos funcionales, prestando mayor atención al sistema reproductor que es el que muestra algunos de los detalles más espectaculares de su biología, entre ellos las enormes potencialidades del quiste o huevo de resistencia. Sigue una descripción de los estados larvarios que llevan hasta el adulto reproductor.
Especial atención merecen las diversas y extremas condiciones fisicoquímicas que caracterizan lo hábitats propios de Artemia en su amplia distribución geográfica que dan paso, finalmente, a unas reflexiones sobre su biogeografía, centrada principal mente en la península ibérica y en su carácter ejemplificador y casi sintetizador de la diversidad propia del área mediterránea a su vez compendio del amplio continente euroasiatico.

EJEMPLO 5- Orden de los elementos del resumen

A Resumen informativo *con un orden convencional de los elementos (objetivo, metodología, resultados y conclusiones)*

Automatización de bibliotecas mediante tratamiento por lotes. Sistema creado en la Facultad de Informática de la Universidad politécnica de Madrid

Se presente el sistema de gestión automatizada del catálogo adoptado por la biblioteca de la Facultad de Informática de la Universidad Politécnica de Madrid, diseñado y realizado íntegramente por personal de la misma. Como característica más destacada del sistema se encuentra el que se realiza por lotes, debido a que el ordenador utilizado sólo admite este tipo de operación.
Se expone el análisis de las necesidades a cubrir, tanto las relativas a la organización del fondo bibliotecario como las de edición de productos impresos de carácter periódico y las de obtención de catálogos e índices generales del fondo completo.

Se indican las características del ordenador utilizado (UNIVAC 9400) y el formato de entrada de los datos Se describe el funcionamiento del sistema y los productos obtenidos, tanto los generados periódicamente (fichas catalográficas, tejuelos y listados de nuevas adquisiciones) como los de carácter acumulativo (catálogo maestro e índices de materiales, de autores y de títulos permutados).
Se concluye indicando el interés del sistema descrito como solución pare pequeñas bibliotecas que sin realizar grandes inversiones pueden llegar a tener informatizado su catálogo utilizando el tratamiento por lotes en ordenadores de menor capacidad que los requeridos pare tratamiento en línea.

B *Resumen informativo con un orden de los elementos orientado hacia los resultados (principales resultados y conclusiones, detalles complementarios, otros resultados y metodología)*

Autorización de bibliotecas mediante tratamiento por lotes. *Sistema creado en la Facultad de Informática de la Universidad Politécnica de Madrid*

Automatización de catálogos para pequeñas bibliotecas por un sistema de gestión automatizada utilizando el tratamiento por lotes en ordenadores de menor capacidad que los requeridos para tratamiento en línea sin realizar grandes inversiones.
Se presenta el sistema de gestión automatizada del catálogo adoptado por la biblioteca de la Facultad de Informática de la Universidad Politécnica de Madrid, diseñado y realizado íntegramente por personal de la misma. Como característica más destacada del Sistema se encuentra el que se realiza por lotes, debido a que el ordenador utilizado sólo admite este tipo de operación.
Se expone el análisis de las necesidades a cubrir, tanto las relativas a la organización del fondo bibliotecario como las de edición de productos impresos de carácter periódico y las de obtención de catálogos e índices generales del fondo completo.
Se indican las características del ordenador utilizado (UNIVAC 9400) y el formato de entrada de los datos, se describe el funcionamiento del sistema y los productos obtenidos, tanto los generadores periódicamente (fichas catalográficas, tejuelos y listados de nuevas adquisiciones) como los de carácter acumulativo (catálogo maestro e índices de materias, de autores y de títulos permutados).

C Resumen indicativo del mismo documento. Este tipo de resumen se ha incluido aquí solamente para demostrar la utilidad de preparar Un resumen informativo, cuando el documento lo permita, como se ha indicado en el apartado 2

Automatización do bibliotecas mediante tratamiento por lotes. sistema creado en la Facultad de Informática de la Universidad Politécnica de Madrid

Se presenta el sistema de gestión automatizada del catálogo adoptado por la biblioteca de la Facultad de Informática do la Universidad Politécnica de Madrid, diseñado y realizado íntegramente por personal de la misma. Se describe el funcionamiento del sistema y los productos obtenidos, tanto los generados periódicamente (fichas catalográficas, tejuelos y listado de nuevas adquisiciones)

como los de carácter acumulativo (catálogo maestro e índices de materias, de autores y de títulos permutados).

EJEMPLO 6 - Diferentes formas de colocar la referencia bibliográfica para los resúmenes que aparecen en las publicaciones secundarias

A *Resumen secundado precedido de la referencia bibliográfica completa. Aunque este orden es el* convencional *el acceso del lector a la información es más lento ya que los títulos de los documentos suelen estar orientados al objetivo del trabajo mis quo a los resultados*

Rodríguez I., Plaza C., Lagunilla M. AUTOMATIZACIÓN DE BIBLIOTECAS MEDIANTE TRATAMIENTO POR LOTES. SISTEMA CREADO EN LA FACULTAD DE INFORMATICA DE LA UNIVERSIDAD POLITECNICA DE MADRID. Rev. esp. Doc. cient. vol.5, no 2, 1982:165-179. Se presenta el sistema do gestión automatizada del catálogo adoptado por la biblioteca de la Facultad de Informática de la Universidad Politécnica do Madrid, diseñado y realizado íntegramente por personal de la misma. Como característica mas destacada del sistema se encuentra el que se real iza por lotes, debido a que el ordenador utilizado solo admite este tipo de operación.
Se expone el análisis de las necesidades a cubrir, tanto las relativas a la organización del fondo bibliotecario como las de edición de productos impresos de carácter periódico y las de obtención de catálogos e índice generales del fondo completo.
Se indican las características del ordenador utilizado (UNIVAC 9400) y el formato do entrada do los datos. Se describe el funcionamiento del sistema y los productos obtenidos, tanto los generados periódicamente (fichas catalográficas, tejuelos y listado do nuevas adquisiciones) como las de carácter acumulativo (catálogo maestro e índices de materias, de autores y de títulos permutados).
Se concluye indicando el interés del sistema descrito como solución para pequeñas bibliotecas que sin realizar grandes inversiones pueden llegar a tener informatizado su catálogo utilizando el tratamiento por lotes en ordenadores de menor capacidad que los requeridos para tratamiento en línea.

B *Resumen secundado seguido de la referencia bibliográfica completa. Esta forma permite la presentación inmediata al lector do los principales resultados del documento, y resulta particularmente adecuada para la colocación de los elementos del resumen orientada hacia los resultados (ejemplo 5 B). Para facilitar el acceso rápido a la referencia bibliográfica se puede destacar ésta utilizando una diferencia de márgenes de distinta tipografía o ambos métodos*

Automatización de catálogos para pequeñas bibliotecas por un sistema de gestión automatizada utilizando el tratamiento por lotes en ordenadores de menor capacidad que los requeridos para tratamiento en *línea sin realizar grandes inversiones*

Se presenta el sistema de gestión automatizada del catálogo adoptado por la biblioteca de la Facultad de Informática de la Universidad Politécnica do Madrid, diseñado y realizado íntegramente por personal de la misma. Como característica más destacada del sistema se encuentra el que se realiza por lotes, debido a que el ordenador utilizado sólo admite este tipo de operación.
Se expone el análisis de las necesidades a cubrir, tanto las relativas a la organización del fondo bibliotecario como las de edición de productos impresos de carácter periódico y las de obtención de catálogos e índices generales del fondo completo.
Se indican las características del ordenador utilizado (UNIVAC 9400) y el formato de entrada de los datos. Se describe el funcionamiento del sistema y los productos obtenidos, tanto los generados periódicamente (fichas catalográficas, tejuelos y listado de nuevas adquisiciones) como los de carácter acumulativo (catálogo maestro e índices de materias, de autores y de títulos permutados).
Rodríguez J., Plaza C., Lagunilla M. AUTOMATIZACION DE BIBLIOTECA MEDIANTE TRATAMIENTO POR LOTES. SISTEMA CREADO EN LA FACULTAD DE INFORMATICA DE LA UNIVERSIDAD POLITECNICA DE MADRID.
Rev. esp. Doc. Cient vol.5, no 2, 1982:165-179.

C *Resumen secundado precedido por el título del documento, pero con indicación de la referencia bibliográfica al final del texto. Esta forma ofrece al lector el tema del documento tal como lo ha presentado el autor y le brinda a continuación la información sobre el contenido del documento. Para facilitar el acceso rápido a la referencia bibliográfica se puede destacar ésta utilizando una diferencia de márgenes, distinta tipografía, o ambos métodos*

Automatización de bibliotecas mediante tratamiento por lotes. sistema creado en la Facultad de Informática de la Universidad Politécnica de Madrid

Se presenta el sistema de gestión automatizada del catálogo adoptado por la biblioteca de la Facultad de Informática de la Universidad Politécnica de Madrid, diseñado y realizado íntegramente por personal de la misma. Como característica más destacada del sistema se encuentra el que se realiza por lotes, debido a que el ordenador utilizado sólo admite este tipo de operación.
Se expone el análisis de las necesidades a cubrir, tanto las relativas a la organización del fondo bibliotecario como las de edición de productos impresos de carácter periódico y las de obtención de catálogos e índices generales del fondo completo.
Se indican las características del ordenador utilizado (UNIVAC 9400) y el formato de entrada de los datos. Se describe el funcionamiento del sistema y los productos obtenidos, tanto los generados periódicamente (fichas catalográficas, tejuelos y listados de nuevas adquisiciones) como los de carácter acumulativo (catálogo maestro e índices de materiales, de autores y de títulos permutados).
Se concluye indicando el interés del sistema descrito como solución para pequeñas bibliotecas que sin realizar grandes inversiones pueden llegar a tener informatizado su catálogo utilizando el tratamiento por lotes en ordenadores de menor capacidad que los requeridos para tratamiento en línea.

Rodríguez J., Plaza C., Lagunilla M. Rev. esp. Doc. cient. vol.5, n02, 1982:165-179.