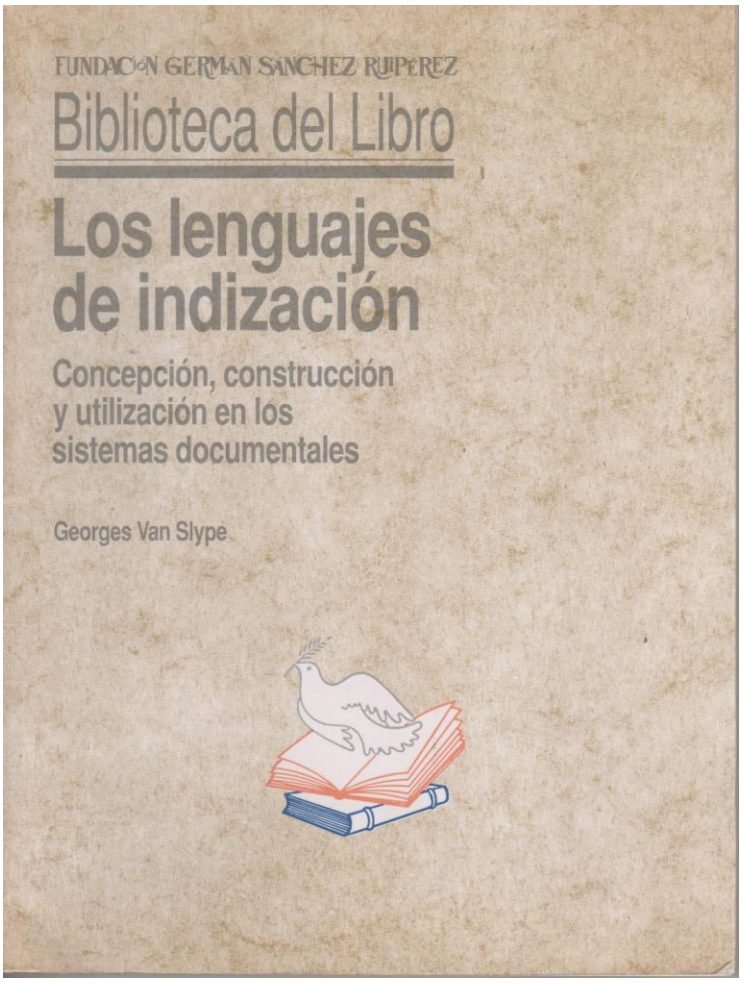
# Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales

Georges van Slype

Pedro Hípola, Félix de Moya (versión española)




http://www.ugr.es/~phipola/Los_lenguajes_de_indizacion.pdf
http://www.amazon.es/Los-lenguajes-indizaci%C3%B3n-construcci%C3%B3n-documentales/dp/8486168600

Georges VAN SLYPE
Doctor en Economía Aplicada
Profesor Asociado de la Universidad Libre de Bruselas
Profesor Asociado de la Universidad de Génova
Director del Bureau Marcel van Dijk,
Ingenieros-consultores en Gestión
Bruselas-París-Londres-Frankfurt

# LOS LENGUAJES DE INDIZACIÓN:

## concepción, construcción y utilización en los sistemas documentales

Traducido por:

Pedro Hípola y Félix de Moya
E. U. de Biblioteconomía y Documentación de Granada

## Prefacio a la traducción española

Los thesaurus construidos y utilizados por empresas privadas y administraciones públicas, nacionales e internacionales, se cuentan por millares en Europa y América. Parece que en España queda por recorrer algún camino para que se generalice el uso de lenguajes controlados de indización.

Agradecemos a la Escuela Universitaria de Biblioteconomía y Documentación de la Universidad de Granada, y más en particular a los profesores Pedro Hípola y Félix de Moya, su iniciativa de traducir al español y publicar la presente obra. Esperamos que esta publicación pueda contribuir a presentar a los lectores, en lengua española, las características de los lenguajes de indización, libres y controlados, a hacer apreciar sus respectivas cualidades, y a exponer las modalidades prácticas de construcción y utilización.

Georges van Slype

## Introduccion a la edicion francesa

La historia de los lenguajes de indización, si bien es breve, está en ebullición, lo cual es síntoma de la rápida evolución de los métodos de almacenamiento y búsqueda de información documental.

Al principio existían los lenguajes de clasificación, que se remontan a la más lejana antigüedad (piénsese, por ejemplo, en la primera clasificación de los conocimientos establecida desde el siglo IV a. C. por Aristóteles) y que habían encontrado su consagración a fines del último siglo, con la construcción de esos monumentos que son la Clasificación Decimal Universal, la Clasificación de Dewey y la Clasificación de la Biblioteca del Congreso.

La aparición de los primeros thesaurus, a principios de los años sesenta, provoca primeramente una especie de guerra de religión, que nosotros personalmente hemos vencido: las grandes clasificaciones provenían de la experiencia acumulada de muchas generaciones de bibliotecarios y ¡resulta que unos documentólogos iconoclastas se permitían poner en cuestión esta arcana pericia!

Hicieron falta algunos años para que los espíritus se apaciguaran y se comprendiera que los dos grandes tipos de lenguajes controlados tenían su lugar:

-los lenguajes de clasificación:

- en las bibliotecas enciclopédicas: para la clasificación de monografías, es decir, para la representación sintética de un tema dentro de los catálogos de materias, y a veces para su clasificación, en sistemas de libre acceso, de acuerdo con las grandes ramas del saber;
- en las bibliotecas especializadas, servicios de documentación y entidades productoras de boletines bibliográficos: para la ordenación de los documentos analíticos (artículos de revistas, comunicaciones a congresos, informes de investigación...) por medio de las entradas de materias de los boletines analíticos y signaléticos;

-los thesaurus: en los servicios de documentación y entidades productoras de boletines de índices, luego en las de bases de datos bibliográficas, para la indización de los documentos, es decir, para la representación analítica de su contenido conceptual por medio de una serie de descriptores, con vistas al almacenamiento y búsqueda de información documental.

Hacia mediados de los años sesenta, un nuevo cambio de escena: ¡el lenguaje natural! ¿Por qué destinar grandes recursos para construir costosos thesaurus y para que los documentalistas indicen documentos, si resulta que basta con almacenar en el ordenador los títulos y los resúmenes (y más tarde los textos completos) de los documentos y realizar la búsqueda documental por medio de las palabras significativas (las palabras clave) contenidas en ellos?

En esto también hicieron falta algunos años para comprender que lenguajes controlados (los thesaurus) y lenguajes libres (las listas de palabras clave) pertenecen al mismo conjunto (los lenguajes de indización) y juegan un papel complementario, más que antagonista:

-el thesaurus, gracias a su concisión, a su falta de ambigüedad y a la posibilidad de ser transferido de una lengua a otra, permite gestionar las búsquedas documentales con una gran precisión, pero a veces en detrimento de la exhaustividad;
-por su parte, la abundancia de lenguaje libre en los títulos, resúmenes y textos permite escapar a las restricciones a veces demasiado rigurosas del thesaurus, y puede asegurar una mayor exhaustividad, al menos en la lengua del que realiza la búsqueda, en detrimento, eso sí, de la precisión.

Así, la lista de palabras clave se añadió al thesaurus de descriptores y al sistema de clasificación, dentro del abanico de los lenguajes documentales.

En el transcurso de los años setenta se produjo una cierta agitación dentro del pequeño mundo de la industria de la información documental: una serie de thesaurus, que habían sido elaborados con grandes esfuerzos, ¡no encontraban quien los usara! Un rápido análisis de la situación mostraría que esa falta de interés no tenía que ver precisamente con los thesaurus. ¿Qué es lo que sucedía? Que determinadas autoridades, nacionales e internacionales, habían decidido crear los thesaurus, es decir, unos instrumentos, con el fin de promover la creación de sistemas de información sectorial... ¡en aquellos sectores en los que los protagonistas no los querían!

Al principio de los ochenta se asiste a una evolución extremadamente curiosa:

-por una parte, el desarrollo considerable de la búsqueda documental a través del acceso en línea a los distribuidores públicos de bases de datos documentales. Ahora bien, los responsables de los centros distribuidores habían tenido como preocupación prioritaria hasta el momento la de rentabilizar sus inversiones vendiendo la mayor cantidad posible de horas de conexión y de referencias suministradas en respuesta a las consultas. Por lo tanto, se interesaban en primer lugar por la exhaustividad de la búsqueda (y en consecuencia, por la búsqueda en lenguaje libre), descuidando un poco la precisión (y por tanto, el uso en línea de los thesaurus de las bases de datos, las cuales habían sido indizadas, sin embargo, por medio de descriptores controlados);
-por otra parte, el desarrollo no menos importante de una serie de bases de datos documentales internas, dentro de las empresas y en la Administración, casi todas basadas en una indización en lenguaje controlado (por un thesaurus «local») y en lenguaje libre: ¡nunca se han construido tantos thesaurus en las organizaciones como en estos últimos cinco años!;

-por último, la aparición de sistemas de indización, automática o asistida, que en algunos casos responden al principio mismo del lenguaje controlado, mientras que en otras ocasiones se basan en un thesaurus. Hasta el momento, la penetración de tales sistemas en el mercado es insignificante. Hoy los esfuerzos se concentran en la aplicación de los sistemas expertos a los sistemas documentales. Al igual que sucedió con los anteriores cambios de escena en el mundo de los lenguajes documentales, parece que los sistemas expertos aportarán una evolución, y no una mutación: los thesaurus constituirán uno de los pilares del sistema de almacenamiento y recuperación documentales: la base de conocimientos, que contendrá la lista de los conceptos evocados en los documentos y en las consultas, bajo una forma normalizada; un segundo pilar, el motor de inferencia, explotará el thesaurus para pasar las peticiones, expresadas en lenguaje libre, a una formulación de las ecuaciones de búsqueda en lenguaje controlado, y posteriormente a la extracción de los documentos pertinentes.

En pocas palabras, los lenguajes de indización, en general, y los thesaurus de descriptores y las listas de palabras clave, en particular, son instrumentos utilizados como tales, según lo que es habitual en casi todos los sistemas documentales de hoy día, o están integrados dentro de instrumentos más sofisticados, como podría llegar a ser la práctica de los sistemas documentales del mañana.

El autor de esta obra, así como la empresa de consultores en la que trabaja desde hace más de veinticinco años, han jugado un papel nada despreciable en esta evolución: nosotros hemos dirigido la elaboración del primer thesaurus multilingüe del mundo (el de la D.I.R.R.: Documentation Internationale de Recherches Routières) desde 1963.

Hemos contribuido ampliamente a difundir uno de los métodos de representación de thesaurus (el de diagramas de flechas, preparado por los autores del thesaurus de E.U.R.A.T.O.M.).

Hemos estado relacionados con la concepción, elaboración, mantenimiento y utilización de decenas de thesaurus y listas de autoridades, mono o multilingües, dentro de los más variados ámbitos de las ciencias humanas y exactas, para empresas privadas y para organismos públicos, nacionales e internacionales.

Hemos realizado, bajo contrato, estudios sobre las características de los thesaurus existentes y sobre las funcionalidades de los programas informáticos de ayuda a la construcción de thesaurus.

Hemos organizado, en fin, coloquios sobre el estado de la cuestión en materia de concepción y utilización de thesaurus. El último se celebró en los locales de la Comisión de las Comunidades Europeas, en Bruselas, durante marzo de 1986.

Hemos pensado que esta experiencia acumulada al cabo de los años debíamos ponerla a disposición de la comunidad de

lengua francesa, que es la nuestra, publicando esta obra sobre la práctica de los lenguajes de indización.

Por otra parte, hemos solicitado a D. Jacques Maniez, profesor del I.U.T. de Dijon, que es una personalidad de nivel universitario y con competencia igualmente reconocida en materia de lenguajes documentales:

-en primer lugar, que prepare una obra sobre los fundamentos teóricos de los lenguajes documentales y sobre la práctica de los lenguajes de clasificación; esa obra será publicada en los próximos meses por la misma editorial, dentro de la misma colección y bajo el mismo título colectivo;

-en segundo lugar, que revise el manuscrito de la presente obra: se ha ocupado de ese trabajo con mucha minuciosidad y nos ha presentado bastantes sugerencias, que nosotros hemos podido aprovechar para mejorar la coherencia y legibilidad de nuestro texto. Le agradecemos profundamente esa preciosa colaboración.

Debemos agradecer así mismo a D. Marcel van Dijk su aliento para la publicación de este libro y a D. Jacques Chaumier su revisión de nuestro manuscrito y sus consejos en el ámbito terminológico.

CAPITULO I

# CONCEPCION DE LOS LENGUAJES DE INDIZACION

## 1. **Definición**

Como muchos términos de la lengua, la palabra «lenguaje» tiene varias acepciones, según los contextos en que aparece:

-para el antropólogo y el lingüista, representa «la función de expresión del pensamiento y de comunicación entre los hombres, realizada por medio de un conjunto de signos vocales (habla) y a veces de signos gráficos (escritura), que constituye una lengua» (Petit Robert 1985);
-para el especialista (informático, estenógrafo..., documentalista) que utiliza una lengua convencional para un uso particular, el lenguaje designa «todo sistema secundario de signos creado a partir de una lengua» (Petit Robert 1985).

Nosotros retendremos aquí esta segunda definición, que se corresponde mejor con el universo que se aborda en esta obra: el del documentalista.
Un lenguaje documental será entonces todo sistema de signos que permita representar el contenido de los documentos con el fin de recuperar los documentos pertinentes en respuesta a consultas que tratan sobre ese contenido. El lenguaje documental no se refiere, pues, a otros criterios utilizados en la búsqueda documental: autor del documento, lengua del texto, fecha de publicación...

Existen dos tipos principales de lenguajes documentales:

-los lenguajes de indización, denominados también lenguajes combinatorios, que permiten representar el contenido de los documentos y de las consultas de forma analítica;
-los lenguajes de clasificación, utilizados más generalmente para representar este contenido de forma sintética.

Por ejemplo, un artículo de cinco páginas de una revista científica o técnica será representado, en general, por:

-un lenguaje de clasificación:
  +de 1 a 3 encabezamientos de materia tomados de un sistema de clasificación;

-un lenguaje combinatorio:
  +de 8 a 12 descriptores tomados de un thesaurus;
  +algunas decenas de palabras no vacías, tomadas de su título y de su resumen, o unas mil palabras no vacías tomadas de su texto completo.

La aproximación analítica propia del lenguaje de indización permite representar el contenido de los documentos y de las consultas a un nivel:

-o bien de los conceptos que son tratados en los documentos o de las informaciones que se buscan por medio de las consultas;
-o bien de las palabras no vacías contenidas en el título, el resumen y a veces en el texto de los documentos, y en el enunciado de las consultas;

mientras que la representación del contenido por medio de un lenguaje de clasificación se realiza a nivel del tema del documento o de la consulta.

*Nota:* Esta definición se aparta de muchos autores, para los que un lenguaje documental es necesariamente un lenguaje artificial; lo cual es cierto para el caso de los sistemas de clasificación y los thesaurus, pero no lo es para las listas de palabras clave extraídas de los títulos, resúmenes y textos de los documentos.

Para nosotros, que nos dedicamos ante todo a la práctica de la documentación, una definición operativa debe fundamentarse sobre esta teoría:

-las palabras clave se utilizan en la búsqueda documental de acuerdo con los mismos principios (post-coordinación, búsqueda booleana) que los descriptores;
-los diversos tipos de lenguajes combinatorios que nosotros describimos (cf. § 3) forman un continuum, desde las listas de palabras, que constituyen los sistemas menos «controlados», hasta los thesaurus de descriptores, que son los sistemas más «controlados».

## 2. **Principio: la post-coordinación**

El principio de funcionamiento de un lenguaje de indización es la indización coordinada.

La indización se define como la actividad que consiste en representar el contenido de un documento o de una consulta de forma analítica, es decir, enumerando los conceptos y/o las palabras.

Cuando se utiliza un lenguaje combinatorio, se dice que la indización es coordinada, en el sentido de que los conceptos y/o las palabras utilizadas para representar el contenido de los documentos podrán, en el momento de la búsqueda documental, ser libremente combinados entre sí para formular las consultas que permitirán recuperar esos documentos.

Por ejemplo, un documento:

-que trata sobre los métodos de detección de ruido radio-eléctrico galáctico por medio de radiotelescopios terrestres sobre satélites;
-indizado: radiotelescopio; observatorio terrestre; estación espacial; detección; ruido radio-eléctrico; galaxia;

-podrá ser recuperado en respuesta a una consulta realizada sobre una combinación cualquiera de los conceptos arriba citados; por ejemplo:
+radiotelescopio y ruido radio-eléctrico;
+estación espacial, radiotelescopio y galaxia,
+observatorio terrestre y estación espacial,
+ruido radio-eléctrico.

Como se puede ver, la coordinación entre los elementos que constituyen la indización se hace a posteriori, en el momento de la indización y de la interrogación, y no a priori, en el momento de la construcción del lenguaje documental, como es el caso de los lenguajes de clasificación. Por este motivo, se dice que la indización a través de un lenguaje combinatorio se efectúa siguiendo el principio de la post-coordinación.

## 3. **Tipología**

Enumeramos a continuación los grandes tipos de lenguajes de indización (§ 3.1). Luego pasamos revista a sus principales características (§ 3.2). Por último describimos su utilización efectiva en los sistemas documentales actuales (§ 3.3).

### .1 Tipos de lenguajes de indización

La tipología de los lenguajes combinatorios se basa esencialmente en el nivel de normalización de su terminología.
Se distingue entre:

-los lenguajes libres, que se constituyen «a posteriori», sobre la base de la indización en lenguaje natural de documentos ya registrados en una colección;
-los lenguajes controlados, construidos «a priori», antes de empezar a indizar los documentos de una colección;
-los lenguajes codificados.

### .1 Lenguaje libre

Existen dos tipos principales de lenguajes combinatorios libres:

-las listas de palabras clave;
-las listas de descriptores libres;

### .1 Lista de palabras clave

Una lista de palabras clave está constituida por una *colección no ordenada* (sino puesta por orden alfabético) *de las palabras significativas, denominadas también no vacías* (es decir, todas la palabras que no son artículos, conjunciones, pronombres, preposiciones, numerales y ciertos verbos y adverbios), extraídas, de forma automática, por el ordenador, a partir del título, del resumen y, cada vez más a

menudo, del texto completo de los documentos registrados dentro de un sistema documental dado.
Ejemplos: biblioteca; servicio; documentación; documental.

La lista de palabras clave es, la mayoría de las veces, monolingüe; igualmente puede ser plurilingüe, es decir, puede contener palabras de dos o más lenguas, pero sin equivalencias entre las palabras de las diferentes lenguas.

.2 Lista de descriptores libres

Una lista de descriptores está constituida por una *colección no ordenada* (sino puesta por orden alfabético) *de conceptos* destacados, por un proceso intelectual, a partir de los documentos registrados dentro de un sistema documental dado; esos conceptos son expresados por palabras o por expresiones extraídas de los documentos, o propuestos por los documentalistas, sin verificar si existen previamente en una lista establecida a priori.
Ejemplos: biblioteca; servicio de documentación; servicio documental.
La lista de descriptores libres es generalmente monolingüe.

.2 Lenguaje controlado

Existen dos tipos principales de lenguajes combinatorios controlados:

-las listas de autoridades;
-los thesaurus de descriptores.

.1 Lista de autoridades

Una lista de autoridades está constituida por una *colección no ordenada* (sino puesta por orden alfabético) *de conceptos* destinados a representar *de manera unívoca* el contenido de los documentos y de las consultas dentro de un sistema documental dado; estos conceptos son expresados por palabras o por expresiones extraídas de una lista finita, establecida a priori; sólo los términos que figuran en esta lista pueden ser utilizados para indizar los documentos y las consultas.
Ejemplos: biblioteca; servicio de documentación.
La lista de autoridades es, la mayoría de la veces, monolingüe.

.2 Thesaurus de descriptores

Un thesaurus es una *lista estructurada* de *conceptos*, destinados a representar *de manera unívoca* el contenido de los documentos y de las consultas dentro de un sistema documental determinado, y a ayudar al usuario en la indización de los documentos y de las consultas(*)1; los

(*)1 Definición de AFNOR: «Lista de autoridades compuesta por descriptores y no-descriptores que obedecen a reglas

conceptos son extraídos de una lista finita, establecida a priori; sólo los términos que figuran en esta lista pueden ser utilizados para indizar los documentos y las consultas; la ayuda al usuario la proporciona la estructura semántica del thesaurus: fundamentalmente las relaciones de equivalencia, de jerarquía y de asociación.

Ejemplos:

-biblioteca
 genérico : sistema de información
 específico : biblioteca pública
 : biblioteca escolar

-servicio de documentación
 equivalente : sistema documental
 genérico : sistema de información
 específico : agencia de resúmenes
 asociado : base de datos bibliográficos.

Un thesaurus de descriptores puede ser monolingüe o multilingüe; en este último caso, el thesaurus incluye así mismo relaciones entre la expresión de los conceptos equivalentes dentro de las diferentes lenguas.

.3 Lenguaje codificado

Prácticamente todos los lenguajes de indización utilizados hoy día están constituidos por elementos léxicos del lenguaje usual: palabras y expresiones del lenguaje natural y descriptores no codificados.

Los sistemas de codificación se utilizan tradicionalmente para los lenguajes de clasificación, donde están totalmente justificados. Las pruebas que se han realizado, a lo largo de los años 1950-60, para representar por códigos los elementos constitutivos de los lenguajes combinatorios, han conducido a un callejón sin salida. En esto también, son las circunstancias que rodean la utilización de los lenguajes las que justifican esta evolución:

-la indización, como es más analítica que la clasificación, exige unos lenguajes más especializados, mientras que la clasificación, más sintética, puede frecuentemente explotar los lenguajes universales, como la C.D.U. (Clasificación Decimal Universal); además, por la misma razón, los lenguajes combinatorios deben evolucionar rápidamente para adaptarse al desarrollo de la terminología científica y técnica. El interés por crear un sistema de notación con códigos universales es entonces más débil;

-la indización exige más tiempo que la clasificación: por tanto, para un volumen idéntico de documentación, existen

---

terminológicas propias, relacionados entre sí por relaciones semánticas (jerárquicas, asociativas, o de equivalencia). Esta lista sirve para traducir a un lenguaje artificial desprovisto de ambigüedad las informaciones expresadas en lenguaje natural» (norma NF 47-100-diciembre 1981).

más documentalistas encargados de la indización que de la clasificación; por otra parte, el progreso de la telemática ha contribuido a incrementar considerablemente el acceso a sistemas documentales especializados, que antes resultaban sólo accesibles para los que los confeccionaban, y que después quedaron abiertos a todos aquellos, documentalistas e incluso usuarios finales, que pudieran estar interesados. La utilización de la jerga propia de un lenguaje documental codificado obstaculizaría indudablemente el acceso masivo de los usuarios.

Sin embargo, muchos lenguajes controlados utilizan un sistema de notación que puede ser denominado topológico: en ellos, la finalidad de la codificación no es sustituir a los descriptores en el momento de la indización de los documentos o de las consultas, sino facilitar la localización de los descriptores dentro de una representación gráfica o clasificatoria del thesaurus.

.2 Características comparativas

Comparamos a continuación los lenguajes combinatorios libres y controlados desde la perspectiva de:

-su contenido (§ 3.2.1);
-su tamaño (§ 3.2.2);
-sus prestaciones (§ 3.2.3);
-los costes de construcción y de utilización (§ 3.2.4).

Hemos excluido de este análisis los lenguajes combinatorios codificados, ya que no son utilizados en los sistemas documentales actuales.

.1 Contenido

.1 Lista de palabras clave

Se trata en general de palabras sueltas (unitérminos), reconocidas automáticamente por el programa informático de gestión documental, que extrae toda cadena de caracteres separada de las demás cadenas por uno o más espacios blancos y/o signos de puntuación, siempre que tal cadena no figure ya dentro de la lista de palabras vacías establecida a priori.
Por consiguiente, en tal lista se encontrarán:

-palabras de todas las categorías gramaticales: nombres comunes, nombres propios, adjetivos, verbos...
-palabras de todas las formas gramaticales: masculino y femenino, singular y plural, declinaciones de los nombres y de los adjetivos, conjugaciones de los verbos;
-todas las variantes ortográficas, comprendidas también todas las faltas de ortografía;
-palabras que designan un concepto preciso (ejemplo: radio-astronomía) y palabras cuya significación es imprecisa (ejemplo: radio; se puede tratar de un modo de comunicación -la radio-, de un aparato de radiorecepción, de un elemento de la geometría);

-palabras en todas las lenguas de los documentos registrados dentro del sistema documental, sin ninguna concordancia entre lenguas;
-palabras que, en los documentos de los que proceden, corresponden efectivamente al tema de esos documentos, y otras que no corresponden en absoluto: ejemplos escogidos de dominios diferentes, disgresiones de los autores, explicaciones sobre cuestiones que no se tratan en sus documentos...

Ciertos sistemas de indización automática permiten normalizar las formas gramaticales: masculino, singular, nominativo, infinitivo.

.2 Lista de descriptores libres

Se trata de una primera forma de control del vocabulario: los documentos son examinados por los documentalistas, que indizan los conceptos principales por medio de palabras o de expresiones, generalmente redactadas en una sola lengua, sea cual sea la lengua del documento.
Se puede establecer una serie de reglas, de las cuales en realidad, según la experiencia, sólo se siguen algunas:

-es bastante fácil retener únicamente los nombres, y no las formas verbales o adjetivas, y registrarlos tan sólo en singular;
-es mucho más difícil conseguir una homogeneidad absoluta en el enunciado de los términos retenidos por los indizadores para designar un mismo concepto:

+quedan numerosas formas sinónimas;
+subsisten variantes ortográficas (ONU, O.N.U.) y errores de ortografía (B.I.T., B.J.T., P.I.T., B.I.D.), ya que, a falta de una lista preestablecida de términos aceptados, el ordenador no puede detectar esas variantes y errores.

Algunos servicios de documentación establecen, a posteriori, una estructura de equivalencias entre las variantes ortográficas de un mismo descriptor, a veces incluso entre los sinónimos de los mismos conceptos: se consigue entonces lo que denominamos «lenguaje de ayuda a la búsqueda».

.3 Lista de autoridades

La lista de autoridades es un verdadero lenguaje controlado:

-al igual que la lista de descriptores libres, reúne conceptos expresados por palabras y expresiones del lenguaje usual bajo una forma canónica: nombres en singular;
-pero, de la misma forma que el thesaurus, se trata de una colección finita, y los términos que en ella se encuentran son los únicos que se pueden utilizar para la indización de los documentos y de las consultas;

-sin embargo, al revés que el thesaurus, la lista de autoridades no contiene relaciones semánticas entre los descriptores; incluye muy raramente relaciones de equivalencia entre no-descriptores y descriptores, y no es multilingüe sino muy excepcionalmente.

Es bastante frecuente que en los sistemas documentales existan listas de autoridades de nombres de personas, de organismos, de instituciones, para las cuales:

-es posible establecer un listado a priori;
-es deseable verificar la ortografía de los términos que figuran en los documentos y en las consultas, para evitar toda pérdida de información.

Pero al hablar de esto nos salimos del dominio de los lenguajes documentales stricto sensu, los cuales, tal y como los hemos definido, se refieren esencialmente al contenido conceptual de los documentos y de las consultas.

.4 Thesaurus de descriptores

El thesaurus de descriptores, al igual que la lista de autoridades, es un lenguaje controlado.
Pero además de los descriptores incluye:

-los no-descriptores, o equivalencias semánticas que facilitan la traducción de los conceptos, expresados en el lenguaje libre de los autores de los documentos y de las consultas, en descriptores del thesaurus;
-las relaciones de jerarquía y asociación, que facilitan la localización de los descriptores más apropiados cuando se realiza la indización de los documentos, y sobre todo en el momento de la formulación de las consultas;
-en el caso, relativamente frecuente, de los thesaurus multilingües, las equivalencias lingüísticas, que permiten obtener, en respuesta a una consulta formulada en una de las lenguas del thesaurus, todos los documentos pertinentes, sea cual sea la lengua del thesaurus con el que han sido indizados.

.2 Tamaño

Una lista de palabras clave puede contener, según la extensión de los dominios cubiertos por el sistema documental y el cuidado que se ha puesto para eliminar los errores ortográficos, desde varias decenas de miles hasta cientos de miles de palabras.
Una lista de descriptores libres puede constar de varias decenas de miles de descriptores; su dimensión varía de acuerdo con el dominio y según el cuidado puesto para eliminar no solamente los errores ortográficos, sino también las variantes ortográficas y, excepcionalmente, los sinónimos más patentes y/o los términos retenidos en el momento de la indización de un documento que no tengan un interés documental real.
Una lista de autoridades incluye de varios cientos a varios miles de descriptores.

Un thesaurus contiene en general algunos miles de descriptores (que se multiplican según el número de lenguas del thesaurus) y de varios cientos a varios miles de no-descriptores.

.3 Prestaciones

Se examinará de forma más especial:

-la univocidad semántica, ligada a la presencia o ausencia de sinonimia y polisemia (§ 3.2.3.1);
-la riqueza de la terminología (§ 3.2.3.2);
-la actualización de la terminología (§ 3.2.3.3);
-la facilidad de utilización:

+para la indización de los documentos (§ 3.2.3.4);
+para la formulación de las consultas (§ 3.2.3.5);

-la coherencia de la indización (§ 3.2.3.6);
-el consumo de memoria (§ 3.2.3.7).

.1 Univocidad semántica

.1 Lista de palabras clave

Las listas de palabras clave se caracterizan por una ambigüedad semántica muy grande: en ellas cada concepto puede ser designado por varias palabras (sinonimia) y cada palabra puede designar varios conceptos (polisemia).

+ *Sinonimia:* un concepto puede expresarse en lenguaje natural por una serie de sinónimos y de perífrasis; la búsqueda de documentos que traten sobre varios conceptos debe, por tanto, realizarse por medio de una consulta que reagrupe, para cada uno de esos conceptos, el conjunto, o un sub-conjunto tan grande como sea posible, de esos sinónimos, con el fin de recuperar la mayor cantidad de documentos pertinentes, sea cual sea la manera en que sus autores hayan decidido expresar los conceptos en cuestión. Por consiguiente, el usuario debe poder reunir todos los sinónimos posibles de los conceptos que le interesan dentro de su consulta; si, por ejemplo, se interesa por la enseñanza de la informática, deberá buscar acerca de: (enseñanza o formación o instrucción o educación o aprendizaje o preparación o escuela o curso...) e (informática u ordenador o computadora o programación o automatización...). La lista de palabras clave le resultará poco útil, en la medida en que cada sinónimo está ordenado alfabéticamente dentro de ella, pero no remite a las demás.

Cuando los conceptos de los documentos están expresados con una perífrasis, es casi imposible recuperarlos; ¿cómo recuperar, por ejemplo, un documento titulado «El PASCAL en la Universidad» si se realiza una consulta sobre la enseñanza de la informática?

Además, la dificultad de enunciar las consultas de forma suficientemente completa, y el consiguiente riesgo de no

recuperar los documentos pertinentes, aumentan en el lenguaje libre por la existencia de:

-equivalentes lingüísticos, en el caso de sistemas documentales multilingües;
-variantes y errores ortográficos: si se examinan listas de palabras significativas presentes en los sistemas documentales, se observa que existen más términos erróneos que términos sin errores ortográficos; sólo los sistemas que dedican importantes recursos humanos o informáticos a detectar esos errores consiguen eliminarlos;
-variantes flexionadas (por ejemplo, el plural: un documento que trata sobre *las* escuelas de informática no será recuperado como respuesta a una consulta sobre informática y escuela -en singular-); los sistemas documentales permiten, sin embargo, superar automáticamente las dificultades de búsqueda ocasionadas por las flexiones, gracias a la utilización del truncamiento: una consulta escrita «escuela$» recuperará, por ejemplo, los documentos en los que la palabra «escuela» aparezca en singular o en plural;
-variantes derivadas (por ejemplo: aprender informática, docencia de la informática, formación en informática); aquí también el truncamiento puede proporcionar una ayuda automática al usuario, pero con un mayor riesgo de respuestas aberrantes; por ejemplo, «form$» permitirá localizar los documentos en los que figuran las palabras formado(s), formación(es), formador(es), formar, pero también otras decenas de palabras, como formato, formalismo, formalidad, formal, formulario...
-siglas, por ejemplo: E.A.O. (Enseñanza Asistida por Ordenador).

+ *Polisemia:* una palabra puede expresar en lenguaje natural varios conceptos; la búsqueda de documentos que tratan sobre una serie de conceptos debe ser necesariamente expresada por palabras, pero cada una de esas palabras puede ser portadora de varios significados; esto significa que determinados documentos obtenidos como respuesta no serán pertinentes. Una consulta sobre la *instruccion$* (de los estudiantes) en *programación* (informática) suministrará, por ejemplo, documentos que traten sobre las *instrucciones* que constituyen los lenguajes de *programación* de ordenador. Igualmente, la consulta sobre la *formación* (enseñanza) en *informati$* suministrará documentos sobre la *formación* (en el sentido de constitución) de equipos de *informáticos*. Si la sinonimia provoca falta de información, la polisemia entraña «ruido».
Los lingüistas (Mounin, 1974) establecen una distinción entre:

-la verdadera polisemia: un término con sentidos diferentes (un significante y varios significados), pero con rasgos semánticos comunes.

Ejemplo: liso (sin asperezas)

y liso (color homogéneo, sin estampado)[*]2

-la homonimia: un significante y varios significados, sin rasgos semánticos comunes, incluso cuando las palabras tienen la misma etimología (es decir, el mismo origen).

Ejemplo: fuga (huida)
y fuga (música).

Se distinguen, dentro de los homónimos:

+los homófonos: una pronunciación y varios significados.

Ejemplo: honda
y onda

+los homógrafos: una grafía y varios significados.

Ejemplo: banco (entidad financiera)
y banco (mueble).

*Nota:* la homofonía no influye por el momento en documentación; evidentemente no seguirá sucediendo lo mismo en los futuros sistemas, en los que la localización de la información ya no se basará en pulsar un teclado o en la lectura óptica, sino en el análisis automático del habla.

En lingüística computacional, la tipología de la polisemia es distinta, pues se distinguen:

-los homógrafos: un significante y varios significados, que pertenecen a formas gramaticales (categorías gramaticales) distintas.

Ejemplo: libro (nombre)
y libro (verbo) (librar).

-los homónimos: un significante y varios significados, de una misma forma gramatical, con o sin rasgos semánticos comunes.

Ejemplo: vela (de un barco)
y vela (de cera).

Esta última tipología interviene en los sistemas de tratamiento automático de lenguas (indización automática, traducción automática), porque los dos tipos de polisemia son tratados por algoritmos de desambiguación diferentes: los homógrafos pueden ser resueltos por un análisis sintáctico, mientras que los homónimos deben ser objeto de un análisis semántico.

---

[*]2 El ejemplo del texto original francés es «uni (conjoint) et uni (même couleur)». Tanto en este caso como en algunos otros a lo largo del libro, ha sido necesario adaptar al castellano los ejemplos aducidos, con el fin de respetar el carácter didáctico del manual [nota de los tr.].

Es necesario resaltar que, para un mismo dominio, el contenido semántico de una lista de palabras clave es muy diferente de una lengua a otra:

-algunas lenguas, como el francés y el inglés, utilizan fundamentalmente palabras simples, muchas de las cuales pueden ser ambiguas (ejemplo: una expresión como «estudio económico» se descompone en ESTUDIO y ECONOMICO), y una pequeña proporción de palabras compuestas, generalmente menos ambiguas;
-mientras que otras lenguas, como el alemán y, en menor medida, el holandés y el danés, cuentan con una gran proporción de palabras compuestas; por tanto, sus listas de palabras clave tendrán un contenido más unívoco (ejemplo: la palabra compuesta WIRTSCHAFTSKUNDE, que significa lo mismo que «estudio económico», figura como tal en una lista de palabras clave alemanas).

.2 Lista de descriptores libres

+ *Sinonimia:* el fenómeno de la sinonimia es aquí menos importante que en las listas de palabras significativas, gracias a la eliminación casi total de las diversas formas flexionadas y derivadas, así como la utilización de una sola lengua. Desgraciadamente, la sinonimia natural del lenguaje subsiste y es más importante todavía que en la lista de palabras clave; en el ejemplo escogido anteriormente, todos los términos enunciados han de ser mantenidos: enseñanza, formación..., automatización; a estas palabras simples se añaden expresiones, como enseñanza programada, tratamiento de la información. Así mismo, las variantes y los errores ortográficos no son eliminados.
El establecimiento de una estructura relacional entre estas variantes y sinónimos conduce, en algunos, pocos, servicios, a la realización de una lengua de ayuda a la búsqueda, que permite al sistema resolver los problemas de sinonimia de forma transparente para los usuarios.

+ *Polisemia:* la polisemia no es suprimida, porque queda en el lenguaje natural; sin embargo, se reduce considerablemente en relación a la lista de palabras significativas, gracias al uso de expresiones, generalmente mucho más significativas que las palabras simples. Por ejemplo, la expresión «servicio militar» es unívoca, mientras que las palabras «servicio» y «ejército» son ambivalentes: pueden referirse al servicio militar, pero también a un servicio del ejército, a un servicio realizado al ejército o por el ejército, o a la realización de un servicio dentro del ejército...

.3 Lista de autoridades

En principio, cada concepto se expresa por un solo descriptor.
En la práctica, la falta de una estructura semántica y, sobre todo, de equivalencias entre no-descriptores y descriptores (ya sea entre conceptos próximos o entre auténticos sinónimos) hace difícil la completa eliminación de

la sinonimia y la polisemia cuando se construye la lista (salvo si se trata, como es frecuente, de una lista de tamaño restringido).

.4 Thesaurus de descriptores

Debido a su concepción (lista controlada y estructura semántica), el thesaurus es el lenguaje combinatorio que menos imprecisión comporta.
Sin embargo, aparecen:

-en los mejores thesaurus, algunos descriptores polisémicos:

+en determinados casos porque quien lo construye ha decidido, con completo conocimiento de causa, agrupar bajo un solo descriptor una serie de conceptos próximos o específicos que no es necesario enumerar en el thesaurus, ya que se sitúan al margen de los dominios cubiertos por el sistema documental.
Ejemplo: en un thesaurus sobre economía en general, el descriptor «horticultura» podrá englobar conceptos como «fruta» y «legumbre».

+en otros casos, puede que un descriptor, aunque esté situado en un contexto semántico que aclara el sentido, conserve dos significaciones distintas; esta situación no suele acarrear consecuencias graves, en la medida en que uno de esos sentidos no pertenece a los dominios cubiertos por el sistema documental.
Ejemplo: en un thesaurus sobre economía, el descriptor «hipoteca» tendrá claramente el sentido de «garantía inmobiliaria del pago de una deuda»; aunque un documentalista encuentre esa palabra con el significado de «dificultad que impide la realización de alguna cosa», no se le ocurrirá utilizar el descriptor «hipoteca», aun cuando éste no estuviera acompañado de una nota de aplicación que precise su uso en el thesaurus.

-en algunos thesaurus, descriptores que designan conceptos cuasi-sinónimos: esto sucede cuando quien construye el thesaurus demuestra una gran laxitud y permite que en su lista se incluyan, como descriptores, términos que deberían ser no-descriptores.
Ejemplo: en un thesaurus sobre política, no es conveniente mantener simultáneamente como descriptores términos como polémica, disputa, querella, controversia, altercado, enfrentamiento, desavenencia, etc., so pena de caer en los inconvenientes de la lista de palabras clave, es decir, en la necesidad de integrar, cada vez que se realiza una consulta sobre uno de estos conceptos, todos los descriptores dentro de la formulación de la consulta. Existen thesaurus en los que al menos una parte de estos descriptores aparecen sin estar acompañados de una nota de aplicación que explicite el sentido particular que el autor asigna a cada uno de estos términos.

.2 Riqueza terminológica

La lista de palabras clave incluye todos los términos tomados de los títulos, de los resúmenes y, a veces, del texto de los documentos, mientras que los otros tres tipos de lenguajes combinatorios son muy selectivos y no retienen más que los conceptos más importantes de los documentos.

Esta riqueza terminológica de la lista de palabras clave es a la vez una ventaja y un inconveniente:

-una ventaja, en la medida en que todas las palabras, las que designan nociones extremadamente específicas y/o las que sólo intervienen en un número muy limitado de documentos, son retenidas, y pueden permitir que se lleven a cabo búsquedas muy refinadas:

+ya sea sobre un término muy frecuente.

Ejemplo: el nombre de una máquina, de un lugar, de un producto, de una marca...

+ya sea sobre un término presente en un gran número de documentos, sin que represente necesariamente en ellos un concepto importante.

Ejemplo: la búsqueda de la palabra «oficio» en documentación jurídica, con el fin de establecer un inventario exhaustivo de todas las disposiciones aplicables, incluso marginalmente, a un tipo de actividad profesional.

-un inconveniente, en la medida en que, aunque se efectúe una búsqueda sobre una palabra:

+se recuperarán todos los documentos que contengan esa palabra, aunque designe:

->un concepto diferente del que se busca (polisemia).

Ejemplo: los textos jurídicos que se refieran a los «oficios» (tipo de documento administrativo).

->un concepto accesorio en un documento;

+y no se encontrarán los documentos en los que el mismo concepto se expresa por otra palabra o conjunto de palabras.

Ejemplo: «actividad laboral de cualquier persona física».

.3 Actualización

En este punto los dos tipos de lenguajes libres son más eficaces que los lenguajes controlados, ya que:

-la lista de palabras clave evoluciona automáticamente, al mismo tiempo que la terminología utilizada en los documentos;
-los indizadores que trabajan a partir de una lista de descriptores libres tienen la posibilidad de añadir nuevos términos a esta lista a medida que los encuentran;

mientras que los lenguajes controlados, compuestos a priori por listas de términos, son puestos al día con un cierto retraso.

Es necesario tener en cuenta, sin embargo, que algunos sistemas documentales cuentan con una «válvula de seguridad», constituida por una lista de posibles descriptores:

-procedentes de la acumulación, por parte de los documentalistas, de términos que ellos habrían querido utilizar para indizar uno o más documentos y que no han sido encontrados en el thesaurus;
-registrados en un campo particular de las referencias bibliográficas, donde son utilizables para la búsqueda documental, de la misma forma que los descriptores ya incluidos en el thesaurus.
-y periódicamente examinados con vistas a poner al día el thesaurus.

### .4 Facilidad de utilización en el momento de indizar los documentos

Los lenguajes libres presentan, a la hora de indizar los documentos, una indudable ventaja en comparación con los lenguajes controlados: no requieren una traducción, a partir de la expresión natural de los conceptos encontrados en los documentos, para obtener una representación de esos mismos conceptos en descriptores del thesaurus o de la lista de autoridades.
Además, el método de la lista de palabras clave tampoco requiere que los documentalistas busquen los conceptos de los documentos, ya que la indización se realiza automáticamente, por reconocimiento de palabras no vacías.
Por tanto permite:

-una importante economía de recursos humanos;
-menor demora entre la llegada de los documentos al servicio de documentación y su integración en la base de datos documental gracias a la no-existencia de un puesto de trabajo de indización ni de la cola de espera correspondiente a ese puesto de trabajo.

### .5 Facilidad de utilización en el momento de la formulación de las consultas

Aquí la superioridad la tienen claramente los lenguajes controlados, y en particular los thesaurus.
La falta de rigor del lenguaje libre hace que la búsqueda documental sea en efecto, muy compleja, por:

-la necesidad de agrupar todas las palabras y expresiones sinónimas de los conceptos de la consulta; la falta de estructura en las listas de palabras clave, así como en las listas de descriptores libres, hace que el usuario no cuente con ninguna ayuda del sistema para realizar esa agrupación;
-la necesidad de utilizar la búsqueda por truncamiento, para no tener que repetir todas las formas gramaticales de una palabra, y la búsqueda por proximidad, para evitar la falsa coordinación entre palabras situadas en contextos diferentes (ejemplo: si las palabras «fabricación» y «cadena» se encuentran en la misma frase, existen muchas posibilidades

de que el documento trate sobre la fabricación en cadena o sobre la cadena de fabricación; si se encuentran en párrafos distintos, esta probabilidad disminuye).

La utilización de los operadores de truncamiento y proximidad es posible en la mayoría de los programas informáticos de búsqueda documental, pero

+hace la búsqueda más difícil, especialmente para un usuario ocasional,
+no produce resultados seguros,

-la necesidad de eliminar una serie de documentos no pertinentes, que aparecerán en respuesta a la consulta, a causa de:

+la polisemia,
+las falsas coordinaciones.

Ejemplo: fabricación de cadenas.

El hecho de que una palabra no vacía aparezca en un documento lo único que muestra es que, en efecto, el autor ha escrito algo sobre esa palabra, mientras que, cuando se usan los otros tres sistemas, la existencia de un descriptor en la indización de un documento significa que un documentalista ha dado fe de que el autor ha escrito efectivamente sobre el concepto representado por ese descriptor:

-la imposibilidad, en el caso de sistemas con lista de palabras clave, de recuperar los documentos que tratan sobre los conceptos de la consulta, pero en los que dichos conceptos son expresados por medio de perífrasis o de manera implícita; en los sistemas con descriptores libres este inconveniente no existe, si el trabajo de indización está bien hecho.

El lenguaje controlado permite eliminar la mayor parte de estos inconvenientes; sin embargo, pueden permanecer algunos casos de falsa coordinación (a no ser que se utilice la técnica de los «vínculos»: ver capítulo III, § 1.4.4.4).

Los thesaurus son más eficaces que las listas de autoridades, gracias a su red de relaciones semánticas, que permite:

-recuperar más cómodamente los descriptores que se van a utilizar, partiendo de cómo se expresaron originalmente los conceptos de la consulta;
-extender la búsqueda a otros conceptos, más específicos, más genéricos, o simplemente asociados a los de la consulta, que representen los diversos puntos de vista previsiblemente adoptados por los autores de los documentos.

.6 Coherencia de la indización

La coherencia de la indización es perfecta en los sistemas de lista de palabras clave: un mismo documento, tratado por un mismo programa informático, en el que serán registradas:

-la misma lista de palabras vacías; y
-la misma lista de separadores de cadenas de caracteres: el espacio en blanco y los signos de puntuación,

será siempre indizado de la misma manera.

Sin embargo, dos documentos diferentes, pero que tratan del mismo tema con términos distintos, serán indizados de forma totalmente heterogénea, ya que la indización por palabras clave se basa exclusivamente en la terminología utilizada por los autores de los documentos.

El thesaurus de descriptores permite una coherencia del orden de un 50 a un 80 %[*]3, según la complejidad de los documentos, la calidad del thesaurus y el nivel de formación de los documentalistas que usan coherentemente el thesaurus.

Esta coherencia, relativamente débil, se debe a que los distintos indizadores perciben de forma diferente:

-el contenido real del documento;
-la parte de ese contenido que será susceptible de responder realmente a las necesidades (inevitablemente futuras) de los usuarios;
-los conceptos importantes que habrán de ser conservados para representar este contenido.
-los descriptores elegidos para representar esos conceptos.

Esta coherencia, relativamente débil, no es un gran inconveniente, en la medida en que la estructura semántica del thesaurus permitirá al usuario, en el momento en el que interroga al sistema documental, enriquecer el contenido de su consulta; por esto el usuario buscará en el thesaurus los descriptores que expresen los conceptos correspondientes a los diferentes ángulos sobre los que los autores pueden haber abordado el tema de su consulta.

Este enriquecimiento de la consulta no resulta siempre algo cómodo, en particular para el usuario ocasional, pero al menos éste dispone de un instrumento (el thesaurus) para hacerlo, mientras que no existe ninguno en el caso de los sistemas de lenguaje libre.

La coherencia de la indización dentro de un sistema de lista de autoridades será, por término medio, peor todavía, a causa de la falta de una estructura semántica; y el enriquecimiento de las consultas, para poder compensar esta incoherencia, será más molesta, por el mismo motivo.

La coherencia de la indización es la peor (de un 20 a un 40 %) en los sistemas de lista de descriptores libres, ya que, al no tener ninguna posibilidad de control automático del vocabulario utilizado, los indizadores son libres de usar los términos que a cada uno de ellos les parezcan más apropiados.

---

[*]3 La tasa de coherencia entre dos indizadores es la ratio existente entre el número de descriptores comunes asignados por los dos documentalistas a un mismo documento, a partir de un mismo thesaurus, y el número total de descriptores, comunes o diferentes, utilizados por esos documentalistas.

Es necesario, por último, hacer notar que la coherencia de la indización de dos documentos diferentes que tratan sobre un mismo tema con términos distintos será, en general, considerablemente mejor en los sistemas fundados sobre la indización manual (thesaurus, lista de autoridades, descriptores libres) que en un sistema basado únicamente en el reconocimiento automático de palabras (lista de palabras clave).

.7 Consumo de memoria

Es sabido que la mayor parte de los sistemas documentales automatizados están organizados en una serie de ficheros. Los más importantes son:

-el fichero diccionario, que contiene:
+la relación tal cual de todos los criterios de búsqueda de los documentos registrados hasta el momento, y especialmente de las unidades léxicas del o de los lenguajes documentales;
+y, si existe un thesaurus, las relaciones semánticas entre sus términos;
+así como los punteros de esas unidades léxicas hacia las entradas del fichero inverso;

-el fichero inverso, del cual cada registro:
+se corresponde con una entrada del fichero diccionario;
+contiene uno o más datos relativos a cada uno de los documentos dentro de los cuales ese criterio de búsqueda interviene; esos datos son:
->un puntero hacia el documento en cuestión dentro del fichero bibliográfico;
->a veces, cuando el sistema permite la búsqueda por proximidad, se indica la o las localizaciones de ese criterio de búsqueda en el interior del documento;

-el fichero bibliográfico, que contiene, para cada documento registrado:
+la descripción bibliográfica: título, autor, fuente, fecha...
+la descripción del contenido: código de clasificación, descriptores libres y/o controlados, resumen;
+en las denominadas bases de datos textuales: el texto completo del documento.

La utilización del sistema de palabras clave conlleva un fichero diccionario y un fichero inverso de gran tamaño, ya que en ellos se recogen todas las palabras no vacías del título, del resumen y, a veces, del texto completo, con indicación, en algunos casos, de su(s) localizacion(es) dentro de cada documento.
La lista de autoridades es, por contra, la que menos memoria consume, ya que tiene un vocabulario limitado y no existe relación entre los términos.

La lista de descriptores libres es más reducida que la lista de palabras clave y no incluye la localización de los descriptores en el interior de cada documento; a menudo es considerablemente más voluminosa que el thesaurus, pero no contiene relaciones semánticas (salvo en el caso de que se haya trabajado sobre la lista para elaborar un lenguaje de ayuda a la búsqueda).

*Nota:* en cualquier caso, la constante reducción de los costes de las memorias informáticas hace que:

-no deba tomarse más en cuenta el factor «consumo de memoria» cuando se evalúa la relación costes/prestaciones de los diferentes lenguajes combinatorios;
-pierda su interés el argumento, cierto, de que no son útiles la mayor parte de las palabras clave, ya que nunca son usadas en el momento de la búsqueda documental.

.4 Costes de construcción y de utilización

.1 Lista de palabras clave

+ *Construcción:*

-coste nulo o casi nulo, dado que basta con registrar al principio una lista de varios cientos de palabras vacías y editar ocasionalmente una lista de palabras clave extraídas por el sistema a partir de los textos registrados.

+ *Indización de los documentos:*

-coste casi nulo, únicamente el que se deriva del proceso de extracción de palabras clave que realiza el ordenador.

+ *Formulación de las consultas:*

-eficiencia reducida, debida a

+un coste elevado para el usuario, que debe arreglárselas él solo para encontrar los sinónimos adecuados;
+una eficacia a menudo poco satisfactoria, por la presencia de polisemias y la no obtención de documentos pertinentes; esta eficacia es a veces, sin embargo, mejor que con un lenguaje controlado, por tener una exhaustividad mayor.

.2 Lista de descriptores libres

+ *Construcción:*

-coste casi nulo, únicamente el que se deriva de la edición, a posteriori, de la lista de los descriptores asignados por los indizadores a los documentos, realizada a veces (en algunos, pocos, servicios) para eliminar los sinónimos y/o los errores ortográficos;

-coste muy elevado (frecuentemente mayor que para realizar un thesaurus) si se trabaja sobre esa lista a posteriori para obtener un lenguaje de ayuda a la búsqueda.

+ *Indización de los documentos:*

-coste casi tan elevado como con una lista de autoridades, ya que el documentalista debe:

+leer el documento para localizar los conceptos importantes;
+intentar representar esos conceptos con la máxima coherencia posible.

+ *Formulación de las consultas:*

-coste casi tan elevado como con una lista de palabras clave, con una eficacia media un poco mayor, gracias a la existencia de un número menor de sinonimias y polisemias, pues se utilizan, además de unitérminos, expresiones, y también gracias a que casi se eliminan las variantes flexionadas y derivadas; la existencia de un lenguaje de ayuda a la búsqueda reduce sensiblemente el coste de formulación de las consultas.

.3 Lista de autoridades

+ *Construcción:*

-coste no despreciable, pero considerablemente menor que en la construcción de un thesaurus.

+ *Indización de los documentos:*

-coste un poco más elevado que con una lista de descriptores libres, ya que es necesario traducir en descriptores controlados los conceptos extraídos del documento durante su examen;
-coste ligeramente más elevado que con un thesaurus, dado que la búsqueda de las equivalencia semánticas es más laboriosa, puesto que la mayor parte de las listas de autoridades no tienen relaciones de equivalencia entre no-descriptores (es decir, sinónimos o cuasi-sinónimos de los mismos conceptos) y descriptores (sinónimos o cuasi-sinónimos utilizados preferentemente para indizar documentos y consultas).

+ *Formulación de las consultas:*

-coste similar al de una lista de descriptores libres, pero con una eficacia mayor, gracias al control del vocabulario;
-coste más elevado que con un thesaurus, pero con una eficacia considerablemente menor, a causa de la falta de estructura semántica en la lista de autoridades.

.4 Thesaurus

+ *Construcción:*

-coste elevado, del orden de varias decenas de meses/hombre.

+ *Indización de los documentos:*

-coste más elevado que con una lista de descriptores libres, a causa de la necesidad de traducir los conceptos en descriptores del thesaurus;
-coste menos elevado que con una lista de autoridades, gracias a la existencia de no-descriptores.

+ *Formulación de las consultas:*

-coste menos elevado y eficacia mucho mayor que con cualquier otro lenguaje, gracias a la existencia de las relaciones semánticas.

.3 Utilización dentro de los sistemas documentales actuales

Ante el balance de las ventajas e inconvenientes de los cuatro tipos de lenguajes combinatorios:

-prácticamente todos los servicios de documentación utilizan la lista de palabras clave:
+por una parte, a causa de su casi nulo coste de:
->construcción,
->indización de los documentos;
+por otra parte, gracias a su eficacia, en ciertos casos mayor, en el momento de la búsqueda documental.

-prácticamente todos los servicios de documentación que disponen de recursos suficientes, y que se preocupan por la eficiencia del sistema que usan, construyen y explotan, además, un thesaurus de descriptores, a pesar de los costes
+en inversiones,
+en la indización de los documentos;

-una cantidad mínima de servicios de documentación utilizan:
+una lista de descriptores libres, que es más costosa y muy poco más eficaz que la lista de palabras clave;
+una lista de autoridades, que es ligeramente menos costosa que un thesaurus, pero claramente menos eficaz tanto en la indización como en la búsqueda.

*Nota:* Algunos servicios de documentación que al principio no disponen de recursos suficientes utilizan el método de la lista de descriptores libres (distinguiéndola bien de la lista de palabras clave), con vistas a utilizar más tarde esa lista de descriptores libres como base de

construcción de un thesaurus. La experiencia muestra que este supuesto es, las más de las veces, falso, porque:

-el vocabulario de los descriptores libres es, por una parte, excesivamente voluminoso, y, por otra, no lo bastante exhaustivo como para constituir la base que permita construir eficientemente (al menos según la relación costes/prestaciones) un thesaurus;
-el fondo documental indizado en lenguaje libre antes de la existencia del thesaurus no se reindiza con ayuda del thesaurus cuando éste está por fin construido;
-la existencia de la lista de descriptores libres es un argumento para oponerse a la construcción del thesaurus o para retrasarla.

## 4. **Estructura de los thesaurus**

En esta obra no examinaremos ni las listas de palabras clave, ni las listas de descriptores libres, ni las listas de autoridades, puesto que estos lenguajes combinatorios se caracterizan por el hecho de que no contienen, entre los términos que las constituyen, ninguna otra estructura aparte de la alfabética, que es totalmente aleatoria.
Los thesaurus de descriptores, por el contrario, cuentan con una fuerte estructura semántica.

### .1 Noción de campo semántico

La semántica es la rama de la lingüística que se ocupa del significado de las unidades lingüísticas y de sus combinaciones en proposiciones y frases.
El análisis documental no se preocupa nada más que de la identificación de los conceptos sobre los que se ocupa un documento o sobre los que trata una consulta: el significado de la combinación de estos conceptos en proposiciones o frases le preocupa poco: un documento debe, en efecto, ser recuperado en un momento dado, por un usuario dado, porque trata sobre una noción dada; poco importa que esta noción intervenga dentro del documento como sujeto, como objeto o como circunstancia, con o sin modificador que matice su alcance.
Lo que cuenta es que el documento aporta una información útil sobre ese concepto. Sin embargo, es necesario hacer notar que en algunos sistemas documentales la combinatoria de los descriptores no es completamente libre y reviste un carácter significativo. Esos sistemas no son muchos, y tal combinatoria interviene, no en la construcción del lenguaje, sino en el momento de su utilización. Por consiguiente volveremos a ello dentro del capítulo sobre la indización y la interrogación.
El análisis documental se interesa esencialmente, pues, por la primera parte de la definición de semántica: el significado de las unidades lingüísticas; le resulta necesario, en efecto, poder atribuir una significación precisa al contenido de un documento o de una consulta, y poder enumerar los conceptos sobre los que tratan los documentos registrados dentro de un sistema documental determinado. Tal enumeración consiste en identificar los

términos -palabras y expresiones- que explican los conceptos, y en determinar el significado que les ha sido asignado.

Esta determinación no se hace (o se hace poco) como en los diccionarios, esto es, por medio de una definición explícita que permita delimitar las características del término. Se realiza sobre todo por medio de la confrontación de los términos en el interior de «campos semánticos».

Un campo semántico es un conjunto de unidades léxicas, ligadas por una estructura de relaciones de significado que permite precisar la significación de cada una de esas unidades(*)4.

Si, por ejemplo, se encuentra la palabra «mesa» dentro del campo semántico «mobiliario» (donde se encontrarán así mismo palabras como butaca, silla, armario), se comprenderá que la palabra «mesa» se utiliza para designar el concepto que se define como «mueble compuesto por una superficie horizontal soportada por una o más patas», y no designa, por tanto, el accidente geográfico.

Esta forma de representar el dominio cubierto por un sistema documental es muy cómoda para el documentalista: la lista de los conceptos presentes en el campo o campos semánticos correspondientes a ese dominio, puede, en efecto, ser establecida de manera que comprenda exactamente la lista de los descriptores que se van a utilizar para indizar los documentos y las consultas, mientras que el diccionario clásico contiene una parte predominante (las definiciones) que no es directamente útil para tal indización.

Veremos más adelante que el thesaurus se diferencia del diccionario en otro punto: en la acepción de los términos contenidos en él, que en un thesaurus es frecuentemente más general y a la vez menos ambiguo que en un diccionario.

.2 Elementos constitutivos

Los elementos constitutivos de un thesaurus son dos:

-las unidades léxicas;
-las relaciones semánticas entre esas unidades.

.1 Unidades léxicas

Un thesaurus puede incluir hasta cuatro categorías de unidades léxicas:

-los títulos que encabezan los conjuntos, en el interior de los cuales se agrupan los términos: campos semánticos

(*)4 Definición de Trier, el lingüista que ha adelantado la noción de campo semántico: «el campo semántico es el conjunto de palabras...que, colocadas una junto a otra, como las piedras irregulares de un mosaico, recubren adecuadamente todo un dominio de significaciones bien delimitado...; dentro de la concepción de los campos conceptuales, existen... mosaicos de nociones asociadas... que la experiencia humana aísla y constituye en unidades conceptuales. Junto a ello existen campos léxicos, formado cada uno por el conjunto de las palabras que recubren, fraccionándolos, los campos conceptuales correspondientes (Mounin - 1963).

(denominados más frecuentemente «temas») o naturalezas de los términos (denominadas «facetas»); estos títulos no son utilizados para indizar los documentos ni las consultas, sino únicamente para agrupar los descriptores;

-los descriptores: palabras o expresiones del lenguaje usual, retenidas más o menos arbitrariamente por quien construye el thesaurus para designar los conceptos que representan el contenido de los documentos y de las consultas, y que efectivamente se utilizan para indizarlos;

-los no-descriptores, llamados también términos equivalentes o términos no preferentes: sinónimos o cuasi-sinónimos de los descriptores, o términos que designan en el lenguaje usual conceptos muy próximos a los que los descriptores representan; los no-descriptores no pueden ser utilizados para indizar los documentos ni las consultas, pero cada uno de ellos reenvía a uno o a dos descriptores que deberán utilizarse para representar los correspondientes conceptos;

-los descriptores auxiliares, que se utilizan en combinación con los descriptores libres para formar los descriptores compuestos que representan los conceptos complejos.

.1 Grupo de descriptores

.1 Utilidad

Un thesaurus contiene, en general, varios miles de descriptores. Para manejar cómodamente tal conjunto interesa subdividirlo en una serie de subconjuntos, cada uno de los cuales agrupará un número limitado (algunas decenas) de descriptores. Gracias a esta limitación, resultará fácil al usuario aprehender los descriptores y sus relaciones semánticas.

.2 Tipología

Esta subdivisión puede concebirse de dos maneras:

+ *la subdivisión por facetas:* al analizar una lengua se observa que todos los términos pertenecen a un número limitado de categorías; en el conjunto de los thesaurus de los que nosotros nos hemos tenido que ocupar hemos consignado la existencia de las siguientes facetas:

-fenómeno: acción natural que escapa a la acción del hombre.
Ejemplo:
MAGNETISMO

-proceso: acción provocada por el hombre.
Ejemplo:
FABRICACION

-materiales: elementos materiales, naturales o no, sobre los que actúan los fenómenos y los procesos.

Ejemplo:
ACERO MAGNETICO

-organización: entidad compleja creada por seres vivos.
Ejemplo:
ACERIA

-ser vivo: microorganismo, vegetal, animal, ser humano considerado en su esencia biológica o en su comportamiento.
Ejemplo:
INGENIERO

-equipamiento: material, edificio, herramienta, vehículo..., construido por el hombre para operar sobre los materiales o sobre los fenómenos a través de procesos.
Ejemplo:
LAMINADORA

-propiedad: característica ligada a uno o más fenómenos, procesos, materiales, seres vivos o equipamientos.
Ejemplo:
RESONANCIA

-disciplina: rama del conocimiento: arte, ciencia, técnica...
Ejemplo:
TECNOLOGIA DEL ACERO.

Por su parte, el Classification Group de Londres propone la siguiente lista de facetas.

-entidades:

+sustancias naturales;
+artefactos (entidades concretas producidas por el hombre);
+mentefactos (entidades abstractas concebidas por el hombre);

-atributos:

+propiedades;

-actividades:

+operación;
+procesos.

Ciertos productores de thesaurus utilizan el método de la agrupación por facetas para distribuir el conjunto de los conceptos en grandes subconjuntos, caracterizados porque sus descriptores pertenecen a una faceta homogénea.

+ *la subdivisión por temas:* al analizar un sistema documental se observa que cada concepto a retener pertenece

al menos a una disciplina, es decir, a un conjunto de materias susceptibles de ser enseñadas y de constituir el cuerpo de conocimientos que permita al hombre aprehender y dominar una parte de su entorno.

Ejemplos:
BIOLOGIA
QUIMICA
SOCIOLOGIA
OCEANOGRAFIA.

Casi todos los productores de thesaurus utilizan este método y reúnen los descriptores en grupos, cada uno de los cuales se corresponde con una disciplina o microdisciplina específica.

*Nota:* Cuando se decide construir un thesaurus, unos mismos descriptores pueden ser agrupados o bien por temas, o bien por facetas, como se ve en el siguiente ejemplo:

| FACETAS | | TEMAS | | |
|---|---|---|---|---|
| | BIOLOGIA | ARTES DECORATIVAS | CONSTRUCION METALICA | TRANSPORTE |
| FENOMENO | DIGESTION | REFLEJO | GRAVEDAD | RETRASO |
| PROCESOS | MANIPULACION GENETICA | PINTURA | ENSAMBLADO | VIAJE |
| ORGANIZACION | MACROMOLECULA | MUSEO | CADENA DE MONTAJE | SOCIEDAD DE TRANSPORTE |
| SER VIVO | BACTERIA | DECORADOR | MECANICO | CAMIONERO |
| MATERIALES | PROTEINA | PAPEL PINTADO | HERRAMIENTA | GAS-OIL |
| EQUIPAMIENTO | MICROSCOPIO ELECTRONICO | PINCEL | ROBOT DE SOLDADURA | LOCOMOTORA |
| PROPIEDAD | BIOMAGNETISMO | FUNCIONALIDAD | FLEXIBILIDAD | RAPIDEZ |
| DISCIPLINA | CITOLOGIA | ESTILISTICA | CONSTRUCCION ASISTIDA POR ORDENADOR | ESTUDIO DEL TRAFICO |

.3 Discusión

El método de agrupar todos los descriptores del thesaurus con el criterio único de las facetas ha de desecharse, y esto por dos motivos:

-la razón fundamental es que tal clasificación es abstracta y no responde a la cultura de los usuarios: estamos

acostumbrados, desde la infancia, a razonar en términos de disciplinas: matemáticas, física, biología, filosofía, estética..., y no en términos de naturaleza de las palabras; la noción de facetas ha sido inventada por un bibliotecario (Ranganathan) y no se enseña en ninguno de los ciclos escolares;

-un motivo añadido es que el número de facetas está limitado a diez, y este número es insuficiente para distribuir el conjunto de descriptores de un thesaurus en grupos suficientemente reducidos como para que sean cómodamente aprehendidos por el usuario.

La agrupación de los descriptores se funda entonces esencialmente sobre la noción de temas.

Esto no significa, sin embargo, que el concepto de facetas no intervenga en la construcción de los thesaurus, sino al contrario:

-desde el nivel de la primera agrupación de descriptores por temas, cabe observar que el número de descriptores a retener para una disciplina sea demasiado grande y que haga falta, por tanto, crear grupos más restringidos; si la naturaleza de la materia lo permite, esos subconjuntos podrán corresponder a microdisciplinas; en otros casos, una agrupación por facetas será más cómoda para el usuario.

Ejemplo: en un sistema documental determinado se incluyen varios cientos de conceptos dentro del dominio de la tecnología de los metales. Se puede pensar en repartir los descriptores correspondientes en una serie de temas, relativos cada uno a la tecnología de un metal en particular. Ahora bien, sucede que los términos utilizados para designar los procedimientos de fabricación de los metales son en un 80 % comunes a todos los metales.

Será entonces más cómodo repartir esos descriptores por facetas; en concreto:

-materiales: los diferentes tipos de metales considerados por el sistema documental correspondiente;

-procesos: las diversas operaciones de fabricación utilizadas en metalurgia y en fabricación metálica;

-propiedades: las características de los metales controlados a lo largo de las operaciones de fabricación.

Dentro de cada jerarquía particular de descriptores, en el interior de los grandes grupos (temas o facetas), se debe tomar la decisión, para cada caso, de organizar la jerarquía sobre el principio de temas o de facetas.

## .2 Descriptor

### .1 Definición

Término (palabra o expresión) que se ha escogido, a partir de un conjunto de sinónimos, de cuasi-sinónimos y de términos emparentados, para representar, de manera unívoca, un concepto susceptible de intervenir en los documentos y en las consultas que se examinan dentro de un sistema documental

dado, e incluido por tanto dentro del thesaurus de descriptores de ese sistema(*)5.

.2 Modalidades de desambiguación

La univocidad de los descriptores (un descriptor designa un solo concepto; un concepto es designado por un solo descriptor) está asegurada según diversas modalidades:

-la sinonimia se elimina gracias a las relaciones de equivalencia (cf. $ 4.2.2.3);
-la polisemia se elimina gracias a que:

+poseen determinadas relaciones semánticas: la pertenencia a un grupo semántico, y las relaciones jerárquicas.
Por ejemplo, si el descriptor «CAJA» se encuentra en el campo «TIPOGRAFIA», tendrá claramente el significado de «tamaño de las letras», y no el de «recipiente»;
+se prefiere una expresión antes que un término aislado que sea insuficientemente preciso.
*Ejemplo:* si se quiere designar el concepto de «niño menor», se escogerá como descriptor el término «menor de edad» antes que el término «menor», demasiado ambiguo;
+se añade una explicación al descriptor cuando su sentido no está suficientemente precisado por su contexto semántico o léxico.

Esta explicación puede ser de cuatro tipos distintos:

+un modificador, añadido entre paréntesis a la derecha del término ambiguo; el descriptor está entonces constituido por ambos: término y modificador; esta técnica se emplea aun cuando no sea posible, lingüísticamente hablando, utilizar una expresión significativa, en vez de una palabra ambigua, para designar un concepto.
Ejemplo: se desea integrar en el thesaurus un descriptor que designe el concepto de menor dentro del dominio del álgebra (menor de un elemento, menor de una matriz); este descriptor se podrá enunciar por:

MENOR (ALGEBRA)

*Nota:* el simple hecho de unir un descriptor a una clase semántica no es suficiente para asegurar su univocidad, cuando el mismo término deba emplearse como descriptor que designe dos o más conceptos.
Ejemplo: dentro de un determinado thesaurus, se debe utilizar el descriptor MERCURIO para designar por una parte

(*)5 Definición de AFNOR: «palabra o grupo de palabras incluidas en un thesaurus y escogidas de entre un conjunto de términos equivalentes para representar sin ambigüedad una noción contenida en un documento o en una petición de búsqueda documental» (norma NZ 47-100-diciembre 1981).

el metal mercurio y por otra el planeta Mercurio; si se incluye el descriptor MERCURIO en el campo semántico «METAL», por una parte, y, por otra, en el campo semántico «SISTEMA SOLAR», sus dos sentidos serán perfectamente distinguidos, pero sucede nada menos que el mismo descriptor, enunciado «MERCURIO», tendrá dos significaciones y será utilizado para indizar dos conceptos diferentes; esto provocará inevitablemente ruido en el momento de la búsqueda documental. Por consiguiente, habrá que distinguir los dos conceptos designándolos por los descriptores: MERCURIO (PLANETA) y MERCURIO (METAL).

+una nota explicativa o definición, añadida a continuación del descriptor, pero sin formar parte de su enunciado; la nota explicativa se revela útil para designar un término de jerga, conocido únicamente por los especialistas de una profesión o de una institución, y poco explícito para otros usuarios potenciales.

Ejemplo: el descriptor «POLITICA AGRICOLA COMUN», en un thesaurus de economía agrícola, será explicitado por una nota «política agrícola de los países miembros de la Comunidad Europea».

La nota explicativa puede así mismo ser útil para precisar la significación particular asignada a un descriptor polisémico que no se utiliza más que con uno de sus significados dentro del thesaurus y si su presencia dentro de un campo semántico no es suficiente para desambiguarlo completamente.

Ejemplo: dentro del dominio del derecho, la palabra «declaración» tiene dos significados:

+testimonio de quien interviene en un interrogatorio oral;

+carta magna, normativa del derecho internacional.

Si dentro del thesaurus debe ser conservado uno sólo (el primero) de estos significados, se utilizará el descriptor «DECLARACIÓN» acompañado de la nota explicativa «testimonio de quien interviene en un interrogatorio oral».

+una nota de aplicación, añadida a continuación del descriptor, sin formar parte de su enunciado, destinada a precisar la utilización que se va a hacer de ese descriptor.

Ejemplo: al descriptor «CONSTRUCCION», se le podrá añadir, dentro de un thesaurus de ingeniería civil, la nota «úsese en el sentido de acción de construir; para designar lo que se construye, úsese el descriptor apropiado: edificio, obra de arte...»;

+una nota histórica, añadida a continuación del descriptor, sin formar parte de su enunciado, destinada a indicar desde cuándo se utiliza el descriptor, y a recordar el descriptor o descriptores que designaban anteriormente el mismo concepto.

Ejemplo: TELEMATICA: hasta 1982, úsese informática y telecomunicación.

*Nota:* también puede añadirse una nota histórica a un no-descriptor que designa un concepto que se encuentra en desuso y que ha perdido por este motivo el estatuto de descriptor.

Ejemplo: mecanografía: a partir de 1970, úsese TRATAMIENTO DE LA INFORMACION

.3 Notación

Las normas nacionales e internacionales prevén diversos tipos de notaciones:

AFNOR:las notas explicativas, de aplicación e históricas se colocan entre corchetes.

ISO:entre el descriptor y la nota explicativa, histórica o de aplicación se coloca una sigla:
NE (= nota explicativa) en francés;
SN (= scope note) en inglés;
D (= Definition) en alemán.

.3 No-descriptor

.1 Definición

Término (palabra o expresión) incluido dentro de un thesaurus, tomado de una lista de sinónimos, de cuasi-sinónimos y de términos emparentados con uno o dos descriptores de ese thesaurus, que está ligado a tal(es) descriptor(es) por una relación de equivalencia semántica susceptible de intervenir en los documentos o en las consultas, pero no utilizable para indizar esos documentos o formular esas consultas(*)6.

.2 Utilidad

Los no-descriptores permiten, por una parte, mejorar la coherencia de la indización entre los documentalistas y, por otra, la coherencia entre la representación de un mismo concepto dentro de un documento y dentro de una consulta.

Contribuyen, pues, a aumentar a la vez:

-la tasa de llamada, es decir, el porcentaje de documentos pertinentes extraídos de la colección en respuesta a una consulta;

-la tasa de precisión, es decir, el porcentaje de documentos que son realmente pertinentes dentro del conjunto de los documentos de la colección extraídos en respuesta a una consulta.

(*)6 Definición de AFNOR: «palabra o conjunto de palabras incluidas en un thesaurus con prohibición de uso y reenvío a uno o más descriptores utilizables» (norma NZ 47-100 - diciembre 1981).

*Ejemplo:*

-se dispone de un thesaurus en el que figuran los descriptores «RECIPIENTE PARA BEBER» y «VASO», y no figuran no-descriptores ni descriptores más específicos que estos dos descriptores;

-consideremos un documento que trata sobre un hanap, es decir, un gran vaso de metal para beber: a falta de no-descriptores, el documentalista dudará sobre el modo de indizar este documento y se arriesga a escoger, a ciegas, el descriptor «VASO», ya que esta palabra está incluida en la definición del término que ha de indizarse;

-cuando un usuario busque los documentos que traten sobre RECIPIENTE PARA BEBER, no encontrará el documento en cuestión, que es pertinente: el sistema habrá producido un «silencio», o «pérdida de información», que origina una disminución de la tasa de llamada;

-cuando un usuario busque los documentos que traten sobre VASO, encontrará el documento en cuestión, a pesar de que no es pertinente: el sistema habrá producido un «ruido», o «parásito», que produce una disminución de la tasa de precisión;

-estos dos escollos se habrían evitado si en el thesaurus figurara la relación:

hanap
úsese RECIPIENTE PARA BEBER.

.4 Descriptores auxiliares

En algunos thesaurus, existe un grupo (en el sentido de § 4.2.1.1) de descriptores particulares, denominados descriptores auxiliares:

-casi todos unitérminos;

-con un sentido relativamente poco preciso (ejemplos de descriptores auxiliares: INFLUENCIA, FACTOR, MOVIMIENTO, TIPO, REPARTO, MONTANTE...);

-que no se ha deseado incluir dentro de un tema específico, a causa de la imprecisión de su significado;

-listados simplemente en orden alfabético, sin relaciones semánticas ni entre sí ni con descriptores pertenecientes a otros grupos.

Según los thesaurus, los descriptores auxiliares tienen el estatuto:

-o bien de descriptores: pueden ser utilizados con independencia de otros descriptores, para indizar el contenido de los documentos o de las consultas en las que los conceptos correspondientes aparecen; en este caso, los descriptores auxiliares sólo existen porque no se ha podido situarlos en un grupo específico: tema o faceta;

-o bien de términos que sólo pueden ser utilizados para indizar un documento o una consulta con la condición de estar ligados (es decir, post-coordinados) a un descriptor elegido de fuera de la lista de descriptores auxiliares. Esta técnica se utiliza para que se pueda indizar un gran número de

conceptos por medio de una amplia combinación de un pequeño número de descriptores y de descriptores auxiliares.

Por ejemplo, los descriptores auxiliares TIPO y CALCULO, combinados con descriptores como CAMBIO, INTERES, FLETE, SALARIO..., permiten designar los conceptos de tipo de cambio, cálculo del cambio, tipo de flete, cálculo del flete, tipo de salario, cálculo del salario..., sin necesidad de crear los descriptores correspondientes en el thesaurus.

Esta práctica era relativamente corriente cuando se comenzaron a elaborar thesaurus; su finalidad era reducir el tamaño del thesaurus. Actualmente se aplica poco:

-por una parte, para evitar las restricciones inútiles cuando se utiliza el thesaurus;
-por otra parte, porque la mayor parte de los programas informáticos de almacenamiento y recuperación documental no permiten hacer la distinción entre descriptores auxiliares y descriptores ordinarios.

Nuestra experiencia en la construcción de thesaurus nos ha mostrado, por otra parte, que casi siempre es posible emplazar las palabras denominadas descriptores auxiliares dentro de la estructura semántica normal del thesaurus, en el interior de temas o de facetas específicos: la edición de una lista separada de descriptores auxiliares resulta, pues, carente de sentido.

.2 Relaciones semánticas

Un thesaurus contiene cuatro (si es monolingüe) o cinco (si es multilingüe) tipos de relaciones semánticas:

-pertenencia a un grupo, entre descriptores y tema(s) o faceta(s) a los que pertenecen esos descriptores;
-equivalencia interlingüística, entre descriptores y, dentro de algunos pocos thesaurus, entre no-descriptores que designan el mismo concepto, o un concepto equivalente, en dos o más lenguas;
-equivalencia semántica intralingüística, entre descriptores y no-descriptores, y viceversa, dentro de la misma lengua;
-jerarquía, entre descriptores de la misma lengua;
-asociación, entre descriptores de la misma lengua.

Destaquemos que esas relaciones se sitúan en planos diferentes:

-la primera enlaza los descriptores con el grupo, o campo semántico, o microdisciplina, al que pertenece, dentro de la misma lengua;
-la segunda enlaza los descriptores de diferentes lenguas;
-la tercera enlaza los descriptores y los no-descriptores, dentro de la misma lengua;
-las dos últimas enlazan entre sí descriptores de una misma lengua.

Se puede además observar que:

-las cuatro últimas relaciones son directamente útiles para la indización;
-mientras que la primera no sirve más que para agrupar los descriptores por «familias» y, por tanto, no es nada más que indirectamente útil para la indización.

Junto a estos cuatro o cinco tipos de relaciones denominadas «clásicas», que aparecen en la mayor parte de los thesaurus utilizados hoy día dentro de los sistemas documentales, existen otros tipos de relaciones, no generalizadas, que son empleadas por equipos de investigación y desarrollo en experiencias de laboratorio que pretenden mejorar las prestaciones de los lenguajes de indización.

.1 Pertenencia

.1 Definición

Relación asimétrica entre un descriptor y el o los campos semánticos a los que ha sido vinculado, dentro de un thesaurus monolingüe o en una versión lingüística de un thesaurus multilingüe(*)7.

En las diversas versiones lingüísticas de un thesaurus multilingüe, los conceptos expresados por descriptores monolingües se vinculan al/a los mismo(s) campo(s) semántico(s); estos campos semánticos son expresados por términos lingüísticamente equivalentes de una lengua a otra: la estructura de pertenencia es, por tanto, la misma en las diferentes versiones de un thesaurus multilingüe.

.2 Utilidad

La relación de pertenencia:

-es de interés para el usuario porque le permite situar el o los campos semánticos a los que pertenece el descriptor y, si fuera necesario, remitirse a ellos para buscar otros descriptores, no directamente ligados al primero por una relación jerárquica o asociativa.

Esta operación es a menudo necesaria para facilitar la formulación de una consulta y, a veces, para ayudar a la indización de un documento;

-es indispensable para poder preparar, automática o manualmente, una lista de descriptores por campos semánticos.

---

(*)7 Muy curiosamente, las normas sobre construcción de thesaurus no definen la relación de pertenencia, a pesar de que recomiendan subdividir el thesaurus en grupos: temas o facetas.

*Nota:* la mayoría de los autores que tratan sobre los thesaurus admiten como relaciones semánticas sólo las de equivalencia interlingüística, equivalencia semántica intralingüística, la jerarquía y la asociación. O bien omiten completamente la noción de pertenencia de los descriptores a uno o más campos semánticos, o bien la tratan únicamente de forma indirecta al definir el concepto de campo semántico (tema o faceta), pero sin referirse específicamente a la relación de inclusión de los descriptores dentro de esos campos. Ahora bien, esta relación no solamente es útil, como acabamos de mostrar, sino que además existe de hecho en la mayor parte de los thesaurus. Sería, pues, lamentable no hablar de ella. El hecho de que los nombres de los grupos, campos semánticos, microdisciplinas, temas o facetas, no sean en general términos utilizados para indizar los documentos y las consultas no constituye un argumento válido para no tener en cuenta las relaciones de pertenencia: lo mismo se puede decir, en efecto, de las relaciones de equivalencia intralingüística, en las que los no-descriptores no son tampoco términos utilizables para la indización. Ahora bien, todos los autores están de acuerdo en considerar que la equivalencia semántica es una auténtica relación semántica.

.3 Tipología

-*Mono-pertenencia:* un descriptor sólo puede pertenecer a un campo semántico.

-*Poli-pertenencia:* un descriptor puede pertenecer a uno, dos o más campos semánticos.

Ejemplo: FILOXERA puede figurar, dentro de un thesaurus de agricultura, en el grupo «insecto» y en el grupo «parasitología».

*Notas:*

1)La mono- y la poli-pertenencia están ligadas a la mono- y a la polijerarquía (cf. § 4.2.2.4).
2)La mayor parte de los thesaurus actuales aceptan la poli-pertenencia.
3)La mayor parte de los sistemas de clasificación sólo admiten la mono-pertenencia y la monojerarquía.

.4 Notación

No existe normalización en materia de relación de pertenencia.

Son posibles dos modalidades:

-o bien un código (por ej.: DOM, para dominio) precedido del enunciado del descriptor y seguido por el nombre del campo semántico.

Ejemplo:

```
FILOXERA
     DOM : insecto;
```

-o bien, si se ha asignado un número de orden o un código mnemónico a cada campo semántico, simplemente se

anota este número o código a continuación del enunciado del descriptor.
Ejemplo:

FILOXERA
(2256).

.2 Equivalencia interlingüística

.1 Definición

Relación bi-unívoca y simétrica entre descriptores de dos o más lenguas diferentes que representan el mismo concepto o un concepto similar(*)8.

Cuatro puntos deben ser esclarecidos dentro de esta definición:

-bi-univocidad (sinónimo: biyectividad): se define como la correspondencia establecida entre dos conjuntos (el thesaurus de la lengua A y el thesaurus de la lengua B), de tal manera que todo elemento de uno es imagen de un solo elemento del otro.

Dicho de otra forma, un thesaurus multilingüe se caracteriza por un número igual de descriptores en cada una de las versiones lingüísticas, y cada descriptor de una versión lingüística tiene obligatoriamente una, y sólo una, equivalencia interlingüística con un descriptor de cada una de las otras versiones lingüísticas;

-simetría: significa que si existe una relación de equivalencia interlingüística entre el descriptor a, dentro de la lengua A, y el descriptor b, dentro de la lengua B, existe la relación inversa entre el descriptor b y el descriptor a;

-similitud: se puede observar que no es cuestión de «traducción» de los descriptores dentro de las diferentes lenguas, sino más bien de «equivalencia» entre conceptos de diferentes lenguas. Parece, en efecto, que la conceptualización del mundo que nos rodea es diferente de una lengua a otra y que, por tanto, la significación de un concepto dentro de una lengua no tiene necesariamente una traducción fiel al 100 % dentro de otra lengua; la correspondencia que entre las lenguas se debe establecer (en virtud de la regla de bi-univocidad) se realiza, pues, a través de las «clases de equivalencia», más que por medio de traducciones literales; este asunto se trata más adelante, dentro de la tipología de las equivalencias interlingüísticas;

-equivalencia interlingüística entre no-descriptores: las lenguas difieren entre sí considerablemente por el tamaño de su vocabulario y por la riqueza de los matices que pueden ser aportados por los diversos sinónimos de los mismos conceptos.

Esto tiene las siguientes consecuencias:

---

(*)8 El concepto de equivalencia interlingüística no está formalmente definido en las normas para la construcción de thesaurus, ni siquiera en las que se dedican especialmente a los thesaurus multilingües.

-los descriptores ligados por una equivalencia interlingüística pueden tener un número de no-descriptores variable según la lengua;
-se puede establecer una equivalencia interlingüística entre algunos no-descriptores de diferentes lenguas, pero no necesariamente en todas las lenguas.

Es por estos dos motivos por lo que, en la mayor parte de los thesaurus multilingües, no se establece la relación de equivalencia interlingüística entre no-descriptores. Se limitan a enumerar, bajo los descriptores de cada lengua, los no-descriptores de la misma lengua.

En un número limitado de thesaurus, sin embargo, esta relación se establece, al menos para las lenguas y los no-descriptores para los que está justificada.

Ejemplo:
el concepto de niebla, en español, se expresa por la palabra:

niebla (gotas minúsculas de agua que flotan en el aire, cerca del suelo).

En un thesaurus sobre agricultura, se enlazarán al descriptor «NIEBLA» los no-descriptores siguientes, que representan:

+un cuasi-sinónimo:
bruma (niebla ligera)
+un término emparentado
llovizna (condensación de la niebla en lluvia muy fina).

Los mismos conceptos se expresan en inglés:

+fog (vapor suspendido cerca del suelo);
+haze (oscurecimiento de la atmósfera, cerca del suelo, debido al calor);
+mist (vapor de agua en forma de gotillas);
+drizzle (lluvia muy fina).

La clase de equivalencia se va a poder establecer entre
niebla, bruma, llovizna, en español
mist, fog, haze, drizzle, en inglés.

El concepto de niebla puede ser representado por el descriptor NIEBLA en español y MIST (cuya definición es la más próxima a la del español: niebla) en inglés; los no-descriptores niebla y llovizna remitirán al descriptor NIEBLA en español; los no-descriptores fog, haze y drizzle remitirán al descriptor inglés MIST; no se podrá establecer equivalencia interlingüística entre bruma, por una parte, y fog y haze, por otra; por el contrario, será posible, si la estructura del thesaurus lo permite, establecer una equivalencia interlingüística entre los no-descriptores llovizna y drizzle.

.2 Utilidad

La relación de equivalencia interlingüística permite:

-en el momento de la construcción de un thesaurus multilingüe, asegurarse de la coherencia de las diversas versiones lingüísticas;

-en el momento de la utilización del thesaurus, indizar los documentos en una lengua (la propia del servicio de documentación) o en varias lenguas (las de los documentos o las de los documentalistas), formular las consultas en la lengua del usuario y recuperar los documentos pertinentes con independencia de cuál sea la lengua en que han sido indizados.

.3 Tipología

Se distinguen los siguientes tipos de equivalencias interlingüísticas:

-equivalencia cierta y puntos de vista similares: identidad de los conceptos designados por los descriptores de dos o más lenguas:

+con la misma etimología.
Ejemplo:
español: metal
inglés: metal;

+con etimologías diferentes.
Ejemplo:
español: acero
inglés: steel;

-equivalencia cierta y puntos de vista diferentes.
Ejemplo:
español: búsqueda documental
inglés: information retrieval (= «recuperación» de información);

-equivalencia parcial, con una fragmentación diferente de la realidad.
Ejemplo:
español: carnero
inglés: sheep (= carnero vivo)
mutton (= carne de carnero).

-existencia de un concepto dentro de una lengua, pero no en otra.
Ejemplo:
inglés de América : highschool
única equivalencia posible en francés: highschool (acompañado de una nota explicativa).

.4 Notación

La norma francesa Z 47-103 - abril 1980, recomienda intercalar entre el descriptor en cuestión y su equivalencia lingüística un símbolo de la lengua del equivalente (según la norma NF X 03-002), seguido del signo : y de un espacio en blanco.

Ejemplo:
en un thesaurus francés:
MER
E : SEA
(E de English).

.3 Equivalencia semántica intralingüística

.1 Definición

Relación asimétrica entre un descriptor y un no-descriptor que expresa un concepto único o conceptos próximos, los cuales serán indizados de manera unívoca por el descriptor, dentro de un thesaurus monolingüe o en una versión lingüística de un thesaurus multilingüe(*)9.

En las diversas versiones lingüísticas de un mismo thesaurus, puede ser distinto el número de no-descriptores vinculados a cada uno de los descriptores equivalentes lingüísticamente de una lengua a otra: la estructura de equivalencia semántica es, por tanto, diferente en las diversas versiones de un thesaurus multilingüe.

La relación es asimétrica, en el sentido de que si un no-descriptor A está ligado por una relación de equivalencia semántica («USESE») a un descriptor B, se sigue necesariamente que B está ligado por una equivalencia semántica («USADO POR») a A.

Un descriptor puede tener ninguno, uno, dos o más no-descriptores. Un no-descriptor reenviará, en general, a un solo descriptor; en algunos thesaurus se admite, sin embargo, que un no-descriptor pueda reenviar a dos o más descriptores; este asunto se trata más adelante en el apartado sobre la tipología de las equivalencias semánticas.

.2 Utilidad

Las relaciones de equivalencia entre no-descriptores y descriptores son utilizadas como una pasarela entre la expresión de los conceptos en lenguaje natural, dentro de los documentos y las consultas, y la expresión de esos mismos conceptos en lenguaje controlado, en el momento de la indización de los documentos y de la formulación de las consultas.

Las relaciones de equivalencia entre descriptores y no-descriptores permiten:

-al usuario: delimitar el campo conceptual del descriptor y traducir a un lenguaje documental, por medio de descriptores, los conceptos que figuran en lenguaje natural dentro de los documentos y las consultas, y que han sido extraídos para la indización, eliminando al máximo los silencios y los ruidos;

-a quien construye el thesaurus: anotar las equivalencias semánticas de cada descriptor, dispersas, por

(*)9 Definición de AFNOR: «relación de sustitución entre descriptores y no-descriptores (sinónimos y no-sinónimos)» (norma NF Z 47-100 - diciembre 1981).

otra parte, dentro del thesaurus, siguiendo la localización alfabética de los no-descriptores. Esta anotación es indispensable porque, si se modifica el descriptor, cuando se realiza una puesta al día del thesaurus, es necesario acceder a cada uno de sus no-descriptores para realizar los cambios oportunos.

.3 Tipología

Se distinguen dos tipos principales de equivalencias semánticas, que contienen cada una dos sub-tipos, que reagrupan a su vez diferentes tipos:

+*Mono-equivalencia, con sinonimia fuerte y reconocida*

*Definición:* a un no-descriptor le corresponde un solo descriptor; esta equivalencia es verdadera para todos los dominios, y por tanto para todos los thesaurus. Lo que puede variar de un thesaurus a otro es el término conservado como descriptor y aquel o aquellos que ha(n) recibido el estatuto de no-descriptor.

*Tipos:*

-sinonimia verdadera
ejemplo: inteligibilidad y comprehensibilidad; concepto y noción;

-variante ortográfica
ejemplo: Méjico y México.

*Nota:* las variantes ortográficas son muy numerosas entre el inglés y el americano; ejemplo: colour y color;

-sigla
ejemplo: AFNOR y Asociación Francesa de Normalización;

-variante de escritura
ejemplo: CEE y C.E.E.;

-equivalencia lingüística considerada como equivalencia semántica cuando el término en lengua extranjera es de uso corriente en una de las lenguas del thesaurus
ejemplo: software y programa.

*Nota:* la equivalencia entre estos términos deberá, además, ser señalada, por medio de una equivalencia interlingüística cuando se trate de un thesaurus multilingüe.

-equivalencia entre idioma contemporáneo e idioma antiguo
ejemplo: avión y aeroplano;

-antonimia
ejemplo: coherencia e incoherencia.

*Nota:* evidentemente dos términos antónimos no son sinónimos en el lenguaje usual. No sucede lo mismo en los lenguajes documentales si existe una gradación entre los dos términos antónimos: cuando un documento trata sobre uno de los conceptos antónimos, trata así mismo, poco o mucho, sobre el otro.

-equivalencia entre el lenguaje usual y el lenguaje científico, en la lengua del thesaurus o en latín
ejemplos: alimentación y nutrición; manzana y poma;

-equivalencia entre lenguaje usual y lenguaje administrativo
ejemplo: dentista y odontólogo;

-equivalencia entre nombre común y nombre de marca
ejemplo: yoghourt y danone.

*+Mono-equivalencia, con sinonimia relativa y convencional*

*Definición:* a un no-descriptor le corresponde un solo descriptor; esta equivalencia es verdadera para algunos thesaurus, pero no para otros, según cuáles sean los dominios principales tomados en cuenta por los thesaurus.
Para dos términos de significación próxima, dentro de un dominio determinado:

-si ese dominio es principal, los dos términos serán, las más de las veces, separados en dos descriptores distintos;
-si es secundario, los dos términos serán asociados, las más de las veces, por una relación de equivalencia semántica, siendo elegido uno como descriptor y el otro como no-descriptor.

*Tipos:*

-términos que pertenecen a la misma familia
ejemplo: cuerda y cordaje (separados dentro de un thesaurus de navegación, donde se encontrarán los dos descriptores «CUERDA» y «CORDAJE», y asociados dentro de un thesaurus textil, en el que se encontrará, por ejemplo, «CUERDA» como descriptor y «CORDAJE» como no-descriptor, asociado a «CUERDA» por una relación de equivalencia semántica);
-términos de niveles jerárquicos diferentes
ejemplo: fruta y manzana (separados en un thesaurus de agricultura y asociados en un thesaurus de economía);
-cuasi-sinonimia
ejemplo: inclinación y pendiente (separados en matemáticas y asociados en geografía).

+*Pluri-equivalencia obligatoria*

*Definición:* a un no-descriptor le corresponden dos descriptores que se utilizan obligatoriamente juntos.

*Tipos:*

-descomposición de un término en sus componentes semánticos fundamentales
ejemplo: lámpara y APARATO + ALUMBRAR.

*Nota:* el «semantic factoring» figura en la mayor parte de las normas sobre construcción de thesaurus; en realidad, este concepto, que procede de la lingüística, se utiliza en algunos sistemas de inteligencia artificial y de traducción automática; pero no interviene en casi ningún thesaurus.

-descomposición de una expresión en sus palabras constituyentes:
ejemplo: recogida de algodón y RECOGIDA + ALGODON

*Nota:* si dentro de un thesaurus existen los descriptores «RECOGIDA» y «ALGODON», es completamente inútil incluir en él el término «recogida de algodón» como no-descriptor; en efecto, no hay ninguna ambigüedad, y todos los usuarios del thesaurus indizarán de la misma manera un documento o una consulta que trate sobre la recogida del algodón; sin embargo, dentro de un thesaurus especializado en producción y/o explotación agraria, se mantendrá en general, además, el descriptor «RECOGIDA DE ALGODON», por ser más específico.

-descomposición de un concepto complejo en descriptores que representen los conceptos más simples
fibrocemento y AMIANTO + CEMENTO (en el caso de un thesaurus sobre la construcción en general; dentro de un thesaurus específico sobre los materiales de construcción, el término «FIBROCEMENTO», será un descriptor, con «AMIANTO» y «CEMENTO» como descriptores genéricos).

*Nota:* en este caso, la relación de pluri-equivalencia es útil, porque los usuarios no saben necesariamente que el fibrocemento es un material compuesto de cemento y de fibras de amianto.

+*Pluri-equivalencia facultativa*

*Definición:* a un no-descriptor le corresponden dos o más descriptores, de los que sólo uno debe ser elegido en el momento de la indización. El único tipo de pluriequivalencia facultativa corresponde a un no-descriptor polisémico
ejemplo: orientador y MAESTRO o TUTOR.
La mayor parte de los thesaurus no utilizan la noción de equivalencia facultativa; sus productores prefieren desambiguar directamente el término a nivel de descriptor:

-ya sea por un modificador

ejemplo:orientador (formador): úsese: TUTOR
orientador (jefe): úsese: MAESTRO.
-ya sea por una nota de aplicación
orientador: úsese: TUTOR
NA: con el sentido de jefe, úsese MAESTRO.

.4 Notación

Existen varias notaciones normalizadas de relaciones de equivalencia.
Nosotros consideraremos aquí:

-la norma experimental francesa de AFNOR: Z 47-100 - diciembre 1973, que ha sido reemplazada por la norma homóloga NF Z 47-100 - diciembre 1981, pero de la que nosotros hemos conservado la codificación por siglas, abandonada en el texto de 1981 a favor de una notación internacional por símbolos gráficos; esta codificación por siglas, más explícita, sigue siendo de hecho usada por numerosos thesaurus, incluso algunos construidos después de 1981;
-la norma experimental francesa Z 47-103 - abril 1980, que proporciona la simbolización por signos gráficos aplicable a todas las lenguas que citamos a continuación;
-la norma inglesa del B.S.I. (British Standards Institution) B.S. 5723: 1979;
-la norma internacional de I.S.O. (International Standardization Organization): I.S.O 2788 - 1974.

Como se puede observar en el cuadro, las normas sólo tienen en cuenta algunas categorías de equivalencias semánticas, de forma no homogénea, por otra parte, de una norma a otra.

| Categorías de equivalencia | Siglas francesas (Z47-100-1973) | Siglas inglesas (BS5723: 1979) | Siglas francesas (e inglesas) (ISO 2788 -1974) | Símbolos gráficos (Z47-103 -1980) |
|---|---|---|---|---|
| *Mono-equivalencia* | | | | |
| entre el no-descriptor(ND) y el descriptor (D) | ND EM D | ND USE D | ND EM (use) D | ND ->D |
| entre el descriptor (D) y el no-descriptor (ND) | D EP ND | D UF ND | D EP (UF) D | D = ND |
| *Pluri-equivalen. obligatoria* | | | | |
| entre el no-descriptor (ND) y | ND EM D1 | ND USE D1 | ND EM (USE) D1 | ND |

| | | | | |
|---|---|---|---|---|
| el descriptor (D1 y D2) | ET D2 | & D2 | + D2 | ->D1 +D2 |
| entre cada uno de los des.(D1 y D2) y el no-des. (ND) (en este cuadro sólo se considera la entrada D1) | D1 EP ND AVEC D2 | no previsto en la norma | D1 EPC(UFC) ND (reenvío a D2 no previsto en la norma) | D1 + D2 = ND |
| *Pluri-equivalen. facultativa* | | | | |
| entre el no-descriptor (ND) y los descriptores (D1 o D2) | ND véase D1 o D2 | no prev. en la norma | no prev. en la norma | ND ->D1? ->D2? |
| entre cada uno de los descriptores (D1 o D2) y el no-descriptor(ND) | no prev. en la norma | no prev. en la norma | no prev. en la norma | no prev. en la norma |

Significado de las siglas:

EM :úsese (employer)
EP :usado por (employé pour)
EPC :usado por en combinación (employé pour en combinaison)
UF :used for
UFC :used for combined.

*Notas:*

1)Los símbolos gráficos de la norma inglesa (no tenidos en cuenta en el cuadro anterior) no son los mismos que los de la norma francesa:

| AFNOR | BSI |
|---|---|
| -> | = |
| = | =,/ |
| + | & |

2)En alemán se usa
-para «úsese»: BS (Benutzen)
-para «usado por»: BF (Benutzt für)

.4 Jerarquía

.1 Definición

Relación asimétrica entre dos descriptores de los que uno es superior a otro por un carácter normativo, dentro de un thesaurus monolingüe o dentro de una versión lingüística de un thesaurus multilingüe(*)10.

En las diversas versiones lingüísticas de un thesaurus multilingüe, los conceptos, expresados por descriptores monolingües, están ligados entre sí por las mismas relaciones jerárquicas: la estructura jerárquica es, pues, la misma para las diferentes versiones de un thesaurus multilingüe.

No existen relaciones jerárquicas entre no-descriptores, ni aun cuando fuera posible establecer tal jerarquía, ni entre no-descriptores y descriptores.

La relación es asimétrica, en el sentido de que si el descriptor A es superior al descriptor B, se sigue necesariamente que B es inferior a A.

Un descriptor puede tener:

-ningún, uno, dos o más descriptores que sean inferiores a él;
-ningún, uno, dos o más descriptores que sean superiores a él:
+si no hay ninguno, significa que el descriptor es «una cabeza de jerarquía» (en inglés: top term);
+si sólo hay uno, el descriptor es monojerárquico;
+si hay dos o más, el descriptor es polijerárquico.

La polijerarquía se puede entender

-en el interior de un mismo grupo de descriptores.
Ejemplo: ACEITE DE OLIVA
genérico: ACEITE PARA ALIMENTACION
genérico: ACEITE VEGETAL
con
ACEITE PARA ALIMENTACION
dom: producción vegetal
ACEITE VEGETAL
dom: producción vegetal.

-o entre grupos distintos.
Ejemplo: FILOXERA
genérico: INSECTO HEMIPTERO
genérico: PARASITO ANIMAL
con
INSECTO HEMIPTERO
dom: insecto
PARASITO ANIMAL
dom: parasitología.

(*)10 Definición de AFNOR: «relación entre dos descriptores de los que uno está subordinado al otro» (NF Z 47-100 - diciembre 1981).

Los descriptores de un thesaurus pueden formar cadenas jerárquicas.

Ejemplo:

ARGAMASA
  ARGAMASA HIDRAULICA
  CEMENTO
    CEMENTO REFRACTARIO.

El número de niveles jerárquicos puede extenderse, según los thesaurus, desde unos pocos hasta unos quince.

En algunos thesaurus la regla es incluir cada descriptor dentro de al menos una cadena jerárquica; los descriptores independientes, si los hay, son automáticamente considerados cabeza de jerarquía.

.2 Utilidad

Las relaciones jerárquicas se utilizan en el momento de:

-la indización de los documentos, en todos los sistemas documentales, para escoger los descriptores que designen de la forma más precisa, y por tanto más específica, los conceptos que se van a representar, de manera que se eliminen lo más posible los ruidos en el momento de la búsqueda;

-la indización de los documentos y, sólo en algunos sistemas documentales, para añadir automáticamente uno o más niveles jerárquicos a cada uno de los descriptores elegidos por el documentalista, de manera que se aumente la llamada (aunque a veces en detrimento de la precisión) durante el momento de la búsqueda;

-la búsqueda documental, y esto, en todos los sistemas documentales, para enriquecer, si es necesario, la formulación de la consulta añadiéndole uno o más descriptores jerárquicamente superiores o inferiores.

Ejemplo: ante una consulta sobre el cemento, podrá ser útil responder suministrando documentos indizados no solamente por «CEMENTO», sino también todos los tipos específicos de cemento (CEMENTO REFRACTARIO...).

Y, a la inversa, si, en respuesta a una consulta específica sobre el «CEMENTO REFRACTARIO», no se encuentran documentos pertinentes, el usuario podrá ampliar su consulta y encontrar respuestas adecuadas en los documentos indizados por «CEMENTO».

Las relaciones jerárquicas aportan, pues, una ayuda al usuario para permitirle que formule su consulta escogiendo de la mejor forma posible los descriptores al/a los nivel(es) jerárquico(s) adecuado(s).

.3 Tipología

Se distinguen en general dos tipos de relaciones jerárquicas:

-*genérico/específico*: el descriptor inferior A es específico del descriptor superior B, lo que se concreta en que se puede responder afirmativamente a la pregunta:

¿el/la A es un(a) B?

Ejemplo: MANZANA y FRUTA.

-*partitivo*: el descriptor inferior A es una parte particular del descriptor superior B, lo que se concreta en que se puede responder afirmativamente a la pregunta:
¿el/la A es una parte de B?

Ejemplos: MOTOR y VEHICULO
FRANCIA y EUROPA.

Un problema práctico que se plantea cuando se concibe un thesaurus es en qué orden se van a colocar los descriptores del mismo nivel jerárquico que dependen de un mismo término.
En la mayoría de los thesaurus, este orden es simplemente alfabético. Por ejemplo, bajo el descriptor genérico FRUTA se encontrarán los descriptores específicos:

CEREZA
FRAMBUESA
FRESA
LIMON
MANZANA
MELON
NARANJA
PERA
SANDIA
(...)

Este tipo de ordenación es el más sencillo; ofrece la ventaja de que puede ser realizada automáticamente por el ordenador en el momento en que se edita el thesaurus. Presenta el pequeño inconveniente de que no es homogénea de una lengua a otra; en inglés, por ejemplo, se encontrará:

APPLE
CHERRY
LEMON
MELON
ORANGE
PEAR
RASPBERRY
STRAWBERRY
WATER-MELON
(..)

Un inconveniente más importante aparece en el caso de una lista de varias decenas de descriptores del mismo nivel jerárquico, cuando se busca un descriptor preciso, sin que se tenga en la cabeza necesariamente su enunciado (esta situación se presenta muy a menudo en la búsqueda documental) y sea necesario para encontrarlo recorrer, como término medio, la mitad de la lista.
Foskett (1982) ha propuesto ocho modos de ordenación para descriptores de un mismo nivel jerárquico que dependan

de un mismo descriptor genérico (llama a esos descriptores «coordinate descriptors»); el modo de ordenación aplicable depende del dominio en el que se sitúen los descriptores en cuestión:

-orden cronológico, aplicable
+en materia histórica y estilística
+para procesos que se desencadenan secuencialmente (ej.: producción, distribución, consumo),
-orden de evolución, aplicable especialmente en biología;
-orden de complejidad creciente, aplicable, por ejemplo, en el dominio de las ciencias (geometría euclídea, geometría no euclídea...);
-orden de dimensión creciente (por ejemplo: composición musical: solo, dúo, trío, cuarteto...) o decreciente (por ejemplo: subdivisión territorial, país, región, provincia, municipio...);
-ordenación en el espacio (por ejemplo, disposición de los países en orden de contigüidad);
-orden de preferencia o de frecuencia de uso en los documentos tratados por el servicio (por ejemplo: religión: catolicismo, protestantismo, judaísmo, islamismo, budismo...);
-orden tradicional (por ejemplo, ordenación de las ciencias: física, química, biología...);
-y, por último, el orden alfabético, aplicable cuando no se pueda utilizar ningún otro modo de ordenación.

Esta aproximación es interesante, en la medida en que un cierto orden en el interior de una lista de descriptores del mismo nivel jerárquico puede facilitar el trabajo del usuario.

Sin embargo, presenta dos inconvenientes principales:

-los descriptores de un mismo nivel jerárquico pueden pertenecer a dos o más modos de ordenación, y no sólo a uno;
-la elección de los descriptores según una o más modalidades puede ser realizada por el hombre, pero no por el ordenador; para automatizar la producción de un thesaurus de este tipo es necesario introducir un código, interpretable por la máquina, que represente el orden de disposición de cada descriptor dentro de la lista.

Esto es precisamente lo que se ha realizado dentro de un tipo particular de thesaurus, denominado thesaurus facetado (Aitchison - 1969): los descriptores inferiores de un descriptor genérico son agrupados por «facetas».

Estas facetas no tienen mucho que ver con las facetas utilizadas para estructurar un thesaurus, tal y como las hemos expuesto en el § 4.2.1.1.2. Aquí no se trata de repartir los términos según su función lingüística (entidades, propiedades...), sino más bien según su significado y su función dentro de los dominios cubiertos por el thesaurus.

Por ejemplo, para repartir en clases homogéneas los descriptores específicos del descriptor «MOTOR», éstos son agrupados según tres facetas:

```
MOTOR
     según la energía
          MOTOR ELECTRICO
          MOTOR HIDRAULICO
          MOTOR TERMICO
          (...)
     según su empleo
          MOTOR DE AVION
          MOTOR DE BARCO
          MOTOR DE COCHE
          (...)
     según la potencia
          MOTOR GRANDE
          MOTOR MEDIANO
          PEQUEÑO MOTOR
          (...)
```

Los términos «según la energía», «según su empleo» y «según la potencia» no son descriptores; son tipos de facetas, que juegan el mismo papel que las siglas de relaciones para agrupar los términos de una misma naturaleza.

El interés de estas facetas es facilitar la consulta del thesaurus; sin facetas, en efecto, la lista de los motores aparecerían como sigue:

```
MOTOR
     MOTOR DE AVION
     MOTOR DE BARCO
     MOTOR DE COCHE
     MOTOR ELECTRICO
     MOTOR GRANDE
     MOTOR HIDRAULICO
     MOTOR MEDIANO
     MOTOR TERMICO
     PEQUEÑO MOTOR
```

lo cual es ciertamente menos cómodo de usar.

*Nota:* en muchos thesaurus no se distingue entre los dos tipos de jerarquías por medio de una notación especial: todas las relaciones jerárquicas son dispuestas en un mismo plano y simbolizadas por una notación indiferenciada.

.4 Notación

| Categorías de relaciones jerárquicas | Siglas francesas (Z47-100-1973) | Siglas inglesas (BS5723-1979) | Siglas francesas (e inglesas) (ISO 2788-1974) | Símbolos gráficos (Z47-103-1980) |
|---|---|---|---|---|
| *Jerarquía indiferenciada* | | | | |

| | | | | |
|---|---|---|---|---|
| entre el descr. superior(A) y el desc.inferior(B) | A<br>TS B | A<br>NT B | A<br>TS (NT) B | no prev. en la norma |
| entre el descr. inferior(B) y el desc.superior(A) | B<br>TG A | B<br>BT A | B<br>TG (BT) A | no prev. en la norma |
| *Genérico/ específico* | | | | |
| entre el descr. genérico(A) y el desc. espec. (B) | A<br>TSG B | A<br>NTG | A<br>TSG (NTG) B | A<br>>B |
| entre el descr. especif.(B) y el desc. genér. (A) | B<br>TGG A | B<br>BTG A | B<br>TGG (BTG) A | B<br><A |
| *Partitivo* | | | | |
| entre el descr. que expresa el todo(A) y el de la parte (B) | A<br>TSP B | A<br>NTB B | A<br>TSP (NTP) B | A<br>>---B |
| entre el descr. que expresa la parte (B) y el del todo (A) | B<br>TGP | B<br>BTP A | B<br>TGP (BTP) A | B<br>---<A |

Significado de las siglas:

TS  :término específico (terme spécifique)
TG  :término genérico (terme générique)
TSG :término específico genérico (terme spécifique générique)
TGG :término genérico genérico (terme générique générique)
TSP :término específico partitivo (terme spécifique partitif)
TGP :término genérico partitivo (terme générique partitif)
NT  :narrower term
BT  :broader term

*Notas:*

1)Los símbolos gráficos de la norma inglesa no son los mismos que los de la norma francesa:

| Tipo | AFNOR | BSI |
|---|---|---|
| indiferenciado | nada | > y < |
| genérico/específico | > y < | nada |

| | | |
|---|---|---|
| partitivo | >--- y ---< | > P y < P |

2)en alemán, se utiliza:

-para «término genérico» : OB (Oberbegriff)
-para «término específico» : UB (Unterbegriff).

3)la designación de facetas puede variar de un thesaurus a otro y, dentro de un mismo thesaurus, de un dominio a otro; por tanto no está normalizada.

.5 Asociación

.1 Definición

Relación simétrica entre dos descriptores que designan conceptos que, aunque no ligados entre sí por una equivalencia semántica o una jerarquía, son susceptibles de evocarse mutuamente, por asociación de ideas, dentro de un thesaurus monolingüe o de una versión lingüística de un thesaurus multilingüe(*)11.

Dentro de las diversas versiones lingüísticas de un thesaurus multilingüe, los conceptos, expresados por descriptores monolingües, están ligados entre sí por las mismas relaciones asociativas: la estructura asociativa es, pues, la misma dentro de las diversas versiones de un thesaurus multilingüe.

No existen relaciones asociativas entre no-descriptores y descriptores, ni entre descriptores ya ligados por una relación jerárquica. La relación es simétrica, en el sentido de que, si el descriptor A está asociado al descriptor B, se sigue necesariamente que B está asociado a A.

Un descriptor puede tener ninguna, una, dos o más relaciones asociativas.

.2 Utilidad

Al igual que las relaciones jerárquicas, las relaciones asociativas aportan una considerable ayuda para la búsqueda documental, gracias a la explicitación de las asociaciones de ideas y de los puntos de vista distintos que pueden haber adoptado los autores de los documentos pertinentes.

Ejemplo: ante una consulta sobre las consecuencias del frenazo de un vehículo, parece evidente responder buscando documentos indizados por FRENAZO.

Pero es posible que los documentos indizados por DERRAPAJE aporten también informaciones útiles. Si dentro del thesaurus figura la relación entre esos dos términos, el usuario dispondrá de un hilo conductor que le guiará de

(*)11 Definición de AFNOR: «relación que indica otras analogías, o lazos de significación entre los descriptores, diferentes de la relación jerárquica o de equivalencia» (NF Z 47-100 - diciembre 1981).

FRENAZO a DERRAPAJE; si por el contrario esta relación no existe, el usuario estará abandonado a la propia inspiración del momento.

.3 Tipología

Existe un gran número de tipos de asociaciones:

-causalidad
ejemplo: ENFERMEDAD e INFECCION
ACCIDENTE y VICTIMA;

-instrumentación
ejemplo: COMERCIO y MERCADO
LUBRICANTE y ENGRASE

-sucesión en el espacio o en el tiempo
ejemplo: TERAPEUTICA y POSOLOGIA
PLANTA y SEMILLA

-concomitancia
ejemplo: SINTOMA y ENFERMEDAD;

-materiales constitutivos
ejemplo: MEDICAMENTO y EXCIPIENTE
CONSTRUCCION y MATERIALES DE CONSTRUCCION;

-similaridad (en el caso de que los dos conceptos similares no hayan sido mantenidos el uno como descriptor y el otro como no-descriptor)
ejemplo: ENSEÑANZA y FORMACION
IMPULSO y EXCITACION;

-antonimia (en el caso de que los dos conceptos antónimos no hayan sido mantenidos el uno como descriptor y el otro como no-descriptor)
ejemplo: INHIBICION y EXCITACION;

-propiedad
ejemplo: TRAFICO y FLUIDEZ
LASER y COHERENCIA;

-objeto de una acción, de un proceso, de una disciplina
ejemplo: ENTOMOLOGIA e INSECTO;

-localización
ejemplo: ENSEÑANZA y ESCUELA.

*Notas:*

1) En general no se crea relación asociativa entre los descriptores de una misma cadena jerárquica:

A
B
C
D

ni entre los descriptores específicos de un mismo descriptor genérico:

A
 B
 C
 D

Efectivamente, en ambos casos la relación entre B y D, por ejemplo, es inútil, porque o bien la cadena jerárquica (en el primer caso) o bien la proximidad física (en el segundo caso) proporcionan ya al usuario una asociación de ideas entre esos términos.

2) No se crea relación asociativa entre un descriptor compuesto (de dos o más palabras) y el o los descriptores unitérminos constituyentes si el thesaurus contiene una presentación permutada, que hace ya aparecer esta relación por un procedimiento puramente tipográfico (cf § 5.4.2.2).

.4 Notación

Se utiliza una única notación para todos los tipos de relaciones asociativas:

-sigla francesa: TA (terme associé)
-sigla inglesa: RT (related term);
-símbolo gráfico internacional: -
-sigla alemana: VB (Verwandter Begriff)

.6 Relaciones no clásicas

Los cinco tipos de relaciones que se han examinado hasta ahora están presentes en la mayor parte de los thesaurus en uso.

Una serie de trabajos de investigación y desarrollo, que pretenden mejorar las prestaciones de los sistemas documentales, se fundamentan sobre otras tipologías de relaciones. Cada equipo de investigación define su propia tipología y la aplica a muestras de algunas decenas o, en raras ocasiones, a varios cientos de documentos y de consultas.

Generalmente los resultados son difícilmente extrapolables a colecciones que contengan decenas o cientos de miles de documentos.

Citemos, como ejemplo de estos trabajos, las investigaciones realizadas sobre el «thesaurus relacional» (Wang - 1985).

A partir del lenguaje natural de un corpus de 222 resúmenes de documentos, se ha construido una serie de cinco «thesaurus»: cada thesaurus integra todos los pares de palabras ligadas por uno de los cinco grupos de relaciones definidos por los autores:

-parte-todo
 ejemplos: parte-todo (cuerno PART vaca)
           jefe-organismo (jefe CAP tribu)
           personal-objeto (equipo EQUIP cañón)

conjunto-elemento (rebaño SET carnero);

-colocación (términos que aparecen frecuentemente en la misma frase)
ejemplos:agente típico (conquistador TAGENT conquistar)
resultado típico (hoyo TRESULT cavar)
instrumento típico (aguja TINST coser)
hábitat-objeto (Africa HOME hiena)
sonido característico (ladrido SON perro)
verbo de destrucción (corregir LIQU falta);

-relación paradigmática (términos ligados por los mismos rasgos semánticos)
ejemplos:hembra-término no marcado (yegua FEMALE caballo)
descendiente-progenitor (potro CHILD caballo)
causa-acción (enviar CAUSE ir)
verbo-adjetivo (blanquear BECOME blanco)
adjetivo-nombre (solar ADJN sol)
adjetivo-verbo (combustible ABLE quemar)
imperfecto-infinitivo (iba PAST ir);

-taxonomía (lo que denominamos relación género/especie) y sinonimia
taxonomía (canario TAX pájaro)
sinonimia (veloz SYN rápido);
-relación antonímica
complementariedad (soltero COMP casado)
antonimia (caliente ANTI frío)
inversión (comprar CONV vender)
parentesco recíproco (marido RECK mujer).

## 5. **Presentación**

### .1 Lista de palabras clave

Las listas de palabras clave se editan solamente ordenadas de forma alfabética, ya que no disponen de estructura semántica sobre la que se pueda basar otra presentación. Estas listas incluyen en general el conjunto de palabras clave de la base de datos, y al lado de cada una se indica su frecuencia dentro de la colección.

Esta frecuencia puede referirse:

-o bien al número de documentos registrados dentro de la colección en los que la palabra clave aparece al menos una vez;

-o bien -lo más habitual- al número de apariciones de la palabra clave en el conjunto de los documentos de la colección; este número es más elevado que el primero, ya que una misma palabra clave puede intervenir más de una vez dentro de un documento.

En algunos servicios de documentación grandes, la lista general está dividida en una serie de listas parciales de palabras clave, según las secciones del servicio o según las fuentes, es decir, grosso modo, por dominio de actividad.

.2 Lista de descriptores libres

Las listas de descriptores libres se editan solamente ordenadas de forma alfabética, por las mismas razones que las listas de palabras clave.

Junto a cada descriptor se indica su frecuencia, que corresponde siempre al número de documentos de la colección indizados por medio de ese descriptor.

La lista puede ser:

-general;
-dividida en listas parciales, según los servicios o las fuentes, o mejor, según el o los códigos de clasificación que figuran en cada documento. Esto permite a cada documentalista seguir mejor la evolución del vocabulario en el dominio o dominios de los que se ocupe;
-permutado, es decir, cada descriptor aparece en la lista tantas veces como palabras significativas contiene; esta presentación es muy útil para agrupar en un mismo sitio los diferentes descriptores que tienen en común una palabra significativa y que pueden, por tanto, designar conceptos con significados próximos: se trata, pues, de un esbozo de estructura asociativa.

Ejemplo: si tenemos una lista que contiene los descriptores:

ACERO RAPIDO
---
---
HERRAMIENTA DE ACERO;

-en una lista alfabética normal, estos descriptores estarán separados el uno del otro por varias páginas;

-mientras que en una lista alfabética permutada cada uno de esos descriptores aparecerá no sólo en su localización alfabética normal, sino que además se encontrará una agrupación de los descriptores en los que la palabra «acero» interviene:

ACERO RAPIDO
HERRAMIENTA DE ACERO
---
---
HERRAMIENTA DE ACERO
---
---
ACERO RAPIDO.

.3 Lista de autoridades

Las presentaciones de las listas de autoridades son las mismas que las de las listas de descriptores libres:

- lista alfabética general
  + no permutada,
  + permutada;
- listas parciales.

.4 Thesaurus de descriptores

La existencia de una estructura semántica permite concebir un gran número de tipos de presentaciones:

- presentación en listas:

  +lista alfabética estructurada (llamada también presentación diccionario):
  ->completa o por grupos
  ->no permutada o permutada,
  +índice alfabético (denominado también presentación léxica)
  ->completa o por grupos,
  +lista jerárquica (llamada también presentación sistemática),

-presentación gráfica:

  +diagrama de flechas de grupo,
  +terminograma de grupo.

.1 Utilidad

En la inmensa mayoría de los servicios de documentación que utilizan un thesaurus de descriptores encontramos al menos dos representaciones:

- una presentación completa:
  +preferentemente, la lista alfabética estructurada completa, con permutación,
  +si no, la lista alfabética estructurada completa sin permutación;
-una presentación de grupo; en orden decreciente de preferencia (es decir, de facilidad de uso):
  +terminograma
  +diagrama de flechas
  +lista alfabética estructurada por grupos, sin permutación
  +lista jerárquica
  +índice alfabético por grupos.

Muchos servicios que utilizan una presentación gráfica editan, para cada página de diagrama de flechas o de terminograma, el índice alfabético del grupo correspondiente.

Algunos (pocos) servicios editan cuatro y hasta cinco presentaciones distintas; la experiencia muestra que el uso de estas presentaciones complementarias es nulo o casi nulo.

La presentación completa se utiliza sobre todo en el momento de la indización de los documentos: los conceptos que se van a indizar aparecen en los documentos de manera explícita en la mayor parte de los casos, y la lista

alfabética es suficiente, muy a menudo, para identificar los descriptores adecuados. Sólo sucede muy raramente, cuando no se encuentra un no-descriptor o un descriptor que corresponda al concepto a representar, que haya que remitirse a una presentación de grupo; esta permite, en efecto, pasar revista rápidamente a todos los descriptores relativos a un campo semántico (microdisciplina) y elegir «al vuelo» los descriptores más representativos de los conceptos que hay que indizar.

Por el contrario, las consultas contienen un porcentaje importante de conceptos implícitos, más o menos próximos, semánticamente hablando, a los conceptos que figuran expresamente en la consulta. La presentación de grupo, y especialmente las presentaciones gráficas, permiten barrer en unos instantes un campo semántico y encontrar los descriptores más pertinentes, es decir, los descriptores que designen los conceptos que intervienen en los documentos más susceptibles de contener informaciones en respuesta a la consulta del usuario.

.2 Presentaciones en listas

Las listas se caracterizan por una estructura por «entradas».

Una entrada es un descriptor o un no-descriptor, acompañado de un cierto número de informaciones semánticas, que varía según el tipo de lista.

Ofrecemos a continuación una descripción general de ese acompañamiento; pero es necesario saber que cada thesaurus difiere, por una o más variantes, de esta presentación general.

.1 Lista alfabética estructurada completa, no permutada

En ella se encuentran dos tipos de entradas alfabéticas:
-los descriptores;
-los no-descriptores.

Dentro de la entrada de cada descriptor se encuentra, con una ordenación idéntica para todo el thesaurus:

-la indicación del grupo o grupos (campos semánticos, microdisciplinas) a que pertenece;
-las equivalencias interlingüísticas (en el caso de un thesaurus multilingüe);
-las notas explicativas, de aplicación e históricas;
-las equivalencias semánticas (no-descriptores);
-los descriptores genéricos según uno, varios o todos los niveles de jerarquía ascendente;
-los descriptores específicos según uno, varios o todos los niveles de jerarquía descendente;
-los descriptores asociados.

En los thesaurus que disponen de una presentación por diagrama de flechas o por lista jerárquica, puede indicarse la localización de los descriptores dentro de alguna(s) de estas presentaciones, bajo la forma de coordenadas (en un diagrama de flechas, designando a qué grupo pertenece) o de un código secuencial (en una lista jerárquica).

Las equivalencias lingüísticas van precedidas por un código que indica la lengua correspondiente y dispuestas según el orden alfabético de ese código, que es el mismo en todas las versiones lingüísticas del thesaurus.

Los no-descriptores y los descriptores asociados van precedidos por la sigla de relación correspondiente y dispuestos por orden alfabético.

Los descriptores genéricos y específicos van precedidos por la sigla de relación correspondiente y dispuestos, dentro de cada nivel jerárquico, por orden alfabético; cada nivel jerárquico adicional está marcado por un sangrado hacia la derecha.

Dentro de la entrada de cada no-descriptor se encuentra:

-el o los descriptores que se han de utilizar;

-la indicación del grupo o grupos a que pertenece cada uno de esos descriptores.

En algunos thesaurus, el carácter de no-descriptor del término de la entrada está marcado tipográficamente:

-bien usando caracteres distintos: letras más pequeñas o itálicas;

-o bien con un signo distintivo, como por ejemplo un asterisco.

.2 Lista alfabética estructurada completa, permutada

Se encuentran cuatro tipos de entradas alfabéticas:

-los descriptores, ordenados según su primera palabra;

-los no-descriptores, ordenados según su primera palabra;

-los descriptores, ordenados según cada una de sus palabras significativas distintas de la primera;

-los no-descriptores, ordenados según cada una de sus palabras significativas distintas de la primera.

Para cada descriptor o no-descriptor ordenado según su primera palabra, se encuentran exactamente las mismas informaciones que en la lista no permutada (cf. § 5.4.2.1).

Para cada descriptor ordenado según las demás palabras significativas, no se encuentra ninguna información, a no ser, en algunos thesaurus, la indicación del grupo o grupos a los que pertenece.

Para cada no-descriptor ordenado según las otras palabras significativas se encuentra:

-el o los descriptores que se han de utilizar;

-y, en algunos thesaurus, la indicación de los grupos a los que pertenece.

.3 Lista alfabética estructurada por grupos

Se crea una lista para cada campo semántico. En el interior de cada uno de esos grupos se incluyen solamente los descriptores (y sus no-descriptores) pertenecientes al grupo, así como toda su información semántica (como en § 5.4.2.1 y 5.4.2.2). Un descriptor que pertenece a varios grupos aparece dentro las listas propias de cada uno de esos grupos.

.4 Indice alfabético completo

En él se encuentra una lista alfabética de los descriptores y (en algunos thesaurus) de los no-descriptores (en este caso con reenvío hacia el descriptor o descriptores correspondientes).

Según los thesaurus, a continuación del enunciado del descriptor pueden o no incluirse:

-sus equivalencias lingüísticas: en tal caso, se tiene un índice multilingüe del thesaurus, y no se integran estas relaciones dentro de la lista alfabética estructurada;
-la indicación del grupo o grupos a que pertenece;
-la indicación sobre su localización (en forma de coordenadas) dentro del o de los diagramas de flechas en los que figura;
-la indicación sobre su localización (en forma de un código secuencial) dentro de la lista jerárquica.

.5 Indice alfabético por grupos

De la misma manera, pero añadiendo un índice para cada uno de los campos semánticos del thesaurus.

.6 Lista jerárquica

En primer lugar, los descriptores se ordenan por campo semántico.

Dentro de cada campo semántico existe una entrada para cada descriptor cabeza de jerarquía, generalmente por orden alfabético.

Para cada entrada se incluye el conjunto de descriptores específicos de cada término cabeza de jerarquía, dispuestos en todos sus niveles jerárquicos descendentes, con:

-un sangrado hacia la derecha para cada nivel jerárquico;
-ordenación alfabética de los descriptores que dependen de un mismo genérico y están colocados en el mismo nivel jerárquico.

La lista jerárquica no contiene los no-descriptores ni ninguna de las informaciones (aparte de la jerarquía descendente de descriptores) que figuran en la lista alfabética estructurada. En algunos thesaurus, sin embargo, cada descriptor va acompañado de un código secuencial, que, reproducido en las demás presentaciones, permite recuperar fácilmente los descriptores dentro de la lista jerárquica. Este código puede ser un código numérico, o un código formado por letras o un código alfanumérico.

.3 Presentaciones gráficas

Las presentaciones gráficas tienen la ventaja de exponer la estructura semántica de cada campo semántico dentro de una hoja de papel, y de valerse, por tanto, de las dos dimensiones de esta superficie, mientras que las presentaciones en listas se editan secuencialmente y no

permiten nada más que explorar la estructura del campo semántico de forma lineal.

Dentro de cada uno de los dos tipos de presentaciones gráficas, se utiliza una hoja por grupo (campo semántico).

Esta hoja contiene dos zonas, delimitadas por un trazado rectangular. En el interior del rectángulo se encuentran los descriptores que pertenecen al campo semántico, así como sus relaciones jerárquicas y asociativas con descriptores situados dentro del mismo grupo.

Fuera del rectángulo se encuentran los descriptores que pertenecen a otros grupos, pero que están ligados a los que figuran dentro del rectángulo por una relación jerárquica (en el caso de polijerarquía) o por una relación asociativa.

.1 Diagrama de flechas

En el interior del rectángulo, en el centro, se encuentran el o los descriptores cabeza de jerarquía. Los descriptores específicos se registran en la parte no central del rectángulo, colocando los más genéricos próximos al centro, y los más específicos próximos a los bordes del rectángulo.

Las relaciones jerárquicas se marcan por medio de flechas que van desde los términos genéricos hacia los términos específicos.

Las relaciones asociativas se indican por líneas rectas, no por flechas, que enlazan los descriptores asociados.

En algunos thesaurus, los no-descriptores están inscritos en caracteres más pequeños o en minúsculas, bajo los descriptores a los que reenvían.

Los descriptores que figuran en el exterior del rectángulo (pertenecientes, pues, a otros campos semánticos) se enlazan con los descriptores interiores por medio de flechas (relaciones jerárquicas) o líneas rectas (relaciones asociativas); cada descriptor va acompañado del código (generalmente un número) del campo semántico al que pertenece.

En muchos thesaurus, el rectángulo interno está reticulado en 100 casillas (diez zonas horizontales - diez zonas verticales), identificadas cada una por un número de 2 cifras: la cifra de su abscisa (de 0 a 9), seguida de la de su ordenada (de 0 a 9).

Esta identificación de la localización:

-se incluye, junto a cada descriptor, en las otras presentaciones del thesaurus, lo que permite situar fácilmente un descriptor dentro del diagrama a partir de su «dirección» obtenida en una lista alfabética;

-es idéntica para los descriptores que representan el mismo concepto dentro de las distintas versiones lingüísticas del mismo thesaurus multilingüe (principio de superponibilidad de los diagramas de diferentes lenguas).

.2 Terminograma

En el interior del rectángulo se encuentra un «cartucho» (pequeño rectángulo) por cada descriptor cabeza de jerarquía. En el interior de este cartucho figura ese descriptor,

acompañado por la jerarquía descendente de los descriptores que están a él ligados, dispuestos en todos sus niveles jerárquicos, y con:

-un sangrado hacia la derecha para cada nivel jerárquico;
-una ordenación alfabética de los descriptores de cada nivel jerárquico.

Las relaciones asociativas, que, recordémoslo, sólo existen entre descriptores de cadenas jerárquicas distintas, son representadas por líneas rectas, no por flechas, entre descriptores asociados. En el exterior del rectángulo se encuentran los descriptores de otros campos semánticos, acompañados por la identificación de su campo, y enlazados con los descriptores del rectángulo interior por medio de flechas o de líneas rectas, igual que en los diagramas de flechas.

Los no-descriptores no figuran, en general, dentro de los terminogramas, de tal manera que los cartuchos correspondientes a un mismo concepto cabeza de jerarquía, en las diferentes versiones lingüísticas de un mismo thesaurus multilingüe, tengan exactamente la misma dimensión y puedan ser dispuestos de la misma forma dentro del rectángulo interno.

En comparación con el diagrama de flechas, el terminograma presenta las siguientes ventajas e inconvenientes:

-es más fácil automatizar parcialmente su producción: el contenido de cada cartucho procede de descomponer la entrada correspondiente a cada descriptor cabeza de jerarquía dentro de la presentación jerárquica, la cual puede ser producida totalmente con el ordenador;

-es posible alojar más descriptores dentro de un terminograma (± 80 como máximo) que dentro de un diagrama de flechas (± 40 como máximo), lo que permite descomponer el thesaurus en menos campos semánticos, y se facilita así la consulta;

-los diagramas de flechas de un mismo grupo en diferentes lenguas son superponibles, lo cual no sucede con los terminogramas, en los que la colocación de los descriptores depende en parte de su ordenación alfabética, variable de una lengua a otra;

-es más molesto codificar la colocación de cada descriptor dentro del terminograma ya que:

+el modo de presentación del terminograma no permite reticular en casillas distintas capaces de contener un solo descriptor,
+la secuencia alfabética de los descriptores de un mismo nivel jerárquico es diferente de una lengua a otra, y la dirección de los mismos conceptos dentro de los terminogramas de un thesaurus en diferentes lenguas no es, por tanto, unívoca;

-por la razón antes indicada, no se registran los no-descriptores dentro de los terminogramas, mientras que es más fácil hacerlo dentro de los diagramas.

6.**Ilustraciones**

A continuación se encontrarán extractos de:

-la lista de las microdisciplinas del thesaurus de la C.N.C.A. (Caisse Nationale de Crédit Agricole); esta lista está estructurada en dos niveles:
+los dominios, o temas generales (ejemplo: management),
+las microdisciplinas, o campos semánticos (ejemplo: gestión de l'enterprise);
-la lista alfabética estructurada completa del thesaurus EUDISED del Consejo de Europa y de la Comisión de las Comunidades Europeas; en él se encuentra:
+para cada descriptor:
->su enunciado,
->el número del o de los terminogramas (= microdisciplina) en los que éste figura,
->sus equivalencias interlingüísticas, precedidas cada una por el código de lengua,
->su nota explicativa,
->su o sus equivalencias semánticas,
->sus relaciones jerárquicas hacia arriba, y luego hacia abajo, hacia todos los niveles jerárquicos,
->sus relaciones asociativas;
+para cada no-descriptor:
->su enunciado
->el descriptor que se ha de utilizar;
-la lista alfabética estructurada por grupos del thesaurus de la O.I.T. (Organización Internacional del Trabajo); los grupos están estructurados en tres niveles:
+los dominios generales (ejemplo: coopération internationale, relations internationales),
+los dominios específicos (ejemplo: coopération internationale);
+las microdisciplinas o campos semánticos designados simplemente por los números (ejemplo: 01.01.4).

Para cada descriptor, se encuentra:

+su enunciado y las equivalencias lingüísticas;
+su nota explicativa;
+su equivalencia semántica;
+sus relaciones jerárquicas hacia arriba y hacia abajo, en un solo nivel jerárquico;
+sus relaciones asociativas.

Los no-descriptores no aparecen dentro de la lista.

-La lista alfabética no estructurada, completa y permutada del thesaurus EUDISED; cada descriptor y no-descriptor aparece en ella ordenado alfabéticamente según cada uno de sus términos significativos:

+para cada descriptor: el número del o de los terminogramas (= microdisciplinas) en los que aparece,
+para cada no-descriptor: el enunciado del descriptor que se ha de utilizar;

-el índice alfabético completo del thesaurus de la C.N.C.A.; en él se encuentra:

+para cada descriptor:
->su enunciado,
->el número de diagrama de flechas (= microdisciplina) en el que figura, así como su localización en forma de sus coordenadas dentro del diagrama,
->sus equivalencias semánticas, en minúsculas;
+para cada no-descriptor:
->su enunciado en minúsculas,
->el descriptor que se ha de utilizar, en mayúsculas, precedido por la palabra «véase» y seguido por su localización dentro de los diagramas de flechas;

-el índice alfabético completo multilingüe (ordenado en francés) del thesaurus VETDOC de la Comisión de las Comunidades Europeas; en él se encuentra, para cada descriptor francés, sus equivalencias lingüísticas en las otras lenguas del thesaurus; los no-descriptores, que pueden no tener equivalencias lingüísticas, no figuran en este índice;

-un diagrama de flechas del thesaurus de E.D.F.; los descriptores figuran en él dentro de la localización indicada en el índice alfabético por grupos; en el interior del diagrama las flechas indican las relaciones jerárquicas, y las líneas rectas, las relaciones asociativas; los reenvíos al exterior del diagrama recogen el enunciado de los descriptores relacionados, así como su número de diagrama de flechas. Los no-descriptores no aparecen;

-el índice alfabético por grupos del thesaurus de E.D.F. (Electricité de France); en él se encuentra:

+para cada descriptor:
->su enunciado,
+->su localización dentro del diagrama de flechas (= microdisciplina) en el que figura, en forma de coordenadas (número entre paréntesis) en el diagrama,
->su frecuencia de indización;
+para cada no-descriptor, en minúsculas:
->el o los descriptores que se han de utilizar;

-un terminograma del thesaurus EUDISED; en el interior del cuadro, se ha dedicado un cartucho para cada descriptor cabeza de jerarquía; en el interior de cada cartucho, la jerarquía descendiente de descriptores, a partir de un término de cabeza, con un sangrado hacia la derecha para cada nivel; la estructura asociativa está marcada por líneas rectas:

+entre descriptores del terminograma (= microdisciplina);

+hacia otros terminogramas: el enunciado de los descriptores se encuentra acompañado del número de esos terminogramas;

-la lista jerárquica del thesaurus ROOT de la British Standards Institution; en ella cada descriptor está registrado en el lugar correspondiente a su localización dentro de la estructura jerárquica del thesaurus; para cada descriptor se encuentra:
+su código alfabético, correspondiente a su localización dentro de la lista, de 1 a 6 letras: una letra para los términos más genéricos; seis letras para los más específicos; se puede resaltar que a determinados niveles jerárquicos los descriptores se caracterizan no por un código único sino por una ramificación (ej.: ENERGY SOURCES: JD/JG), de manera que podrán admitir más de 26 términos específicos;
+su enunciado, con caracteres de grosor (= espesor) decreciente para los cuatro primeros niveles jerárquicos y con un sangrado hacia la derecha para cada nivel jerárquico;
+a veces precedido por la indicación de una faceta (ejemplo: (By device));
+a veces seguido por:
->una nota de aplicación entre corchetes [...],
->no-descriptores, precedidos del signo =,
->descriptores genéricos procedentes de otra jerarquía, precedidos de los signos *<,
->descriptores específicos que proceden de otra jerarquía, precedidos de los signos *>,
->descriptores asociados, procedentes de los signos *-.

*Nota:* los descriptores genéricos, específicos y asociados van seguidos de su código, el cual permite localizarlos fácilmente dentro de la lista jerárquica.

-la lista alfabética estructurada completa del thesaurus ROOT de la B.S.I.; se encuentra:
+para cada descriptor:
->su enunciado,
->su código, que constituye en realidad su dirección dentro de la lista jerárquica (ej.: ENERGY SOURCES JD/DG),
->su nota de aplicación eventual (símbolo [...]),
->sus equivalencias semánticas (símbolo = en caso de mono-equivalencia, o + ... = ** ... en caso de pluri-equivalencia obligatoria),
->sus relaciones jerárquicas para un solo nivel, hacia arriba, luego hacia abajo, después sus relaciones asociativas, en la misma jerarquía en la que el descriptor

aparece como entrada dentro de la lista jerárquica (símbolos <,> y -),

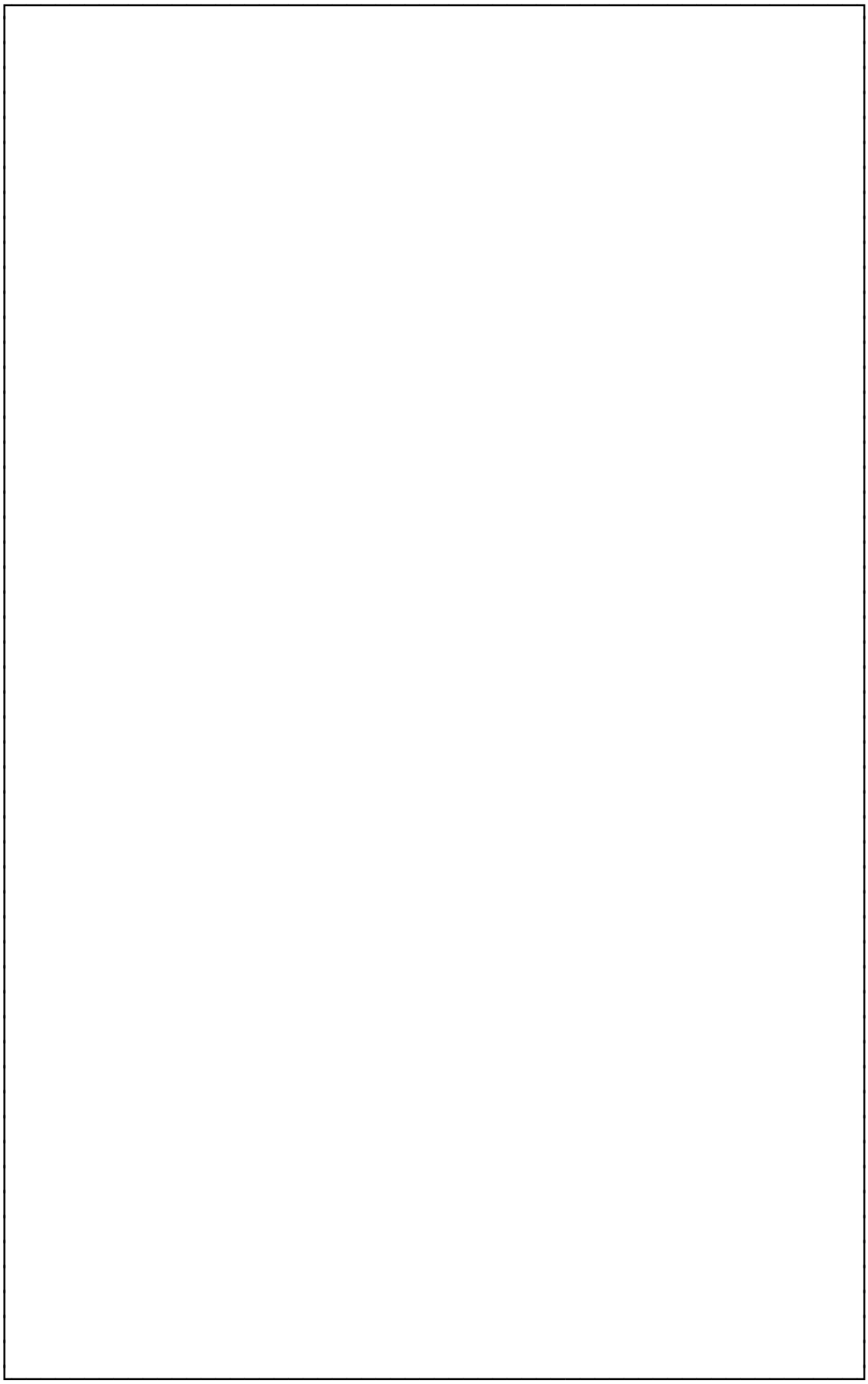
->sus relaciones jerárquicas a un solo nivel, hacia arriba, luego hacia abajo, después sus relaciones asociativas, en otras partes de la jerarquía (símbolos *<, *>, *-), con su código;

+para cada no-descriptor:

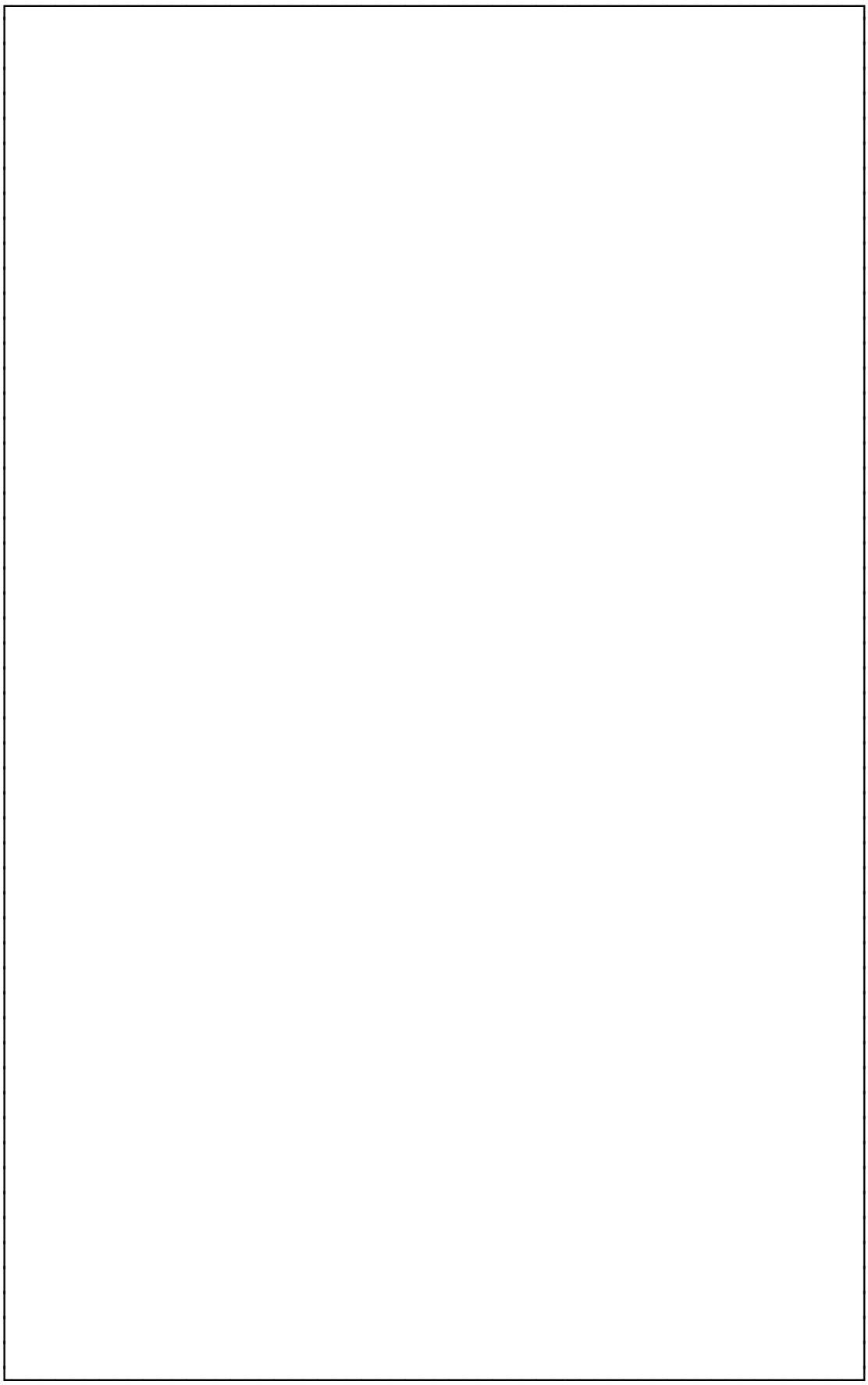
->su enunciado,

->el descriptor que se ha de utilizar y su código (símbolo ->),

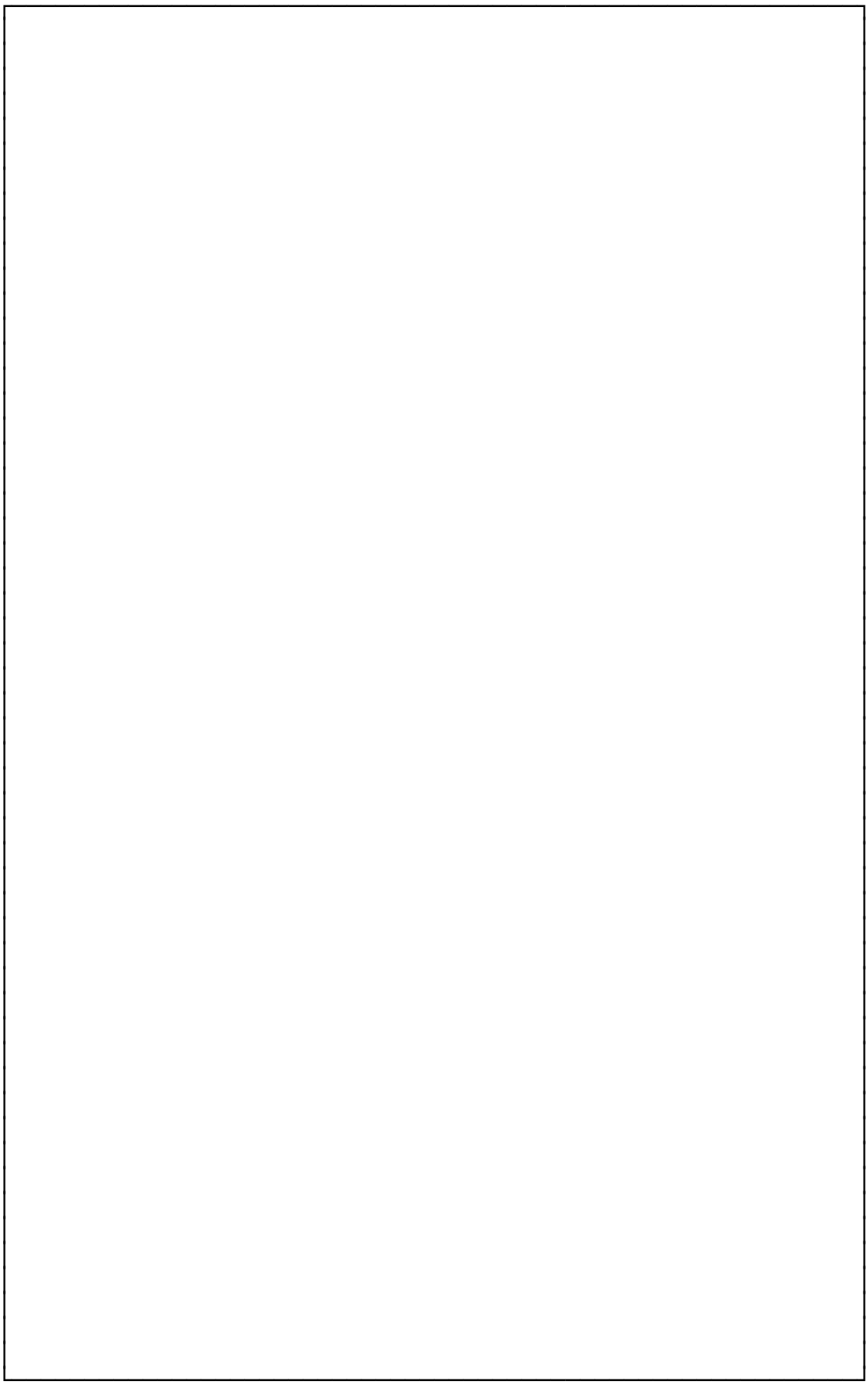
->o los descriptores que se han de utilizar juntos obligatoriamente y su código (símbolo -> ... + ...).



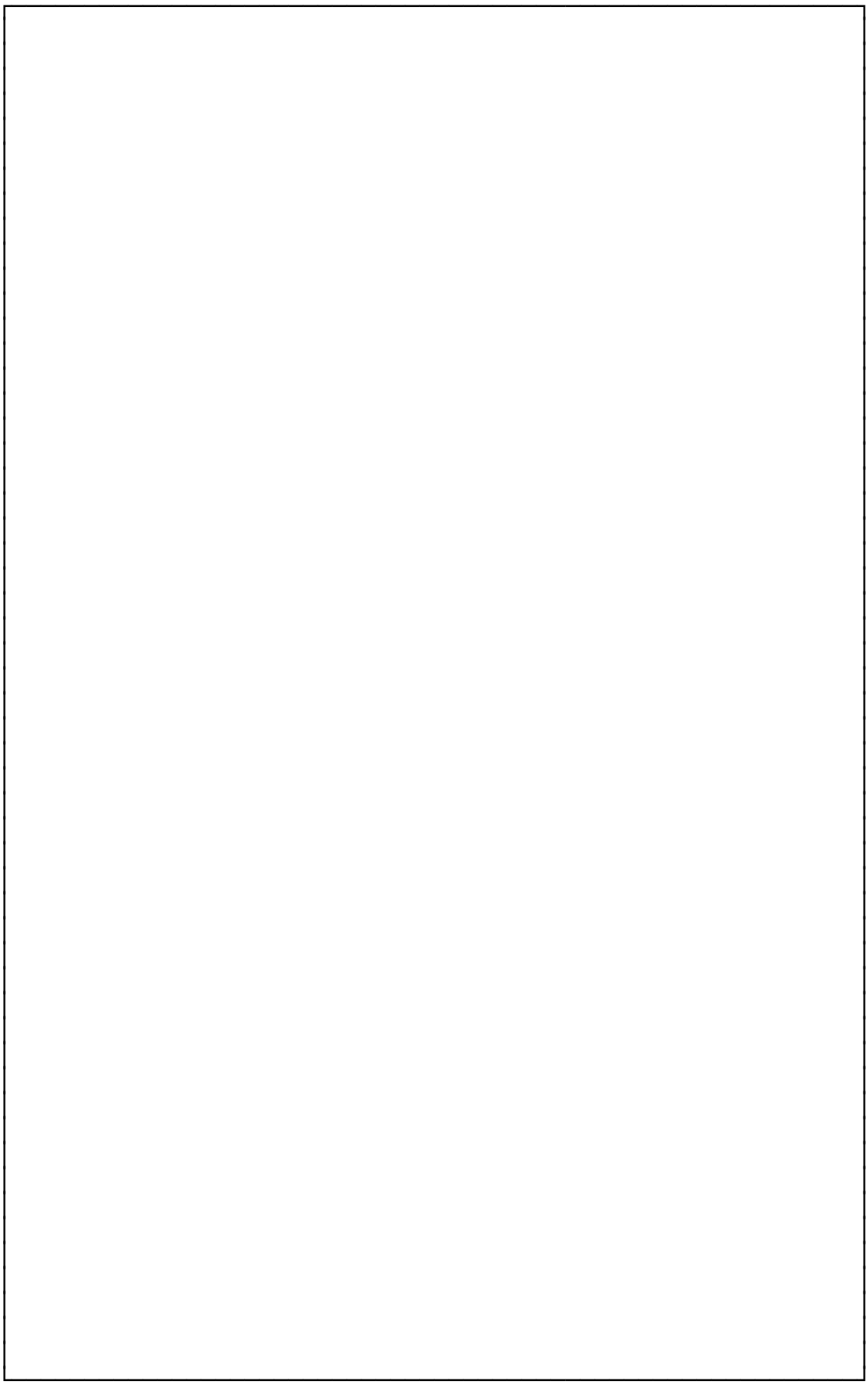
(lista de microdisciplinas)



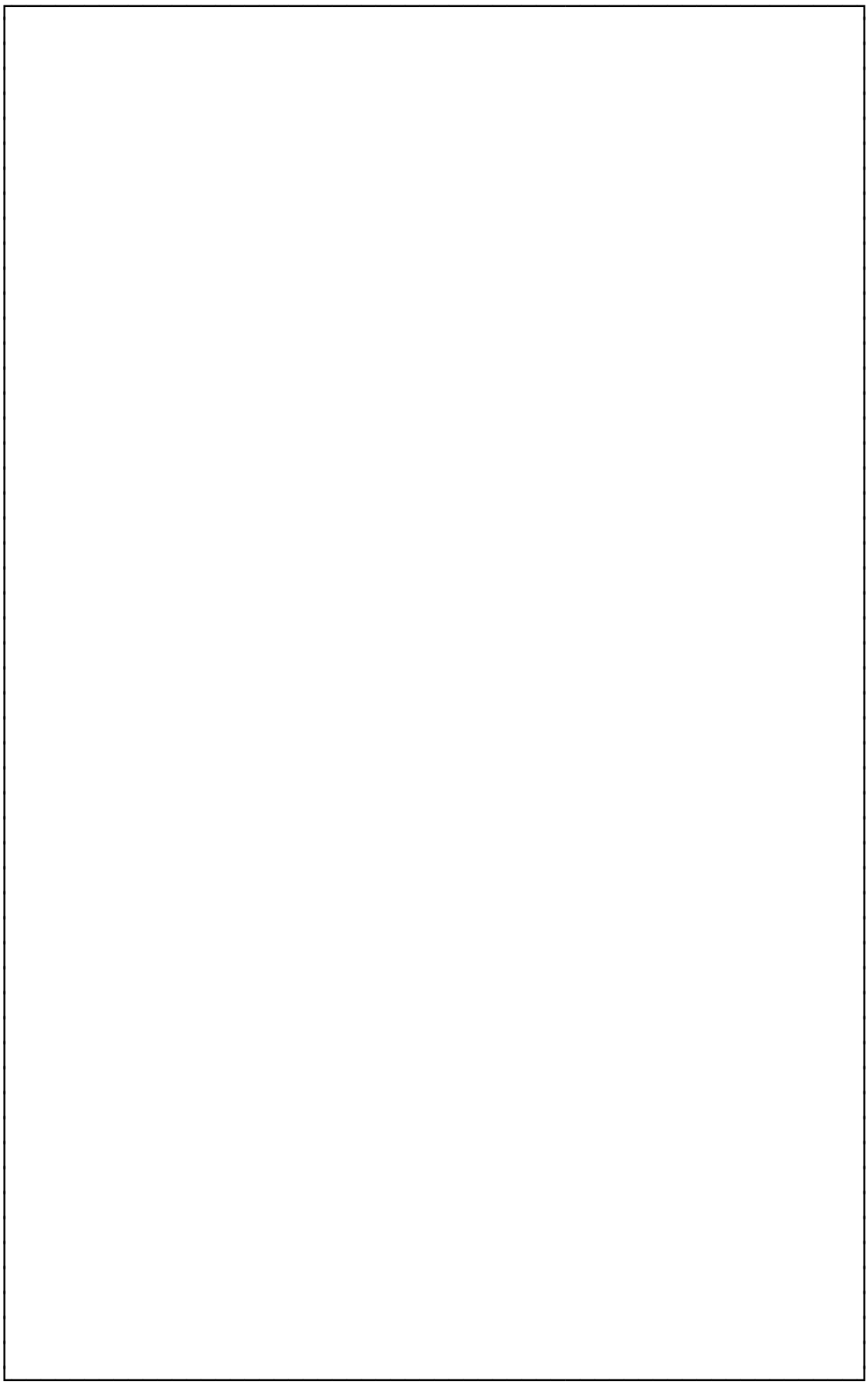
Thesaurus de la CNCA
(lista de microdisciplinas - continuación)



Thesaurus EUDISED
(extracto de la lista alfabética estructurada completa)

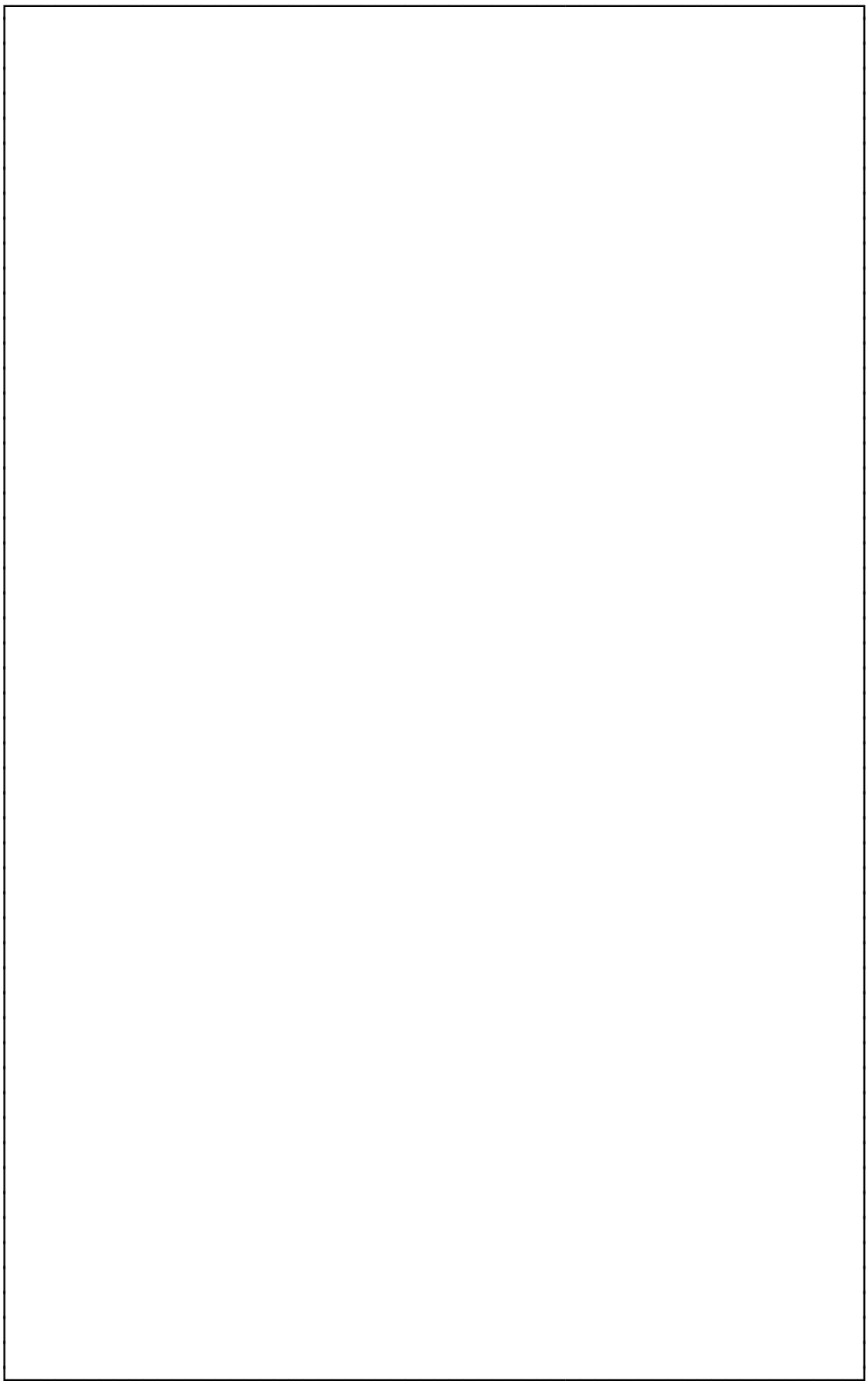


Thesaurus de la OIT
(extr. de la lista alfabética estructurada por grupos)

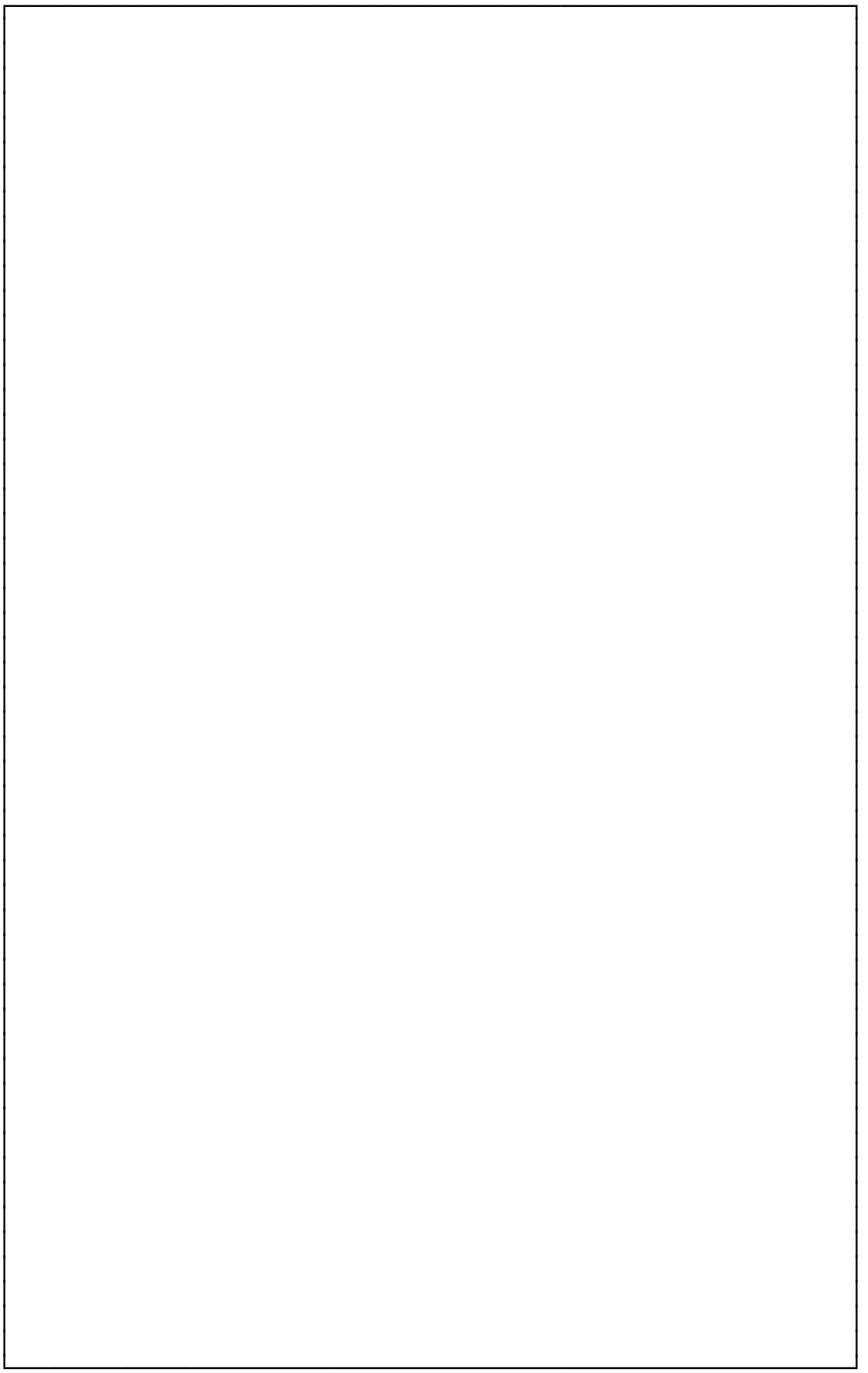


Thesaurus EUDISED
(extracto de la lista alfabética no estructurada,

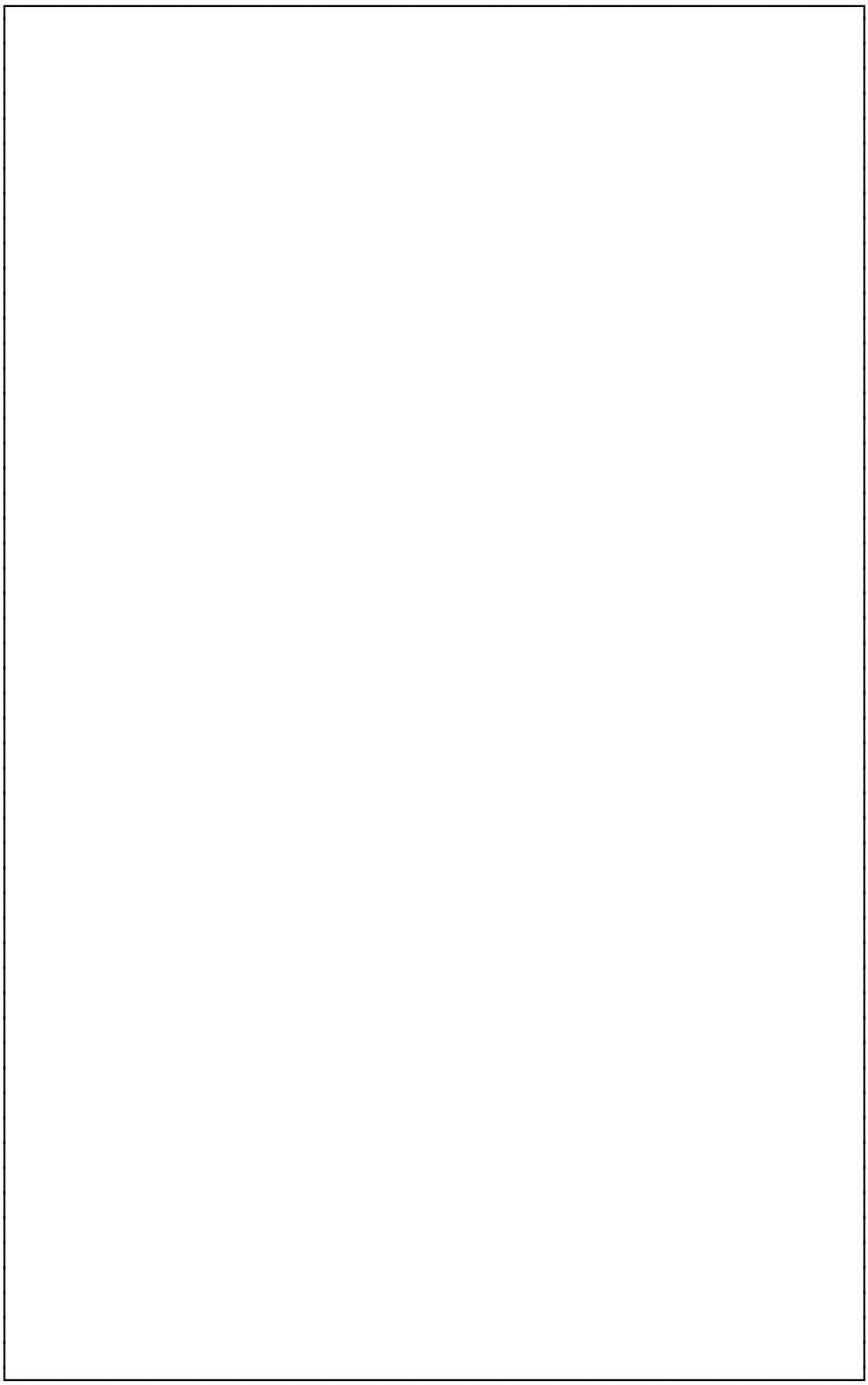
completa y permutada)



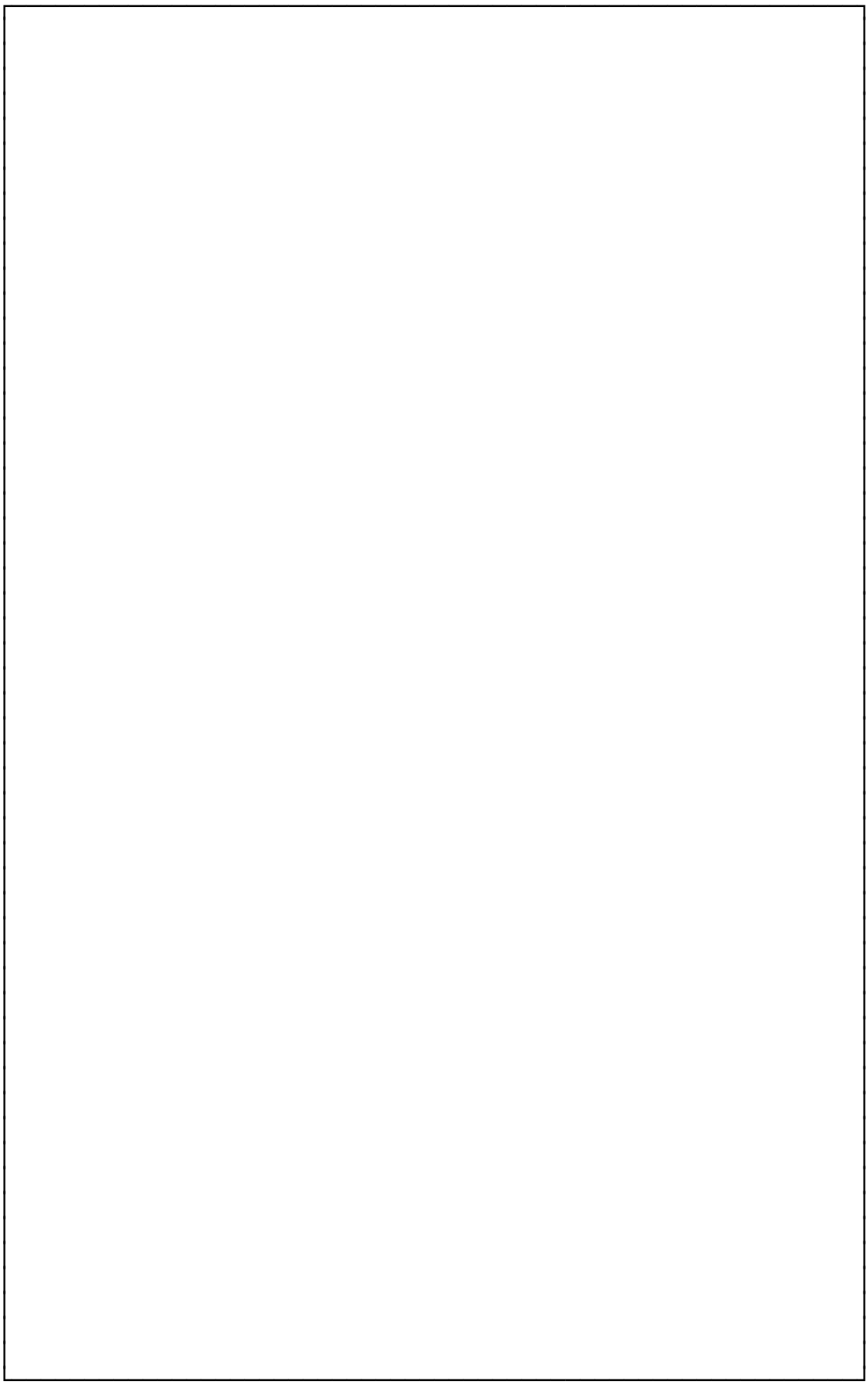
Thesaurus de la CNCA
(extracto del índice alfabético completo)



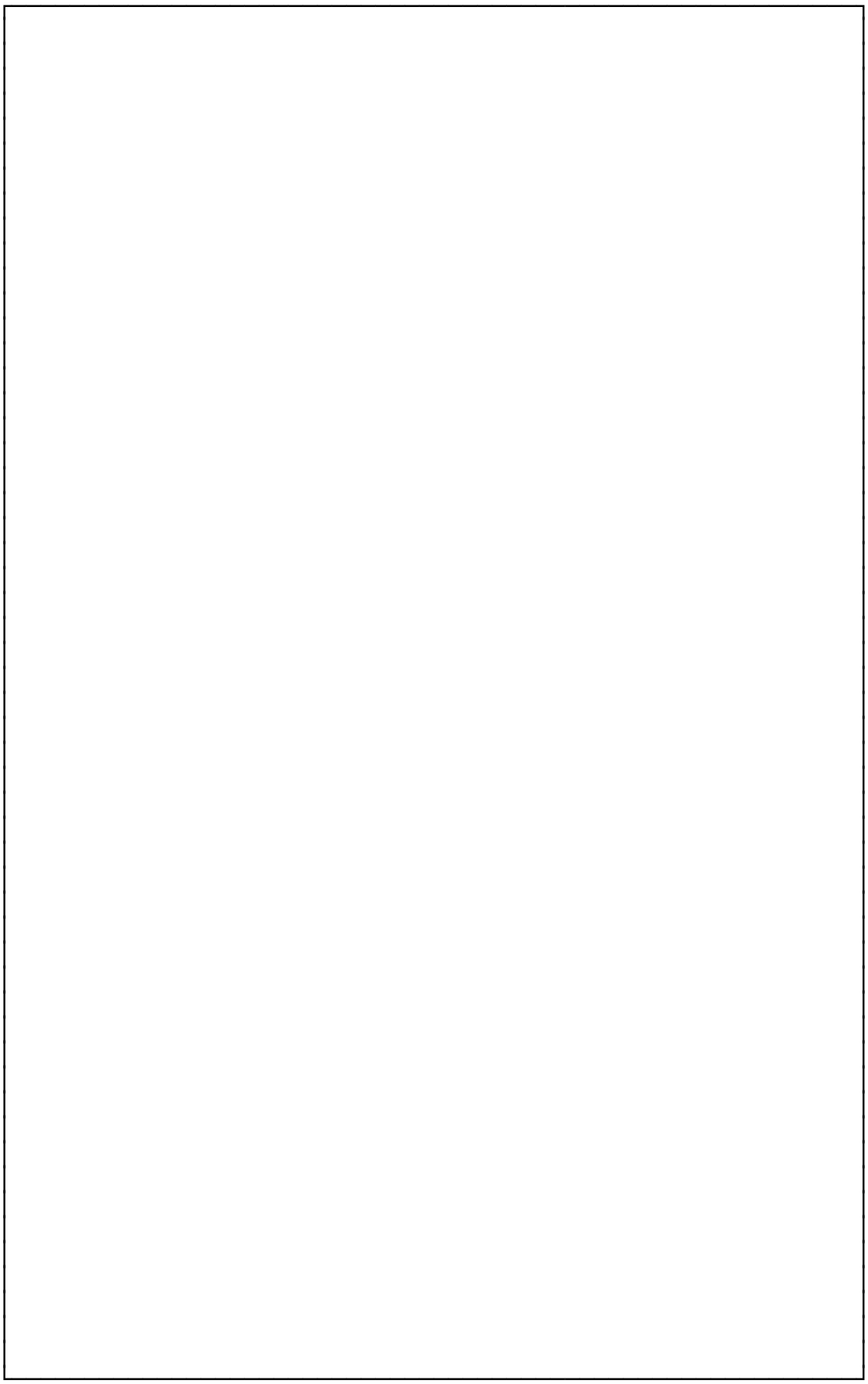
Thesaurus VETDOC
(extracto del índice alfabético completo multilingüe)



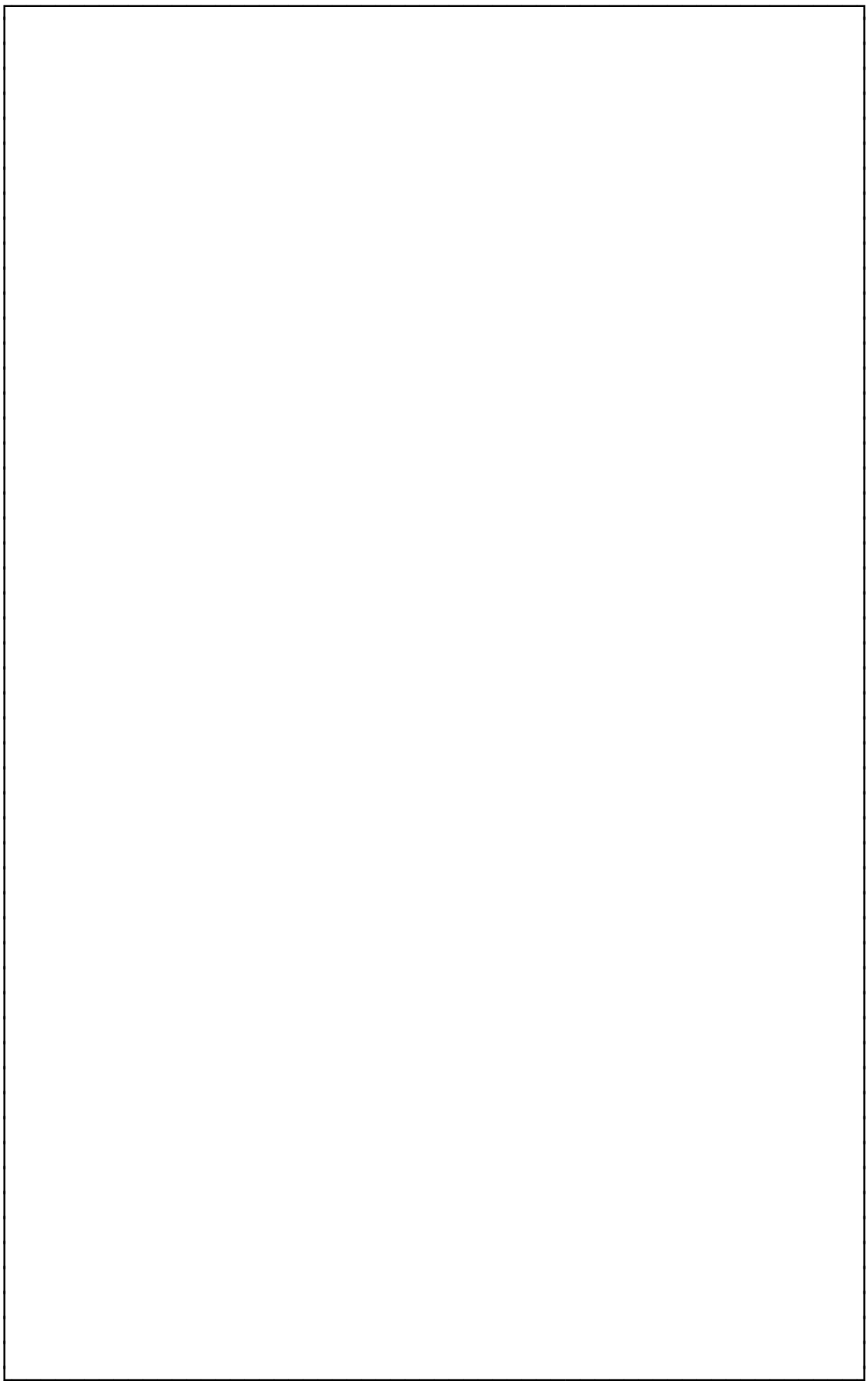
Thesaurus de EDF
(diagrama de flechas de una microdisciplina)



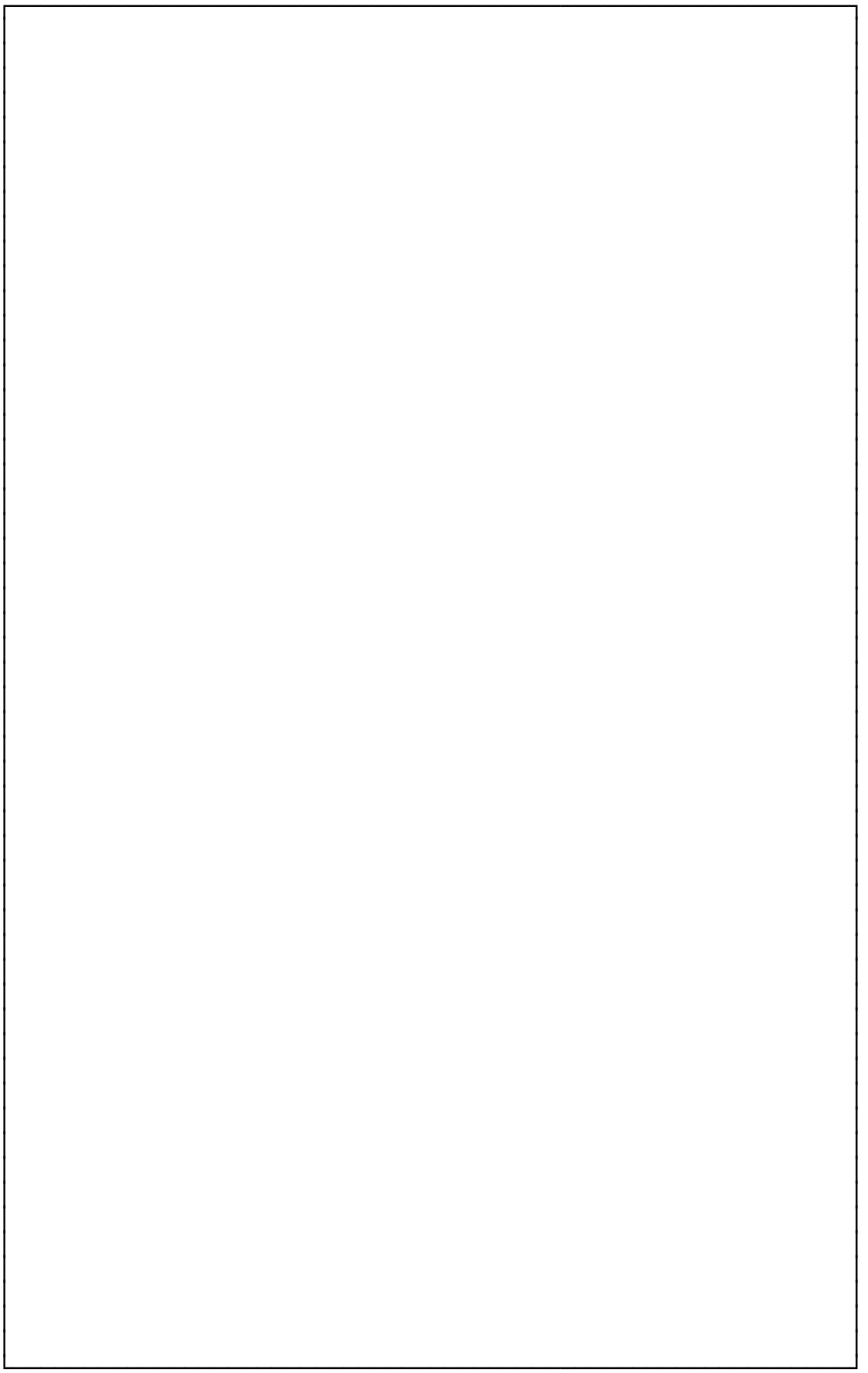
Thesaurus de EDF
(índice alfabético de un grupo de descriptores)



Thesaurus EUDISED
(terminograma de una microdisciplina)



Thesaurus del BSI
(extracto de la lista jerárquica)



Thesaurus del BSI
(extracto de la lista alfabética estructurada completa)

## 7. **Bibliografía**

Se incluyen aquí las obras generales sobre los lenguajes combinatorios. Los documentos, y especialmente las normas, en materia de construcción y de utilización son citadas en las bibliografías de los capítulos II y III.


AITCHISON (J.) et al. -
Thesaurofacet: a thesaurus and faceted classification for engineering related subjects -, English Electric Company, 1969.

AITCHISON (J.) -
Indexing languages: classification schemes and thesauri, in ANTHONY (L.J.), ED. - *Handbook of special librarianship and information work* - London: Aslib, 1982, p. 207-261.

CHAUMIER (J.) -
*Les langages documentaires* - Paris: Enterprise Moderne d'Edition, 1978, 148 p., ISBN: 2-7044-0594-8.

COATES (E.) et al. -
*Broad System of Ordering (BSO), schedule and index* - The Hague: FID, 1978, 198 p., ISBN: 92-66-00564-9.

COATES (E.) et al. -
*Système général de classement, tables e index* - La Haye: FID, 1981

COYAUD (M.) -
*Linguistique et documentation* - Paris: Larousse, 1972, 173 p.

COYAUD (M.) -
*Introduction à l'etude des langages documentaires* - Paris: Klincksieck, 1966, 148 p.

DE GROLIER (E.) -
*Etude sur les catégories générales applicables aux classifications et codifications documentaires* - Paris : Unesco, 1962, 262 p.

DEWEZE (A.) -
*Traitement de l'information linguistique* - Paris Dunod, 1966, 211 p.

DIETSCHMANN (H.J.), ed. -
*Representation and exchange of knowledge as a basis of information process* - Amsterdam: North Holland, 1984, 433 p., ISBN: 0-444-87563-8.

FOSKETT (A.C.) -
*The subject approach to information* - London: Clive Bingley, 1982, 574 p. ISBN: 0-85157-313-4.

HUTCHINS (W.J.) -
*Languages of indexing and classification: a linguistic study of structures and functions* - Stevenage: Peter Peregrinus, 1975, 148 p., ISBN: 0-90122-368-9.

Institut Gustave Roussy -
*Le macrothésaurus des sciences et des techniques* - Paris: CILF, 1979.

LANCASTER (F.W.) -
*Vocabulary control for information retrieval* - Washington, D.C.: Information Resources Press, 1986, 270 p., ISBN: 0-87815-053-6.

LAUREILHE (M.T.) -

*Le thésaurus: son rôle, sa structure, son élaboration* - Villeurbanne: Presse de l'ENSB, 1981, 88 P.

MOUNIN (G.) -
*Les problèmes théoriques de la traduction* - Paris: Gallimard, 1963, 296 p.

MOUNIN (G.) -
*Dictionnaire de la linguistique* - Paris: P.U.F., 1974 340 p.

ROBERTS (N.) -
*The prehistory of the information retrieval thesaurus* - *Journal of Documentation*, vol. 40, nº 4, December 1984, p.271-285.

VICKERY (B.C.) -
*Faceted classification* - London: Aslib, 1961, 70 p.

WALKEDR (D.E.), KARLGREN (H.) and KAY (M.) -
*Natural language in information science; perspectives and direction for research* - The Hague: FID., 1977, 194 p., ISBN: 91-7282-131-0.

WANG (Y.C.) et al. -
Relational thesauri in information retrieval - *Journal of the American Society for Information Science*, vol. 36, nº 1, January 1985, p. 15-27.

WILLETS (M.) -
An investigation of the nature of the relation between terms in thesauri - *Journal of Documentation*, vol. 31, nº 3, September 1975, p.158-184.

CAPITULO II

# CONSTRUCCION Y MANTENIMIENTO

En este capítulo se estudiarán las modalidades de puesta a punto de lenguajes combinatorios utilizados dentro de los servicios de documentación y las bibliotecas especializadas. No nos ocuparemos, sin embargo, de la preparación de listas de palabras clave, ya que de hecho consiste en una simple acumulación, a medida que se registran los documentos, de las palabras clave extraídas por el ordenador a partir de los títulos, resúmenes y textos de los documentos.

Con un procedimiento contrario al utilizado en el capítulo I, comenzaremos por describir la construcción del lenguaje más complejo: el thesaurus de descriptores; y, como quien puede lo mucho puede lo poco, nos ocuparemos a continuación de los lenguajes menos sofisticados: lista de autoridades y lista de descriptores libres.

## 1. **Construcción de un thesaurus de descriptores**

Cuando se crea un nuevo servicio de documentación, o cuando un servicio de documentación ya existente

- -crea una nueva base de datos documental,
- -amplía sustancialmente la cobertura disciplinar de una base de datos ya existente,
- -o decide mejorar la eficacia de la búsqueda utilizando a partir de ese momento un thesaurus,

se deben abordar sucesivamente dos tipos de actuaciones:

- -un estudio de oportunidad, con el fin de fijar las grandes orientaciones del proyecto,
- -el propio proceso de construcción, en tanto que, por supuesto, las conclusiones del estudio de oportunidad sean positivas y hayan sido aceptadas.

Por motivos didácticos, no vamos a seguir el orden cronológico: nos ocuparemos en primer lugar del proceso de construcción (§ 1.1); a continuación se analizará el contenido del estudio de oportunidad (§ 1.2).

### .1 Proceso de construcción

La construcción de un thesaurus comprende ocho (thesaurus monolingüe) o nueve (thesaurus multilingüe) etapas:

- -recolección del vocabulario en lenguaje natural dentro de los dominios que cubrirá el thesaurus;
- -subdivisión del conjunto de los dominios que se van a cubrir en una serie de microdisciplinas;
- -transformación progresiva del vocabulario libre en un lenguaje controlado, estableciendo las relaciones

de pertenencia, de equivalencia semántica, de jerarquía, y redactando las notas explicativas;
-búsqueda de las equivalencias interlingüísticas (si se trata de un thesaurus multilingüe);
-enriquecimiento del thesaurus por medio de relaciones asociativas;
-realización de la edición nº 0 del thesaurus;
-formación de los indizadores;
-test del thesaurus;
-revisión final y edición número 1.

.1 Recolección del vocabulario

Esta fase consiste en buscar y registrar todas las palabras y expresiones significativas que intervienen en la disciplina o disciplinas cubiertas por el thesaurus.
Pueden utilizarse algunas fuentes terminológicas alternativas o complementarias:

-un lenguaje documental ya existente en el servicio de documentación: sistema de clasificación, lista de palabras clave o de descriptores libres, lista de autoridades; esta fuente terminológica es especialmente valiosa en la medida en que refleja la experiencia adquirida hasta el momento por el servicio: términos utilizados, frecuencia de utilización de esos términos;
-thesaurus ajenos al servicio, pero dedicados a los mismos dominios o a dominios próximos, descubiertos gracias a alguno de los repertorios de thesaurus disponibles en el mercado, de los cuales el más importante es el de la G.I.D. (Gesellschaft für Information und Dokumentation); un thesaurus ya existente resulta útil porque aporta una colección organizada de conceptos de la que se podrá extraer lo que convenga, pero rara vez se adoptadará en su totalidad para organizar el propio patrimonio informativo;
-terminología utilizada en los tratados, manuales, léxicos especializados, reglamentaciones técnicas publicadas o específicas del organismo: esas fuentes tienen la gran ventaja de ofrecer una terminología generalmente admitida y a menudo ya desambiguada y estructurada por sus autores;
-bases de datos terminológicas;
-indización intelectual, en lenguaje natural, por medio de descriptores libres, de una muestra de varios cientos de documentos representativos para cada uno de los grandes dominios que va a cubrir el thesaurus: se obtiene de esta manera un inventario no ordenado de la terminología efectivamente utilizada por los autores para designar los conceptos que ellos manejan habitualmente, puede que en varias lenguas, si se trata de una muestra multilingüe para crear un thesaurus multilingüe;
-indización automática, en lenguaje natural, por medio de palabras clave, de esos mismos documentos; la experiencia muestra que desgraciadamente este método es poco eficaz porque:
+la lista obtenida sólo incluye palabras aisladas (unitérminos), que sólo en algunos casos, cuando están esclarecidas por su contexto, sirven para encontrar los conceptos;

+aunque se utilice una gran lista de palabras vacías, el ordenador proporciona un porcentaje extremadamente elevado de términos sin ningún interés discriminante;

-consulta a especialistas vinculados con el organismo al que pertenece el sistema documental: esta fuente es indispensable para obtener la terminología propia de las diversas prácticas profesionales dentro del organismo, así como una explicitación de la acepción específica de ciertos términos para estos especialistas.

.2 Listado de las microdisciplinas

Un examen rápido del vocabulario recogido durante la primera etapa permite establecer una lista, todavía muy provisional, de los grandes temas, y a veces de facetas específicas para los dominios que va a cubrir el thesaurus.

El número de grupos que se va a establecer está en función directa del tamaño del thesaurus que ha sido aprobado al realizar el estudio de oportunidad (cf. § 1.2.2.3); una regla empírica consiste en dividir por 50 el número correspondiente a ese tamaño; por ejemplo, si se pretende crear un thesaurus de 3.000 descriptores, habrá que intentar establecer durante esta fase una lista de aproximadamente 60 microdisciplinas. Cada microdisciplina podría así contener unos 50 descriptores como media. Normalmente la lista de las microdisciplinas se ordenará en torno a unas disciplinas, o dominios, más generales, de 6 a 12, definidos durante el estudio de oportunidad (cf. § 1.2.2.1).

La lista de microdisciplinas así dispuesta será utilizada directamente durante la siguiente etapa (cf. § 1.1.3); se revisará continuamente a lo largo de todo el proceso de construcción del thesaurus:

-las microdisciplinas que contengan más de 40 (en el caso de una presentación por diagrama de flechas) u 80 (en el caso de una presentación por terminogramas) descriptores serán subdivididas;

-las microdisciplinas que incluyan un número demasiado pequeño de descriptores serán reagrupadas.

La lista de las microdisciplinas tiene como única finalidad preparar un sistema de agrupación de los descriptores que sea aceptable para el usuario y cómodo de consultar; no sirve en modo alguno para establecer una clasificación científica de los descriptores.

Las microdisciplinas corresponden tanto a temas como a facetas, según el modo de agrupación que parezca más natural (cf. capítulo I, § 4.2.1.1).

.3 Reducción a un lenguaje documental

Esta etapa comprende tres fases:

-estructura de equivalencia semántica,
-estructura jerárquica,
-tratamiento de los doblecartes usos.

.1 Elaboración de la estructura de equivalencia semántica

-Se distribuyen los términos en lenguaje natural, recopilados durante la primera etapa, entre las microdisciplinas listadas en el transcurso de la segunda etapa.
-Se examina cada microdisciplina, con el fin de:
+enumerar los conceptos presentes y establecer para cada uno de ellos la lista de términos sinónimos o cuasi-sinónimos (esta lista se denomina, según algunos autores, clase de equivalencia);
+dentro de cada lista:
->se elige la designación de los descriptores:
#o bien de la lista de términos disponibles,
#o bien de fuera de esta lista, si parece que otra denominación distinta es más adecuada porque corresponda al vocabulario más habitual entre los usuarios del futuro thesaurus;
->se eligen los términos que tendrán el estatuto de no-descriptores y se enlazan, por medio de una relación de equivalencia semántica, con el/los descriptor(es) ya seleccionado(s).
El problema práctico que se plantea en este momento es el de elegir el término que tendrá el estatuto de descriptor. Los criterios a utilizar son los siguientes: dentro de la lista de términos de significación próxima se elegirá como descriptor el que sea
#más «neutro», es decir, el menos ambiguo,
#más usado en la profesión, o dentro del organismo para el que se construye el thesaurus: el thesaurus será tanto más apreciado por sus usuarios cuanto más respete su «jerga» o su «cultura de empresa»;
->se eliminan los términos vacíos de significado que se hayan conservado hasta el momento;
->se reagrupan o dividen, si es necesario, las microdisciplinas de acuerdo con su contenido, según se ha indicado en § 1.1.2.
-se redactan las notas explicativas

.2 Elaboración de la estructura jerárquica en el interior de cada microthesaurus

Esta estructura constará de cadenas jerárquicas, de una a doce por disciplina; cada cadena se presenta como un árbol invertido cuyo tronco es un descriptor cabeza de jerarquía, y las sucesivas ramificaciones corresponden a descriptores cada vez más específicos.
Durante esta fase aparecerán:

-dobles usos, que habrán de ser tratados por eliminación o por polijerarquía;
-lagunas, que se rellenarán añadiendo los descriptores correspondientes a los conceptos que faltan.

Esta estructura jerárquica estará fundada, al igual que la estructura de las microdisciplinas, sobre la dialéctica de temas y facetas: para cada nivel jerárquico de cada cadena se elegirá un tema o una faceta como criterio de ubicación, de acuerdo con lo que parezca más cómodo o más natural para el usuario.

.3 Tratamiento de los dobles usos

-Después del tratamiento de todas las disciplinas, se examina el conjunto de términos conservados.
-Se toman en cuenta las polijerarquías: el mismo concepto, expresado hasta el momento de forma idéntica o distinta dentro de dos o más microthesaurus, deberá desde ahora ser designado por un único descriptor (las otras designaciones conservadas hasta el momento pasan a ser no-descriptores), que tendrá una relación de pertenencia con cada una de las microdisciplinas a las que pertenece.
-Tratamiento de las anomalías:
+descriptores idénticos que designan conceptos diferentes en dos o varias microdisciplinas: se desambiguan añadiendo un modificador o adoptando otra designación;
+no-descriptores idénticos ligados a descriptores diferentes: se desambiguan de la misma manera;
+términos conservados a la vez como descriptor en un campo y como no-descriptor en otro campo: se elimina uno de alguno de los dos campos, sustituyéndolo, si se presenta el caso, por un término preciso que no tenga doble uso;
+relaciones jerárquicas entre descriptores y no-descriptores: se excluyen.

.4 Búsqueda de equivalencias interlingüísticas

Es importante situar correctamente en el tiempo la producción de las diferentes versiones lingüísticas del thesaurus:

-después de la realización de la estructura jerárquica en la lengua de base:

La realización de un thesaurus multilingüe necesita una adaptación recíproca de cada una de las versiones lingüísticas, de manera que se obtenga la máxima coincidencia entre los conceptos representados por los descriptores de las diferentes lenguas.
La búsqueda de las equivalencias lingüísticas, por tanto, no se puede hacer más que a partir del momento en que el contenido semántico de los descriptores de la lengua de base esté fijada, gracias a las relaciones de pertenencia, equivalencia semántica y jerarquía, y a las notas explicativas

-antes de la realización de la estructura asociativa en la lengua de base:

Las relaciones de asociación no contribuyen en general a precisar el sentido de los descriptores; por tanto, no son necesarias en el momento de la búsqueda de las equivalencias lingüísticas; por el contrario, su presencia entorpecería el trabajo de revisión del thesaurus en la lengua de base, necesario para buscar las coincidencias de significado entre los descriptores de las diferentes lenguas

-paralelamente en las diferentes versiones lingüísticas diferentes a la lengua de base:

La búsqueda de las equivalencias lingüísticas conduce, en efecto, a situaciones diferentes de unos pares de lenguas a otros, y es importante advertir en cada microdisciplina los problemas presentes dentro de las diferentes lenguas para buscar una solución satisfactoria en todas las lenguas.

Algunos productores de thesaurus prefieren construir paralelamente todas las versiones lingüísticas del thesaurus y armonizarlas después. Este modo de proceder tiene la ventaja de que elimina la noción de lengua de base, lo cual es a veces psicológica o políticamente difícil de hacer admitir; presenta el inconveniente de aumentar considerablemente (más del doble) los recursos humanos dedicados a la construcción del thesaurus.

La búsqueda de los equivalentes lingüísticos puede conducir a las siguientes situaciones:

*+No hay retroacción sobre la versión en la lengua de base*

-*equivalencia cierta* (afortunadamente es el caso más frecuente, con diferencia, para la mayoría de los thesaurus).

Ejemplo:«CARRETERA» en español y «ROAD» en inglés;

-*ligera diferencia de significado.*

Ejemplo:«SERVICIO DE DOCUMENTACION» en español y «SPECIAL LIBRARY» en inglés.

-*completa falta de equivalencia*

+designación por una perífrasis.
Ejemplo:«CARTE SCOLAIRE» en francés y «SCHOOL DISTRIBUTION» en inglés;
+adopción de un término extranjero.
Ejemplo: «HALL» en inglés y en español.

*Nota:* el término español «distribuidor» es mucho menos usado que «hall»; se tomará, por tanto, «hall» como descriptor y «distribuidor» como no-descriptor.

*+Retroacción necesaria sobre la versión en la lengua de base*

-Superposición de significados, debido a la existencia de un término implícitamente polisémico en la lengua de base (implícitamente, porque la polisemia sólo aparece cuando se compara con otra u otras lenguas).

Ejemplo: se incluye el descriptor «DERRAPAJE» en la versión española, considerada como lengua de base de un thesaurus sobre circulación vial. La traducción al alemán muestra que esa lengua distingue dos conceptos:

+el derrapaje hacia adelante: rutschen,
+el derrapaje hacia un lado: schleudern.

Se pueden adoptar diferentes soluciones:

1)se toma «DERRAPAJE HACIA ADELANTE» y «DERRAPAJE HACIA UN LADO» en español y «RUTSCHEN» y «SCHLEUDERN» en alemán;
2)se adopta esta solución y además se conserva «DERRAPAJE» como término genérico español y «RUTSCHEN UND SCHLEUDERN» en alemán;
3)sólo se conserva «DERRAPAJE» en español y «RUTSCHEN UND SCHLEUDERN» en alemán y además se toma «rutschen» y «schleudern» como no-descriptores en alemán.

En las Directrices para el establecimiento y desarrollo de los thesaurus multilingües (ISO - 1985) se encontrará una gran cantidad de situaciones difíciles y propuestas de solución.

.5 Elaboración de la estructura asociativa

Esta fase, bastante tediosa, consiste en:

-examinar uno a uno cada descriptor del thesaurus;
-pasar revista rápidamente a los descriptores localizados en otros campos semánticos con los que el descriptor que se está examinando pueda tener una relación de asociación; si se presenta el caso, las relaciones de asociación podrán ser creadas así mismo entre descriptores que pertenezcan a cadenas jerárquicas diferentes dentro de la misma disciplina; en ningún caso se crearán relaciones asociativas:

+entre descriptores de la misma cadena jerárquica,
+entre descriptores y no-descriptores.

.6 Realización de la edición experimental

El thesaurus se edita de forma sumaria, pero con las diferentes presentaciones previstas.
La edición número 0, en principio, no está destinada a un uso operativo, sino fundamentalmente a la formación de los indizadores y al test del thesaurus.

.7 Formación de los indizadores

El ejercicio dura varios días; consiste en:
-presentar el thesaurus,
-hacer que los indizadores indicen la misma muestra de documentos,
-tras la indización de cada documento:
+se dialoga sobre los puntos de vista adoptados por cada indizador, con el fin de hacer que progresivamente coincidan;
+se calcula la tasa de coherencia (ratio entre el número de descriptores comunes y el número total de descriptores distintos utilizados por dos personas o dos grupos para indizar el mismo documento); al principio del ejercicio la tasa será escasa (± 20-30 %); al terminar la formación tenderá hacia su valor ideal (± 50-80 %).

.8 Test del thesaurus

El test del thesaurus tiene por objeto obtener, de los indizadores y, si es el caso, de los documentalistas encargados de formular las consultas, sus impresiones, en forma de propuestas para:

-añadir descriptores y no-descriptores olvidados durante la construcción,
-modificar las estructuras jerárquica y asociativa,
-explicitar por medio de notas explicativas descriptores todavía ambiguos.

Para organizar este test,

-se seleccionan varios cientos de documentos y, si es posible, de consultas;
-se hace que los documentalistas los indicen usando el nuevo thesaurus;
-se pide a los documentalistas que preparen un informe especial en el que aparezcan separados los diversos tipos de propuestas que acabamos de enunciar;
-se realiza, por último, una estadística de estas propuestas.

*Nota:* teniendo en cuenta el escaso número de documentos indizados, es inútil a estas alturas contabilizar:

-la frecuencia de los descriptores efectivamente utilizados,
-la ratio entre el número de descriptores utilizados al menos una vez y el número total de descriptores del thesaurus.

.9 Revisión final y realización de la edición operativa

El trabajo consiste en:

-revisar el thesaurus en función de las propuestas aportadas por los documentalistas durante el test, en la medida en que un examen rápido de cada una de las sugerencias demuestre su interés;

-redactar un prefacio, que generalmente contiene:

+los nombres del autor o autores, el de quien realizó el encargo, el del servicio o servicios que lo usarán,
+el objetivo,
+la lista de los dominios y microdisciplinas,
+a veces, la lista de las fuentes terminológicas,
+el proceso histórico de construcción,
+la presentación de la estructura semántica del thesaurus, ilustrada con ejemplos,
+las características de thesaurus: version(es) lingüística(s), número de dominios y de microdisciplinas, número de descriptores, de no-descriptores (por lengua), de notas explicativas (por lengua), de relaciones jerárquicas y asociativas,
+las presentaciones utilizadas,
+las instrucciones para la indización de los documentos y la formulación de las consultas,
+información sobre el mantenimiento;

-tirar o hacer imprimir tantos ejemplares del thesaurus como necesiten todos sus usuarios; es útil prever una encuadernación consistente, ya que el thesaurus será un instrumento de trabajo manipulado a diario por los documentalistas;

-notificar la existencia del thesaurus a uno de los centros internacionales que llevan el recuento:
para Europa y
+para el registro de la Unesco:

Centralny Instytut Informacji Naukowo Technicznej Economicznej (CIINTE)
Clearinghouse
Al. Niepodleglosci 188
Varsovia - Polonia

+para el registro de la Comisión de las Comunidades Europeas

Gesellschaft für Information und Dokumentation (GID)
Lyoner Strasse 44-48 (Arabella Center)
Frankfurt am Main - RFA.

.2 Estudio de oportunidad

El estudio de oportunidad consiste en:

-definir el mercado del futuro thesaurus,
-adoptar las grandes opciones relativas a las características cualitativas y cuantitativas del thesaurus que se va a construir,
-localizar las fuentes terminológicas más importantes,
-elegir el proceso de tratamiento más adecuado en función del tamaño del thesaurus, del plazo de realización y de los medios disponibles,

-designar el equipo encargado de la elaboración, o al menos definir sus funciones,
-estimar el coste de la operación,
-establecer el calendario de realización.

.1 Definición del mercado

.1 La demanda

Un thesaurus se hace para que sea utilizado por un servicio de documentación, una red de servicios de documentación, un productor de bases de datos documentales, documentalistas... Además, es preciso que se pueda designar nominalmente a estos usuarios y tener seguridad de que podrán tener una motivación para emplear el thesaurus que se va a construir.

Esta afirmación parece evidente; sin embargo, la experiencia muestra que muchos thesaurus, construidos a veces con grandes gastos, se han rechazado desde su lanzamiento: ¡quienes los construyeron sencillamente no se habían preocupado de asegurarse una «clientela» para su producto!

La definición del mercado permitirá así mismo estimar:

-el precio que los usuarios estarían dispuestos a pagar para adquirir el thesaurus;
-el número de ejemplares a imprimir.

.2 La oferta

Existe en el mercado una serie de inventarios de lenguajes documentales, el más importante de los cuales es el de la GID (1985).

Uno de los primeros puntos que cabe preguntarse es si existe previamente un lenguaje documental para el dominio o dominios del futuro thesaurus; tal lenguaje puede existir en forma de thesaurus de descriptores; en ese caso es necesario analizar:

-su riqueza léxica: ¿contiene una representación de todos los conceptos que han de ser cubiertos dentro del sistema documental para el que se trata de adquirir o elaborar un thesaurus?
-su vocabulario: los descriptores del thesaurus existente ¿respetan la terminología habitual de los usuarios del sistema documental?
-sus relaciones semánticas: las relaciones jerárquicas y asociativas ¿reflejan los puntos de vista generalmente adoptados por los usuarios del sistema documental?

La experiencia muestra que, aparte de los sistemas documentales para una red de cooperación, existen pocos thesaurus que, construidos para las necesidades de una documentación concreta, puedan ser utilizados tal cual para otra documentación.

En efecto, cada organismo que utiliza un thesaurus tiene su propia historia, sus propios centros de interés, sus

propias prácticas profesionales, su propia cultura de empresa, y, por tanto, su propia terminología.

Un thesaurus ya existente no parece muchas veces adecuado porque descompone la realidad según la apreciación de sus productores y porque esta apreciación no es la misma de un organismo a otro.

Esto se traduce en divergencias en:

- -la elección de los descriptores y de los no-descriptores,
- -el reparto de los descriptores entre las microdisciplinas,
- -las relaciones jerárquicas y asociativas.

Si los imperativos económicos o políticos prescriben, sin embargo, recurrir a un thesaurus ya existente, la elección del mejor thesaurus se podrá realizar de la manera siguiente:

- -se adquieren los thesaurus disponibles para los dominios tratados por el sistema documental en cuestión;
- -se prepara una muestra de documentos para indizar (del orden de unos mil documentos);
- -se indizan esos documentos en lenguaje natural, en la lengua del sistema documental, por medio de palabras y expresiones que designen los conceptos encontrados en los documentos;
- -se intenta traducir los descriptores libres de la indización en descriptores controlados extraídos de cada uno de los thesaurus examinados, uno tras otro;
- -el thesaurus que haya proporcionado la mayor cantidad de descriptores que respondan a las necesidades del sistema documental será el candidato para la elección:
  - +si proporciona más del 95 % de los descriptores necesarios, podrá ser adoptado como thesaurus para el sistema;
  - +si proporciona del 80 al 95 % de los descriptores necesarios, podrá ser completado por los descriptores que falten y llegar a ser, tras esta adaptación, el thesaurus del sistema;
  - +si proporciona del 50 al 80 % de los descriptores necesarios, sólo podrá ser usado como una de las fuentes, quizá las más importante, de terminología incorporable al thesaurus del sistema, el cual deberá ser construido;
  - +si proporciona menos del 50 % de los descriptores necesarios, sólo podrá servir como fuente secundaria de terminología del thesaurus del sistema, el cual deberá ser construido.

.2 Definición de las características del thesaurus

Las características que hay que definir son fundamentalmente:

-dominios a cubrir,
-compatibilidad con un thesaurus ya existente,
-tamaño del thesaurus,
-lenguas,
-tipos de relaciones que se usarán,
-características formales,
-características cualitativas.

.1 Lista de los dominios

El listado de las materias a cubrir por el thesaurus se realiza a partir del conocimiento que debe tener quien lo construye acerca de los dominios que ha de abordar el sistema: disciplinas ya tratadas por un servicio existente o que han de ser cubiertas por un servicio que se pone en marcha y cuyo dominio de actividad ha sido fijado tras un estudio sobre las necesidades de los usuarios o tras una orden recibida.

Esta lista de dominios debe ser muy sintética; incluye de cinco a cincuenta (según la importancia del thesaurus) temas fundamentales sobre los que versa(rá) el sistema documental.

Por ejemplo:

-thesaurus de la Caisse Nationale de Crédit Agricole:

+economía general,
+economía agrícola,
+economía financiera,
+gestión,
+derecho;

-thesaurus de l'Electricité de France:

+biología,
+ciencias de la tierra,
+economía,
+derecho,
+ciencias humanas,
+entorno social,
+empresa,
+actividad comercial,
+información,
+gestión de personal,
+regulación del trabajo,
+reglamentación técnica,
+matemáticas,
+informática,
+física,
+mecánica de fluidos,
+física de la materia,
+neutrónica,
+térmica,
+química,
+metrología,
+producto básico,
+propiedad de los materiales,
+estudio de los materiales,
+manipulación y operación,

+utilización de los equipos,
+utilización del sistema,
+seguridad protección,
+utillaje mecánico,
+maquinaria no eléctrica,
+material eléctrico,
+utillaje eléctrico,
+producción de energía,
+distribución - transporte de electricidad,
+construcción,
+transporte,
+técnicas - gas,
+utilización de la energía;

-thesaurus EUDISED, para el tratamiento de la información sobre educación, del Consejo de Europa y de la Comisión de las Comunidades Europeas:

+principios y sistemas de educación,
+política de enseñanza,
+centros de enseñanza, profesorado, alumnado,
+edificios escolares y materiales pedagógicos,
+programas de enseñanza y materias enseñadas,
+investigación en materia de educación, información pedagógica,
+psicología de la educación,
+sociología de la educación,
+economía de la educación,
+administración pública,
+entidades geopolíticas.

.2 Compatibilidad con un thesaurus existente

Hemos visto antes que es raro que un organismo adopte tal cual un thesaurus creado por otro organismo. Se puede intentar al menos que los thesaurus construidos para el/los mismo(s) dominio(s) sean compatibles.

La compatibilidad de los thesaurus es un tema delicado, y quienes realizan el encargo de elaborar nuevos thesaurus están siempre muy preocupados por que resulten «conciliables» con los thesaurus ya existentes.

Stricto sensu, la compatibilidad de dos thesaurus consiste en que los documentos indizados con los descriptores de un thesaurus puedan ser recuperados a través de consultas formuladas por medio de otro thesaurus, y viceversa.

Así definida, la compatibilidad entre dos thesaurus sólo puede existir si uno de los thesaurus es un sub-conjunto del otro, al menos desde el punto de vista léxico; las relaciones de pertenencia, de jerarquía y asociativas pueden ser diferentes, a condición de que la significación de los descriptores sea la misma. Con este espíritu había sido construido el thesaurus enciclopédico americano del E.J.C. (Engineering Joint Council), que pretendía proporcionar una lista de descriptores, de la cual cada productor de thesaurus específico podía extraer libremente los términos que le interesasen para construir su propio lenguaje documental.

Muy pocos productores de thesaurus han utilizado esta posibilidad, y la experiencia muestra que incluso los

thesaurus denominados por sus autores compatibles no lo son en realidad, o lo son muy poco.

Esto no tiene nada de extraño si se recuerda que un thesaurus está construido por un colectivo de usuarios que, siempre, tiene su propia cultura y, por tanto:

-su propia parcelación de la realidad, lo cual significa que los conceptos seleccionados para ser incluidos en el thesaurus no son los mismos; esta divergencia se refleja especialmente en la forma en que son precoordinados los descriptores (véase más adelante la noción de precoordinación de los descriptores, en el § 1.2.2.7).

-su propio punto de vista sobre las entidades constitutivas de esta realidad; lo que se traduce en relaciones de pertenencia y de jerarquía diferentes de un thesaurus a otro y, por tanto, en una disconformidad de sentido entre descriptores idénticos dentro de thesaurus diferentes;

-frecuentemente su propia jerga para designar esas entidades; lo que conduce a divergencias en la elección de descriptores y no-descriptores que designen los mismos conceptos.

Así las cosas, han aparecido en la literatura una serie de tentativas de «reconciliación» de thesaurus; esta reconciliación se obtiene añadiendo relaciones de equivalencia entre los thesaurus que se van a reconciliar, lo que permite convertir automáticamente la indización realizada con un thesaurus en una indización por medio del otro thesaurus; desgraciadamente esta conversión va acompañada en cada caso de una pérdida de información (salvo en el caso excepcional de dos thesaurus que contengan un sub-conjunto común de conceptos designados por descriptores diferentes).

.3 Tamaño

Un thesaurus es un vocabulario controlado y ha de constar, pues, de un número limitado de descriptores.

Es deseable calcular ese número a priori:

-por una parte, para poder evaluar los recursos necesarios para su elaboración;

-por otra parte, para poner un límite al a veces excesivo entusiasmo de los realizadores, que a menudo tienden a cargar los thesaurus con conceptos específicos al margen de los dominios cubiertos por el thesaurus.

Se puede pensar que el tamaño del thesaurus depende fundamentalmente de la riqueza léxica de los dominios a cubrir.

En realidad no es así: un thesaurus no es un diccionario, cuya calidad depende de ser lo más exhaustivo posible, sino un instrumento que ha de ofrecer las mayores prestaciones posibles y que ayude a indizar los documentos y a formular las consultas para recuperar los documentos pertinentes.

Su tamaño depende entonces de:

-el número de documentos que se ha de registrar dentro del correspondiente sistema documental;
-la profundidad de la indización de esos documentos (es decir, el número medio de descriptores asignados a cada documento);
-las modalidades habituales de las interrogaciones y, especialmente, el número medio de conceptos que intervienen en cada consulta;
-el número medio de documentos que se pretende suministrar a cada usuario en respuesta a sus consultas (no demasiados, para evitar el «ruido»; no demasiado pocos, para evitar el «silencio»).

C. Vernimb (1968) ha propuesto una fórmula:

$$T = a.k \sqrt[r.k]{\Phi v}$$

donde:
T =número de descriptores del thesaurus;
a =profundidad de la indización (en general, de 8 a 12 descriptores por documento);
k=número medio de descriptores ligados por una unión en el interior de un grupo de descriptores dentro de una ecuación de búsqueda (número de descriptores dentro de un grupo: en general, de 3 a 4);
n =número medio de grupos de descriptores dentro de una ecuación de búsqueda (número de grupos ligados por una intersección: en general, de 2 a 3);
v =número de documentos dentro del sistema;
r =número medio deseado de documentos recuperados por consulta.

Ejemplos:
1) a = 10; n = 3; v = 100.000; r = 20; k = 3,5
T = 400
2) a = 10; n = 2; v = 100.000; r = 20; k = 3,5
T = 1300
3) a = 10; n = 2; v = 1.000.000; r = 50; k = 3,5
T = 2.600

La experiencia muestra que la aplicación de esta fórmula conduce a un thesaurus muy reducido, e insuficientemente específico. Por nuestra parte preferimos una aproximación más empírica, basada únicamente en el volumen de crecimiento anual de la colección:

-para 10.000 documentos/año : T = de 500 a 1.500 descriptores,
-para 20.000 documentos/año : T = de 1.000 a 2.000 descriptores,
-para 50.000 documentos/año : T = de 2.000 a 4.000 descriptores,
-para 100.000 documentos/año : T = de 3.000 a 6.000 descriptores.

Es muy importante proponerse como objetivo al principio cuál será el tamaño del thesaurus que se va a construir, y atenerse a él estrictamente. Esta «barrera» será muy valiosa

para poner límite al perfeccionismo de quien lo construye, y más todavía al de los especialistas a los que está destinado con el fin de contribuir a su trabajo, y para validar los resultados.

A menudo, la elaboración del thesaurus le produce, en efecto, a su autor una especie de exaltación, que podríamos calificar, por comparación con el mundo de la inmersión submarina, como una «embriaguez de las profundidades». Es necesario recordar constantemente que el thesaurus sólo es un instrumento documental, y que no debe ambicionar la profundidad propia de una enciclopedia.

.4 Lenguas

La determinación de las lenguas que se van a cubrir depende esencialmente de las competencias lingüísticas de los futuros usuarios del thesaurus:

-si todos utilizan la misma lengua, no tendrá ningún interés elaborar un thesaurus multilingüe; en todo caso, se podrán incluir dentro del thesaurus, como no-descriptores, los términos extranjeros encontrados frecuentemente dentro de la literatura;

-si utilizan lenguas diferentes y la política del organismo no es obligar a todos a emplear una sola lengua internacional, y si los recursos del sistema documental son suficientes, el thesaurus podrá ser multilingüe, es decir, disponer de una versión lingüística en la lengua de todos o de los principales grupos de usuarios.

Los thesaurus multilingües dentro de los dominios científicos incluyen como mínimo tres lenguas (alemán, inglés, francés); los thesaurus económicos incorporan en general además el español.

Los thesaurus de instituciones internacionales cubren:

+o bien las lenguas oficiales de esos organismos,
+o bien las lenguas de todos los países miembros; es el caso, por ejemplo, de los thesaurus encargados por la Comisión de las Comunidades Europeas.

.5 Tipos de relaciones

Casi todos los thesaurus incluyen un número limitado de relaciones, escogidas a partir de la gama, ligeramente más amplia, que ofrecen las principales normas nacionales e internacionales(*)12:

-nota explicativa: NE,
-mono-equivalencia semántica: EM.../EP...,
-pluri-equivalencia semántica obligatoria: EM... ET.../EP... AVEC...,
-equivalencia interlingüística (para thesaurus multilingüe),

(*)12 Se mantienen aquí, y en el resto del libro traducido, las abreviaturas francesas, que han sido explicadas en el capítulo I, § 4.2.2.3.4. [nota de los tr.].

-jerarquía: TG.../TS...,
-asociación: TA.../TA...

Muy pocos thesaurus
-distinguen
+los dos tipos de jerarquías: específico/genérico y todo/parte;
+los tres tipos de notas explicativas: definición, nota de aplicación, nota histórica;
-utilizan:
+la pluri-equivalencia semántica facultativa: VOIR... OU.

Por el contrario, muchos thesaurus utilizan la relación de pertenencia (de un descriptor a una, dos o más microdisciplinas), que no está formalizada por ninguna norma.

.6 Características formales

+*Forma*

Usese, en la medida de lo posible:

-la forma nominal antes que la forma verbal o adjetiva.
Ejemplo:
«AUTOMATIZACION», en lugar de «automatizar» o «automático»;
-el número singular

*Nota:* en inglés se prefiere utilizar el plural para los descriptores que designan entidades (seres vivos, organizaciones, equipamientos), como respuesta a la pregunta «how many?», y el singular para los descriptores que designan materiales y propiedades, como respuesta a la pregunta «how much?». Nosotros nunca hemos entendido muy bien el interés de esta regla. Los ingleses no la interpretan todos de la misma forma. Por otra parte, algunos productores de thesaurus ingleses no la aplican.

-la secuencia normal de los términos (si bien el orden invertido corresponde a una larga tradición de los bibliotecarios y de algunos productores de índices);

Ejemplo:
«RECOGIDA DEL ALGODON», mejor que «algodón, recogida».

+*Longitud máxima*

Algunos programas informáticos imponen una limitación a la longitud de los descriptores y de los no-descriptores.
Con independencia de esta restricción, limitar la longitud es beneficioso para quien construye el thesaurus, pues en caso contrario podría tender a precoordinar demasiado los conceptos y a mantener descriptores que representen temas complejos más que conceptos simples.
Según los programas informáticos, la longitud máxima es de 30 a 255 caracteres (letras, cifras, signos de puntuación,

espacios en blanco) por descriptor o no-descriptor. Por nuestra parte, recomendamos una longitud máxima de 50 caracteres (si el programa lo permite).

+*Riqueza tipográfica*

Según los programas, cabe utilizar de 63 a 255 caracteres distintos. Hace mucho tiempo, los productores de thesaurus se contentaban con emplear las letras mayúsculas. En la actualidad exigen cada vez más, y con razón, poder utilizar así mismo las minúsculas y, en consecuencia, los signos diacríticos (acentos, cedillas...). En el caso de los thesaurus multilingües y cuando deban aparecer las equivalencias lingüísticas dentro de una misma edición, se deben definir especificaciones particulares: el sistema ha de permitir procesar todas las letras y signos diacríticos dentro de todos los alfabetos (latín, cirílico...) de las lenguas del thesaurus. La codificación que rige los alfabetos latinos está cubierta por la norma ISO 6937.

.7 Características cualitativas

+*Significación de los descriptores*

Los descriptores de un thesaurus deben representar conceptos y han de tener una significación por sí mismos. Ahora bien, dentro de algunos dominios, puede parecer interesante, en teoría, utilizar los «infraconceptos» (denominados por algunos productores de thesaurus «descriptores auxiliares») para limitar el número de descriptores dentro del thesaurus, permitiendo un número de entradas muy elevado dentro de los ficheros de búsqueda: estos descriptores auxiliares, limitados en número, pueden en efecto ser añadidos a algunos descriptores, lo que permite representar una cantidad importante de conceptos compuestos por post-coordinación por medio de una lista limitada de términos.

Ejemplo: «super» puede formar una gran cantidad de descriptores compuestos cuando es unido a descriptores simples como «conductor», «fosfato», «estructura»...

Nosotros opinamos que se debe evitar esta práctica, a pesar de sus ventajas: un thesaurus ha de poder ser usado no sólo por documentalistas, sino también por usuarios ocasionales; por tanto, interesa que la utilización del thesaurus sea lo más cómoda posible y que se integren los descriptores compuestos útiles, evitando recurrir a reglas de composición: aunque esas reglas sean simples, complican el trabajo del usuario.

Algunos autores han propuesto una técnica similar, denominada análisis por factores semánticos (semantic factoring) o también por componentes semánticos (componential analysis), pero no se ha usado prácticamente nunca, por el mismo motivo de «amigabilidad» del thesaurus. Esta técnica consiste en representar un concepto no por su enunciado habitual dentro de la lengua, sino por sus rasgos semánticos;

el interés de esta aproximación es que permite limitar considerablemente el tamaño del thesaurus.

Por ejemplo, el concepto «mujer» será representado por la post-coordinación de los componentes semánticos «humano», «femenino» y «adulto»; así mismo, el concepto de «termómetro» lo formarán «aparato», «medida», y «temperatura». Un inconveniente adicional de este método es que, como los conceptos están descompuestos en rasgos semánticos o en «infra-conceptos», sus enunciados no figuran dentro del thesaurus, y entonces no pueden ser ligados por relaciones de equivalencia, de jerarquía y de asociación, que ayuden a los usuarios a recuperarlos.

Así mismo, una práctica que es corriente dentro de los lenguajes de clasificación no puede ser usada dentro de los lenguajes combinatorios: nos referimos a la enunciación parcialmente implícita de un concepto, que resulta posible por su localización dentro de la clasificación. Por ejemplo, se puede tener dentro de una clasificación:

OBRERO
    CUALIFICADO
    ESPECIALIZADO

mientras que dentro de un thesaurus, se tendrá

OBRERO
    OBRERO CUALIFICADO
    OBRERO ESPECIALIZADO

de manera que cada entrada del thesaurus, considerada aisladamente, tenga una significación precisa cuando los descriptores estén presentados en orden alfabético.

+*Ratio de precoordinación*

Esta ratio procede de dividir el número de veces que aparecen las palabras que constituyen las palabras del thesaurus por el número de descriptores.

Lo ideal es que esta ratio se sitúe entre 1,5 y 2 para un thesaurus francés; un descriptor debe, por tanto, contener como media de 1,5 a 2 palabras significativas (VAN SLYPE - 1976).

Una ratio demasiado elevada produciría un thesaurus de tamaño demasiado grande: en efecto, si se combinan entre sí los conceptos simples para construir descriptores compuestos, el resultado corre el riesgo de ser demasiado elevado.

Por el contrario, una ratio demasiado reducida haría que dentro del thesaurus hubiera:

-unitérminos sin significación precisa (ejemplo: la palabra «servicio»), que no corresponde a un concepto definido;

-descriptores de utilización muy frecuente, y por tanto insuficientemente discriminantes para la búsqueda (por ejemplo, dentro de un thesaurus económico un descriptor como «economía» no tiene ningún interés; por el contrario, descriptores como «economía financiera», «economía industrial», «economía agrícola» serán útiles).

La ratio de precoordinación es:

-más elevada dentro de los dominios centrales del sistema documental cubierto por el thesaurus (para evitar una frecuencia de indización excesiva);
-menos elevada dentro de los dominios marginales (para obtener descriptores más generales y evitar una frecuencia de indización demasiado reducida).

Por ejemplo:

-dentro de un thesaurus sobre la economía en general se podrá tener como descriptores distintos
POLITICA
TIPO DE INTERES

-dentro de un thesaurus sobre economía financiera se tendrá como descriptor adicional:
POLITICA DE TIPO DE INTERES

destinado a obtener una indización más precisa por medio de la utilización de descriptores específicos para designar conceptos específicos: un documento indizado por «POLITICA DE TIPO DE INTERES» y «BANCO DE FRANCIA» se recuperará con una consulta que contenga estos descriptores; y no se recuperará, si no se desea, con una consulta sobre «POLITICA» y «BANCO DE FRANCIA».
La noción de precoordinación corresponde a lo que los gramáticos denominan la composición: «una palabra es compuesta desde el momento en que evoca en la mente no ya imágenes distintas correspondientes a cada una de las palabras que la integran, sino una imagen única» (Grévisse). Los ejemplos proporcionados por los gramáticos de la lengua usual (estrella de mar, arco del triunfo) se aplican en el ámbito de los lenguajes documentales. Pero quien construye el thesaurus debe ir más allá que el lenguaje corriente y ha de reflejar el lenguaje especializado de los dominios que trata: dentro de un diccionario se encuentran como entradas, en la mayoría de los casos, palabras simples, y, las menos veces, palabras compuestas; por el contrario, dentro de un thesaurus las palabras compuestas (que nosotros denominamos precoordinadas) son mucho más frecuentes, y a menudo suponen la mayor parte de las entradas.

+*Tasa de equivalencia*

La tasa de equivalencia es la ratio entre el número de no-descriptores y el número de descriptores del thesaurus.
Según los dominios, esta tasa se sitúa entre 0,5 y 2. Una tasa elevada (una gran cantidad de no-descriptores) mejora la coherencia y la precisión de la indización, ya que tiene como resultado el aumentar el número de entradas alfabéticas del thesaurus, y en consecuencia la probabilidad de que un término en lenguaje natural que aparece dentro de un documento o de una consulta coincida con una entrada del thesaurus.

+*Tasa de enriquecimiento*

La tasa de enriquecimiento es la ratio entre el número de relaciones jerárquicas y asociativas, por una parte, y el número de descriptores, por otra.

Si contamos como una unidad cada relación y su inversa (A TG B y B TS A; o C TA D y D TA C), esta ratio debería situarse entre 1 y 3: cada descriptor debería estar ligado al menos a otro descriptor, de cara a disponer de una red semántica eficaz (VAN SLYPE - 1976).

Una ratio demasiado reducida es lo característico de un thesaurus semánticamente poco estructurado (téngase en cuenta que esta ratio es igual a cero dentro de una lista de autoridades), en el que el usuario encontrará demasiado pocos reenvíos de un descriptor a otro.

Una ratio demasiado elevada significa que las relaciones son demasiado numerosas como para que resulte cómodo examinarlas.

+*Flexibilidad*

La flexibilidad es la proporción de palabras significativas simples utilizadas para constituir los descriptores compuestos (es decir, precoordinados) y que se encuentran así mismo dentro del thesaurus como descriptores no compuestos o como no-descriptores.

Ejemplos:

+si un thesaurus incluye el descriptor «POLITICA DE TIPO DE INTERES», es deseable que los dos conceptos que lo componen, «POLITICA» y «TIPO DE INTERES» figuren igualmente como descriptores, con el fin de permitir indizar específicamente estos tres conceptos.

+si, por el contrario, se encuentra dentro del thesaurus el descriptor «CONTROL DEL GASTO PUBLICO», es interesante incorporar así mismo los descriptores «GASTO PUBLICO» y «GASTO», pero no los términos «control» ni «público», que son demasiado imprecisos.

Parece que una ratio de 0,6 se corresponde con la flexibilidad óptima dentro de un thesaurus en francés (VAN SLYPE - 1976).

.3 Elección de las fuentes terminológicas

Hemos citado en el § 1.1.1 los diversos tipos de fuentes terminológicas que se pueden utilizar para elaborar un thesaurus.

Durante el estudio de viabilidad será necesario localizar las fuentes que serán efectivamente utilizadas:

-lenguajes documentales ya existentes dentro del servicio que pretende crear un nuevo thesaurus;
-lenguajes documentales ya existentes fuera de ese servicio;
-tratados, diccionarios técnicos...;
-expertos.

Para cada una de esas fuentes será necesario:

-determinar su accesibilidad;
-evaluar sus aportaciones.

.4 Proceso de tratamiento

Distinguiremos cinco posibles procesos de tratamiento:

-tratamiento completamente manual;
-utilización de un programa de tratamiento de textos;
-utilización de un sistema de gestión de thesaurus en proceso por lotes;
-uso mixto de un programa de tratamiento de textos y de un sistema de gestión de thesaurus por lotes;
-utilización de un sistema de gestión de thesaurus en modo interactivo.

La elección entre estos métodos depende de:

-el tamaño del thesaurus: un thesaurus de varios cientos de términos no exige la utilización de un sistema automatizado; un thesaurus multilingüe de varios miles de términos sólo se puede hacer utilizando un programa informático especializado;
-el o los tipos de presentación del thesaurus: elaborar una presentación gráfica y una simple lista alfabética no estructurada se puede muy bien realizar a mano;
-el programa informático de almacenamiento y recuperación documental utilizado dentro del sistema documental: si este programa dispone de un módulo de gestión de thesaurus, tendrá mucho interés utilizarlo incluso para un thesaurus pequeño, ya que en cualquier caso el thesaurus deberá ser registrado dentro de los ficheros gestionados por el programa.

.1 Tratamiento completamente manual

Entre los productores de thesaurus predominan dos métodos:

-el método de fichas;
-el método gráfico.

.1 Método de fichas

-El vocabulario natural recogido se registra en fichas a razón de un término por ficha.
-Las fichas se ordenan por microdisciplinas.
-La conversión a lenguaje documental se realiza, dentro de cada microdisciplina, examinando las fichas y:
    +eliminando las fichas de los términos que definitivamente no se conservan para ser incluidos dentro del thesaurus;
    +anotando un código «D» junto a los términos conservados como descriptores, un código «ND» junto a los términos conservados como no-descriptores;

+anotando en las fichas de los no-descriptores el descriptor o descriptores que se han de utilizar;
+anotando en las fichas de los descriptores:
->las equivalencias semánticas;
->el código de la microdisciplina;
->si es necesario, una nota explicativa.

-Las relaciones jerárquicas se establecen para cada disciplina:
+ordenando las fichas de los descriptores según cada cadena jerárquica;
+anotando en cada ficha el o los descriptores genéricos y específicos dentro de un solo nivel jerárquico en ambos sentidos: ascendente y descendente;
+cuidando que sean anotadas las relaciones invertidas en todas las fichas afectadas;
+añadiendo nuevas fichas de descriptores si al examinar las cadenas que se constituyen aparecen lagunas.

-Se verifican los dobles usos ordenando las fichas de todo el thesaurus por orden alfabético: cuando un descriptor aparece dentro de dos o más microdisciplinas:
+si se decide mantenerlo así (polijerarquía), se anota el código del nuevo o nuevos microthesaurus, así como las relaciones jerárquicas dentro de ese o esos nuevos microthesaurus, en una de las fichas, y la otra u otras fichas del mismo descriptor adquieren un estatuto de fichas de referencia (rappel);
+si se observa que la presencia de un descriptor dentro de dos o más microdisciplinas corresponde a una polisemia, se modifica el enunciado del descriptor en la(s) ficha(s) adecuada(s) y se introducen las modificaciones correspondientes en todas las fichas en las que ese descriptor aparece con relación jerárquica o de equivalencia semántica;
+cuando un no-descriptor aparece dos o más veces, se desambigua (por un modificador o un calificativo) o se le vincula a dos o tres descriptores (poliequivalencia) y se introduce esta modificación en todas las fichas afectadas;
+cuando un término aparece como descriptor dentro de una microdisciplina y como no-descriptor en otra, se suprime o se desambigua uno de los dos enunciados y se introduce esta modificación en todas las fichas afectadas.

-Se establecen ahora las relaciones asociativas para cada nueva microdisciplina, después de haber ordenado las fichas por microdisciplinas:
+anotando en cada ficha el o los descriptores asociados;
+cuidando que sean anotadas las relaciones invertidas en todas las fichas afectadas.

-La edición del thesaurus se lleva a cabo tras ordenar alfabéticamente de nuevo las fichas y entregar el resultado:
+o bien a dactilografía;
+o bien al impresor.

Se obtiene así una presentación alfabética estructurada completa, con un solo nivel jerárquico en cada sentido. Si se desea, las fichas (o una copia de las fichas) pueden ahora ser ordenadas por microdisciplinas para preparar la edición de una lista alfabética, estructurada o no, por microdisciplinas.

*Apreciación*

Este método es el más clásico; presenta numerosos inconvenientes:

-falta una visión de conjunto sobre el contenido de una microdisciplina;
-se tienen que ordenar tres veces todas las fichas;
-persisten una serie de errores ilocalizables, fundamentalmente:
+relaciones semánticas no reflejadas, por olvido, en todas las fichas que corresponden;
+divergencias ortográficas en los enunciados de un mismo descriptor dentro de varias fichas, debidas a errores de copia;
+términos o relaciones que faltan, por omisiones en el momento de la dactilografía o del mecanografiado para la impresión.

.2 Método de los grafos

-El vocabulario natural recogido se registra en hojas de papel, a razón de una hoja por microdisciplina.
-La conversión en lenguaje documental se realiza, para cada microdisciplina, examinando el contenido de cada hoja y:
+tachando los términos que definitivamente no se conservan para ser incluidos en el thesaurus.
+subrayando los términos conservados como descriptores;
+manteniendo tal cual los términos conservados como no-descriptores.
-Las relaciones jerárquicas y las equivalencias semánticas se establecen para cada disciplina:
+volviendo a copiar en una nueva hoja los descriptores, una cadena jerárquica tras otra, y señalando su nivel jerárquico por la posición en la hoja:
->o bien colocando en el centro de la hoja los descriptores más genéricos y hacia los bordes los descriptores más específicos, ligando unos con otros por flechas, en el caso del diagrama de flechas;
->o bien con un sangrado hacia la derecha para cada nivel jerárquico, en el caso del terminograma;

+volviendo a copiar los no-descriptores bajo los descriptores correspondientes, con caracteres más pequeños;
+añadiendo nuevos descriptores y no-descriptores a la hoja, si la jerarquía así elaborada revela lagunas.

-Se verifican los dobles usos esporádicos:
+volviendo a copiar en fichas cada descriptor, acompañado por el código de la microdisciplina, y cada no-descriptor, junto al que se anota cuál es el descriptor equivalente y el código de la microdisciplina de ese descriptor;
+ordenando las fichas por orden alfabético; los procesos se realizan de forma similar al método de las fichas, pero los resultados se señalan a la vez en la hoja (diagrama o terminograma) y en las fichas afectadas; es necesario destacar, sin embargo, que hay que modificar muchas menos fichas que en el método anterior, ya que aquí las fichas no incluyen las relaciones jerárquicas.

-Se establecen las relaciones asociativas para cada microdisciplina, y por tanto para cada hoja:
+enlazando por una línea los descriptores asociados dentro la misma microdisciplina;
+volviendo a copiar en el margen de la hoja cuál es el descriptor y el código de la microdisciplina de los descriptores de otras microdisciplinas que han de ser asociados a los descriptores de la microdisciplina que se está examinando, y enlazando descriptores externos e internos por una línea;
+volviendo a copiar las relaciones invertidas en las hojas afectadas.

-La edición del thesaurus se realiza:
+dactilografiando o imprimiendo los diagramas de flechas o los terminogramas, en el caso de la presentación esquemática;
+dactilografiando o imprimiendo las fichas de control de doble uso, tal cual, para una presentación alfabética no estructurada completa (considerada a menudo suficiente, en la medida en que las relaciones jerárquicas y asociativas aparecen en la presentación gráfica);
+ordenando las fichas por microdisciplinas, volviendo a copiar en ellas las relaciones jerárquicas y asociativas que figuran en la presentación gráfica, y volviéndolas a ordenar por orden alfabético y dactilografiándolas o imprimiéndolas para obtener una presentación alfabética estructurada completa, si se considera necesario.

*Apreciación*

Este método es superior, con mucho, al método de las fichas, ya que presenta como ventajas:

-un tratamiento de cada microdisciplina teniendo una visión de conjunto sobre su contenido;
-una sola ordenación, en vez de tres, si nos conformamos con una presentación alfabética no estructurada;
-una edición gráfica con la estructura jerárquica para todos los niveles, la estructura asociativa y la estructura de equivalencia semántica;
-una disminución de los riesgos de error, en la medida en que la señalización de una jerarquía o de una asociación se realizan respnto de la elaboración;

desde el punto de vista técnico, todo esto se puede realizar directamente en un tratamiento de textos, pero ello nos obliga a duplicar los gastos de personal, a prolongar el tiempo de elaboración, sin que desaparezcan ninguno de los inconvenientes de los métodos puramente manuales.

### .2 Utilización de un sistema automatizado de gestión de thesaurus en proceso por lotes

Estos sistemas pueden, como los tratamientos de textos, intervenir dentro del proceso de elaboración manual del thesaurus. Su intervención en paralelo es posible, técnicamente hablando, pero a costa de aumentar considerablemente los gastos y dilatar el tiempo de elaboración, a causa de lo fastidiosos que son los procedimientos de supresión y las modificaciones de términos y/o de relaciones dentro de los sistemas de proceso por lotes.

Sin embargo, las aportaciones de estos sistemas son muy importantes y su utilización es muy eficaz cuando ya se está finalizando la elaboración de un thesaurus.

Existen dos tipos de sistemas de gestión de thesaurus:

-el más simple es el que se incluye como uno de los módulos de la mayor parte de los programas informáticos de almacenamiento y recuperación documental (ejemplos: MISTRAL de BULL, BASIS de BATELLE, GOLEM de SIEMENS);
-el más sofisticado es un programa concebido especialmente para la gestión de thesaurus (ejemplo: ASTUTE, de la Comisión de las Comunidades Europeas).

### .1 Módulo de gestión de thesaurus

Los módulos de gestión de thesaurus asociados a programas de almacenamiento y recuperación documental permiten:

-registrar los términos y sus relaciones directas (por ejemplo: de un no-descriptor hacia un descriptor, de un descriptor específico hacia el genérico);
-producir automáticamente las relaciones invertidas (sin olvidos ni errores ortográficos);
-ordenar los términos alfabéticamente;
-editar una lista alfabética completa, estructurada para uno (ej.: GOLEM, BASIS) o para todos (ej.: MISTRAL) los niveles jerárquicos.

Algunos de ellos ofrecen aún más prestaciones (especialmente BASIS) y permiten además:

-gestionar automáticamente un thesaurus multilingüe, introduciendo las relaciones jerárquicas y asociativas en un sentido y dentro de una lengua, así como las relaciones lingüísticas entre lenguas, y editando automáticamente la lista alfabética estructurada completa dentro de cada una de las lenguas del thesaurus;
-ordenar los términos por microdisciplinas y editar una lista alfabética, estructurada o no, por microdisciplinas.

El mayor inconveniente de estos módulos (además del funcionamiento en proceso por lotes, que acabamos de «condenar») es que no realizan ninguna verificación de verosimilitud: se pueden introducir dos términos ligados por una relación de jerarquía, por una relación de asociación y por una relación de equivalencia semántica y ver cómo el sistema acepta esos errores generando impasible las relaciones invertidas. Existe un solo programa de almacenamiento y recuperación documental, que nosotros conozcamos, que permita evitar esta contrariedad, gracias a que realiza una validación de las relaciones registradas y rechaza las que son incompatibles: se trata del programa ADLIB (en mini-ordenadores PRIME).

.2 Programa especializado de gestión de thesaurus

Los programas especializados asumen todas las funcionalidades de los módulos más potentes e incluyen la validación automática de los términos y relaciones registrados, emitiendo un mensaje de error para cada incompatibilidad detectada. Estos programas permiten además ordenar los descriptores por orden jerárquico y editar tanto las listas jerárquicas como los índices multilingües.
Estas listas jerárquicas descendentes pueden ser producidas por microdisciplinas: así resulta posible recortar cada cadena jerárquica y pegarla en una hoja de papel para organizar el contenido interno del terminograma; se obtiene entonces una compatibilidad total entre el contenido de las listas y el contenido de la presentación gráfica.

.3 Uso mixto de un programa de tratamiento de textos y de un sistema de gestión de thesaurus en proceso por lotes

Esta agrupación presenta una ventaja importante si se cumplen las condiciones siguientes:

-se elabora el thesaurus según el método gráfico de los terminogramas;
-hay seguridad de que los terminogramas ya construidos sufrirán pocas modificaciones (menos del 20 %) tras la realización de los terminogramas y la verificación final. En la mayoría de los casos esta seguridad sólo se adquiere cuando se han terminado todos los terminogramas.

En estas condiciones, se puede definir el procedimiento así:

-tras haber realizado la conversión a lenguaje documental y haber establecido las relaciones jerárquicas y las equivalencias semánticas, en una hoja por microdisciplina o grupo de microdisciplinas:
+se mecanografía el contenido de las hojas en cuestión y se incluye un código delante de cada no-descriptor y un sangrado para cada nivel jerárquico (sin necesidad de una previa ordenación alfabética manual de los descriptores para cada nivel jerárquico);
+se editan, con el tratamiento de textos, los borradores de terminogramas;
+se verifica inmediatamente, releyéndolo, y se corrige con el tratamiento de textos;
-se prepara un programa informático que permitirá transformar el formato «tratamiento de textos» en formato de entrada para el módulo o para el programa de gestión de thesaurus; este reformateo debe añadir ante todo:
+la relación de pertenencia (a la microdisciplina) de cada uno de los descriptores;
+la relación entre los descriptores y su(s) no-descriptor(es);
+la relación entre los descriptores genéricos y su(s) específico(s) dentro de cada nivel jerárquico;
-se convierte el formato y se cargan automáticamente las microdisciplinas, los descriptores, los no-descriptores y las relaciones de equivalencia semántica y de jerarquía dentro del fichero de entrada del módulo o del programa de gestión de thesaurus;
-se hace el tratamiento de validación;
-se examinan los mensajes de error (incompatibilidad de relaciones), se realizan tratamientos intelectuales y se abordan las correcciones, simultáneamente en:
+el programa de tratamiento de textos;
+el módulo o programa de gestión de thesaurus;
-se realiza el tratamiento de inversión de las relaciones;
-se lleva a cabo la edición
+por microdisciplinas;
+del conjunto de las microdisciplinas procesadas hasta el momento;
-cuando haya sido tratada la totalidad de las microdisciplinas, se podrá:
+hacer el mecanografiado, directamente sobre el fichero de entrada del sistema de gestión de thesaurus, o vía tratamiento de textos, de:
->las notas explicativas,
->las relaciones asociativas,
->las relaciones de equivalencia interlingüística,
->y las relaciones de equivalencia semántica en las lenguas no de base;
+hacer que se lleven a cabo los tratamiento de validación, de inversión de las relaciones asociativas y de generación de la estructura jerárquica, asociativa y de equivalencia semántica dentro de las lenguas no de base;

+hacer que se edite el thesaurus con las diferentes presentaciones deseadas.

*Apreciación*

El procedimiento que se ha descrito permite que se ganen varias semanas en la realización del thesaurus y que se simplifiquen considerablemente los procedimientos habituales de detección y de corrección de errores. Sin embargo, sólo es interesante comenzar la carga en el tratamiento de textos cuando ya se haya terminado sobre papel una parte apreciable de los terminogramas y se esté bastante seguro de que no habrá que volver sobre su contenido.

## .4 Utilización de un sistema de gestión de thesaurus en modo interactivo

Desde 1985 han aparecido en el mercado una serie de programas (en Francia: ALEXIS, de la sociedad ERLI; en el Reino Unido: STRIDE, de B.N.F. Metals Technology Centre; en Israel y Alemania: DOMESTIC, del National Center of Scientific and Technological Information de Tel-Aviv y de KTS-Informationssysteme de Munich) que gestionan un thesaurus de forma interactiva permitiendo:

-registrar los términos y las relaciones en el terminal;
-validar inmediatamente cada entrada, emitiendo un mensaje de error cuando sea necesario;
-actualizar inmediatamente el fichero, para cada una de las entradas correctas, realizando automáticamente las inversiones de las relaciones y transfiriendo las relaciones directas e inversas a las diferentes versiones lingüísticas;
-llevar a cabo las supresiones y las modificaciones de términos y de relaciones en la misma forma interactiva;
-visualizar e imprimir, cuando se desee, de forma inmediata, la parte del thesaurus sobre la que se desea trabajar.

Alexis, por ejemplo, permite gestionar:

-las unidades léxicas o lexías (descriptores o no-descriptores), compuestas obligatoriamente de:
+una grafía o forma gráfica (ejemplo: la cadena de caracteres INFORMATICA);
+un tipo o categoría gramatical (ejemplo: nombre común).

Una unidad léxica contiene 256 caracteres como máximo y puede ser:

+o bien una palabra simple
ejemplo: INFORMACION/nombre
+o bien una palabra compuesta, que se representa dentro del sistema por un diagrama arbóreo en el que
->las hojas contienen una grafía,
->los nodos contienen un tipo,
->el tronco resulta de la concatenación de todas las hojas.

Ejemplo:

(SISTEMA DE INFORMACION DOCUMENTAL)

O/nombre

(SISTEMA DE INFORMACION) O/nombre

O/nombre O/preposición O/nombre O/adjetivo
SISTEMA DE INFORMACION DOCUMENTAL

-los textos, formados por una cadena de caracteres (4.000 como máximo), sin distinción de componentes (nota explicativa);
-relaciones no invertidas entre una lexía y un texto.

Ejemplo:

NE

INFORMACION úsese con el sentido de hacer informes y no con el de acción de informar.

-relaciones invertidas entre dos lexías, simples o compuestas.

Ejemplo:

SISTEMA DE INFORMACION/ nombre

TECNOLOGIA DE LA INFORMACION/nombre

INFORMACION SOBRE EL SISTEMA/nombre

PARA CADA CAMPO SEMANTICO
CAMPO SEMANTICO

TS

TG

TA

TA

SISTEMA DE INFORMACION DOCUMENTAL/nombre

COMPONENTE COMPUESTO
SISTEMA/nombre

COMPONENTE COMPUESTO
INFORMACION/nombre

COMPONENTE COMPUESTO
DOCUMENTAL/adjetivo

EM

EP

TA

TA

INFORMARSE/nombre

DOCUMENTACION/nombre

Alexis permite realizar las funciones siguientes:

-creación de un modelo de thesaurus; declarando, en modo interactivo:

+los tipos de unidades léxicas (512 como máximo);

+los nombres de las relaciones (512 nombres de relaciones como máximo, es decir, bastante más de lo que exigen las normas en materia de thesaurus);

+las reglas de integridad.

Ejemplo: una sola relación semántica entre dos lexías.

-actualización:

+al introducir en el terminal

->una nueva lexía;

->una nueva relación entre dos lexías;

->la supresión de una relación, sin suprimir las lexías correspondientes;

->una modificación de lexía;

->la supresión de una lexía y de sus relaciones;

+al introducir en el terminal correcciones en respuesta a los mensajes de error emitidos tras la aplicación automática de las reglas de integridad;

+se muestran las relaciones invertidas creadas automáticamente por el sistema

-consulta:

+al introducir en el terminal un esquema, es decir, una lista de tipos de lexías y de nombres de relaciones que se desea visualizar;

+al introducir en el terminal una lexía;

+se muestra en el terminal esa lexía y las lexías que le quedan ligadas, de acuerdo con el esquema; cada lexía recibe un nº de orden dentro de la lista mostrada, lo que permite invocarla por ese nº;

+navegación a través del thesaurus, introduciendo el nº de orden de una lexía que aparece en la pantalla, para visualizar las lexías relacionadas con ella, siguiendo el mismo esquema.

La consulta está facilitada por la existencia de una serie de programas:

+corrección ortográfica.

Ejemplo: SISTEM va a recuperar SISTEMA;

+gestión de truncamiento.

Ejemplo: SISTEM... va a recuperar SISTEMA, SISTEMICO, SISTEMATICO;

+ extensión morfológica.

Ejemplo: SISTEMATIZAR va a recuperar SISTEMATIZACION;

+gestión de proximidad y de secuencia.

Ejemplo: SISTEMA... INFORMACION va a recuperar SISTEMA DE INFORMACION, pero no INFORMACION SOBRE EL SISTEMA;

+búsqueda booleana

->intersección.

Ejemplo: SISTEMA e INFORMACION va a recuperar SISTEMA DE INFORMACION e INFORMACION SOBRE EL SISTEMA;

->unión.

Ejemplo: SISTEMA o INFORMACION va a recuperar SISTEMA, INFORMACION, SISTEMA DE INFORMACION, INFORMACION SOBRE EL SISTEMA, TECNOLOGIA DE LA INFORMACION, SISTEMA DE INFORMACION DOCUMENTAL, TECNOLOGIA DE LA INFORMACION;

->combinación de la búsqueda booleana y de la búsqueda de nombres de relaciones.

Ejemplo: EM INFORMARSE y TA DOCUMENTACION va a recuperar SISTEMA DE INFORMACION DOCUMENTAL;

-edición alfabética del thesaurus con o sin todas las relaciones o con una parte.

.5 Definición de las responsabilidades

Hay que prever las siguientes responsabilidades dentro del equipo de construcción:

-responsable(s) de la construcción del thesaurus (en la lengua de base);
-responsable(s) de la validación técnica del contenido del thesaurus;
-responsable(s) de la búsqueda de equivalencias interlingüísticas;
-responsable(s) de la dactilografía (fichas, diagramas de flechas) y de la mecanografía (términos y relaciones).

Junto al equipo de construcción, a menudo es útil designar un comité director.

.1 Comité director

Constituido por:

-un representante a alto nivel de la Dirección del organismo en el que se construye el thesaurus;
-el responsable del sistema de información para el que se construye el thesaurus;
-el jefe del proyecto de construcción del thesaurus;
-y, si es el caso,
+una personalidad externa: consultor especializado, responsable de algún otro servicio de documentación;
+delegados a alto nivel de los principales Departamentos del organismo usuario del sistema documental relacionados con la construcción del thesaurus.

El comité director tiene como responsabilidades:

-hacer ejecutar y aprobar el estudio de oportunidad;
-designar los realizadores y distribuir las tareas (cf. más adelante);

-recibir un informe periódico del progreso de los trabajos, preparado por el jefe del proyecto;
-adoptar en el momento deseado las medidas necesarias para una buena ejecución o para la reorientación de los trabajos.

.2 Responsable de la construcción

En lo que se refiere al perfil y al número de responsables de la construcción, se puede optar entre numerosas posibilidades:

+*Perfil:*

-documentalista: buena solución si el documentalista conoce muy bien las materias que van a ser cubiertas por el thesaurus y puede dedicarse a tiempo completo a la construcción;
-especialista del tema, destinado durante el tiempo de la construcción: buena solución si ese especialista tiene dominio sobre la totalidad de las materias que va a cubrir el thesaurus y si puede hacer el esfuerzo de manejar las técnicas relacionadas con el almacenamiento y recuperación documental;
-generalista, destinado durante el tiempo de la construcción: buena solución si ese generalista puede manejar las técnicas documentales y apoyarse en un equipo sólido de validación y si las tareas de validación abarcan también la preparación de la lista de posibles descriptores.

+*Personal:*

-una persona: solución ideal, en la medida en que interesa que todas las microdisciplinas sean tratadas de forma homogénea y que cada decisión (añadir o suprimir un término o una relación) se aplique a varias microdisciplinas;
-varias personas: solución más costosa, que sólo reduce el plazo de construcción con la condición expresa de que sea una sola persona la responsable de la calidad del conjunto del thesaurus, y las otras personas desempeñen tareresponsas de preparación y actúen de intermediarios con los encargados de la validación, dentro de cada grupo de microdisciplinas.

.3 Responsable de la validación

+*Perfil:*

-responsable de la construcción: solución ideal si está así mismo preparado en todos los dominios del thesaurus, lo que es de hecho muy poco frecuente;
-especialista en los dominios del thesaurus, cuya función será, según el nivel de preparación del responsable o responsables de la construcción:
+proponer listas de posibles descriptores para cada dominio;
+definir los términos desconocidos o mal conocidos por el responsable de la construcción;

+aprobar los proyectos preparados por el responsable de la construcción;
->elegir los descriptores y los no-descriptores,
->relaciones de pertenencia, de jerarquía y de asociación, así como las notas explicativas.

+*Personal:*

-Si se trata de un responsable que pertenece al segundo tipo de perfil, es necesario prever tantas personas como sean necesarias para cubrir el conjunto de los dominios del thesaurus; dentro de las empresas en las que la documentación se genera de forma muy descentralizada, y en colaboración con los usuarios, se tratará de designar un responsable por cada servicio usuario.

.4 Responsable de las equivalencias interlingüísticas

+*Perfil:*

-traductor que pertenezca a la empresa: buena solución en la medida en que se podrán aprovechar los conocimientos de la terminología de la casa;
-traductor independiente: solución que sólo se puede adoptar cuando es posible que las equivalencias lingüísticas sean validadas por especialistas de los dominios que se tratan;
-especialista bilingüe en los dominios del thesaurus: es la mejor solución, pero la más difícil de aplicar por motivos de disponibilidad y, a veces, de motivación:
-documentalista o especialista monolingüe (para la lengua de la versión de base) en los dominios del thesaurus, encargado de buscar las equivalencias lingüísticas dentro de diccionarios: mala solución generalmente, a no ser que resulte posible que se valide su trabajo con profundidad;
-documentalista o especialista monolingüe (para cada una de las lenguas destino): solución generalmente mejor que la precedente.

+*Personal:*

-al menos un traductor para cada lengua destino;
-o un especialista por lengua y por dominio.

.5 Distribución de las responsabilidades: recomendaciones

El método más eficaz consiste en reunir ambas competencias (en documentación y en los dominios cubiertos) en una persona muy motivada, para cada una de las lenguas del thesaurus; a la persona más preparada en los dominios del thesaurus, o en la construcción de thesaurus, se le nombra jefe de proyecto, y se elige su lengua materna como lengua de base del thesaurus: se obtiene así rápidamente un thesaurus de calidad, que puede ser explotado inmediatamente y cuyas primeras revisiones serán poco importantes.
A falta de poder constituir un equipo de estas características, será necesario:

-designar a un documentalista responsable del trabajo de construcción del thesaurus monolingüe o de la versión en la lengua de base (que debe ser la lengua materna del documentalista) de un thesaurus multilingüe;
-hacer que intervengan activamente los usuarios que conozcan bien la materia de cada uno de los dominios del thesaurus, y que se encarguen de:
+proporcionar las listas de posibles descriptores dentro de sus dominios;
+explicitar aquellos términos que parezcan ambiguos a los ojos del documentalista;
+aprobar las elecciones realizadas por el documentalista en materia de descriptores y de relaciones semánticas para las microdisciplinas de sus dominios de competencia;
-obtener la colaboración de especialistas para cada lengua y dominio con el fin de producir las otras versiones lingüísticas del thesaurus, si es multilingüe.

El tiempo de construcción será considerablemente mayor que con el primer método, pero el resultado podrá ser de muy buena calidad.

Se puede intentar un tercer método, que se denomina por comisionado, y que consiste en hacer que trabaje un equipo de documentalistas, del mismo rango, vinculados o no con especialistas de los dominios cubiertos; el tiempo de construcción será largo y la calidad del thesaurus obtenido no estará necesariamente garantizada.

.6 Estimación del coste

Es necesario considerar tres cuestiones:

+*Personal:*
-responsable de la construcción de un thesaurus monolingüe o de la versión de un thesaurus multilingüe en la lengua de base:
de 6 a 24 meses/hombre, según
+el tamaño del thesaurus,
+la organización del trabajo,
+la preparación de la persona.

A este número de meses/hombre hay que añadir aproximadamente un 10 % más por cada lengua tratada, si el thesaurus es multilingüe, para tener en cuenta la retroacción necesaria entre lenguas

-responsables de la validación:

de 5 a 30 días por dominio, según
+el tamaño del thesaurus, dentro del dominio considerado,
+el tipo de intervención solicitada: preparación o no de lista de posibles descriptores o simple validación a posteriori;

-responsables de las equivalencias interlingüísticas:

de 3 a 12 meses/hombre por cada lengua destino, según
+el tamaño del thesaurus,
+la preparación de la persona;

-responsable de la dactilografía y/o de la mecanografía:

de 6 a 24 meses/hombre
para un thesaurus monolingüe o la versión de base de un thesaurus multilingüe,
de 2 a 6 meses/hombre por cada lengua destino.

+*Informática:*

Cuando se utiliza un programa de ordenador especializado para la construcción de thesaurus, el volumen de gastos en tratamientos es del 25 al 35 % del coste del personal.
Cuando se usa el módulo de edición de thesaurus de un programa de almacenamiento y recuperación documental, el coste informático sólo representa del 5 al 10 % del coste de personal.

+*Impresión:*

El coste de impresión varía considerablemente según si se hacen varias fotocopias a partir del thesaurus impreso en papel por el ordenador, o se hace una tirada de unos miles de ejemplares con un cliché offset a partir de la reproducción del thesaurus en cinta magnética tras un proceso de fotocomposición.

.7 Establecimiento del calendario

La realización de un thesaurus, desde el comienzo de las operaciones hasta la edición número 0 del thesaurus, exige:

-de seis meses a dos o tres años, según
+el tamaño del thesaurus,
+la organización del trabajo;
-de cuatro a doce meses adicionales, en el caso de un thesaurus multilingüe, para el conjunto de las versiones en lenguas no de base tratadas en paralelo.

Los tests y la edición final pueden llevar de tres a seis meses.

Interesa concentrar la construcción del thesaurus en un período lo más breve posible:

-para poder rentabilizar rápidamente la importante inversión que representa su construcción;
-porque, en el lanzamiento de un nuevo sistema documental, o en su modernización, dentro del conjunto de las operaciones (adquisición y puesta en funcionamiento de un programa informático, designación y formación de documentalistas...) la etapa más larga es la realización del

thesaurus: por sí misma constituye el «punto crítico» dentro del planning del nuevo sistema, es decir, lo que condiciona la fecha de lanzamiento;

-para que no se dispersen los esfuerzos del equipo encargado de la construcción y que su motivación se mantenga a un nivel elevado.

## 2. **Mantenimiento de un thesaurus de descriptores**

Un thesaurus debe ser puesto al día regularmente:

-por una parte, para corregir los errores y lagunas detectadas tras su construcción;
-y, por otra parte, para seguir la evolución de
+las ciencias, técnicas y reglamentaciones cubiertas por el thesaurus;
+los dominios cubiertos por el sistema documental, cuando varíen los intereses de sus usuarios.

Esta puesta al día se realiza en dos etapas:

-seguimiento del uso;
-revisión.

### .1 Seguimiento del uso

El seguimiento se ocupa de tres elementos:

-*la frecuencia de uso de los descriptores:* prácticamente todos los programas informáticos de almacenamiento y recuperación documental permiten mantener al día y editar, cuando se desee, la lista de descriptores, junto con su frecuencia de indización (es decir, indicando el número de documentos de la base de datos que han sido indizados por medio de ellos); algunos programas permiten ordenar los descriptores por microdisciplinas.

Algunos programas (BASIS, por ejemplo) permiten así mismo señalar la frecuencia de utilización de los descriptores en el enunciado de las consultas.

Por último, se pueden desarrollar programas «a medida» para editar las listas parciales, ordenadas a veces por microdisciplinas, de:

+únicamente aquellos descriptores cuya frecuencia sobrepase un determinado umbral (por ejemplo: frecuencia superior a 100);
+únicamente aquellos descriptores cuya frecuencia se sitúe por debajo de otro umbral (por ejemplo: frecuencia menor de 2).

La experiencia muestra, en efecto, que la frecuencia de utilización de los descriptores es muy diversa: una gran cantidad de descriptores no se usan, o se usan muy poco; una pequeña cantidad de descriptores se usan con gran frecuencia; a grandes rasgos, y en general, el 20 % de los descriptores más usados representan el 80 % de las indizaciones. Es muy importante conocer esta distribución estadística para fundamentar el mantenimiento del thesaurus;

-*la ausencia de conceptos dentro del thesaurus* que sí aparecen en los documentos: tal caso se puede producir como consecuencia de una laguna de construcción o por la aparición de un nuevo concepto en la literatura.

Algunos sistemas documentales contienen, además del campo «descriptores» (se sobreentiende: extraídos del thesaurus), un campo «posibles descriptores», en el que los indizadores podrán registrar descriptores libres correspondientes a esos conceptos que faltan.

Periódicamente se puede editar la lista de posibles descriptores, acompañados de su frecuencia de indización.

-*las dificultades vinculadas a la utilización del thesaurus* (aparte de la ausencia de determinados conceptos):

+descriptores cuyo sentido es demasiado próximo y ->entre los que se duda frecuentemente cuando se indiza un documento;

->que deben ser incluidos simultáneamente dentro de la formulación de las consultas;

+descriptores cuya acepción es poco clara;

+divergencias entre la diferentes presentaciones del thesaurus;

+relaciones semánticas ausentes, que se echan en falta en el momento de la indización de un documento (lo más frecuente es que se trate de una equivalencia semántica) o cuando se formula una consulta (lo más frecuente es que se trate entonces de una relación jerárquica o asociativa).

Se debe organizar el trabajo de los indizadores de forma que se registren estas dificultades a medida que surgen; desgraciadamente la experiencia muestra que esto es mucho más difícil de conseguir que realizar el seguimiento de la frecuencia de los descriptores (que la lleva a cabo el ordenador) o que añadir los posibles descriptores (¡lo que a menudo resulta, sin embargo, superfluo!).

.2 Puesta al día

Periódicamente (tras seis meses, en el caso de un nuevo thesaurus; después, todos los años o cada dos años) se centralizan y editan los datos registrados a lo largo del seguimiento de uso del thesaurus.

Se puede examinar entonces cada caso separadamente y tomar una decisión:

-descriptor demasiado frecuente:

+mantenerlo tal cual,

+o disgregarlo en dos o más descriptores más específicos, obtenidos bien por precoordinación de palabras simples, o bien incluyendo dentro del thesaurus un nivel de especificidad adicional.

Ejemplos:

1er caso: el thesaurus contiene el descriptor POLITICA FINANCIERA; se añade: POLITICA FINANCIERA DEL ESTADO, POLITICA FINANCIERA DEL BANCO NACIONAL, POLITICA FINANCIERA DE EMPRESAS, etc.

2º caso: el thesaurus contiene el descriptor MANZANA; se añade GOLDEN, REINETA, VERDE DONCELLA, etc.;

-descriptor demasiado poco frecuente:

+mantenerlo tal cual;
+o suprimirlo del thesaurus;
+o «degradarlo» al rango de no-descriptor (lo más frecuente, como no-descriptor de un descriptor más genérico existente);

-posible descriptor:

+no incluirlo, si sólo se ha solicitado una vez y no parece que tenga interés;
+incluirlo como no-descriptor, si parece que enriquecerá el léxico, pero no corresponde a un concepto suficientemente diferenciado del que designa un descriptor existente como para hacerlo nuevo descriptor: es el caso más frecuente;
+incluirlo como descriptor;

-dificultades encontradas:

+añadir o suprimir descriptores o no-descriptores;
+añadir o suprimir relaciones jerárquicas y asociativas;
+añadir o revisar las notas explicativas.

*Notas:*

1) es necesario evitar, en la medida de lo posible, suprimir descriptores que hayan sido ya utilizados para indizar los documentos; si esta supresión se revela absolutamente indispensable (especialmente en el caso de un descriptor demasiado frecuente) será necesario reindizar todos los documentos en los que esos descriptores intervienen, para mantener la coherencia entre el fichero de búsqueda y el thesaurus;

2) si el seguimiento de uso del thesaurus lo han de realizar todos los documentalistas que lo utilizan, la revisión del thesaurus en sí debe ser obligatoriamente centralizada en una sola persona, a menudo designada como «administrador del thesaurus», que asegurará la coherencia; esta persona debería ser preferentemente la que haya tenido la responsabilidad de la construcción del thesaurus;

3) en algunos sistemas documentales se organiza una revisión del thesaurus con una periodicidad muy corta (por ejemplo, todos los meses), realizando notas modificativas y

reeditando, tras un intervalo más largo (por ejemplo, de 2 a 5 años), el thesaurus revisado.

4) se debe adjuntar una nota histórica precisa a los descriptores específicos añadidos al thesaurus, que precise

-la fecha de introducción;
-y, cuando esto no sea evidente, el descriptor utilizado hasta esa fecha para representar el correspondiente concepto.

Ejemplo: POLITICA FINANCIERA DEL ESTADO (desde 1/6/85; antes úsese POLITICA FINANCIERA y ESTADO).

### 3. **Construcción y mantenimiento de una lista de autoridades**

La lista de autoridades se presenta como un conjunto no estructurado de descriptores, generalmente de tamaño restringido.
Los pocos servicios que las usan las han creado:

-o bien durante un proceso de construcción de un thesaurus, pero quedándose en las primeras fases:
+se recoge el lenguaje natural;
+se seleccionan los términos que representan los conceptos más utilizados dentro de los dominios cubiertos por el sistema;
-o bien utilizando una aproximación más empírica:
+se adopta el índice de materias de una o varias obras de referencia;
+un especialista elabora una lista de materias;
+se analiza un sistema de clasificación ya existente.

El mantenimiento está asegurado por medio del seguimiento de la literatura, a través de la indización de los documentos: cuando un nuevo concepto aparece, se elige el descriptor correspondiente y se añade a la lista.

### 4. **Construcción y mantenimiento de una lista de descriptores libres**

En la mayoría de los casos, una lista de descriptores libres procede de la simple acumulación de descriptores asignados libremente por los indizadores a los documentos, a medida que se registran estos últimos: por tanto, aquí no hay «construcción» ni «mantenimiento» sistemáticamente organizados.
En los pocos sistemas documentales en los que a posteriori se trabaja sobre la lista de descriptores libres para conseguir un lenguaje de ayuda a la búsqueda, se realiza un verdadero esfuerzo continuado para estructurar el vocabulario.
Este trabajo se puede organizar en tres fases:

-se examina la lista de descriptores libres utilizados hasta el momento;
-se establecen relaciones semánticas;

-se mecanografían esas relaciones y se edita el lenguaje de ayuda a la búsqueda.

.1 Examen de la lista de descriptores libres

La lista se edita, junto con la frecuencia de indización de los descriptores, en forma de:

-o bien una lista completa;
-o bien listas parciales de descriptores ordenados por grupos en función del o de los códigos de clasificación asignados a los documentos indizados por medio de ellos; esta modalidad es evidentemente la más interesante, porque es más cómodo trabajar con listas parciales, y por tanto menos voluminosas que la lista completa; presenta sin embargo cierto inconvenientes:
+pueden figurar los mismos descriptores dentro de varias listas parciales;
+las variantes semánticas de los mismos conceptos y las variantes ortográficas de los mismos descriptores están dispersas en varias listas parciales.

Este examen tiene por objeto detectar:

-en un primer paso, las variantes ortográficas de los mismos descriptores.
Ejemplos: Alemania, República Federal de Alemania, República Federal alemana, RFA, R.F.A.;

así como las falta de ortografía y de mecanografía.

Ejemplos: Alemania, Alemana, Alemani, Almania...

-en un segundo paso, las equivalencia semánticas entre descriptores.

Ejemplos: autoridad(es) alemana(s), estado alemán...

-y muy excepcionalmente, en el caso en que los recursos humanos disponibles permitan tratar las demás relaciones semánticas: las jerarquías y asociaciones.

Ejemplos: land, gobierno alemán, ejército alemán...

.2 Establecimiento de relaciones semánticas

Todos los descriptores libres que designan el mismo concepto constituyen una clase de equivalencia y están enlazados por una relación de equivalencia semántica.
Esto permitirá que, cuando los usuarios introduzcan una de las formas del concepto dentro de la formulación de su consulta, el sistema extienda automáticamente la consulta a todo el conjunto de la clase de equivalencia.
Si se ha decidido crearlas, las relaciones de jerarquía y de asociación deben entonces ser introducidas entre clases

de equivalencia y no entre descriptores que designen los conceptos de forma unívoca, como en un thesaurus ordinario.

*Notas:*

1) en teoría es posible elegir un descriptor preferente para designar cada concepto y dar el estatuto de no-descriptor a todas las variantes y errores ortográficos y a todos los equivalentes semánticos ciertos. En la práctica no se hace porque sería necesario entonces corregir la indización de todas las referencias registradas hasta el momento;

2) la construcción de relaciones de equivalencia y, si es el caso, de jerarquía y de asociación, es un trabajo mucho más ingente que en los thesaurus clásicos, construidos a priori, en los que la existencia de un lenguaje pre-controlado permite poner límite al crecimiento del vocabulario.

.3 Mecanografiado y edición

El mecanografiado se lleva a cabo como en un thesaurus clásico, pero con muchos menos gastos. Así mismo la edición es mucho más voluminosa.

## 5. **Bibliografía**

.1 Normalización

AFNOR -
*Règles d'établissement des thésaurus monolingues* - Paris: AFNOR, décembre 1981, 20 p., NF Z 47-100

AFNOR -
*Principes directeurs pour l'établissement des thésaurus multilingues* - Paris: AFNOR, avril 1980, 21 p. Z 47-101.

AFNOR -
*Thésaurus monolingues et multilingues, symbolisation des relations* - Paris: AFNOR, avril 1980, Z 47-103.

ISO -
*Principes directeurs pour l'établissement et le développement de thésaurus monolingues* - Genève: ISO, 1974, 14 p., ISO 2788-1974 (F).

ISO -
*Principes directeurs pour l'établissement et le développment de thésaurus multilingues* - Genève: ISO, 1985, 61 p., ISO 5964-1985 (F).

ISO -
*Traitement de l'information, jeux de caractères codés pour la transmission de texte* - Genève: ISO, 1983, 12 + 35 p., ISO 6937/1 et 2.

ISO -
*Codes pour la représentation des noms de pays* - Genève: ISO, 1981, 50 p., ISO 3166-1981 (E/F).

.2 Repertorios de thesaurus

Bibliographic Bulletin of the Clearinghouse at Jinte, Warsaw, Institute for scientific, technical and economic information: publica anualmente un complemento con una lista de thésaurus y sistemas de clasificación.

GID (Gesellschaft für Information und Dokumentation) -
*Thesaurus guide: analytical directory of selected vocabularies for information retrieval* - Amsterdam: North Holland, and Luxembourg: Office for Official Publications of the European Communities, 1985, 749 p., ISBN: 0-444-87736-3/92-825-4897-X.

.3 Bibliografía general sobre la construcción de lenguajes documentales

AITCHISON (J.) and GILCHRIST (A.) -
*Thesaurus construction: a practical manual* - London: Aslib, 1972, 95 p., ISBN: 0-85142-0427.

ERLI -
*Alexis - Manuel d'utilisation* - Paris: ERLI, 1985, 164 p., réf. AL 1.2/MU 85.01.

LAUREILHE (M.T.) -
*Le thésaurus: son rôle, sa structure, son élaboration* - Villeurbanne: Presse de l'ENSB, 1981, 88 P.

MACCAFFERTY (M.) -
*Thesauri and thesaurus construction* - London: Aslib, 1977, 191 p., ISBN: 0-85412-102-4.

SOERGEL (D) -
A universal source thesaurus as a classification generator - *Journal of ASIS*, vol. 23, nº 5, september-october 1972, p. 299-305.

SOERGEL (D) -
*Indexing languages and thesauri: construction and maintenance* - Los Angeles: Melville Publishing Co, 1974, 632 p., ISBN: 0-471-81047-9.

TOWNELY (H.) and GEE (R.D.) -
*Thesaurus making: grown your own word-stock* - London: Deutsch, 1980, 206 p., ISBN: 0-233-97225-0.

VAN SLYPE (G.) -
*Définition des caractéristiques essentielles des thesauri* - Luxembourg: Commission des Communautés européennes, 1976, vol. 1,52 p.; vol. 2, 60 p.

VAN SLYPE (G.) -

*Definition des grandes orientations d'ASTUTE II* - Luxembourg: CCE, 1985, 64 p., BR 20.516.

VERNIMB (C.) -
*The importance of a thesaurus in retrieval* - Luxembourg: Euratom, 1968, 17 p., informe interno.

VICKERY (B.C.) -
*Faceted classification: a guide to the construction and use of special schemes* - London: Aslib, 1960, 70 p., ISBN: 0-85142-010-9.

.4 Thesaurus «universales»

COATES (E.) et al. -
*Broad System of Ordering (BSO), Schedule and index* - The Hague: FID, 1978, 198 p., ISBN: 92-66-00564-9.

COATES (E.) et al. -
*Système général de classement, tables et index* - La Haye: FID, 1981.

Institut Gustave Roussy -
*Le macrothésaurus des sciences et des techniques* - Paris: CILF, 1979.

CAPITULO III

# UTILIZACION DE LOS LENGUAJES DE INDIZACION

La cadena de tratamiento de la información documental consta de doce «puestos de trabajo» o actividades:

- -elección de las fuentes;
- -adquisición de los documentos;
- -selección de los documentos que se van a registrar dentro del sistema documental;
- -catalogación descriptiva, o enumeración de los datos identificativos: título, autor(es), fuente, fecha...;
- -clasificación: se asigna una o varias materias, extraídas de un sistema de clasificación, para caracterizar el tema del documento;
- -indización humana: se asignan descriptores, generalmente extraídos de un thesaurus, para describir el contenido conceptual del documento;
- -condensación: se redacta un resumen;
- -mecanografiado y validación;
- -indización automática, en la mayoría de los casos por medio de palabras clave, para completar la descripción del contenido del documento;
- -actualización de los ficheros de edición;
  - +de los boletines signaléticos y analíticos, estructurados fundamentalmente sobre la base de clasificaciones;
  - +de los boletines de índices, estructurados fundamentalmente sobre la base de la indización humana;
- -actualización de los ficheros de búsqueda, estructurados fundamentalmente sobre la base de las indizaciones humana y automática;
- -búsqueda documental y difusión selectiva a partir de perfiles, que se realizan en los ficheros de búsqueda y en los boletines de índices.

De estas doce etapas, seis están directamente relacionadas con la utilización de un tipo u otro de lenguaje de indización:

- -la indización humana;
- -la validación;
- -la indización automática;
- -la actualización de los ficheros de edición de los boletines de índices;
- -la actualización de los ficheros de búsqueda;
- -la búsqueda documental.

## 1. **La indización humana**

### .1 Definición

La indización humana es la operación que consiste en enumerar los conceptos sobre los que trata un documento y representarlos por medio de un lenguaje combinatorio:

-lista de descriptores libres;
-o lista de autoridades;
-o thesaurus de descriptores.

### .2 Objetivo

Su finalidad es la búsqueda documental, ya se realice ésta a partir de los boletines de índices o a partir de los ficheros de búsqueda.

### .3 Rasgos distintivos

La indización humana es una actividad fundamentada en la apreciación de un ser humano.
Esta apreciación se ejerce en dos planos, el primero de los cuales permite diferenciar la indización humana de la automática, y el segundo permite distinguir la indización humana de la clasificación:

-el plano de las unidades significativas reconocidas: mientras que la indización automática reconoce, ante todo, cadenas de caracteres que constituyen palabras no vacías, la indización humana distingue conceptos, es decir, representaciones mentales de objetos de conocimiento;

-el plano de la selectividad: mientras que la clasificación se sitúa al nivel más sintético, el de la expresión más general del contenido (*el tema* del documento), la indización humana analiza el documento, es decir, se reconocen los elementos constitutivos: *los conceptos*; este análisis se realiza de forma crítica; el indizador:
+selecciona, a partir de los conceptos explícitamente presentes dentro del documento, aquellos sobre los que el documento aporta una información susceptible de interesar a los usuarios del sistema documental;
+busca los conceptos implícitos sobre los que, aunque no estén designados nominalmente en el documento, se aporta información dentro de él.

### .4 Etapas

La indización humana se realiza en cuatro etapas:

-se revisa el contenido del documento;
-se seleccionan los conceptos;
-se traducen los conceptos en descriptores;

-se establecen enlaces sintácticos entre los descriptores.

.1 Toma de contacto

El documentalista lee rápidamente:

-el título;
-el sumario;
-el resumen, si lo hay;
-la introducción general;
-las introducciones y conclusiones de los principales capítulos y/o parágrafos;
-los enunciados de las tablas y figuras;
-las conclusiones generales;

y recorre muy rápidamente, «en diagonal», el texto mismo del documento. El fin perseguido en esta lectura rápida no es «dominar» el contenido del documento, sino solamente ver de qué trata.

.2 Elección de los conceptos

A medida que realiza la lectura, el documentalista identifica los conceptos sobre los que trata el documento.
De entre las nociones explícitamente presentes, no retiene aquellas:

-de las que el autor mismo dice que no va a tratar.
Ejemplo:en esta obra, la catalogación descriptiva;
-que sólo figuran dentro del documento a título de ejemplo.
Ejemplo:en esta obra, la educación: si bien éste es el tema de uno de los thesaurus de las ilustraciones);
-que están presentes en la argumentación del autor dentro del documento, pero sobre los que no se aporta ninguna información suficientemente completa que pueda interesar al usuario.
Ejemplo: en esta obra, la clasificación.

Por el contrario, retiene:

-las nociones explícitamente presentes, sobre las que se aporta información susceptible de responder a una necesidad de un usuario del sistema documental.
Ejemplo: en esta obra: la indización;
-las nociones implícitamente contenidas dentro del documento, que, si bien no son designadas como tales por el autor, son tratadas suficientemente en detalle como para interesar ocasionalmente a los usuarios.
Ejemplo: en esta obra, la informática documental.

.3 Traducción de los conceptos en descriptores

Una vez enumerados los conceptos, su expresión en lenguaje natural debe ser traducida en descriptores.

Esta operación se realiza en varias etapas, la primera de las cuales es independiente del tipo de lenguaje combinatorio utilizado, y el resto sí depende de ese lenguaje:

-lista de descriptores libres;
-lista de autoridades;
-thesaurus de descriptores.

La primera etapa consiste simplemente en que, cuando el documento está en una lengua distinta de la del lenguaje combinatorio usado, se traduce la expresión de los conceptos a la lengua, o a una de las lenguas, del sistema: en este último caso, preferentemente a la lengua del documentalista encargado de la indización.

.1 Indización por medio de una lista de descriptores libres

Esta operación se puede hacer, según los servicios de documentación y, en cierta medida, según los indizadores de un mismo servicio:

-o bien con el mínimo control del vocabulario: el documentalista toma simplemente las designaciones de los conceptos:
+tal y como las encuentra en el documento, para los conceptos explicitados en la lengua de la lista;
+tal y como los ha traducido, para los conceptos explícitos que figuran dentro de un documento en una lengua distinta a la de la lista;
+tal y como los ha enunciado él mismo, para los conceptos implícitos.

No compara estas designaciones con los enunciados que figuran ya dentro de las listas de descriptores libres; sin embargo, procede a normalizarlos:
+transformando las formas verbales y adjetivas en formas nominales (por ejemplo: MEDIDA en vez de: medir; LONGITUD en vez de: largo);
+poniendo en masculino singular las formas en femenino y/o en plural (por ejemplo: DOCTOR en vez de: doctora(s));

-o bien con un control del vocabulario más potente: para esto el documentalista compara las designaciones normalizadas de los conceptos, tal y como los ha elaborado, según acabamos indicar, con la lista existente de descriptores libres; dado que esta lista es puramente alfabética, que no dispone de relaciones semánticas y que no puede servir como instrumento de validación para el ordenador en el momento del mecanografiado, la profundidad de este control dependerá únicamente de la voluntad del documentalista y del tiempo disponible.

Ejemplo: si un documento trata sobre «documentación automatizada» y la lista de descriptores no contiene ese descriptor, sino que incluye la designación «informática documental», perdido entre varias decenas de miles de descriptores, hay una probabilidad muy alta de que el documento sea

indizado por «documentación automatizada» y de que ese término quede añadido desde entonces por el sistema a la lista de descriptores.

.2 Indización por medio de una lista de autoridades

Hay que distinguir tres hipótesis:

-la expresión de un concepto del documento (o de su traducción a la lengua de la lista, tras la normalización léxica), figura efectivamente dentro de la lista de autoridades: hay una probabilidad muy alta de que el documento sea indizado correctamente;
-la expresión de un concepto del documento no figura dentro de la lista, pero se encuentra un término equivalente.
Ejemplo: el documento trata sobre una «intervención sobre metal» y la lista de autoridades incluye el descriptor «tratamiento de metales».

Son posibles tres casos:

+el documentalista conoce bien su lista (lo que sucede a menudo, ya que las listas de autoridades tienen frecuentemente un tamaño limitado a varios cientos de descriptores): encontrará el término adecuado y el documento será correctamente indizado;
+el documentalista conoce mal su lista, o está poco motivado como para explorarla, pero el sistema incluye una «válvula de seguridad», en forma de un campo de «descriptores libres» además del campo «descriptores controlados»: la tendencia natural del documentalista será aprovechar esta posibilidad y anotar la designación del concepto, en lenguaje natural normalizado, dentro del campo «descriptores libres». El contenido de ese campo será periódicamente acumulado para todas las referencias, y se verá cómo, junto a la lista de autoridades, crece una lista de descriptores libres no validados;
+en el tercer caso, la situación se presenta como en el segundo caso citado, pero el sistema no incluye campos para «descriptores libres» (esta situación es bastante frecuente en los sistemas de lista de autoridades): hay una probabilidad muy alta de que el concepto no sea conservado dentro de la indización del documento; por lo cual este último no podrá ser recuperado cuando se realice una consulta sobre ese concepto (salvo si se utilizan las palabras clave procedentes de la indización automática, y en el caso de que esas palabras clave expresen correctamente el concepto no conservado dentro de la indización humana);

-la expresión de un concepto del documento no figura dentro de la lista, y ésta no contiene ningún término

equivalente: se vuelve a producir la situación de alguno de los dos últimos casos de la hipótesis anterior.

*Nota:* en algunos servicios, la actividad del indizador está facilitada por una preselección a partir de la lista de autoridades, completa, para cada listado de descriptores usados en la indización: basta entonces con que el documentalista anote dentro del listado los descriptores que representan el contenido del documento.

.3 Indización por medio de un thesaurus de descriptores

Se presentan tres casos:

-la expresión del concepto corresponde a un descriptor que lo representa: éste se transcribe en el listado de indización;
-la expresión del concepto corresponde a un no-descriptor que lo representa: este último reenvía al descriptor equivalente que se ha de usar, y se vuelve a producir la situación anterior;
-no existe una entrada dentro del thesaurus que corresponda exactamente a la designación del concepto. Se pueden utilizar entonces dos métodos:

+o bien hacer que trabaje su imaginación: evocar otra formulación del concepto en lenguaje natural y ver si corresponde dentro de la lista alfabética del thesaurus a una entrada, descriptor o no-descriptor: se vuelve a producir entonces una de las situaciones de los dos primeros casos;
+o bien definir la clase o clases generales (microdisciplinas) que engloban el concepto que se ha de traducir, consultar la presentación por grupos del thesaurus (listas sectoriales, diagramas de flechas, terminogramas o, con el mayor rigor, listas jerárquicas), revisar rápidamente todos los descriptores que allí se encuentren, o los que compongan la cadena o cadenas jerárquicamente próximas al concepto que se ha de representar, y buscar, entre los descriptores disponibles, el que represente mejor ese concepto:
->ya sea un descriptor específico;
->o bien, a falta de descriptor específico, el descriptor genérico que mejor reproduzca el concepto; en este último caso se mantendrá ese descriptor genérico y, si el sistema lo permite, se añadirá la designación del concepto, en lenguaje natural normalizado, dentro del campo «posibles descriptores».

## .4 Establecimiento de enlaces sintácticos entre los descriptores conservados

El tipo de enlaces sintácticos que se puede establecer entre los descriptores asignados por indización humana depende el sistema de almacenamiento y recuperación documental utilizado; es independiente del tipo de lenguaje de indización, salvo en los sistemas, muy pocos, en los que el uso de la sintaxis exige una división previa de los términos en categorías predeterminadas (ejemplo: el sistema PRECIS).

En los sistemas documentales se pueden utilizar seis tipos de enlaces, de los cuales el primero es, con mucho, el más frecuente:

- -la yuxtaposición;
- -la ponderación;
- -la especificación de un punto de vista;
- -la especificación de un vínculo;
- -la especificación de un rol;
- -la integración dentro de un resumen.

### .1 Yuxtaposición

Se registran los descriptores unos a continuación de otros:

- -o bien en el orden en el que se han encontrado los conceptos correspondientes dentro del documento: es el método más habitual;
- -o bien con un orden significativo, que es el que se utilizará al editar un boletín de índices en papel (cf. § 4.2.3).

En ambos casos los descriptores se separan unos de otros por un «separador», es decir, por un signo de puntuación reconocido como tal por el programa informático; este signo es generalmente único (por ejemplo: el punto y coma), mientras que en la indización automática el sistema podrá distinguir varios signos para separar las cadenas de caracteres unas de otras y reconocer así las palabras clave: el espacio en blanco, los puntos, las comas, los punto y coma, etc.

Ejemplo: considérese un documento que trata sobre la epidemiología de la varicela en el adolescente con relación a los deportes que éste practica y a la ocupación profesional de sus progenitores.

En simple yuxtaposición, su cadena de indización se presentará como sigue:

EPIDEMIOLOGIA; VARICELA; ADOLESCENTE; DEPORTE; OCUPACION PROFESIONAL; PROGENITOR;

### .2 Ponderación

En algunos sistemas documentales se requiere a los indizadores que distingan dos grupos de descriptores para cada documento:

-los descriptores principales, que caracterizan el contenido principal del documento;
-los descriptores secundarios, que representan los conceptos sobre los que el documento aporta menos información.

Este método es muy útil para poder «calibrar» la estrategia de la búsqueda documental en función de las necesidades del usuario:

-si éste solicita una bibliografía muy precisa de documentos que traten a fondo el tema que le incumbe, la búsqueda sólo se deberá realizar sobre los descriptores principales;
-si, por el contrario, solicita una bibliografía exhaustiva, que englobe incluso los documentos que sólo tratan de forma accesoria su tema de interés, la búsqueda se realizará sobre la totalidad de los descriptores, principales y sencudarios.

Conviene hacer notar que la elección de la categoría (principal o secundaria de los descriptores) depende de su peso dentro del documento, y no de una clasificación a priori dentro del thesaurus en dos listas de descriptores: principales o secundarios. Dicho de otra forma, el sistema utiliza un solo thesaurus, o una sola lista de descriptores, libres o controlados, y cada descriptor, según su importancia dentro de un documento, podrá ser registrado en el campo «descriptor principal» o en el campo «descriptor secundario» de ese documento. Un mismo descriptor será primario o secundario según el documento.

Ejemplo:

-descriptores primarios: EPIDEMIOLOGIA; VARICELA; ADOLESCENTE;
-descriptores secundarios: DEPORTE; OCUPACION PROFESIONAL; PROGENITOR.

Esta ponderación puede ser más refinada todavía, y algunos sistemas permiten asignar a cada descriptor un peso, o coeficiente de ponderación que varía, por ejemplo, de 1 a 3 o incluso a 4, destinado a marcar la importancia de cada concepto dentro del documento. Cuanto mayor es el refinamiento de la ponderación, la subjetividad del indizador va a desempeñar un papel evidentemente mayor; éste es el motivo por el que se recomienda poco ir más allá de estos dos niveles de peso: principal y secundario.

.3 Especificación de un punto de vista

Algunos sistemas permiten acoplar dos descriptores, o un descriptor del thesaurus y un «aspecto», extraído de una pequeña lista de autoridades, o un descriptor auxiliar, que especifica desde qué punto de vista está considerado el descriptor dentro del documento.
El indizador enlazará el aspecto, por un signo de puntuación particular (por ejemplo, los dos puntos), sólo al descriptor que corresponda

Este método se utiliza especialmente en el sistema MEDLARS (documentación médica de la National Library of Medicine, en USA); en él se encuentra una lista de aspectos que precisan el punto de vista bajo el que está considerado cualquier descriptor que designa una enfermedad: diagnóstico, terapéutico, epidemiológico..., así como otra lista de aspectos que permiten precisar el ángulo bajo el está considerado un medicamento: indicación, posología, efecto secundario...

Ejemplo: antineurálgico : efecto secundario; reúma.

Tal segmentación permite evitar que se suministre este documento en respuesta a una consulta sobre los efectos secundarios del reúma. Este método se utiliza bastante en los sistemas de información médica y farmacéutica.

.4 Especificación de un vínculo

Algunos documentos tratan sobre varios temas distintos, y cada tema queda representado por la coordinación de dos o más descriptores. Para evitar, en el momento de la búsqueda, falsas coordinaciones y, por tanto, ruido, algunos programas informáticos (por ejemplo: BASIS, de Batelle) permiten establecer tantos grupos de descriptores como temas tratados, enlazando todos los descriptores de cada uno de los grupos por un índice, esto es, un número de vínculo diferente de un grupo a otro. Cuando se realiza la búsqueda documental, se podrá, dentro de la formulación de la consulta:

-o bien declarar que los descriptores de los documentos contenidos en la consulta deben tener el mismo índice de vínculo (téngase en cuenta que el valor preciso de este índice no nos es conocido y no debe ser declarado; basta con que ese valor sea el mismo para todos los descriptores de la consulta);

-o bien no tener en cuenta los vínculos y combinar los descriptores sin prestar atención a su número de vínculo.

Este método consiste, de hecho, en crear dentro del registro del documento tantos sub-registros como cadenas de descriptores caracterizadas por el indizador por medio de un mismo índice.

Ejemplos:

EPIDEMIOLOGIA (1); VARICELA (1); ADOLESCENTE (1); DEPORTE (1); OCUPACION PROFESIONAL (2); PROGENITOR (2).

Indizado de esta manera, este documento no será considerado pertinente por el sistema como respuesta a una consulta con vínculo que trate sobre la influencia de las ocupaciones profesionales de los adolescentes sobre la varicela; por el contrario, será suministrado como respuesta a una consulta con vínculo sobre la influencia del deporte sobre la varicela de los adolescentes; y también como respuesta a una consulta sin vínculo sobre la influencia de las ocupaciones profesionales de los progenitores sobre la varicela de los adolescentes.

.5 Especificación de un rol

En algunos casos pueden establecerse falsas coordinaciones entre conceptos que intervienen dentro de los mismos documentos, pero desempeñando en ellos una función diferente. Algunos programas informáticos (por ejemplo: GOLEM, de Siemens) permiten evitar esos ruidos especificando el rol que desempeñan tales conceptos dentro de cada documento. Al descriptor correspondiente se le añade un indicador de rol, escogido de una pequeña lista de posibles roles y expresado bajo la forma de un código. Así resulta posible limitar una búsqueda, especificando el rol que ha de corresponder a un descriptor dentro de un documento para que tal documento responda a lo que se requiere, o no tomar en consideración los roles.

Este método consiste, de hecho, en crear sub-registros no por documento, como en la técnica de los vínculos, sino por descriptor.

Ejemplo: considérese un documento indizado

IMPORTACION; ENDIVIA; FRANCIA (D); BELGICA (O);

Los dos roles definidos son O para origen y D para destino.

Indizado de esta manera, este documento no será considerado pertinente por el sistema como respuesta a una consulta sobre el movimiento de endivias entre Bélgica, considerada como importadora, y Francia, considerada como exportadora. Si se plantea la misma consulta, pero sin interesarse por la dirección del movimiento comercial y por tanto sin precisar los roles dentro de la consulta, el documento será considerado pertinente.

Este método se usa mucho dentro de los sistemas documentales para el dominio de la industria química, en la que el mismo producto puede intervenir como materia prima, como producto semi-terminado o terminado, o como catalizador.

Desde el punto de vista conceptual, se trata de un método muy próximo al que consiste en especificar los puntos de vista (§ 1.4.4.3); se diferencia fundamentalmente por la cantidad de especificadores: muy reducida en el caso de los roles (menos de diez), y más numerosa en el caso de los puntos de vista (varias decenas).

.6 Integración dentro de un enunciado estructurado

Algunos sistemas documentales están concebidos de manera que integran, dentro de un campo único, la indización y la condensación. En tales sistemas el documentalista debe, tras haber elegido una serie de descriptores para expresar el contenido conceptual de un documento, redactar un texto estructurado, dentro del cual todas las palabras y expresiones significativas -o parte de ellas- sean esos descriptores.

Existen tres variantes, según si el texto estructurado es:

-un resumen redactado con una sintaxis libre, a gusto del documentalista, con los descriptores marcados por signos especiales (por ejemplo: comillas), para que sean reconocidos

como tales por el sistema y ser tenidos en cuenta en el fichero de búsqueda.

La documentación de la OIT (Oficina Internacional del Trabajo) se organiza de acuerdo con este principio;

-una fórmula de indización redactada con una sintaxis controlada (es decir, en la que todas las formas gramaticales posibles están definidas a priori) y muy limitada; esta sintaxis se utiliza, sobre todo, para producir las entradas de un boletín de índices y, secundariamente, para obtener una traducción automática de esas entradas a una o a varias lenguas. El sistema PRECIS, utilizado por la British National Bibliography, se basa en este principio (cf. § 4.2.4).

-un resumen redactado con una sintaxis controlada muy compleja, utilizada para producir una «estructura canónica de frase» y para permitir una traducción automática del resumen de muy buena calidad. El sistema TITUS del Institut Textile de France funciona según este modelo.

Tal tipo de indización proporciona resultados similares a los producidos por una indización con vínculo; aquí el vínculo está implícito: se sobreentiende por el hecho de que dos o más descriptores pertenezcan a una misma frase dentro del resumen.

.5 Características de la indización humana

Nosotros distinguimos:

-la profundidad de la indización;
-el tiempo de indización;
-la coherencia de trabajo de los indizadores;
-las características cualitativas: exhaustividad y especificidad.

.1 Profundidad de la indización

En la mayor parte de los sistemas documentales actuales, la profundidad de la indización humana se sitúa en un intervalo máximo de 6 a 30 descriptores; la inmensa mayoría de los documentos se indizan por medio de 8-12 descriptores como media.

.2 Tiempo de indización

La indización humana requiere de 5 a 15 minutos de trabajo, según el tamaño del texto que se va ha de indizar, su complejidad, la profundidad de la indización y lo familiarizado que esté el indizador con el tema y la lengua del documento y con el thesaurus.

.3 Coherencia de la indización

La coherencia de la indización de un mismo documento por dos documentalistas, es decir, la ratio entre el número de descriptores comunes y el número total de descriptores, comunes o diferentes, va del 50 al 80 %, según la calidad del

manual de indización, la formación recibida por los documentalistas y la meticulosidad con que realicen su trabajo.

La coherencia de la indización se mide de la manera siguiente:

-dos documentalistas (o dos equipos de documentalistas) indizan el mismo documento (o el mismo conjunto de documentos) por medio de un mismo thesaurus, trabajando independientemente uno de otro;
-se cuenta separadamente, para cada documentalista (o equipo):

+por una parte: el número de descriptores idénticos utilizados por los dos documentalistas;
+por otra parte: el número total de descriptores, idénticos o distintos, utilizados por los dos documentalistas;
+la tasa de coherencia es la ratio entre estos dos números.

Ejemplo:

El documentalista 1 ha asignado los descriptores: A B C D E F
El documentalista 2 ha asignado los descriptores: A C D F G H
Hay 4 descriptores idénticos (A-C-D-F),
y un total de 8 descriptores (A-B-C-D-E-F-G-H).
Tasa de coherencia = 4/8 = 50 %.

.4 Cualidades de la indización

-La exhaustividad mide la calidad en la elección de los conceptos realmente significativos, es decir, que contienen información pertinente para los usuarios:

+una exhaustividad demasiado reducida hará que no se recuperen documentos pertinentes, y, por tanto, que disminuya la tasa de llamada y que aumenten los silencios;
+una exhaustividad demasiado elevada hará que se recuperen documentos que no contengan información pertinente sobre los conceptos de la consulta; por tanto, hace que disminuya la precisión y aumenten los ruidos.

La exhaustividad depende fundamentalmente de:

+la política de indización;
+la calidad del trabajo de los documentalistas, y especialmente de su capacidad de juzgar lo que es importante y lo que no lo es, y su «olfato» para detectar los conceptos implícitos.

-La especificidad mide la calidad en la elección de los descriptores que corresponden efectivamente a los conceptos incluidos dentro del documento; distinguimos:

+la especificidad vertical: el descriptor debe situarse en el mismo nivel de especificidad que el concepto o, por defecto, en el nivel jerárquico inmediatamente superior existente en el thesaurus.

Ejemplo: el concepto «micro-ordenador» se ha de indizar por MICRO-ORDENADOR, en tanto que este descriptor existe dentro del thesaurus, y no por EQUIPO INFORMATICO; si MICRO-ORDENADOR no existe dentro del thesaurus, y sí existe ORDENADOR, se utilizará este último descriptor antes que el descriptor EQUIPO INFORMATICO; por último, si el thesaurus incluye EQUIPO INFORMATICO, y no MICRO-ORDENADOR ni ORDENADOR, se empleará EQUIPO INFORMATICO.

Una buena especificidad vertical hace aumentar a la vez la precisión y la tasa de llamada;

+la especificidad horizontal: un concepto compuesto debe ser traducido por un descriptor precoordinado, si existe, antes que por la asociación de descriptores simples.

Ejemplo: el concepto «cultivos de huerta» será indizado por el descriptor HORTICULTURA, si existe, antes que por los descriptores CULTIVO y HUERTA.

Una buena especificidad horizontal hace disminuir el riesgo de falsas coordinaciones y, por tanto, que aumente la precisión.

La especificidad, tanto la horizontal como la vertical, depende de:

+la riqueza del thesaurus: evidentemente sólo se pueden indizar los conceptos presentes dentro del lenguaje documental;
+la calidad del trabajo de los documentalistas, y especialmente de su minuciosidad.

## 2. **Validación de la indización**

La validación de la indización se puede hacer con referencia a una lista pre-registrada. Esto significa que tal validación sólo es posible para los términos que aparecen en los campos de «descriptores controlados» (controlados por un thesaurus o una lista de autoridades); esta validación no se puede hacer en los campos de «descriptores libres» o de «posibles descriptores», ni, a fortiori, en el campo de «palabras clave».

La validación de la indización la realiza el programa informático de almacenamiento y recuperación documental, tras haberse mecanografiado la noticia bibliográfica, comparando los descriptores asignados a los documentos con los descriptores que existen dentro del thesaurus o dentro de la lista de autoridades pre-registrada. El sistema emite un mensaje de error cada vez que no encuentra dentro de la lista controlada un término existente dentro de una noticia.

El error puede provenir:

-o bien de un fallo de copia en el momento de la indización o del mecanografiado de la noticia;
-o bien de que el documentalista conoce deficientemente el thesaurus y ha utilizado un descriptor que no existe.

Algunos sistemas rechazan sistemáticamente todos los términos no reconocidos, es decir, los que no figuran dentro del thesaurus; otros permiten que el documentalista decida la acción que se ha de realizar:

-se rechaza;
-o el documentalista introduce correcciones;
-o se fuerza que un término se integre dentro de un fichero de descriptores libres;
-o que el término se integre dentro del thesaurus o dentro de la lista de autoridades y quede aceptado en el fichero de búsqueda.

Esta última posibilidad, aunque técnicamente posible en algunos programas informáticos, debería prohibirse: en efecto, un thesaurus no ha de poder ser modificado en cualquier momento, en caliente, al registrar un documento, con el consiguiente riesgo de no tomar las suficientes precauciones.

Algunos sistemas permiten registrar no sólo los descriptores, sino también los no-descriptores, que se traducen automáticamente en los descriptores equivalentes.

Por último, algunos sistemas, pocos, corrigen automáticamente los errores ortográficos:

-o bien recurren a incluir sistemáticamente, a priori, dentro del thesaurus los errores encontrados más frecuentemente, en forma de no-descriptores, previendo su conversión automática en el correspondiente descriptor correcto, gracias a una equivalencia semántica;
-o bien lo hacen por medio de un algoritmo que calcula la proximidad ortográfica entre los términos no reconocidos en la indización de una noticia y los descriptores del thesaurus (ejemplo: ALEXIS).

Considérese, por ejemplo:

+un extracto alfabético de un thesaurus

->ABASTO
->ABSTRACCION
->ABSENTISMO
->ACCIDENTE
->ACELERACION

+una noticia en la que figura, por error, el término ABSCENTISMO. El sistema, al no encontrar el término «abscentismo» dentro del thesaurus, va a calcular la ratio entre, por una parte, el número de caracteres idénticos del término erróneo y cada uno de los descriptores cuyas dos primeras letras sean AB

y, por otra parte, el número total de caracteres del término erróneo; elegirá automáticamente cómo descriptor aquel cuya ratio sea más elevada:

| | A B A S T O | A B S T R A C C I O N | A B S E N T I S M O |
|---|---|---|---|
| A | 1 0 0 0 0 0 | 1 | 1 |
| B | 0 1 0 0 0 0 | 1 | 1 |
| S | 0 0 0 1 0 0 | 1 | 1 |
| C | 0 0 0 0 0 0 | 1 | |
| E | 0 0 0 0 0 0 | | 1 |
| N | 0 0 0 0 0 0 | | 1 |
| T | 0 0 0 0 1 0 | | 1 |
| I | 0 0 0 0 0 0 | 1 | 1 |
| S | 0 0 0 1 0 0 | | 1 |
| M | 0 0 0 0 0 0 | | 1 |
| O | 0 0 0 0 0 1 | 1 | 1 |
| RATIOS: | | | |
| | 10/12 = 0,83 | 5/12 = 0,42 | 6/12 = 0,50 |

## 3. **La indización automática**

La indización automática es la operación que consiste en que el ordenador reconoce los términos que figuran dentro del título, del resumen, del texto completo (si éste ha sido almacenado junto con la descripción documental) y a veces también dentro de la indización humana, y emplea estos términos, o bien tal cual, o bien después de transformarlos en otros términos, equivalentes o conceptualmente próximos, con el fin de convertirlos en elementos que se incorporan al fichero de búsqueda y quedan disponibles para recuperar el documento.

Existe en el mercado una gran variedad de sistemas de indización automática o semi-automática:

- -enriquecimiento automático de la indización humana por autoreenvío genérico;
- -indización automática no selectiva, es decir, teniendo en cuenta todas las palabras no vacías del documento;
  - + sin normalización del vocabulario natural,
  - + o con normalización del lenguaje natural;
- -indización automática selectiva, es decir, teniendo en cuenta solamente algunos términos, seleccionados por el algoritmo del sistema como los más representativos del contenido del documento:
  + en lenguaje natural,
  + o en lenguaje controlado;
- -indización asistida por ordenador.

Es importante señalar que:

-solamente los dos primeros métodos (enriquecimiento de la indización e indización totalmente automática, no selectiva, sin normalización de vocabulario) son utilizados de forma generalizada por la gran mayoría de los programas informáticos de almacenamiento y recuperación documental existentes en el mercado;
-los otros métodos son todavía objeto de numerosos trabajos de investigación y desarrollo; no obstante, cada uno de ellos está representado en el mercado por al menos un productor, y algunos desde hace más de diez años. A pesar de esto, es necesario señalar que no tienen todavía una verdadera penetración en el mercado, que la gran mayoría de sus aplicaciones son puramente experimentales y que muchas experiencias se han finalizado sin resultados.

Esta situación tiene su origen en la naturaleza esencialmente no algorítmica de los fenómenos lingüísticos: para indizar adecuadamente un texto es necesario empezar por comprenderlo; ahora bien, los ordenadores actuales no han sido construidos para comprender datos tan poco formalizados como lo son los textos.

No es aceptable compararlos con los sistemas de traducción automática, los cuales sí se van abriendo paso en el mercado: la traducción trata de trasladar un texto completo a la lengua destino; si una parte de este texto está mal traducido, la redundancia del lenguaje natural permite, sin embargo, comprender, casi totalmente, una traducción automática.

La indización, por el contrario, consiste en una condensación muy intensa del texto fuente: el contenido de un artículo de cinco páginas pasa a ser representado por sólo una decena de descriptores, como término medio; para llegar a este resultado es necesario poner en práctica facultades de apreciación y de valoración que sobrepasan la capacidad de los ordenadores de la generación actual.

.1 Enriquecimiento de la indización por autoreenvío

Un cierto número de programas de almacenamiento y recuperación documental disponen de una opción de autoreenvío genérico. Si se hace uso de esta opción, cuando se empieza a usar una base de datos, todos los descriptores que asigna el indizador se verán automáticamente completados por la máquina gracias a otros descriptores que están vinculados con éstos en el thesaurus por medio de una relación jerárquica ascendente.

Por ejemplo, a un documento indizado humanamente:
PRUEBA NO DESTRUCTIVA - TRACCION - ACERO AUSTENITICO

se le asignarán por autoreenvío los descriptores adicionales siguientes:

PRUEBA DE CONTROL - COMPROBACION MECANICA - ACERO REFRACTARIO

gracias al conjunto de relaciones jerárquicas de un thesaurus, del que aquí se muestra un extracto:

```
ACERO REFRACTARIO
     ACERO AUSTENITICO
     ACERO FERRITICO
PRUEBA DE CONTROL
     PRUEBA DESTRUCTIVA
     PRUEBA NO DESTRUCTIVA
COMPROBACION MECANICA
     COMPRESION
     FLEXION
     TORSION
     TRACCION
```

Este autoreenvío corresponde, dentro de los thesaurus de descriptores, al mecanismo que se encuentra automáticamente incorporado en las clasificaciones: cuando se asigna a un documento un código de clasificación específico es posible recuperar el documento por medio de una consulta planteada a ese nivel específico o a cualquier nivel genérico en la jerarquía ascendente a partir de ese nivel específico.

En algunos casos, el autoreenvío resulta útil: si, por ejemplo, se realiza una consulta acerca de las pruebas de control efectuadas sobre aceros refractarios, el documento del ejemplo siguiente es verdaderamente pertinente, y se recupera fácilmente gracias al autoreenvío genérico realizado en el momento de su inclusión en el fichero de búsqueda. En muchos casos, por el contrario, el autoreenvío tendrá como efecto ahogar al usuario con una masa ingente de documentos que tratan, sí, sobre el tema de su consulta, pero a un nivel demasiado específico; por ejemplo, en el caso de una consulta acerca de las pruebas de control realizadas sobre comprobaciones mecánicas, el sistema responderá proporcionando:

-por una parte, algunos documentos que fueron indizados humanamente: PRUEBA DE CONTROL y COMPROBACION MECANICA, y que corresponden exactamente al nivel de generalidad solicitado;
-por otra parete, un gran número de documentos indizados humanamente:

```
+ o bien PRUEBA DESTRUCTIVA
       y COMPRESION
       o FLEXION
         etc...
+ o bien PRUEBA NO DESTRUCTIVA
       y COMPRESION
       o FLEXION
         etc...
```

y que aportan informaciones a un nivel que corre el riesgo de ser demasiado específico.

El inconveniente fundamental del autoreenvío es que fija la indización genérica e impide al usuario decidir por sí mismo si han de recuperarse:

-o los documentos genéricos;

-o los documentos específicos;
-o ambos.

en respuesta a una consulta genérica.

No recomendamos tampoco el uso de esta opción. Los programas, en efecto, permiten a los usuarios ejercer su libertad de elección en el momento de la búsqueda; éstos pueden recuperar:

-o bien los documentos genéricos, haciendo su consulta a nivel genérico.

Ejemplo: PRUEBA DE CONTROL y COMPROBACION MECANICA.

-o bien los documentos específicos, haciendo la consulta a nivel específico.

Ejemplo: PRUEBA NO DESTRUCTIVA y TRACCION

-o bien documentos genéricos para uno o más conceptos de la consulta, y específicos para otros, enunciando cada descriptor al nivel adecuado.

Ejemplo: PRUEBA DE CONTROL y TRACCION

-o bien todos los documentos susceptibles de responder a una consulta, tanto si son genéricos como específicos:
+haciendo la consulta a nivel genérico;
+e introduciendo una instrucción puntual (ya que se aplica solamente a una consulta o a un grupo de consultas, y no a todas) de extender la consulta a uno o a x niveles específicos.

Ejemplo: PRUEBA DE CONTROL y COMPROBACION MECANICA con extensión va a generar:
(PRUEBA DE CONTROL o PRUEBA DESTRUCTIVA o PRUEBA NO DESTRUCTIVA) y (COMPROBACION MECANICA o COMPRESION o FLEXION o TORSION o TRACCION).

Como se ve, este método ofrece al usuario cuatro grados de libertad (cuatro posibilidades de elección), mientras que el autoreenvío genérico no le proporciona más que uno (solamente una alternativa impuesta, que corresponde a la cuarta de las soluciones que son posibles si no se practica el autoreenvío genérico en el momento de la carga de la base de datos).

Otra cosa distinta es el autoreenvío marcando con un signo distintivo los descriptores genéricos autoreenviados, lo que permite diferenciar los descriptores genéricos asignados:

-o bien por el indizador humano (ejemplo: PRUEBA DE CONTROL);
-o bien por la máquina (ejemplo: TODAS LAS PRUEBAS DE CONTROL o PRUEBA DE CONTROL +).

En este caso la libertad del usuario en el momento de la búsqueda documental está asegurada: si el usuario desea

recuperar documentos que traten sobre un concepto genérico en un solo nivel jerárquico, utilizará el descriptor no marcado (ejemplo: PRUEBA DE CONTROL) y no obtendrá como respuesta más que los documentos indizados humanamente: PRUEBA DE CONTROL; si, por el contrario, desea recuperar documentos que traten sobre un concepto genérico (ejemplo: comprobación mecánica) y sobre todos los conceptos específicos que tiene asociados, utilizará únicamente el descriptor marcado (ejemplo: COMPROBACION MECANICA +), que le permitirá encontrar los documentos indizados humanamente: COMPRESION, FLEXION, TORSION o TRACCION.

.2 Indización automática, no selectiva, en lenguaje libre no normalizado

Es el método de la lista de palabras clave, utilizado prácticamente por todos los programas de almacenamiento y recuperación documental:

- -se almacena a priori en el sistema una lista de palabras vacías, o antidiccionario, con varios cientos de artículos, pronombres, conjunciones, adverbios..., así como una lista de separadores: espacio en blanco, apóstrofe, comillas, paréntesis, punto, coma, punto y coma, signo de exclamación, de interrogación...;
- -cada documento almacenado será objeto de un análisis automático para reconocer las palabras: se considera palabra toda cadena de caracteres flanqueada por uno, dos o más separadores;
- -cada una de las palabras reconocidas es comparada con la lista de palabras vacías;
- -las palabras del texto que no figuran en la lista de palabras vacías se utilizan para actualizar la lista de palabras clave:

+si estaban ya en la lista de palabras clave, su frecuencia dentro de esta lista se incrementa según el número de veces que aparecen en el texto;

+si no figuran todavía, serán añadidas con una frecuencia inicial igual al número de veces que aparecen en el texto;

+en ambos casos se añade un puntero en el fichero inverso, que remite al documento o a su referencia bibliográfica en el fichero bibliográfico (cf. organización de los ficheros de búsqueda en el § 5).

En resumen, la lista de palabras clave contiene:

- -únicamente palabras (también llamadas unitérminos), simples (ejemplo: psicología) o compuestas (ejemplo: psicopedagogía), pero no expresiones (ejemplo: psicología del adulto);
- -todas las formas derivadas que pertenecen a:

+categorías gramaticales (nombres, verbos, adjetivos) diferentes; ejemplo: documento, documentar, documental;

+la misma categoría gramatical; ejemplos: documento, documentación, documentalista;

-todos las formas flexionadas de estas palabras: conjugaciones, declinaciones, géneros, números; ejemplos: documentamos, documentales, documentos...;
-todas las variantes y errores ortográficos; ejemplos: docmento, documeto...

Pero, por otra parte, un cierto número de términos significativos no aparecerán en la lista:

-ciertos nombres que son homógrafos de palabras vacías; ejemplo: sobre (nombre común o preposición);
-siglas en las que cada letra está seguida de un punto; ejemplo: O.N.U. (como el punto es interpretado por el sistema como separador, la lista de palabras clave contendrá para este término las letras O, N y U, muy alejadas la una de la otra en el orden alfabético, y no la palabra ONU).

Este método es muy sencillo de poner en funcionamiento para almacenar los documentos en la base de datos. Sin embargo, en el momento de la búsqueda documental, ocasionará grandes problemas al usuario, debido principalmente a:

-la no-selectividad de este tipo de indización;
-la existencia de sinónimos y palabras polisémicas;
-la no-normalización de las formas gramaticales;
-la desmembración de expresiones que representan un solo concepto en varias palabras no significativas o poco significativas.

El lenguaje de instrucciones de los programas documentales permite, en cierta medida, ayudar al usuario a resolver una parte de estos problemas (cf. § 6.4); podemos citar sobre todo:

-el truncamiento, que facilita la búsqueda de términos con las mismas raíces, y de diferentes formas derivadas y flexionadas;
-la proximidad (adyacencia, misma frase...), que resuelve en parte el problema de las expresiones.

.3 Indización automática, no selectiva, en lenguaje libre normalizado

Este método respeta la riqueza de la información textual del documento, pero la canaliza un poco eliminando, según cual sea el sistema:

-las variantes y errores ortográficos;
-las formas flexionadas;
-las formas derivadas, cuando éstas pertenecen a muchas categorías gramaticales.

Ejemplo: entre las diversas formas procedentes de la raíz «document» y enumeradas en el anterior apartado, este método eliminará:

-docmento, documeto...

-documentamos, documentales, documentos...
-documentar, documental...

y conservará:

-documento, documentación, documentalista...

Algunos sistemas, pocos, permiten también reconocer las expresiones (ejemplo: servicio de documentación), o, por el contrario, descomponer las palabras compuestas (ejemplo: psicopedagogía, que podrá ser recuperada como respuesta a una búsqueda sobre la pedagogía).
Este método permite:

-limitar notablemente el tamaño de los ficheros de búsqueda;
-simplificar el trabajo del usuario, pues ya no tendrá que introducir todas las formas derivadas de las mismas raíces cuando formule una consulta, ni operar con truncamientos, pero tendrá todavía que pensarse todas las equivalencias semánticas y lingüísticas, operar con la proximidad de las palabras para reconstruir las expresiones que representan ciertos conceptos y eliminar los ruidos causados por la polisemia de los términos y por la no-selectividad de la indización.

A continuación describimos el sistema PASSAT, basado en el uso de un léxico completo de una lengua natural y una lista de sufijos de género, número, declinación y conjugación.
PASSAT (Programm zur automatischen Selektion von Stichwörterm aus Texten) es el módulo de indización automática de Golem (Grosspeicherorientierte, listenorganisierte Ermittlungsmethode), el programa de almacenamiento y recuperación documental de SIEMENS.
PASSAT esta formado por:

-dos ficheros de base:
  +un fichero «Vergleichswörternlist», o fichero diccionario;
  +un fichero «Endunngs- und Bindungslist», o fichero de desinencias y enlaces;
-los programas de gestión y utilización de estos ficheros.

Los diccionarios pueden estar en letra mayúscula y en minúscula, pero sin signos diacríticos (acentos, diéresis, cedilla...). La explotación se lleva ha cabo como si todo estuviera almacenado en tipografía pobre (sólo mayúsculas).

.1 Ficheros

+*Fuentes*

Los ficheros de base son los propios de una lengua dada. Podrían ser elaborados y suministrados por SIEMENS, y que cada usuario los ajustara con su vocabulario específico.

En realidad, SIEMENS sólo ha llevado a cabo esto de manera parcial para el alemán (28.000 entradas en el fichero diccionario, lo cual no representa más que una pequeña parte de un diccionario corriente de la lengua, que tiene como mínimo de 40.000 a 50.000 entradas), para el inglés (7.500 entradas) y para el holandés, en menor medida.

Cada cliente usuario de PASSAT debe, por tanto, elaborar su propio fichero diccionario.

+*Fichero diccionario*

Un registro para cada palabra o expresión del lenguaje natural.
Cada registro contiene los siguientes datos:

-radical de palabra (Stammwort), es decir:
+la parte más larga, común a todas las formas gramaticales (flexiones) de una misma palabra (ejemplo: internac, inver);
+o la expresión compuesta en singular (ejemplo: «estrella de mar»);

-estatuto, representado por un código:
+D: para una palabra que ha de ser conservada como descriptor cuando se encuentre en un texto (ejemplo: justicia);
+N: para una palabra que no ha de ser conservada como descriptor cuando se encuentre en un texto (palabra vacía) (ejemplo: su).

-datos de indización complementaria (sustitución):
+E (=Ersatz) significa que, si la palabra está presente en el texto, aunque deba ser conservada como descriptor (D), PASSAT deberá reemplazarla obligatoriamente por otra palabra (cuyo enunciado figura al final del registro);
+V (=Verweiss): significa que la palabra, si está presente en el texto y es conservada como descriptor (D), deberá:
->por una parte, ser conservada por PASSAT;
->por otra parte, ser completada con otras palabras, de 1 a 50, cuya lista figura al final del registro (autoreenvío);
+K (=Kein): la palabra, si está presente en el texto y ha de ser conservada como descriptor (D), será tomada en su forma canónica (proporcionada por el fichero de desinencias) por PASSAT, sin ser sustituida ni complementada;

-naturaleza de la palabra: código de dos cifras que designa:
+o bien la parte de la oración a la que corresponde esa palabra, y su género.

Ejemplos:
01 adjetivo, pronombre (ej.: alto)
02 nombre propio (ej.: Euratom)
03 numeral (ej.: tres)
04 nombre trivial, preposición, artículo...
05 sigla (ej.: MOPU)
06 verbo (ej.: preparar)

07 sustantivo masculino (ej.: uso)
08 sustantivo femenino (ej.: urgencia).

*Nota:* PASSAT, sin embargo, no realiza análisis sintáctico

+o bien un código que indica que la «palabra» que encabeza el registro es una expresión compuesta, o que corresponde a un plural alemán con diéresis (XX); en estos casos el código es ciertamente explotado por PASSAT para las necesidades de la indización;
-número de orden (2 cifras hexadecimales, es decir, 256 posibilidades) de una de las listas del fichero de desinencias, en el que se encontrarán todas las desinencias posibles de la palabra;
-número de orden (2 cifras hexadecimales, es decir, 256 posibilidades) de uno de los registros de los ficheros de enlaces;
-eventualmente (si sustitución E): el enunciado del término preferido (todos los datos útiles sobre éste se encontrarán en el registro reservado a tal palabra);
-eventualmente (si sustitución V): el enunciado de los términos que se utilizarán para complementar la palabra que encabeza el registro;
-matriz de asociación (sólo cuando se use el módulo opcional de PASSAT, que convierte los términos de indización controlados en descriptores de un thesaurus: cf. § 3.5): de cero a treinta pares de coordenadas de 4 posiciones hexadecimales cada una.

+*Fichero de desinencias y enlaces*

->Desinencias

Un registro por cada lista de desinencias, con un máximo de 256 listas.
Cada registro contiene:

-su número de orden (2 cifras hexadecimales)
-el enunciado del radical de una palabra modelo.

Ejemplo: +capata
+carg
+trabajador

-la lista de desinencias susceptibles de ser añadidas a los radicales referenciados por el número de orden de la lista de desinencias en el fichero diccionario; la primera de las desinencias de la lista es aquella que deberá ser considerada por PASSAT para formar el descriptor que se ha de conservar para la indización (forma canónica).

Ejemplo: capata
z
ces
el descriptor será: capataz
carg

ar
o
as
a
amos
áis
an
aba
etc.

el descriptor será: cargar.

Cuando el enunciado de la palabra diccionario coincide con su propia forma canónica, la primera desinencia se indica con «-» en la lista de desinencias.

Ejemplo: trabajador

-
a
es
as

el descriptor será: trabajador.

La longitud de las desinencias está limitada a 15 caracteres.

->Enlaces

Un registro por cada lista de enlaces, con un máximo de 254 listas.
En algunas lenguas (sobre todo en alemán) las palabras compuestas se pueden construir yuxtaponiendo palabras simples, enlazándolas o no con una letra.

Ejemplo: bildung (educación)
zugang (acceso)
bildungszugang (acceso a la educación)
land (tierra)
arbeiter (trabajador)
landarbeiter (trabajador agrícola).

La entrada «bildung» en el fichero diccionario contendrá un reenvío hacia la letra s en el fichero de enlaces, mientras que «land» remitirá a «-».

.2 Indización

La indización automática bajo PASSAT se realiza documento a documento, en proceso por lotes:

-se introduce el documento (título, resumen y/o texto completo),
-se reconocen las palabras (todas las cadenas de caracteres entre espacios blancos y/o signos de puntuación),
-se ordenan alfabéticamente las palabras (con punteros que permitirán reconstruir los conceptos compuestos por varias palabras, si esto se considera necesario después de consultar el diccionario),

-para cada palabra del texto, se busca en el fichero diccionario la entrada más pequeña que tenga el mayor número de caracteres en común con ella.

Ejemplo: palabra del texto : «ocupan»
palabras del diccionario : «ocup»
«ocupación».
palabra del diccionario
conservada : «ocup»;

-se accede a la lista de desinencias propia de cada palabra;
-se buscan diversas flexiones posibles hasta que se obtenga una coincidencia completa: «ocup-an», o hasta que se llegue al final de la lista de desinencias;
-a veces se accede a la lista de enlaces propia de cada palabra;
-en caso de que esté presente un código «palabra compuesta», se busca si coincide con las palabras del texto alguna de las palabras compuestas que figuran en esta entrada de diccionario; el sistema sólo trabaja por constituyentes inmediatos (es decir, con términos contiguos); si, por ejemplo, el texto incluye: «...persona física o moral ...» y en el diccionario existen: «persona física» y «persona moral», el sistema va a indizar «persona física» y no indizará «persona moral»; si, por el contrario, el texto que se indiza contiene «...persona física o persona moral...», el sistema indizará con los dos descriptores;
-cuando se encuentra una palabra y su flexión exacta, el sistema:
+indizará el texto (si estatuto D de la palabra), por medio de:
->la forma canónica de esta palabra (sustitución K)
->tal forma, así como de la forma canónica de todos los términos complementarios designados en el fichero diccionario (sustitución V)
->la forma canónica del término de sustitución designada en el fichero diccionario (sustitución E)
+no indizará el texto con esa palabra (si estatuto N);
-cuando una palabra que está presente en el texto no figura en el fichero diccionario, o si su forma flexionada no puede ser reconstruida a partir del fichero de desinencias, el sistema la imprimirá dentro de una lista de palabras no transmitidas.

.4 Indización automática selectiva, en lenguaje libre

El comienzo del proceso de indización es el mismo que en la indización automática no selectiva (cf. § 3.2); un algoritmo suplementario procede a la ponderación de los términos reconocidos; únicamente los términos que sobrepasen un determinado peso, correspondiente a un umbral prefijado, serán conservados para la indización del documento.

Se pueden utilizar diferentes algoritmos:

-cálculo de la frecuencia absoluta de una palabra dentro de un texto;
-cálculo de la frecuencia relativa, es decir, de la ratio existente entre el número de apariciones de una palabra, por una parte, dentro del texto del documento que se va a indizar y, por otra, dentro de la colección, es decir, en el conjunto de documentos indizados hasta entonces.

Describimos a continuación el sistema SPIRIT, basado en la utilización de un léxico completo de una lengua natural, seguida de un análisis lingüístico y de un análisis estadístico.

SPIRIT (Système Syntaxique et Probabiliste d'Indexation et de Recherche d'Informations Textuelles) es un programa de almacenamiento y recuperación documental, concebido por la sociedad SYSTEX y comercializado por la sociedad CISI.

Incluye programas de indización automática de documentos (descrita anteriormente) y de consultas (véase § 6.5.5) y está basado en una aproximación mixta, lingüística y estadística, en la que un modelo estadístico simple se apoya sobre un análisis lingüístico bastante potente.

Funciona en francés, en inglés y pronto en árabe.

Se basa en dos ficheros diccionarios fundamentales:

-el fichero diccionario principal;
-el fichero diccionario de expresiones idiomáticas.

Los diccionarios y los textos que se van a procesar pueden estar:

-o bien en tipografía «rica» (mayúsculas, minúsculas y signos diacríticos: acentos, diéresis, cedilla...);
-o bien en tipografía «pobre» (sólo mayúsculas);
-o bien en tipografía mixta.

.1 Diccionarios

El sistema incluye dos diccionarios:

+*Diccionario principal:*

-un registro para cada una de las flexiones gramaticales (singular y plural de los nombres; singular, plural, masculino y femenino de los pronombres y adjetivos; todas las personas de todos los tiempos y de todos los modos de los verbos) de las palabras de la lengua usual: ± 450.000 ENTRADAS (=formas de palabras) en francés: 270.000 formas de palabras acentuadas y 170.000 formas de palabras no acentuadas, correspondientes a 40.000-45.000 formas canónicas (=palabras fuente): singular de los nombres, masculino singular de los adjetivos y pronombres, infinitivo de los verbos;
-cada registro del diccionario incluye:
+una entrada = forma flexionada de una palabra

->o una palabra acentuada,
ejemplo: guías
->o una palabra no acentuada (sólo accesible en caso de mecanografiado en tipografía pobre)
ejemplo: ira.

Para las palabras acentuadas el registro incluye además:
+una serie de propiedades lingüísticas y, sobre todo, la parte o partes de la oración (=categoría gramatical).
Ejemplo: guías, nombre o verbo.
+la palabra fuente correspondiente a una propiedad lingüística.
Ejemplo: guías : nombre : guía
guías : verbo : guiar
+a veces, un sinónimo.
Ejemplo: guía (nombre) : indicador.

Para las palabras no acentuadas, el registro incluye un puntero hacia las palabras acentuadas correspondientes.

Ejemplo: IRA:ira (sustantivo)
o irá (verbo).

+*Diccionario de expresiones idiomáticas:*

Contiene un número limitado (± 2.500) de formas desarrolladas de siglas y de expresiones hechas, como por ejemplo: «a pesar de», «poner por obra», «estrella de mar», etc. Las palabras compuestas no aparecen, en sí mismas, en este diccionario: su número es ilimitado y deben ser reconocidas por los algoritmos.

*Nota:*

Téngase en cuenta así mismo la intervención particular del fichero inverso, por otra parte clásico en todos los sistemas de búsqueda documental: la frecuencia de indización de cada forma-fuente es el valor que se utiliza para calcular el peso de esos términos en el momento del tratamiento estadístico de la indización de los documentos: el peso es tanto más elevado en tanto en cuanto la forma fuente es localizada en pocos documentos.

.2 Indización

La indización automática de los documentos (títulos, resúmenes y, lo más frecuente, textos completos) se realiza en once o doce etapas, que pueden ser agrupadas en dos clases:

-tratamientos lingüísticos: el análisis lingüístico tiene como fin determinar las unidades de lenguaje: palabras, grupos de palabras y palabras en relación sintagmática;

-tratamientos estadísticos: el análisis estadístico opera sobre las unidades de lenguaje detectadas en el momento del análisis lingüístico.

+*Tratamientos lingüísticos*

-Después de haber compuesto el documento, o haber convertido a formato SPIRIT un texto ya compuesto, se reconocen las palabras (toda secuencia de caracteres entre espacios blancos y/o signos de puntuación) y los nombres propios (palabras cuya primera letra es una mayúscula);
-se ordenan alfabéticamente;
-se analizan morfológicamente consultando el diccionario principal:
+se asocian los datos del diccionario a las palabras encontradas;
+las palabras del documento no encontradas en el diccionario se imprimen en una lista de errores; se trata de:
->o palabras todavía no incorporadas al diccionario;
->o errores ortográficos;
-se ordenan las palabras en la secuencia del documento;
-se reconocen las locuciones consultando el diccionario de expresiones idiomáticas;
-se realiza el análisis sintáctico, con el fin de:
+despejar la ambigüedad de términos homógrafos, es decir, de términos cuya ambigüedad procede de que corresponden a dos o más categorías gramaticales diferentes;
+reconocer las expresiones.

Este análisis se realiza de manera totalmente automatizada por medio de un programa de análisis sintáctico.
Este programa trata de:

+resolver las clases de homógrafos:
->verbo/sustantivo (caso de la palabra: guías)
->adverbio/adjetivo
->etc.

+reconocer los tipos de enlaces (es decir, las relaciones de dependencia) entre palabras contiguas (prácticamente vacías):
->sujeto + verbo;
->verbo + complemento directo;
->verbo + complemento indirecto;
->sustantivo + complemento de sustantivo;
->etc.

Las expresiones compuestas de palabras contiguas así reconocidas son conservadas para la indización, con el mismo tratamiento que las palabras simples constituyentes.

Ejemplo:

si en un texto figura «... la puesta en marcha...», el sistema conservará:
+las palabras simples:
->puesta
->marcha
+la expresión:
->puesta marcha.

+*Notas:*

1) por su concepción, el sistema no debe resolver las ambigüedades polisémicas no homográficas (una forma = dos o más significados correspondientes a la misma categoría gramatical; ejemplo: conductor -de automóviles o de flujo eléctrico-; la palabra fuente sigue siendo en efecto la misma; aquí: conductor). Sí resuelve, por el contrario, los casos de polisemias homográficas (una forma = dos o más significados correspondientes a dos o más categorías gramaticales distintas; ejemplo: guías, plural del nombre «guía» y guías, segunda persona del singular del presente de indicativo del verbo «guiar»);
2) el sistema conserva todas las expresiones que su programa le permite reconocer, aunque éstas no sean realmente significativas en la lengua.

Ejemplo: en la frase:

El economista modeliza el tráfico de vehículos.
SPIRIT va a conservar:
economista
modelizar
tráfico
vehículo
economista modelizar
modelizar tráfico
tráfico vehículo;

-se reconocen las expresiones compuestas, basándose en criterios sintácticos y semánticos; el sistema puede reconocer expresiones cuyos términos constituyentes no estén contiguos.

Ejemplo:

la expresión «servicio de documentación» podrá ser reconocida en un texto donde se diga: «...el servicio automatizado de documentación...»; este tratamiento, de base algorítmica (no basado en un diccionario), no funciona más que de manera experimental;

-explicitación: el sistema puede reconocer conceptos implícitos, es decir, conceptos que intervienen en un documento sin estar explícitamente representados en él (esto no funciona más que de manera experimental).

Ejemplo:

en una ley, un artículo puede mencionar que «el infractor será condenado con una pena de...». El sistema de explicitación, basado en criterios semánticos (la existencia de ciertos términos, aquí: infractor, condenado, pena...) va a reconocer el concepto implícito de «sanciones» y se lo va a proponer al documentalista (proceso semi-automático);

-se eliminan las palabras vacías, a partir de criterios:
+gramaticales: lista de partes de la oración vacías (los pronombres, por ejemplo);
+morfológicos: lista de palabras vacías.

-normalización, gracias a la información aportada por el diccionario principal; esta normalización se puede hacer a tres niveles:
->forma flexionada -> forma canónica (=palabra fuente).

+Ejemplo: guías
->si nombre : guía
->si verbo : guiar
+ortografías diferentes.

Ejemplo: I.V.A.: IVA
+sinónimo -> término preferencial.

Ejemplo: guía (nombre): indicador.

Existe el proyecto de que, en una etapa ulterior, el sistema pueda asimismo restablecer las formas de base a partir de las formas derivadas.

Ejemplo:modelaron : modelo
modelado : modelo.

+*Tratamientos estadísticos:*

-se calcula una función de peso para cada palabra (o expresión) normalizada. Este peso mide la carga informativa (=discriminante) de la palabra para la búsqueda documental. El peso es tanto más elevado cuanto menos frecuente es la palabra en la base de datos (en el límite, una palabra presente en todos los documentos tiene un peso nulo);
-llegado el caso, se eliminan las palabras o expresiones de peso inferior a un umbral preestablecido.

.5 Indización automática selectiva, en lenguaje controlado

Es el tipo de indización automática más refinado, en tanto que

-selectivo
-en lenguaje controlado: normalmente por una lista de autoridades.

El sistema PASSAT, descrito anteriormente (§ 3.3) incluye un módulo opcional de conversión automática de los términos conservados como descriptores de una lista de autoridades, con selección de los descriptores susceptibles de precisar mejor los temas fundamentales de los documentos indizados.
Este sistema se basa en:

-un fichero llamado «de asociaciones»;
-un programa de indización.

.1 Fichero

Se establece una lista de los descriptores considerados más importantes.
Esta lista puede incluir desde varios cientos hasta varios miles (máximo 65536) de descriptores; es específica para cada servicio de documentación y para la combinación de materias que se trate.
Esta lista está creada por medio de una utilidad de PASSAT, pero la definición de los «descriptores importantes» que contiene, es:

-puramente intelectual;
-y enteramente empírica: no se suministra ninguna regla que precise la forma de construirla.

Cada descriptor de esta lista está identificado por un número de cuatro dígitos hexadecimales.
Cuando se desea utilizar esta lista, se asocia a cada radical del diccionario conservado como descriptor (D) (cf. § 3.3.1) uno o varios pares (máximo 30) de estos descriptores en un fichero llamado «de asociaciones».

Ejemplos:

| Descriptor del fichero diccionario | Pares de descriptores asociados dentro de la lista de descriptores importantes |
|---|---|
| azada | herramienta - agricultura |
| pico | herramienta - agricultura; boca - ave |
| cavar | azada - tierra |

Evidentemente los mismos términos pueden aparecer en la columna de la derecha y en la columna de la izquierda para entradas diferentes (ejemplo: azada).
Es necesario hacer este trabajo intelectualmente. Una utilidad del programa crea dentro del fichero diccionario los enlaces considerados útiles.

.2 Indización

La indización de un texto por PASSAT ha consistido en establecer la lista de los términos del texto reconocidos por el sistema, en su forma canónica (cf. § 3.3.2).

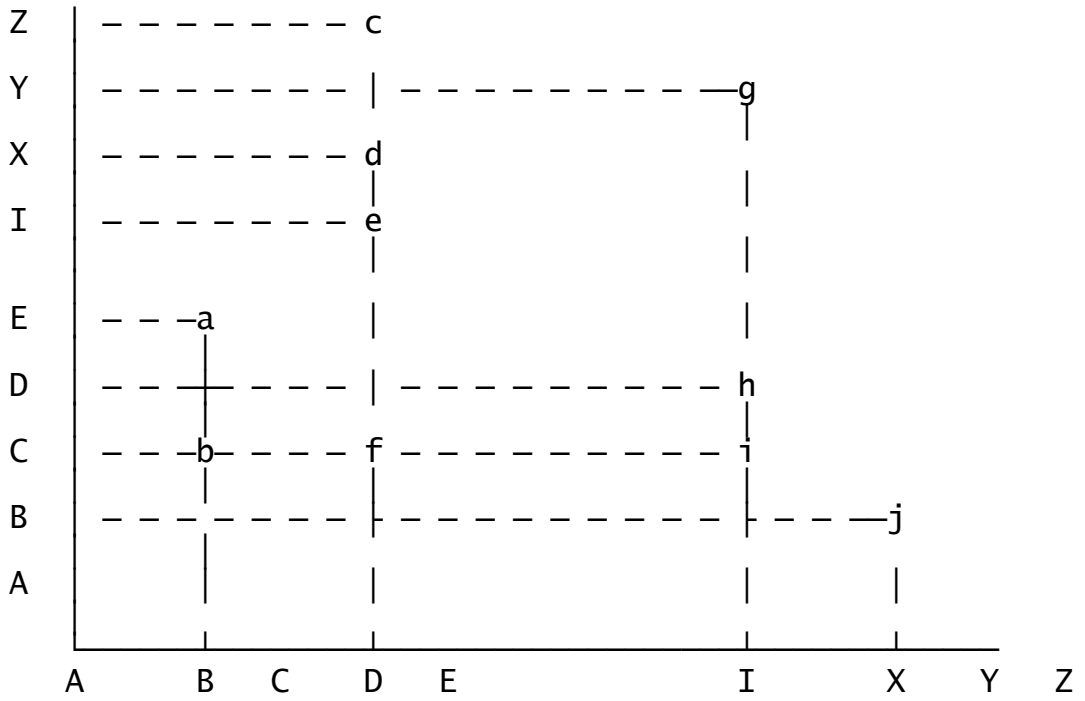
La matriz de asociación permitirá asignar un peso relativo a cada término y no conservar más que los descriptores cuyo peso sea superior a un umbral fijado a priori.

Ejemplo:

sea A - B - C - D - E ... I ... X - Y - Z la lista de los descriptores asociados (de una colección), es decir, de los descriptores considerados importantes y que figuran dentro del fichero de descriptores.

sea a - b - c - e = f - g - h - i - j la lista de los términos de indización (de un documento) extraídos del fichero diccionario.

La matriz de asociación de un documento se presenta como sigue:




El sistema cuenta las frecuencias de aparición de los descriptores asociados:

| en abscisas: | en ordenadas: |
|---|---|
| A: 0<br>B: 2 | A: 0<br>B: 1 |

| | |
|---|---|
| C: 0<br>D: 4<br>E: 0<br>.<br>.<br>.<br>I: 3<br>.<br>.<br>.<br>X: 1<br>Y: 0<br>Z: 0 | C: 3<br>D: 1<br>E: 1<br>.<br>.<br>.<br>I: 1<br>.<br>.<br>.<br>X: 1<br>Y: 1<br>Z: 1 |

A cada término de la indización se le asigna un número igual al total de las frecuencias de sus descriptores asociados; la cifra más elevada es equivalente al 100 %; se calcula el peso, en %, de cada término de indización:

| Descriptor de Indización | Descriptor asociado y suma de sus pesos (respectivamente en abscisas y ordenadas) | Peso Total | % |
|---|---|---|---|
| a | B + E --> 2 + 1 | 3 | 42, 9 |
| b | B + C --> 2 + 3 | 5 | 71, 4 |
| c | D + Z --> 4 + 1 | 5 | 71, 4 |
| d | D + X --> 4 + 1 | 5 | 71, 4 |
| e | D + I --> 4 + 1 | 5 | 71, 4 |
| f | D + C --> 4 + 3 | 7 | 100 |
| g | I + Y --> 3 + 1 | 4 | 57, 1 |
| h | I + D --> 3 + 1 | 4 | 57, 1 |
| i | I + C --> 3 + 3 | 6 | 85, 7 |
| j | X + B --> 1 + 1 | 2 | 28, 6 |

Inicialmente se ha de fijar un umbral, por ejemplo en el 66 % (3 valores posibles: 50 %, 66 % u 80 %). Todos los términos de indización con pesos relativos inferiores a esa cifra serán eliminados: sólo los términos b - c - d - e - f - i serán mantenidos, con el estatuto de descriptores.

.6 Indización asistida por ordenador

La indización asistida por ordenador comprende dos fases:

-una fase de preindización automática, durante la cual el ordenador analiza el texto que recibe y le asocia una serie de descriptores, generalmente extraídos de una lista de autoridades, que propone al documentalista.
-una fase de diálogo entre el documentalista y el ordenador, durante la cual la lista propuesta en la fase precedente se afina por la acción del hombre.

A continuación se describen tres programas informáticos de indización que funcionan según este modelo, uno de manera determinista simple, y los otros dos según un proceso muy elaborado, con carácter probabilista el segundo y con carácter lingüístico el tercero.

.1 Indización asistida determinista

Este sistema se puede utilizar con la mayoría de los programas informáticos tradicionales de almacenamiento y recuperación documental. En primer lugar se registra una lista de autoridades.
Se realiza un resumen en lenguaje libre de los documentos, pero exigiendo a los documentalistas:

-por una parte, que utilicen una terminología lo más normalizada posible para designar los conceptos que intervienen en los resúmenes;
-por otra parte, que destaquen los descriptores del resumen enmarcándolos con un signo especial (por ejemplo: comillas).

Una vez compuestos los resúmenes, el ordenador compara cada uno de los descriptores, más o menos controlados, de la indización (palabras o expresiones) con los descriptores de la lista de autoridades.
Los descriptores reconocidos se introducen automáticamente en los índices. Los descriptores del resumen que no figuren en la lista de autoridades se presentarán en pantalla, y el documentalista podrá decidir:

-o bien integrarlos en el thesaurus, en cuyo caso serán también incorporados a los índices;
-o bien no conservarlos como descriptores, en cuyo caso los signos especiales que los destacan serán elmininados.

En algunos sistemas se puede indizar con los términos no reconocidos dentro de una lista de descriptores libres.

.2 Indización asistida con orientación probabilista

SINTEX (Système d'INdexation de TEXtes) es un programa de indización semi-automático, concebido y puesto en el mercado por la sociedad SIE (Société Informatique Européenne), de Versalles.

Está operativo para el francés, inglés y español. Se encuentran en fase de estudio las versiones para indizar el holandés y el alemán.

Permite extraer de un texto los términos de indización, explícitos o implícitos, que representan el contenido del documento. Esos términos de indización pertenecen a una lista de autoridades.

Sólo puede funcionar tras una fase de preaprendizaje, durante la cual el sistema trata un corpus de muestra de ± 4.000 documentos indizados y clasificados humanamente.

Puede funcionar para cualquier sistema de almacenamiento y recuperación documental.

Incluye ficheros diccionarios, programas de preaprendizaje (a partir del corpus de muestra indizado humanamente) y programas de indización automática.

Utiliza tipografía pobre (letras mayúsculas solamente).

.1 Diccionarios

El sistema contiene ocho o nueve ficheros:

+*Léxico de palabras simples (unitérminos)*

Este fichero se construye automáticamente durante la fase de preaprendizaje, por reconocimiento de:

- -todas las secuencias de caracteres entre espacios blancos y/o signos de puntuación;
- -extraídas de las zonas: títulos, resúmenes, descriptores y/o textos;
- -de un corpus de muestra representativo del fondo documental que se va a tratar, introducido en el sistema tras una indización y clasificación humanas previas;
- -y que no pertenezcan a una lista de palabras vacías (pronombres, preposiciones, conjunciones...) registrada previamente.

Este fichero puede contener algunas decenas de miles de palabras simples, en todas sus flexiones gramaticales; sin embargo, en lo que se refiere al francés, se realiza un proceso de normalización plural-singular.

+*Léxico de descriptores*

Esta lista, que constituye un lenguaje controlado de descriptores (lista de autoridades compuesta por palabras simples y expresiones compuestas), de algunos miles de términos, debe ser elaborada a priori; el sistema la registra cuando se ejecutan los programas de preaprendizaje, extrayendo los descriptores asignados humanamente a los documentos del corpus de muestra.

Cada registro contiene la designación íntegra del descriptor.

*Nota:*

El autor del lenguaje controlado no ha de crear las relaciones jerárquicas o asociativas. El sistema va a crear automáticamente las asociaciones, que no serán fijas.

*+Léxico de dominios*

Este fichero está formado por una lista de 64 términos generales como máximo, denominados «dominios», «epígrafes», «temas» o «sectores».
Esta lista debe ser elaborada a priori; el sistema la registra cuanto se ejecutan los programas de preaprendizaje, extrayendo los «dominios» asignados humanamente a los documentos del corpus de muestra.

*+Fichero de correspondencia formal descriptores/unitérminos constituyentes*

Este fichero se construye automáticamente a partir de los descriptores suministrados por el corpus de muestra.
Cada registro:

-está formado por un descriptor estraído de la zona «descriptor» de uno de los documentos del corpus de muestra;
-por la lista de unitérminos (= palabras simples) constituyentes de ese descriptor.

*+Fichero de correspondencia formal unitérminos constituyentes/descriptores*

Este fichero es el inverso del anterior.
Cada registro consta:

-de un unitérmino que forma parte de uno o más descriptores;
-del descriptor o descriptores que contienen ese unitérmino.

*Nota:*

SINTEX no hace ninguna diferencia entre las diversas polisemias de una misma palabra (FABRICAS puede ser tanto un verbo como un sustantivo).

*+Fichero de correspondencia medida de unitérminos/descriptores*

Este fichero se construye automáticamente por análisis de coocurrencia de los descriptores dentro de la zona «descriptores» y de los unitérminos dentro de las zonas «títulos», «resúmenes», y/o «textos», en el corpus de muestra.
Cada registro:

-está formado por un unitérmino (extraído del texto);
-reenvía a cada uno de los descriptores encontrados dentro del corpus de muestra, en los documentos en los que se produce coocurrencia del unitérmino con esos descriptores;

Cada enlace tiene asignado un peso, entre cero y uno, denominado «medida de correspondencia», que está en función del número de coocurrencias.

*Notas:*

-Solamente se tienen en cuenta las coocurrencias comprendidas entre un mínimo y un máximo.

-El sistema es del tipo «por aprendizaje», en el sentido de que la lista de unitérminos y la ponderación de descriptores ligados a los unitérminos asociados son revisadas tras cada remesa de actualizaciones del fondo documental.

*+Fichero de correspondencia medida de descriptores/descriptores*

Este fichero se construye automáticamente por análisis de las coocurrencias entre descriptores, dentro del corpus de muestra.

Cada registro:

-está formado por un descriptor;
-reenvía a cada uno de los descriptores coocurrentes, al interior de una pareja de frecuencias de coocurrencias considerada significativa;

Cada enlace tiene asignado un peso, entre cero y uno, llamado «norma de correspondencia».

*+Fichero de correspondencia medida de descriptores/dominios*

Este fichero se construye automáticamente por análisis de coocurrencia entre «descriptores» y «dominios», dentro del corpus de muestra.

Cada registro:

-está formado por un descriptor;
-reenvía a cada uno de los dominios coocurrentes, al interior de una pareja de frecuencias de coocurrencias considerada significativa;

Cada enlace tiene asignado un peso, entre cero y uno.

*+Léxico de no-descriptores (opcional)*

Lista de términos no aceptados, con reenvío de cada no-descriptor al descriptor adecuado.

.2 Preaprendizaje del sistema

Se registra el corpus de muestra de ± 4.000 documentos; cada documento contiene:

-de 200 a 10.000 caracteres de títulos, resúmenes y textos en lenguaje natural;
-de 1 a 50 descriptores, procedentes de una indización humana, a partir de una lista de autoridades (a priori) que contiene de varios cientos a varios miles de descriptores;
-1 o más dominios, pertenecientes a una clasificación humana, tomados de una lista de 64 dominios como máximo construida a priori.

Se construyen automáticamente los ficheros del § 3.6.2.1

.3 Indización

La indización se realiza a partir del campo o campos especificados por el documentalista: título, resumen...; tiene lugar tras haber sido registradas las referencias, que en principio no son ni indizadas ni clasificadas humanamente; comprende las fases siguientes:

+*Indización automática*

-se reconocen los unitérminos del texto dentro del campo o campos que se van a analizar, comparándolos con los unitérminos del fichero de unitérminos;
-se consulta el fichero de correspondencia medida de unitérminos/descriptores y:
+para cada unitérmino se extraen todos los descriptores asociados y sus pesos de coocurrencia con el unitérmino en cuestión;
+para cada uno de los descriptores asociados así extraídos (que pueden haber sido extraídos varias veces por estar ligados a varios unitérminos del texto): se calcula una ponderación global y se eliminan los descriptores recuperados una sola vez o demasiadas veces; los descriptores conservados conforman la indización potencial: lista de «descriptores potenciales»;
-se buscan, dentro de este conjunto de «descriptores potenciales», los «descriptores formales», es decir, aquellos cuyas palabras constituyentes se encuentran formalmente, con normalización singular-plural, dentro del texto, y próximas:
+o bien dentro de la misma frase y en la misma secuencia;
+o bien en cualquier lugar del texto y en cualquier secuencia;
-a partir de este sub-conjunto de descriptores formales:
+se consultan los ficheros de correspondencia medida de descriptores/descriptores y descriptores/dominios;
+se determina el eje del documento para esos diferentes dominios, es decir, la suma de los pesos del conjunto de «descriptores formales» para cada dominio; los valores inferiores al 10 % no se declaran; un documento puede, por ejemplo, tener el 25 % para el eje del dominio

«política extranjera» y el 15 % para el eje «relación países terceros». Un documento puede también tener un 100 % para un sólo eje.

+*Diálogo hombre-máquina*

-se muestra este eje, es decir, la lista de los dominios y los porcentajes de pertenencia del documento a cada dominio;
-el documentalista selecciona el dominio o dominios del documento:
+o bien la lista mostrada;
+o un sub-conjunto de la lista mostrada;
+o uno o varios dominios distintos de los de la lista;
+o bien dominios de la lista y dominios de fuera de la lista;
-se aplica un nuevo «eje» al conjunto complementario de los «descriptores potenciales» aislados en la primera selección;
-se calcula un nuevo valor de asociación tomando en cuenta sólo aquellos dominios que corresponden de modo principal al eje del documento;
-se calcula un nuevo factor de plausibilidad de cada descriptor (formal o potencial) conservado en función de la distribución de los valores de asociación;
-los descriptores potenciales y formales se disponen en orden decreciente de plausibilidad;
-se muestra la lista ordenada;
-el documentalista selecciona los descriptores o un nuevo eje (= lista de dominios y %) y se vuelve a repetir el cálculo de plausibilidad.

.3 Indización asistida con orientación lingüística

ALEXDOC es uno de los niveles de aplicación del programa de almacenamiento y recuperación documental ALEXIS, concebido y comercializado por la sociedad ERLI, de Charenton.

ALEXDOC, que se encuentra en estado de prototipo avanzado (en 1985), contiene programas de ayuda a la indización de documentos (descritos a continuación) y de consultas (véase § 6.5.4).

Está basado en:

-la explotación de un thesaurus de descriptores gestionado por ALEXIS (cf. capítulo II, § 1.2.4.5);
-la aplicación de reglas lingüísticas sofisticadas.

Permite gestionar una base de datos documental en lengua francesa.

Explota:

-el conjunto de los datos bibliográficos y el conjunto de los datos «thesaurus» propios de ALEXIS;
-el conjunto de datos «diccionario de base de la lengua» específico de ALEXDOC;

-reglas de transformación y de sustitución, de naturaleza sintáctica, que explotan las relaciones semánticas del thesaurus.

Los diccionarios y los textos a tratar pueden ser:

-o bien, preferentemente, en tipografía rica (mayúsculas, minúsculas y signos diacríticos);
-o en tipografía pobre (mayúsculas solamente);
-o bien en tipografía mixta.

.1 Conjuntos de datos

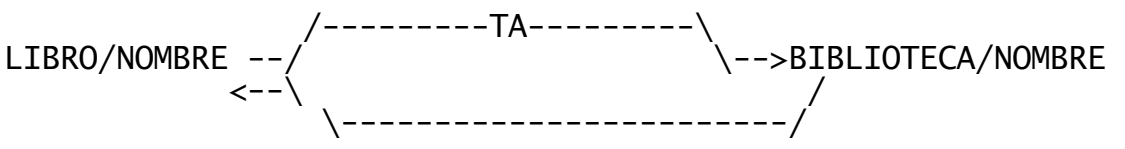
+*Conjunto de datos «thesaurus»*

El registro de cada descriptor (denominado, por el productor, unidad léxica o lexía) consta de:

-el enunciado del descriptor (forma gráfica, o grafía, de 256 caracteres como máximo); ejemplo: LIBRO;
-un código de categoría gramatical (tipo); ejemplo: NOMBRE, o VERBO; el sistema permite definir hasta 512 tipos diferentes;
-un puntero hacia cada una de las lexías con las que el descriptor tiene una relación semántica (es decir, se indica el lugar donde se encuentra la lexía en cuestión dentro del fichero); ejemplo: XYZ (siendo XYZ la dirección de la lexía BIBLIOTECA/NOMBRE);
-un código que precisa la naturaleza de esa relación; ejemplo: genérico, específico, asociado, equivalente, pertenencia a un campo semántico o microthesaurus...; el sistema permite gestionar hasta 256 tipos de relaciones diferentes; la mayoría de las relaciones están invertidas, lo que significa que si existe una relación entre una lexía A y una lexía B, hay automáticamente una relación inversa entre la lexía B y la lexía A: esta relación puede ser:

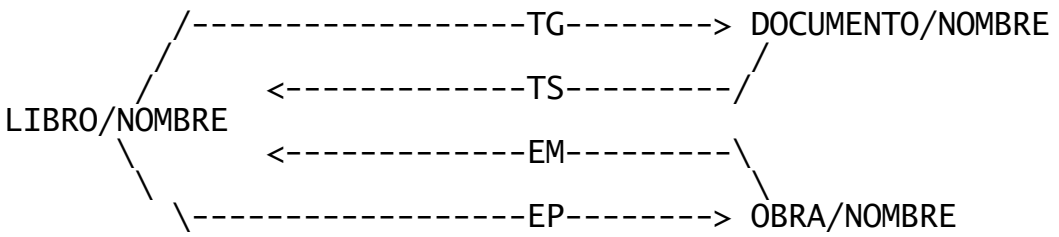
+o bien simétrica: la relación directa y la relación inversa contienen el mismo nombre.

Ejemplo (el puntero se simboliza aquí por una flecha):

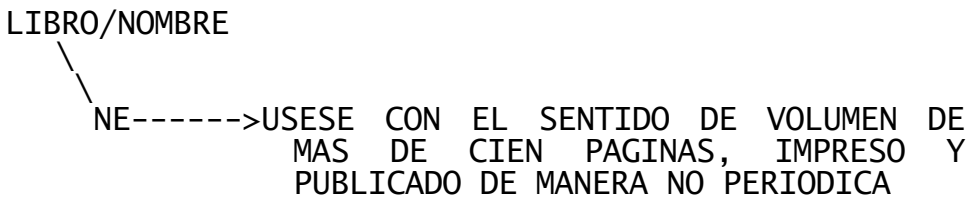



+o bien disimétrica;: la relación directa y la relación inversa contienen un nombre diferente.

Ejemplo:




Algunas relaciones no están invertidas: las que existen entre una lexía (256 caracteres como máximo) y un texto (4.000 caracteres como máximo) (ejemplo: una nota explicativa).




Los descriptores compuestos están representados en el sistema por un diagrama arbóreo en el que:

- -las hojas solas contienen una grafía;
- -los nodos (simbolizados, en el ejemplo que se incluye a continuación, por el signo O) contienen un tipo;
- -el tronco resulta de una concatenación virtual (es decir, no físicamente presente en el fichero, para economizar espacio en él, pero reconstruible por el sistema a través de punteros, sin que el usuario se tenga que preocupar de ello) de todas las hojas; esta concatenación virtual se simboliza a continuación poniendo entre paréntesis las palabras compuestas;
- -pueden existir uno o más niveles.

Ejemplos:

+un nivel

SERVICIO/NOMBRE DE/PREP DOCUMENTACION/NOMBRE

O/NOMBRE
(SERVICIO DE DOCUMENTACION)

+dos niveles

AUTOMATIZACION DE SERVICIO DE DOCUMENTACION
/NOMBRE /PREP /NOMBRE /PREP /NOMBRE

O/NOMBRE
(SERVICIO DE DOCUMENTACION)

O/NOMBRE
(AUTOMATIZACION DEL SERVICIO DE DOCUMENTACION)

*Nota:* el sistema introduce automáticamente una relación entre las palabras compuestas y las palabras simples que las componen: COMPUESTO POR/COMPONENTE DE

+*Conjunto de datos bibliográficos*

En la organización de tipo relacional de ALEXIS, el fichero bibliográfico no se distingue del fichero thesaurus: se integran los dos ficheros.

Cada categoría de información (autores, título, fecha, descriptores, lengua, resumen) está caracterizada por un «tipo» (/AUTOR, /TITULO, /FECHA, /DESCRIPTOR) y está enlazada con el nº del documento por una relación que es

-o bien invertida y disimétrica, para aquellas informaciones que sean criterios de búsqueda y que estén tratadas por el sistema cómo lexías (256 caracteres como máximo por criterio, simple o compuesto).

10.352/Nº Ejemplo (documento nº 10.352).

TIENE POR AUTOR
PEREZ
AUTOR DE /AUTOR

TIENE POR FECHA 1986
/FECHA
FECHA DE

TIENE POR LENGUA FRANCES
/LENGUA
LENGUA DE

TIENE POR TITULO (ORDENADOR Y BIBLIOTECA)

TITULO DE O/TITULO

ORDENADOR Y BIBLIOTECA
/NOMBRE /CONJ /NOMBRE

INDIZADO POR BIBLIOTECA UNIVERSITARIA/DESCR

FRANCIA/DESCR

INDIZADO POR

INFORMATICA DOCUMENTAL/DESCR

PROGRAMA INFORMATICO/DESCR

PRESTAMO/DESCR

-o bien no invertida, para aquellas informaciones que no sean criterios de búsqueda y que sean tratadas por el sistema como textos (4.000 caracteres como máximo por texto).

Ejemplo (documento 10.352)

10.352/Nº
\
\---->TIENE POR RESUMEN -> La informática para los sistemas documentales ha producido progresos muy grandes en las bibliotecas de facultad; los programas de gestión de préstamo son muy utilizados en la Universidad francesa.

+*Conjunto de datos «diccionario de base de la lengua»*

Este diccionario, específico de ALEXDOC, contiene términos que se pueden encontrar en los títulos y en los resúmenes de los documentos que se van a indizar, pero que no son necesariamente descriptores; aquí se encuentran:

-verbos;
-palabras gramaticales: artículos, preposiciones, conjunciones, pronombres...;
-locuciones;
-términos que, aunque no se encuentran dentro del thesaurus, están ligados a un término del thesaurus por una relación semántica;
-una relación, denominada «rasgo», para marcar aquellos términos que son descriptores.

Ejemplo:

(BIBLIOTECA UNIVERSITARIA)
RASGO O /NOMBRE

BIBLIOTECA/NOMBRE UNIVERSITARIA/ADJETIVO

/DESCRIPTOR NOMINALIZACION DE

RASGO UNIVERSIDAD/NOMBRE

RASGO

TS

FACULTAD/NOMBRE

Este diccionario, suministrado con ALEXDOC, contiene aproximadamente 40.000 términos.

.2 Reglas

-Regla de derivación morfológica, de forma sufijo 1, sufijo 2/tipo 1, tipo 2.

Ejemplos:

AR, O/VERBO
significa que un verbo en AR puede existir con el sufijo O, y viceversa; esta regla permite pasar automáticamente, por ejemplo,
de LIBRO/VERBO
a LIBRAR/VERBO

AR, ANZA/VERBO, NOMBRE
significa que de un verbo en AR puede derivarse un nombre que termine en el sufijo ANZA, y viceversa; esta regla permite pasar automáticamente, por ejemplo,
de LIBRAR/VERBO
a LIBRANZA/NOMBRE

ESA, IA/ADJETIVO, NOMBRE
permite pasar automáticamente, en el documento10.352,
de FRANCESA
a FRANCIA

AL, ACION/ADJETIVO, NOMBRE
permite pasar automáticamente, en el documento 10.352,
de DOCUMENTAL
a DOCUMENTACION

,S/NOMBRE, NOMBRE
permite pasar automáticamente, en el documento 10.352,
de PROGRAMAS
a PROGRAMA

-Reglas de reestructuración: permiten recuperar las palabras compuestas cuyos componentes están dispersos en una expresión más extensa.

Ejemplo: la indización de la expresión «informática para los sistemas documentales», que figura en el resumen del

documento 10.352 permitirá recuperar el descriptor INFORMATICA DOCUMENTAL, si ese descriptor figura en el thesaurus y se ha incorporado al sistema la regla NOMBRE1 - PREP - ART - NOMBRE2 - ADJ/NOMBRE1 - ADJ.

-Reglas de desambiguación, para permitir determinar el tipo exacto de lexías homográficas, es decir, de lexías formadas por una grafía idéntica, pero que pertenecen a tipos diferentes.

Ejemplos:

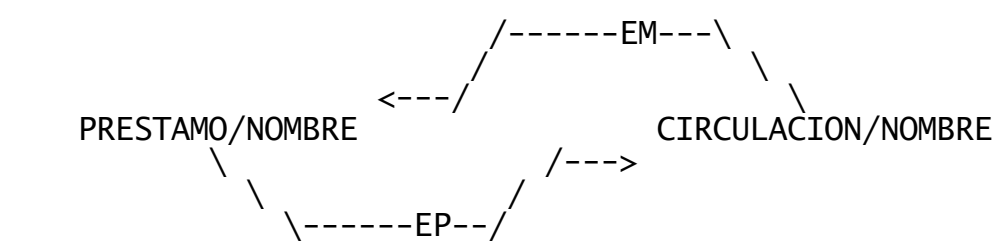
SON/NOMBRE
SON/VERBO
Esta desambiguación se realiza por análisis del contexto a la izquierda y a la derecha de las grafías ambiguas.

Ejemplo:

en el resumen del documento 10.352, el sistema reconocerá SON como un verbo, porque el término que le precede (préstamo) es un nombre y el término que le sigue (muy) es un adverbio(*)13.

-Reglas de transformación semántica, que explotan las relaciones semánticas del thesaurus y del diccionario.

Ejemplos: en el documento 10.352,
+la relación de equivalencia




permite indizar con el descriptor
CIRCULACION

+la regla

NOMBRE - PREPOSICION - NOMBRE ESPECIFICO DE UNA NOMINALIZACION DE UN ADJETIVO / NOMBRE - ADJETIVO

(*)13 Al traducir esta parte del libro se ha realizado una adaptación del ejemplo. La obra original se ocupa del texto «...les progiciels de gestion du prêt sont très utilisés...». PRET es palabra ambigua: puede tratarse de un nombre (= «préstamo») o de un adjetivo (= «prestado»). El sistema reconoce que es un nombre porque va precedido por un artículo (du) y seguido por un verbo (sont) [nota de los tr.].

permite indizar con el descriptor
BIBLIOTECA UNIVERSITARIA
a partir de la expresión
BIBLIOTECA DE FACULTAD
que figura en el resumen.

-Reglas de reconocimiento de las frases, por reconocimiento de los signos de puntuación; el sistema establece distinción entre los puntos de una sigla (ej. O.N.U.), que no serán identificados como separadores de frases, y los puntos al final de una frase. Sin embargo, en un texto en tipografía pobre, el sistema no sabe identificar un final de frase si el último nombre es una sigla.

*Ejemplo:*

LA MOCION SE HA VOTADO EN LA O.N.U. EL DELEGADO FRANCES HA PUESTO EL VETO. será considerada como una sola frase.

.3 Indización

El sistema reconoce los descriptores analizando el título y el resumen en diferentes etapas.

-Identificación de las frases dentro del texto (y especialmente dentro del resumen) por reconocimiento de los signos de puntuación.

*Ejemplo:* documento 10.352:

1) Ordenador y biblioteca.
2) La informática para los sistemas documentales ha producido progresos muy grandes en las bibliotecas de facultad.
3) Los programas de gestión de préstamo son muy utilizados en la Universidad francesa.

-Identificación de cada una de las palabras verificando su existencia dentro del thesaurus y dentro del diccionario de base de la lengua, aplicando las reglas de derivación morfológica.

*Ejemplo:*

ORDENADOR/NOMBRE DESCR
Y/CONJ
BIBLIOTECA/NOMBRE DESCR
LA/ART
INFORMÁTICA/NOMBRE DESCR
PARA/PREP
LOS/ART
SISTEMA/NOMBRE - NO DESCR
DOCUMENTAL/ADJ
DOCUMENTACION/NOMBRE DESCR
HACER/VERBO
PRODUCIR/VERBO

PROGRESO/NOMBRE DESCR
MUY/ADV
GRANDE/ADJ
EN/PREP
LAS/ART
BIBLIOTECA/NOMBRE DESCR
DE/PREP
FACULTAD/NOMBRE NO DESCR
LOS/ART
PROGRAMA/NOMBRE DESCR
DE/PREP
GESTIÓN/NOMBRE NO DESCR
DE/PREP
PRESTAMO/NOMBRE NO DESCR
SON/NOMBRE
SON/VERBO
MUY/ADV
UTILIZADO/ADJ
UTILIZACION/NOMBRE DESCR
EN/PREP
LA/ART
UNIVERSIDAD/NOMBRE DESCR
FRANCES/ADJ.

-Desambiguación de los términos ambiguos explorando su contexto dentro de la frase y determinando su tipo.

*Ejemplo:*

del par

SON/NOMBRE
SON/VERBO
sólo se conserva el segundo, ya que la palabra SON está precedida por el nombre PRESTAMO y seguido por el adverbio MUY.

-Identificación de las palabras compuestas, gracias a las relaciones COMPUESTO POR/COMPONENTE DE y a la aplicación de las reglas de reestructuración.

*Ejemplo:*

se reconocen:

INFORMATICA DOCUMENTAL/NOMBRE DESCR
SISTEMA DOCUMENTAL/NOM DESCR

-Aplicación de las reglas de transformación semántica

*Ejemplo:*

+se reemplaza

PRESTAMO/NOMBRE NO DESCR
por
CIRCULACION/NOMBRE DESCR

+se añade

DOCUMENTACION/NOMBRE DESCR
a
SISTEMA DOCUMENTAL/NOMBRE DESCR

+se reemplaza

BIBLIOTECA DE FACULTAD/NOMBRE - NO DESCR
por
BIBLIOTECA UNIVERSITARIA/NOMBRE DESCR

-Eliminación de los términos no-descriptores.

*Ejemplo:*
La lista queda:

1ORDENADOR
2BIBLIOTECA
3INFORMÁTICA
4DOCUMENTACION
5INFORMATICA DOCUMENTAL
6SISTEMA DOCUMENTAL
7DOCUMENTACION
8BIBLIOTECA
9BIBLIOTECA UNIVERSITARIA
10PROGRAMA
11CIRCULACION
12UTILIZACION
13UNIVERSIDAD
14FRANCIA.

-Eliminación de los descriptores simples que forman parte de la composición de descriptores pluritérminos, así como los dobles usos.

*Ejemplo:*
La lista queda:

1ORDENADOR
2DOCUMENTACION
3INFORMATICA DOCUMENTAL
4SISTEMA DOCUMENTAL
5BIBLIOTECA UNIVERSITARIA
6PROGRAMA
7CIRCULACION
8UTILIZACION
9UNIVERSIDAD
10FRANCIA.

-Se añaden los genéricos, al primer nivel, de cada descriptor conservado, explotando las relaciones jerárquicas del thesaurus.

*Ejemplo:*
La lista queda:

1ORDENADOR
2INFORMATICA
3DOCUMENTACION

4SISTEMA DE INFORMACION
5INFORMATICA DOCUMENTAL (que tiene el mismo genérico que 1: INFORMATICA)
6SISTEMA DOCUMENTAL (mismo genérico que 3: SISTEMA DE INFORMACION)
7BIBLIOTECA UNIVERSITARIA
8PROGRAMA (mismo genérico que 1)
9CIRCULACION (genérico: SISTEMA DOCUMENTAL, ya citado)
10UTILIZACION
11UNIVERSIDAD
12ESCUELA
13FRANCIA
14EUROPA.

-Validación por parte del documentalista de la indización propuesta por el sistema: en efecto, el documentalista puede:

+aceptar;
+rechazar;
+completar

los descriptores conservados por el sistema.

*Ejemplo:*

para la lista anterior, el documentalista conservará, por ejemplo, los descriptores nº 5 - 7 - 8 - 9 - 13.

## 4. **Organización de los boletines de índices por materias**

Un boletín de índices por materias es un boletín bibliográfico editado periódicamente (por ejemplo, todos los meses), que contiene las referencias de los documentos registrados en un sistema documental durante el tiempo que pasa entre dos ediciones sucesivas del boletín, y que es usado para búsquedas documentales sin la intervención de un ordenador.

Cada entrada del índice contiene:

-uno o más términos de indización, puestos como encabezamiento y localizables gracias a su ordenación alfabética;
-una o más referencias de documentos

La referencia de un documento está formada, según el tipo de índice, por:

-o bien un número de orden que remite a un resumen del documento dentro de un boletín analítico o a la referencia completa del documento (autor, título, fuente...) dentro de un boletín signalético;
-o la referencia parcial de ese documento: autor y título, por ejemplo;
-o la referencia completa: autor, título, fuente;
-o bien la referencia (autor, título, fuente), completada por la lista de descriptores asignados al documento en el momento de su indización.

Cada referencia aparece dentro del boletín de índices tantas veces cuantos términos de indización contiene, en la posición alfabética de esos términos. En muchos índices sólo se conservan como entradas alfabéticas del boletín algunos términos de indización seleccionados.

Se distinguen:

-los índices de vocabulario libre;
-los índices de vocabulario controlado.

.1 Indices de vocabulario libre

Un índice de vocabulario libre es aquel en el que las entradas están formadas:

-en todos los casos: por palabras clave extraídas del título de los documentos en el momento de su indización automática;
-a veces, en pocos sistemas, se completan con descriptores libres asignados por los indizadores en el momento de la indización humana de los documentos, para enriquecer el título, cuando este último tiene un contenido informativo muy pobre.

El índice de vocabulario libre se presenta bajo una de las formas siguientes:

-*índice KWAC* (*K*ey *W*ord *A*nd *C*ontext)

El título de cada documento se ordena alfabéticamente, de forma sucesiva para cada una de sus palabras no vacías; estas palabras significativas se aíslan de su contexto, y se ponen como «encabezamiento», es decir, al comienzo de la referencia o del grupo de referencias que caracterizan.

Ejemplo: el documento 22.347, titulado «La utilización de los scanners en los hospitales» se encontrará en tres lugares del índice:

```
HOSPITALES
 22.347  La utilización de los scanners en los hospitales
..
..
SCANNERS
 22.347  La utilización de los scanners en los hospitales
..
..
UTILIZACION
 22.347  La utilización de los scanners en los hospitales
```

-*índice KWIC* (*K*ey *W*ord *I*n *C*ontext)

El título del documento se centra alrededor de la palabra sobre la que se realiza la ordenación alfabética. Al quedar alineadas en columna las palabras sobre las que se realiza la ordenación, a veces destacadas tipográficamente,

es posible suprimir la línea en la que se sitúa el encabezamiento dentro de un KWAC. Cuando un título es demasiado largo, se corta a la derecha o a la izquierda, o a ambos lados, siempre con el fin de economizar el espacio, dedicando una sola línea a cada aparición de cada referencia dentro del índice:

Ejemplo:

```
 22.347  scanners en los hospitales
..
..
 22.347  lización de los scanners en los hospitales
..
..
 22.347              la utilización de los scanner
```

-*índice KWOC* (*K*ey *W*ord *O*ut of *C*ontext)

Se presenta como el KWAC, pero la palabra que figura como encabezamiento se extrae realmente del título y no aparece en él.

Ejemplo:

```
HOSPITALES
 22.347  La utilización de los scanners en los ...
..
..
SCANNERS
 22.347  La utilización de los ... en los hospitales
..
..
UTILIZACION
 22.347  La ... de los scanners en los hospitales
```

.2 Indice de vocabulario controlado

Un índice de vocabulario controlado es aquel cuyas entradas están formadas por descriptores extraídos de un thesaurus o de una lista de autoridades, que caracterizan el contenido de los documentos cuyas referencias figuran en el boletín.

Se distinguen cuatro tipos de índices de vocabulario controlado:

-los índices basados en la indización completa de los documentos;
-los índices basados sólo en los descriptores principales de los documentos:

+con entradas simples: cada entrada está formada por un descriptor único;
+con entradas complejas, sin sintaxis: cada entrada está formada por dos o más descriptores coordinados, simplemente yuxtapuestos;
+con entradas complejas con sintaxis: cada entrada está formada por dos o más descriptores coordinados, ligados entre sí por indicadores de rol.

.1 Indice completo

El modelo más corriente de índice completo es el índice diccionario o índice de columnas.

Los términos de encabezamiento son los descriptores del thesaurus o de la lista de autoridades, ordenados alfabéticamente.

Bajo cada descriptor figuran las referencias de todos los documentos indizados con ese descriptor, en indización principal o secundaria, durante el período de tiempo cubierto por el índice.

Cada una de las referencias contiene:

-en todos los índices diccionario, el número de orden del documento dentro de un boletín analítico o signalético en el que figura la descripción completa del documento: título, autor, fuente, a veces resumen. Los números de los documentos se registran en seis columnas: en la primera se encuentran los números que terminan en 0; en la segunda los números cuya cifra de las unidades es 1; y así hasta la décima columna, que contiene los números de documentos que terminan en 9. El documento 10.345, por ejemplo, se incluye en la columna 5;
-en algunos boletines de índice diccionario, una o más informaciones complementarias:

+el peso del descriptor en la indización del documento: principal o secundario; notación: por ejemplo, P para principal y S para secundario.;
+el número de vínculo asignado a veces al descriptor en la cadena de indización; por ejemplo: (1) para los descriptores de la primera sub-cadena;
+el índice de rol asignado a veces al descriptor dentro del documento; ejemplo: O para origen, D para destino;
+un código de lengua para indicar la lengua del documento en la que figura el descriptor.

El índice diccionario presenta, en comparación con todas las demás formas de índices publicados en papel, dos ventajas destacadas:

-se basa en la indización completa de los documentos por medio de descriptores controlados;

-se organiza de manera que permite una búsqueda booleana de los documentos pertinentes: basta con examinar, columna por columna, las referencias descritas por los descriptores de la consulta, para recuperar los documentos indizados con uno y otro (intersección), con el uno o el otro (unión), o con el uno pero no con el otro (negación) de los descriptores que se requieren.

En cambio, es extremadamente desagradable si se compara con los otros índices, ya que es el único en el que la información sobre los documentos se limita a números, a veces seguidos de algunas informaciones complementarias codificadas; por otra parte, permite una búsqueda mucho menos ágil que con las bases de datos documentales de uso conversacional.

.2 Indice parcial de entradas simples

Los términos de encabezamiento son los descriptores del thesaurus o de la lista de autoridades, ordenados alfabéticamente.

Bajo cada descriptor figuran sólo las referencias de los documentos indizados por medio de esos descriptores como indización principal durante el período de tiempo cubierto por el índice.

Cada una de esas referencias contiene:

-como mínimo, el título del documento indizado por medio del descriptor del encabezamiento;
-y, además, en la mayor parte de los boletines de índices:
+el autor o los autores,
+la fuente: lugar, editorial, o título de revista, volumen, número,
+la fecha de publicación,
-y, por último, pero muy raramente:
+listado de los demás descriptores principales,
+o un resumen muy breve.

Estas informaciones ocupan un espacio relativamente amplio (de una a varias líneas de texto), mientras que cada uno de los números de los documentos que figuran en el índice de columnas sólo ocupa la décima parte de una línea. Esto explica que no resulta posible basar estos índices en la indización completa de los documentos, pues el boletín resultaría de un tamaño exagerado.

Estos índices son mucho más amigables que los índices de columnas, en la medida en que la información que proporcionan es directamente explotable; sin embargo, no permiten combinar varios descriptores para formar una ecuación booleana. Por este motivo se utilizan más para la puesta al día en conocimientos (current awareness) que para la búsqueda retrospectiva propiamente dicha.

.3 Indice parcial de entradas compuestas sin sintaxis

Este tipo de índice combina, en cierta medida, las ventajas de:

-los índices de columnas, en la medida en que permite combinar dos o más descriptores cuando se realiza una búsqueda, pero con ciertas limitaciones en comparación con los índices de columnas:

+solamente indización principal;
+sólo son posibles algunas combinaciones de descriptores;

-los índices parciales de entradas simples, de los que toman la (relativa) riqueza informativa de las referencias publicadas.

Estos índices se basan en tres niveles (tres pesos) de indización de los documentos referenciados:

-indización mayor: los dos, tres o cuatro descriptores que expresan el tema principal del documento y que se situarán uno a uno en encabezamientos alfabéticos dentro del boletín de índices;
-indización intermedia: de uno a tres descriptores complementarios, que expresan los conceptos importantes del documento, varias combinaciones de los cuales se emplazarán conjuntamente en entradas alfabéticas, bajo los descriptores mayores del encabezamiento;
-indización menor: de cinco a diez descriptores suplementarios, que expresan conceptos relativamente poco importantes dentro del documento (pero acerca de los cuales el documento aporta, sin embargo, informaciones útiles); estos descriptores no aparecen ni como encabezamientos ni como entradas; en algunos índices, pueden aparecer dentro de las referencias.

Consideremos, a título de ejemplo, la indización completa de un documento titulado «La formación de los gestores y usuarios de información en la enseñanza secundaria»:

descriptores mayores,
o encabezamientos:FORMACION; SISTEMA DE INFORMACION; ENSEÑANZA SECUNDARIA;
descriptores intermedios:DOCUMENTALISTA; DOCENTE; ALUMNO;
descriptores menores:BUSQUEDA DOCUMENTAL; MICRO-INFORMATICA.

Los descriptores de las dos primeras categorías, que constituirán las entradas alfabéticas del índice, son ante todo dispuestas en una secuencia determinada, que, según los sistemas, será:

-o bien en cualquier orden, no significativo por tanto (orden de aparición dentro del documento, u orden alfabético);
-o bien significativo: una secuencia significativa facilitará la búsqueda, en la medida en que constituya una breve notación del contenido del documento, que el usuario podrá comprender sin

necesidad de dirigirse a los títulos de los documentos bajo las entradas del índice.

El carácter significativo de una secuencia de descriptores se puede obtener de dos maneras: situando los descriptores:

+ya sea en un contexto creciente: los primeros descriptores de la cadena son muy específicos, y cada uno de los descriptores siguientes está considerado dentro de un contexto más amplio que el que le precede;

+ya sea en un contexto decreciente, del descriptor más general, al principio de la cadena, hasta el descriptor que representa el punto de vista más específico dentro del documento.

La mayor parte de los productores de índices adoptan el método del contexto creciente.

Ejemplo:

-secuencia no significativa:
ALUMNO; DOCENTE; DOCUMENTALISTA; ENSEÑANZA SECUNDARIA; FORMACION; SISTEMA DE INFORMACION.

-secuencia en contexto creciente:
FORMACION; ALUMNO; DOCUMENTALISTA; DOCENTE; SISTEMA DE INFORMACION; ENSEÑANZA SECUNDARIA.

-secuencia en contexto decreciente:
ENSEÑANZA SECUNDARIA; SISTEMA DE INFORMACION; DOCENTE; DOCUMENTALISTA; ALUMNO; FORMACION.

Como puede verse, el contexto creciente es más «elocuente»; por otra parte, se puede hacer notar que el orden de los descriptores es casi el de los conceptos dentro del título del documento.

Por último, algunos autores prefieren organizar la secuencia de descriptores poniéndolos según el orden de las facetas que son relevantes; por ejemplo: fenómeno - proceso (FORMACION) - materiales - organización (ENSEÑANZA SECUNDARIA; SISTEMA DE INFORMACION) - ser vivo (DOCUMENTALISTA; DOCENTE; ALUMNO) - equipamiento - propiedad - disciplina.

La siguiente etapa en la construcción del índice consiste en permutar los descriptores de la secuencia, es decir, en obtener bajo cada encabezamiento una serie de entradas que corresponden a todas las formulaciones posibles de consultas del usuario para las que el documento es pertinente; la ventaja de este método es que, si el usuario formula una consulta (post) coordinando una serie de descriptores y no encuentra en el índice las entradas (pre) coordinadas correspondientes, puede estar casi seguro de que no existen documentos que respondan a su consulta.

Desgraciadamente, el número de permutaciones que se puede obtener a partir de una cadena de n descriptores se expresa por el factorial de n, [n!=1x2x3...x (n-1) x n], en el caso de que todas las entradas contengan todos los descriptores.

En el ejemplo anterior, en el que la cadena contiene 6 descriptores mayores e intermedios, el número de permutaciones, y por tanto de entradas alfabéticas posibles, es de 720. Evidentemente no resulta posible conservar todas estas posibilidades por motivos de espacio y de facilidad de uso. Los productores de índices sólo admiten una pequeña parte de las posibles permutaciones; como es natural, se esfuerzan por conservar las permutaciones que parecen más significativas, produciéndolas por ordenador.

Keen (1977) ha enumerado una serie de algoritmos de manipulación de índices corrientes:

+*Rotación*

Se crea una entrada por cada descriptor de encabezamiento; cada entrada recoge todos los descriptores mayores e intermedios de la cadena inicial, según la secuencia de esta cadena.

*Aplicación:* el Indice del Bulletin Signalétique nº 101 del CNRS.

Ejemplo:
ENSEÑANZA SECUNDARIA
...
FORMACION - ALUMNO - DOCUMENTALISTA - DOCENTE - SISTEMA DE INFORMACION - ENSEÑANZA SECUNDARIA
...
...
FORMACION
...
FORMACION - ALUMNO - DOCUMENTALISTA - DOCENTE - SISTEMA DE INFORMACION - ENSEÑANZA SECUNDARIA
...
...
SISTEMA DE INFORMACION
...
FORMACION - ALUMNO - DOCUMENTALISTA - DOCENTE -SISTEMA DE INFORMACION - ENSEÑANZA SECUNDARIA

+*Ciclo*

Se crea una sola entrada por cada descriptor de encabezamiento; cada entrada recoge todos los descriptores de la cadena inicial; dentro de cada entrada, la secuencia se inicia a partir del encabezamiento, formando un bucle, al principio de la cadena, alrededor del primer descriptor de la secuencia inicial.

*Aplicación:* el índice de Oceanic Abstracts.

Ejemplo:
ENSEÑANZA SECUNDARIA
...
(ENSEÑANZA SECUNDARIA) - FORMACION - ALUMNO - DOCUMENTALISTA - DOCENTE - SISTEMA DE INFORMACION
...
...

FORMACION
...
(FORMACION) - ALUMNO - DOCUMENTALISTA - DOCENTE -SISTEMA DE INFORMACION - ENSEÑANZA SECUNDARIA
...
...
SISTEMA DE INFORMACION
...
(SISTEMA DE INFORMACION) - ENSEÑANZA SECUNDARIA - FORMACION - ALUMNO - DOCUMENTALISTA - DOCENTE

*Nota:* en este tipo de índice se puede suprimir el primer descriptor de la cadena, pues aparece ya en el encabezamiento; es la razón por la que lo hemos puesto entre paréntesis.

+*Articulación*

Se crea una sola entrada por cada descriptor de encabezamiento; cada entrada recoge todos los descriptores de la cadena inicial. La entrada cuyo descriptor de encabezamiento corresponde al primer descriptor de la cadena recoge exactamente la secuencia inicial. En las otras entradas, el descriptor de encabezamiento está seguido por el descriptor que le precede en la cadena inicial (se invierte, pues, la secuencia de estos dos descriptores); tras este descriptor la secuencia continúa como en el ciclo.

*Aplicación:* el índice de Chemical Abstracts.

Ejemplo:
ENSEÑANZA SECUNDARIA
...
(ENSEÑANZA SECUNDARIA) - SISTEMA DE INFORMACION - FORMACION - ALUMNO - DOCUMENTALISTA - DOCENTE
...
...
FORMACION
...
(FORMACION) - ALUMNO - DOCUMENTALISTA - DOCENTE -SISTEMA DE INFORMACION - ENSEÑANZA SECUNDARIA
...
...
SISTEMA DE INFORMACION
...
(SISTEMA DE INFORMACION) - DOCENTE - ENSEÑANZA SECUNDARIA - FORMACION - ALUMNO - DOCUMENTALISTA

+*Truncamiento progresivo*

Se crea una sola entrada por cada descriptor de encabezamiento; las entradas no recogen todos los descriptores de la cadena inicial: cada entrada recoge la palabra del encabezamiento, seguida de la lista invertida de sólo aquellos descriptores que le preceden en la cadena inicial

*Aplicación:* British Technology Index.

Ejemplo:
ENSEÑANZA SECUNDARIA
...
(ENSEÑANZA SECUNDARIA) - SISTEMA DE INFORMACION - DOCENTE - DOCUMENTALISTA - ALUMNO - FORMACION
...
...
FORMACION
...
(FORMACION)
...
...
SISTEMA DE INFORMACION
...
(SISTEMA DE INFORMACION) - DOCENTE - DOCUMENTALISTA - ALUMNO - FORMACION

+*Permutación selectiva, de dos en dos*

Se crean tantas entradas por cada descriptor de encabezamiento como descriptores hay en la cadena, menos uno; cada entrada recoge un solo descriptor de la lista; cada descriptor de la cadena aparece por turno bajo el encabezamiento.

*Aplicación:* Permuterm Index del Science Citation Index.

Ejemplo:
ENSEÑANZA SECUNDARIA
...
ALUMNO
...
DOCENTE
...
DOCUMENTALISTA
...
FORMACION
...
SISTEMA DE INFORMACION.
...
...
FORMACION
...
ALUMNO
...
DOCENTE
...
DOCUMENTALISTA
...
ENSEÑANZA SECUNDARIA
...
SISTEMA DE INFORMACION.

SISTEMA DE INFORMACION
...
ALUMNO
...

DOCENTE
...
DOCUMENTALISTA
...
ENSEÑANZA SECUNDARIA
...
 FORMACION
...
...

+*Combinación selectiva*

La primera entrada recoge exactamente la cadena inicial; las otras entradas recogen una parte variable de la cadena inicial, en la misma secuencia.

Un algoritmo particular de combinación selectiva es SLIC (Selective Listing in Combinaison):

-las cadenas están limitadas a 5 descriptores como máximo;
-se ordenan alfabéticamente esos descriptores;
-la selección de las combinaciones es la siguiente:

ABCDE (cadena inicial)
ABCE
ABDE
ABE
ACDE
ACE
ADE
AE
BCDE
BCE
BDE
BE
CDE
CE
DE
E

Este sistema presenta todas las ventajas de una permutación completa de todos los descriptores de la cadena, evitando además la gran cantidad de entradas de una permutación normal (5!=120), pues limita a 16 el número de entradas.

La búsqueda en tal índice se realiza ordenando, en primer lugar, por orden alfabético los descriptores de la consulta. Si estos descriptores están presentes en una cadena del índice, se encontrará una entrada correspondiente a la consulta, y esto sea cual sea el resto de los descriptores presentes en la cadena.

*Aplicación:* el índice del volumen 25 (1974) del Journal of the American Society for Information Science.

.4 Indice parcial de entradas compuestas con sintaxis

En comparación con los índices anteriormente descritos, los índices con sintaxis presentan la ventaja de designar el rol de los descriptores dentro de las cadenas de indización que figuran como entradas, lo que aumenta notablemente su inteligibilidad.

En el British Technology Index, por ejemplo:

- una coma entre dos descriptores indica que el segundo se encarga de precisar el primero:

welding, arc (soldadura por arco voltaico)

- dos puntos entre dos descriptores indican su independencia recíproca:

streets: decoration

- un punto y coma separa dos descriptores, de los cuales el primero representa un material y el segundo un producto:

steel; pipelines.

En el sistema Precis (*PRE*served *C*ontext *I*ndex *S*ystem), de la British Library, el número y la significación de roles son más importantes.

Ejemplo:

(2), situado delante de un descriptor, significa que ese descriptor designa una acción que es el núcleo alrededor del cual se articula la cadena de indización.
(1) designa el descriptor que es objeto de esa acción.
(3) designa el descriptor agente de la acción.
$V y $W introducen locuciones prepositivas que permiten deshacer las ambigüedades.

La cadena de indización
(1) interpretación del actor
(2) enseñanza $W de
(s) importancia de la $V del $W en la
(3) improvisación
producirá especialmente las entradas siguientes:

- Interpretación del actor. Enseñanza. Importancia de la improvisación
- Improvisación. Importancia en la enseñanza de la interpretación del actor

que son evidentemente mucho más significativas que una simple yuxtaposición de esos descriptores.

Esta ventaja tiene evidentemente como precio un coste más elevado en la indización humana.

## 5. **Organización de los ficheros de búsqueda**

De forma esquemática, el núcleo de una base de datos documental clásica está formada por tres ficheros o conjuntos de datos:

- un *fichero diccionario*, que contiene la lista de los criterios de búsqueda introducidos hasta el momento dentro de los documentos de la base (autores, organismos, palabras clave, descriptores libres o

posibles descriptores) y los descriptores controlados registrados a priori. Se crea en él un registro lógico para cada criterio o grupo de varios criterios. Este registro contiene:
+el enunciado del criterio de búsqueda tal cual;
+un puntero hacia el registro correspondiente a ese criterio, que se encuentra en el fichero inverso (ver más adelante);
+en algunos sistemas, además, cuando se ha registrado un thesaurus en el fichero diccionario:
->un puntero hacia cada uno de los demás descriptores, dentro del fichero diccionario, con los que el descriptor del registro tiene una relación semántica; este puntero se completa con una información sobre la naturaleza de la relación: jerárquica, asociativa...

*Nota:* en otros sistemas la red de relaciones semánticas se registra:
+o bien en un fichero especializado (ejemplo: fichero de relaciones de MISTRAL);
+o bien en el fichero diccionario y, además, parcialmente en el fichero inverso, para acelerar la consulta (ejemplo: BASIS);

-un *fichero inverso* de los criterios de búsqueda y especialmente de las palabras clave y descriptores. En él se crea un registro lógico para cada criterio de búsqueda. Este registro contiene:
+un puntero hacia cada una de las noticias caracterizadas por ese criterio de búsqueda, ordenadas en el fichero bibliográfico (ver más adelante);
+en algunos sistemas, junto a cada puntero:
->el número de vínculo y/o el índice de rol asignado a ese criterio dentro del documento correspondiente al puntero;
->una información topológica sobre la localización del criterio dentro del documento; nº de campo, nº de orden de frase o de sub-campo dentro del campo, nº de orden de palabra dentro de la frase o dentro del sub-campo.

*Nota:* en otros sistemas, la localización de los criterios se registra en un fichero especializado (ejemplo: fichero topológico de STAIRS).

-Un *fichero bibliográfico*, que contiene el texto de las noticias documentales registradas en la base: autores, títulos, fuente..., descriptores, resumen y a veces el texto completo del documento. A cada noticia le corresponde un registro lógico. Cada registro lógico se encuentra en un emplazamiento bien determinado del fichero bibliográfico; a ese emplazamiento le corresponde una «dirección», que es el valor del puntero contenido dentro del

registro lógico de todos los criterios de búsqueda que caracterizan el documento y figuran en el fichero inverso.

La actualización de la base de datos se hace tras:

-el mecanografiado de la noticia que contiene la descripción bibliográfica del documento, los descriptores asignados por el documentalista y a veces su resumen y/o su texto completo;
-la validación, especialmente de la existencia de descriptores controlados dentro del fichero diccionario (cf. § 2);
-la indización automática de palabras clave del título, del resumen y a veces del texto (cf. § 3.2)

Esta actualización se realiza de la forma siguiente:

-se añaden las nuevas noticias dentro del fichero bibliográfico;
-se extraen los criterios de búsqueda: palabras clave, descriptores libres y controlados, autores..., a los que se añade el valor de la dirección de la noticia dentro del fichero bibliográfico, así como, según los sistemas, los datos topológicos y los índices de vínculo y de rol;
-se consulta el fichero diccionario, para encontrar en el fichero inverso la dirección (puntero) de los criterios de búsqueda ya existentes dentro del fichero diccionario;
-se actualiza el fichero diccionario:
    +se añaden nuevos registros para los criterios de búsqueda nuevos;
-se actualiza el fichero inverso:
    +se añaden nuevos punteros, con los datos topológicos y los roles y los vínculos asociados, dentro de los registros relativos a los criterios de búsqueda ya existentes;
    +se añaden nuevos registros para los criterios de búsqueda nuevos.

*Nota:* la organización que acabamos de describir se encuentra presente en prácticamente todos los sistemas documentales de hoy; se apoya fundamentalmente en el principio del fichero inverso. Un número, todavía limitado, de sistemas documentales se basan en los SGBD (Sistemas de Gestión de Bases de Datos); la organización de ALEXDOC (cf. § 3.6.3.1) es típica de esta aproximación.

## 6. **Búsqueda Documental**

Una búsqueda documental se realiza a través de una serie de pasos y puede comprender hasta seis etapas:

-el usuario toma conciencia de una necesidad de información sobre un punto particular, más o menos preciso;

-comunica esta necesidad, en forma de una «pregunta» al documentalista, en el caso, que es el más frecuente, de que el usuario delegue la búsqueda documental en un documentalista;
-el documentalista, con la ayuda del usuario, enuncia, en lenguaje natural, los conceptos sobre los que se busca la información;
-el documentalista selecciona el sistema o sistemas documentales que va a interrogar: bases de datos y/o boletines de índices;
-para cada uno de los sistemas documentales que se interrogan:

+se accede al sistema;
+se traduce la expresión de los conceptos de la consulta a una formulación por palabras clave y descriptores libres y/o controlados propios del sistema documental;
+se realiza una ecuación con la consulta, es decir, se establecen entre las palabras clave y los descriptores las relaciones sintácticas permitidas por el sistema: lógica de Boole, proximidad...;
+la búsqueda se extiende, con la ayuda del ordenador, a otros documentos, que no corresponden directamente con la ecuación de búsqueda, pero son susceptibles, sin embargo, de responder a la consulta del usuario;
+se extraen las referencias de los documentos que responden a la ecuación de búsqueda y a sus extensiones;

-para la totalidad de los sistemas documentales que se interrogan:

+se eliminan los duplicados, es decir, las referencias localizadas dos veces, porque están referenciadas por dos o más sistemas documentales de los que se han interrogado;
+se eliminan las referencias que, a la vista de su título, de su resumen y a veces de su texto, no responden a la petición del usuario;
+se ajusta la bibliografía, poniendo en un mismo orden los datos bibliográficos obtenidos de los diferentes sistemas interrogados y organizando las referencias de acuerdo con un criterio único (autor, fecha...);
+se comunica la bibliografía al usuario;
+el usuario examina la bibliografía y selecciona los documentos que efectivamente va a consultar;
+se adquieren los documentos que se van a consultar, ya sea en la biblioteca del servicio de documentación del usuario, o bien encargando un original o una copia a un librero, al editor, a una biblioteca pública o a los distribuidores de las bases de datos consultadas;

+el usuario examina los documentos recibidos y emite una apreciación final de pertinencia sobre la bibliografía obtenida.

En el transcurso de estos procesos, los lenguajes de indización pueden intervenir en seis momentos:

-selección de los sistemas documentales que se van a interrogar;
-enunciado de los conceptos de la pregunta, en lenguaje natural;
-traducción a un lenguaje de indización;
-formulación de la ecuación;
-extensión asistida por el ordenador;
-apreciación final de pertinencia.

.1 Selección de los sistemas documentales que se van a interrogar

Para determinar, entre las miles de bases de datos accesibles en el mundo, las que pueden ser más susceptibles de responder a su consulta, el usuario(*)14 puede acudir a varios métodos, especialmente a:

-los repertorios, en papel o en ordenador, de bases de datos, generalmente organizados siguiendo el principio de un índice de materias selectivo: cada base de datos está indizada por medio de un número limitado de descriptores genéricos (de uno a diez), que describen su contenido de manera muy sintética.
-los índices de materias analíticas en ordenador, creados por algunos centros distribuidores y que están formados por una acumulación de todas las palabras clave y descriptores que existen en todas las bases de datos gestionadas por el distribuidor: el usuario introduce en su terminal los términos que caracterizan su necesidad, y el sistema le responde señalando la frecuencia de indización de cada uno de esos términos en cada una de las bases de datos en las que intervienen al menos una vez dentro del fichero inverso; cuanto más elevada es la frecuencia de indización de un término dentro de una base de datos, mayor interés tiene consultar esa base.

.2 Elección de los conceptos de la consulta

La formulación de una pregunta es la operación simétrica de la indización de un documento; esta última consistía en enumerar los conceptos tratados dentro del documento; esos conceptos estaban entonces identificados por palabras clave y/o descriptores ensamblados en una cadena de indización. La

(*)14 Designaremos a continuación «usuario» (de sistemas documentales, se sobreentiende) a aquella persona, ya sea usuario final o documentalista, que de hecho lleva a cabo la búsqueda documental. Hoy día, en la inmensa mayoría de los casos, esa persona es el documentalista.

formulación de la pregunta comienza también por establecer una lista, tan completa como sea posible, de los conceptos que constituyen la pregunta; a continuación estos conceptos son traducidos en palabras clave y/o descriptores que se reúnen en una ecuación de búsqueda.

Sin embargo, la enumeración de los conceptos que componen una pregunta exige una aproximación muy diferente de la que se lleva a cabo cuando se hace la lista de los conceptos que caracterizan el contenido de un documento. La indización de un documento es una operación reductora: el contenido de un documento de cinco a diez páginas queda considerablemente condensado cuando se representa por un número de ocho a doce descriptores, conforme a una tasa del 1/100 al 1/1.000. Por el contrario, lo más frecuente es que la formulación de una pregunta produzca una extensión del enunciado final: una petición, expresada por el enunciado de un número de dos a cinco conceptos, se enriquece de forma que se obtienen, como media, de cinco a doce palabras clave y/o descriptores, de acuerdo con una tasa de expansión del 100 al 300 %.

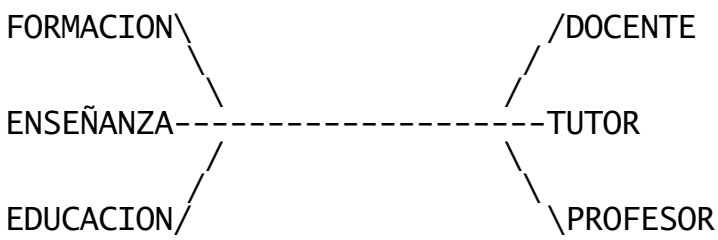
Este enriquecimiento se explica por el hecho de que un documento puede responder a una consulta abordando el tema de diversas formas posibles.

Ejemplo:

A una pregunta sobre la formación de los docentes pueden responder documentos indizados de diferentes formas:

-documento A:FORMACION; DOCENTE;
-documento B:ENSEÑANZA; TUTOR;
-documento C:EDUCACION; PROFESOR.

El análisis de la pregunta debe llevar, pues, a ampliarla, según lo que se denomina en lingüística estructural «eje paradigmático» del discurso. El paradigma es el conjunto de unidades lingüísticas situadas en el mismo lugar de una cadena de términos susceptibles de ser conmutados, a elección del hablante, para expresar diferentes puntos de vista:




El conjunto de términos «formación, enseñanza y educación» y el conjunto de términos «docente, tutor y profesor» constituyen cada uno un paradigma. En el interior de cada uno de esos conjuntos, cada término puede ser sucesivamente sustituido por otro para reconstruir una de las cadenas de indización pertinentes que caracterizan los documentos que responden a la pregunta:

FORMACION - DOCENTE
FORMACION - TUTOR
FORMACION - PROFESOR
ENSEÑANZA - DOCENTE
ENSEÑANZA - TUTOR
ENSEÑANZA - PROFESOR
EDUCACION - DOCENTE
EDUCACION - TUTOR
EDUCACION - PROFESOR.

En este ejemplo cada paradigma está formado por una faceta:

-un proceso, para el primero (FORMACION-ENSEÑANZA-EDUCACIÓN);
-una entidad personal, para el segundo (DOCENTE, TUTOR, PROFESOR).

En muchos casos una pregunta puede contener numerosos ejes paradigmáticos, aun cuando la consulta sólo trate en principio sobre un número limitado de conceptos, incluso sobre un sólo concepto.
Ejemplo:
A una pregunta sobre la descentralización administrativa, pueden responder:

-el documento A:DESCENTRALIZACION ADMINISTRATIVA;
-el documento B:AUTONOMIA REGIONAL;
-el documento C:ESTATUTO; REGION; PROVINCIA;
-el documento D:PODER DE DECISION; REGION;
-el documento E:COLECTIVIDAD DESCENTRALIZADA;
-el documento F:PODER LOCAL; AYUNTAMIENTO;

Este ejemplo muestra que el mismo tema, la descentralización del poder del Estado hacia colectividades locales, puede ser abordado en distintos documentos, todos pertinentes, según puntos de vista variados:

-el del proceso: la descentralización;
-el del resultado del proceso: el poder de decisión, el estatuto;
-el de la propiedad adquirida como fruto del proceso: la autonomía regional;
-el de las entidades relacionadas con el proceso: el poder local, la región, la provincia, el ayuntamiento.

Como se puede ver en estos ejemplos, la mayor parte de los conceptos que constituyen una pregunta están implícitos y deben ser explicitados cuando se formula la búsqueda; por el contrario, en el momento de la indización de un documento, sólo ha de ser explicitada una reducida minoría de los conceptos que se van a representar.
En principio, para buscar los conceptos implícitos en una pregunta se hace uso del razonamiento: el usuario experimentado sabe que el contenido de la mayoría de las preguntas debe ser enriquecido para cubrir todos los puntos de vista presentes dentro de los documentos que responderán a

esas preguntas; y trata de reconstruir mentalmente esos puntos de vista.

Durante la siguiente etapa de la búsqueda, todavía se podrá ampliar el número de conceptos de la pregunta por medio de una consulta a los ficheros diccionarios y al thesaurus, haciendo uso del ordenador y/o de las ediciones en papel de los thesaurus de los sistemas documentales interrogados (cf. § 6.3); se podrá enriquecer de nuevo la pregunta tras obtener las primeras referencias pertinentes, a la vista de su indización, durante una operación denominada «bucle de pertinencia» (cf. § 6.5.3).

.3 Traducción de los conceptos a un lenguaje de indización

Los conceptos en lenguaje natural, listados durante la etapa anterior, se comparan uno a uno con las entradas alfabéticas del fichero diccionario. Se pueden producir muchos casos, según que la expresión en lenguaje natural del concepto:

-corresponda exactamente con una entrada: palabra clave, descriptor libre, descriptor controlado, no-descriptor que reenvía a un descriptor: en este caso no hay problema, y el término puede ser conservado tal cual (a veces, si el sistema no lo hace por sí mismo, introduciendo manualmente el enunciado de un descriptor que corresponde a un no-descriptor).

Ejemplos:

| Conceptos en lenguaje natural | Entradas del fichero diccionario | Estatuto de la entrada | Términos de formul. de la consulta |
|---|---|---|---|
| región | región | palabra clave o descriptor | REGION |
| autonomía regional | autonomía regional | necesariamente descriptor | AUTONOMIA REGIONAL |
| municipalidad | municipalidad EM Ayuntam. | no-descriptor | AYUNTAMIENTO |

-corresponda parcialmente con una entrada, por el hecho de que se han dividido expresiones compuestas en palabras clave unitérminos; en este caso es necesario buscar dentro del diccionario el otro o los otros términos constituyentes; si son encontrados, se conservan todos los términos; si no, interesará no conservar ninguno.

Ejemplos:

| Conceptos en lenguaje natural | Entradas del fichero diccionario | Estatuto de la entrada | Términos de formul. de la consulta |
|---|---|---|---|
| poder de decisión | poder<br>decisión | palabra clave<br>palabra clave | PODER<br>DECISION |
| poder local | poder<br>localidad | palabra clave<br>palabra clave | nada |

-no corresponda con ninguna entrada: en este caso es necesario, como en el momento de la indización de un documento, intentar encontrar, mentalmente, una expresión equivalente, o consultar el thesaurus.

En los tres casos interesa enriquecer la formulación de la pregunta así obtenida consultando el thesaurus de la base de datos, y más en particular las relaciones jerárquicas asociativas de los descriptores ya encontrados: los términos suplementarios así descubiertos serán automáticamente expresados en lenguaje controlado, ya que provendrán del thesaurus.

Ejemplo:

Se podrá encontrar:

-al mirar AYUNTAMIENTO: el descriptor genérico: ORGANO AUTONOMO LOCAL;
-al mirar COLECTIVIDAD DESCENTRALIZADA: el descriptor específico: COLECTIVIDAD LOCAL;
-al mirar DESCENTRALIZACION ADMINISTRATIVA: el descriptor asociado TRANSFERENCIA DE COMPETENCIAS.

.4 Establecimiento de enlaces sintácticos entre los términos de indización

Hemos visto que en la inmensa mayoría de los sistemas documentales los descriptores utilizados para indizar los documentos son simplemente yuxtapuestos unos junto a otros.

La formulación de las preguntas exige la intervención de relaciones entre los diferentes términos (palabras clave y/o descriptores) de la consulta, destinadas a precisar la naturaleza de las combinaciones: «post-coordinación».

Mientras que la elaboración de la lista de términos de una pregunta se sitúa sobre el eje paradigmático del discurso, la fijación de las relaciones entre estos términos se emplaza en el eje sintagmático, es decir, en el eje sobre el que se combinan las unidades lingüísticas para dar un sentido a su agrupación. Es el motivo por el que esas relaciones se denominan «sintácticas».

Ejemplo:

Los dos descriptores FORMACION y DOCENTE pueden ser combinados en una pregunta, de manera que los documentos obtenidos en respuesta traten, según lo que desee el usuario:

-ya sea sobre la formación de los docentes;
-o bien sobre la formación de cualquiera, pero no de los docentes.

En el primer caso el operador de relación que se colocará entre los dos descriptores será: Y, y en el segundo caso: NO.

Según los sistemas de búsqueda, se pueden utilizar diferentes tipos de relaciones sintácticas, o bien ellas solas, o bien en combinación con otras:

-enlaces unitivos, siguiendo la lógica de Boole;
-truncamiento;
-proximidad topológica dentro del texto de la noticia o del documento;
-comparación.

.1 Lógica booleana

Es la lógica más clásica, usada en casi todos los sistemas documentales; procede de aplicar el álgebra de Boole.
Examinaremos:

-los operadores lógicos;
-la noción de ecuación de búsqueda;
-la noción de estrategia de búsqueda;
-las características cualitativas de la búsqueda.

.1 Operadores lógicos

Los enlaces entre los términos (descriptores o palabras clave) se denominan operadores.
Se distinguen tres operadores principales:

-el operador de intersección: Y: enlaza dos términos que deben estar obligatoriamente presentes en la indización de un documento para que éste sea considerado pertinente.

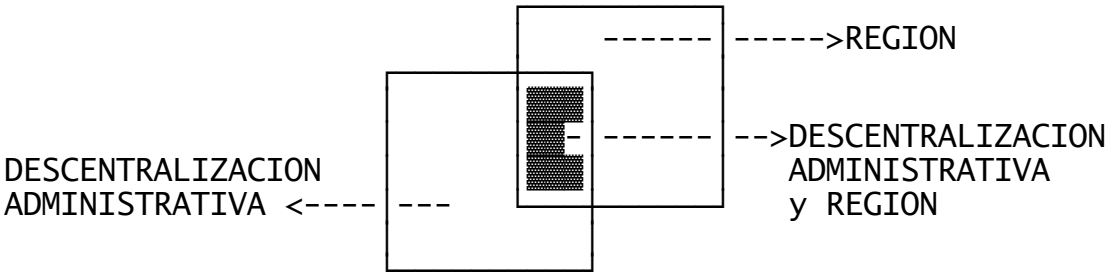
Símbolos: H,*,.

Ejemplo:
consulta:busco una documentación sobre la descentralización administrativa en las regiones
formulación:DESCENTRALIZACION ADMINISTRATIVA y REGION.

Esta formulación se puede representar por medio de un diagrama de Venn: si se representa por un círculo el conjunto de los documentos indizados por «DESCENTRALIZACION ADMINISTRATIVA» y por otro círculo el conjunto de los documentos indizados por «REGION», la intersección de los dos

círculos contiene el subconjunto de documentos que tratan a la vez sobre descentralización administrativa y región.

*Representación por diagrama de Venn:*




Sólo son pertinentes los documentos indizados a la vez por «DESCENTRALIZACION ADMINISTRATIVA» y «REGION» (zona punteada)

-el operador de unión: O: enlaza dos términos de los que o uno u otro, o los dos, deben estar presentes en la indización de un documento para que este último sea considerado pertinente.

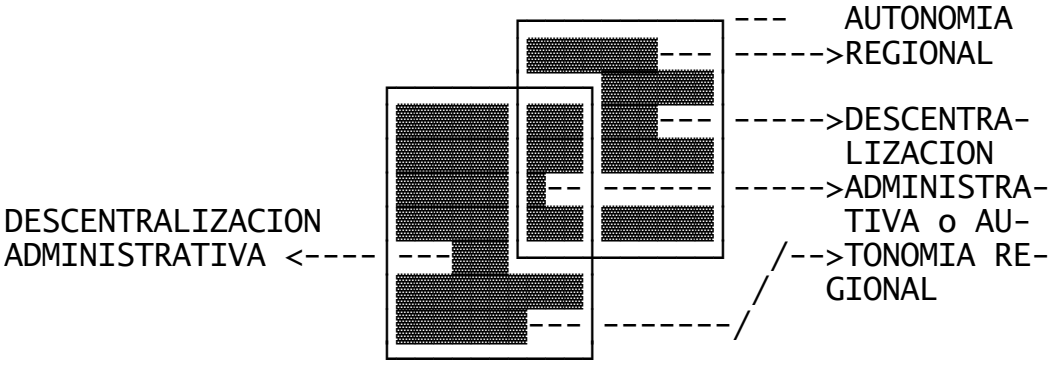
Símbolos: U, +

Ejemplo:

consulta:busco una documentación sobre la descentralización administrativa o sobre la autonomía regional

formulación:DESCENTRALIZACION ADMINISTRATIVA o AUTONOMIA REGIONAL.

*Representación:*




Todos los documentos indizados, ya sea por «DESCENTRALIZACION ADMINISTRATIVA», o por «AUTONOMIA REGIONAL» o a la vez por «DESCENTRALIZACION ADMINISTRATIVA» y «AUTONOMIA REGIONAL» son pertinentes.

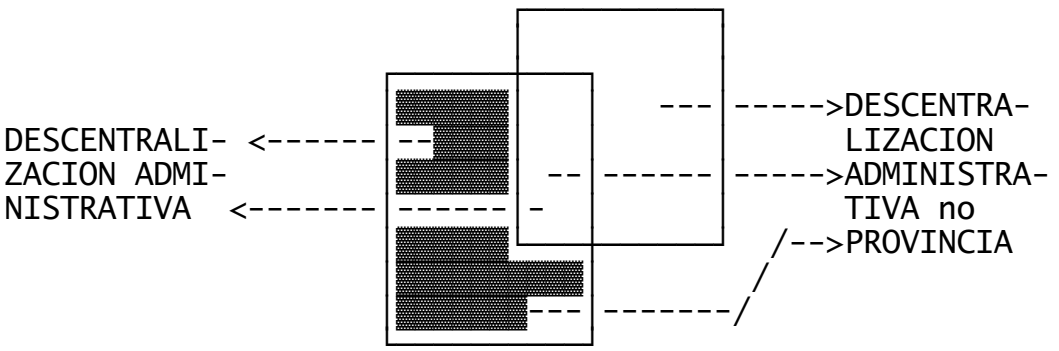
*Nota:* cuando se formula la consulta, es necesario atender al hecho de que la conjunción de coordinación *Y* en el lenguaje usual se traduce a veces por el operador *O* en lógica booleana.

Ejemplo:
consulta:busco una documentación sobre las regiones *y* los ayuntamientos
formulación:REGION o AYUNTAMIENTO

-el operador de complementación: NO: enlaza dos términos de los que el primero debe estar presente y el segundo ausente, en la indización de un documento, para que éste sea considerado pertinente.
Símbolo: -

Ejemplo:
consulta:busco una documentación sobre la descentralización administrativa, pero no a nivel de provincias
formulación:DESCENTRALIZACION ADMINISTRATIVA no PROVINCIA.

*Representación:*

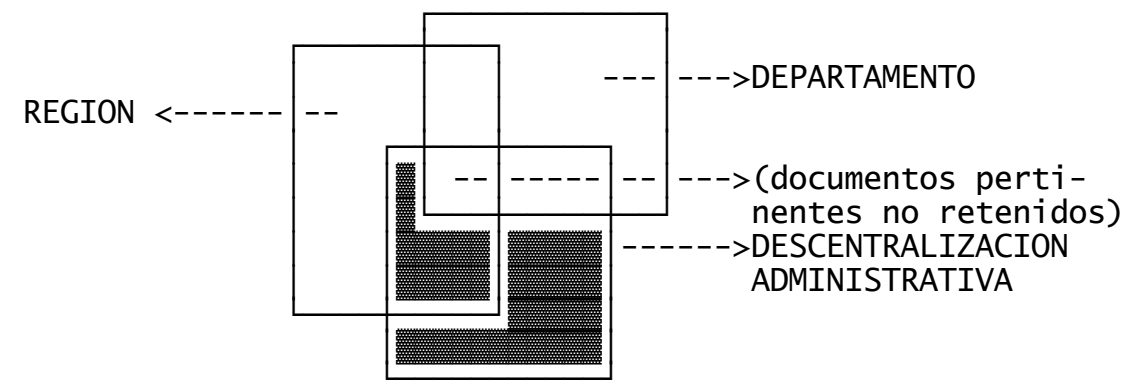



Sólo son pertinentes los documentos indizados por «DESCENTRALIZACION ADMINISTRATIVA» excluyendo los que están indizados a la vez por «PROVINCIA».

*Notas:*

1)Al utilizar este operador hay riesgo de descartar documentos pertinentes.

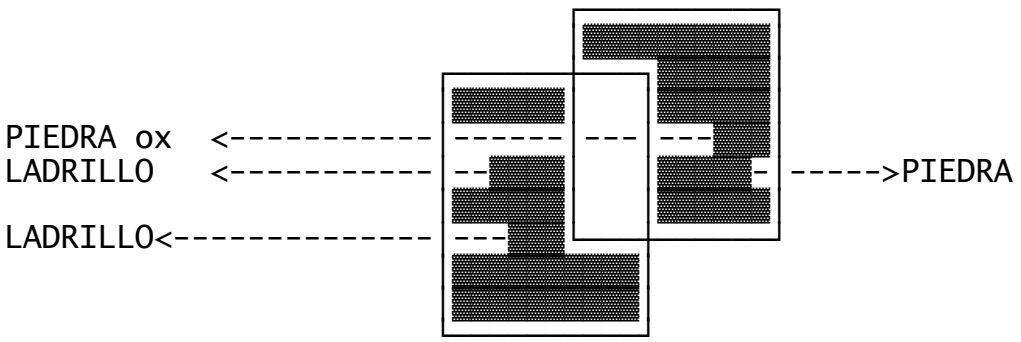
Ejemplo:en el conjunto precedente, aquellos que traten de provincia (y que, por tanto, serán eliminados por ello) y a la vez de región (tema para el que son pertinentes):



2)algunos sistemas, pocos, utilizan el O EXCLUSIVO, que enlaza dos términos de los que o uno u otro, pero no los dos, debe estar presente en la indización de un documento para que este último sea considerado pertinente.

Símbolo: U

Ejemplo:
consulta:busco una documentación sobre las construcciones en piedra o en ladrillo
formulación:PIEDRA ox LADRILLO




.2 Ecuación de búsqueda

Para gestionar una búsqueda utilizando la lógica booleana, se establece una ecuación de búsqueda que enlaza con los operadores adecuados los diferentes términos escogidos:

-se forman grupos de términos enlazados por O para representar cada uno de los temas de la consulta y expresar los diferentes puntos de vista que pueden representar esos temas en el eje paradigmático;
-se separan los grupos por paréntesis;
-se enlazan esos grupos por Y para expresar la necesaria coocurrencia de diferentes temas de la consulta dentro de los mismos documentos, en el eje sintagmático.

Ejemplo:

(DESCENTRALIZACION ADMINISTRATIVA o AUTONOMIA REGIONAL) y (REGION o AYUNTAMIENTO) no DEPARTAMENTO.

*Notas:*

1)En algunos casos, más raros, se enlazan los términos en el interior de los grupos por una intersección, y los grupos se enlazan por una unión.

Ejemplo:

(REGION y ESTATUTO) o (AYUNTAMIENTO y PODER DE DECISION).

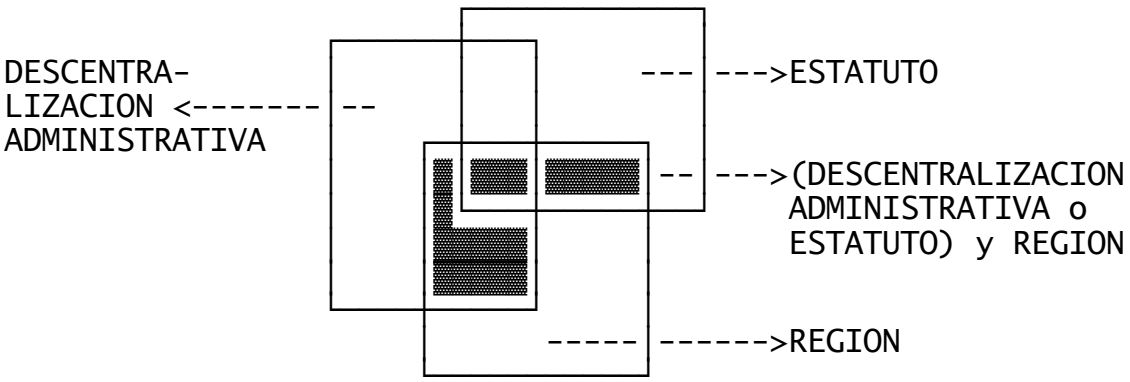
2)Es fundamental colocar correctamente los paréntesis para que la ecuación se corresponda adecuadamente con la consulta.

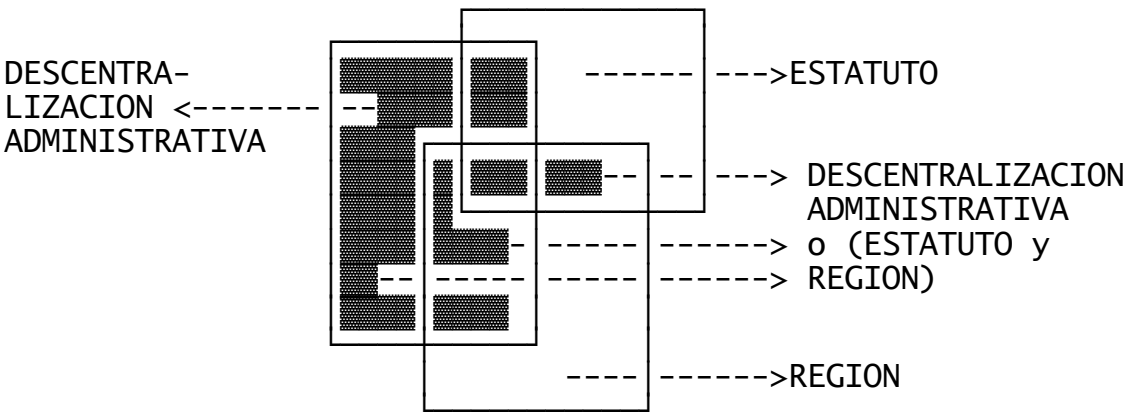
Ejemplo:

(DESCENTRALIZACION ADMINISTRATIVA o ESTATUTO) y REGION

recuperará documentos distintos que

DESCENTRALIZACION ADMINISTRATIVA o (ESTATUTO y REGION)




3)en una ecuación de búsqueda pueden intervenir varios niveles de paréntesis

Ejemplo:

DESCENTRALIZACION ADMINISTRATIVA y ((REGION y ESTATUTO) o (AYUNTAMIENTO y PODER DE DECISION))

4)aunque la lógica booleana trate sobre palabras clave (extraídas del título, del resumen y/o del texto de los documentos), las expresiones compuestas de varias palabras deben ser expresadas en la ecuación

por la intersección de esas palabras cuando la búsqueda se realiza en el fichero inverso.

Ejemplo: (en un sistema en el que no hubiera indización por medio de descriptores, o en un sistema con descriptores pero donde no se buscaran, además, las veces que aparecen las expresiones dentro del texto libre, fuera de los campos de indización controlada).

((DESCENTRALIZACION y ADMINISTRATIVA) o (AUTONOMIA y REGIONAL)) y (REGION o AYUNTAMIENTO) no DEPARTAMENTO.

.3 Estrategia de búsqueda

La lógica booleana permite llevar a cabo una verdadera «estrategia en la búsqueda»: el manejo de Y y de O permite, en efecto, modular el número de respuestas pertinentes:

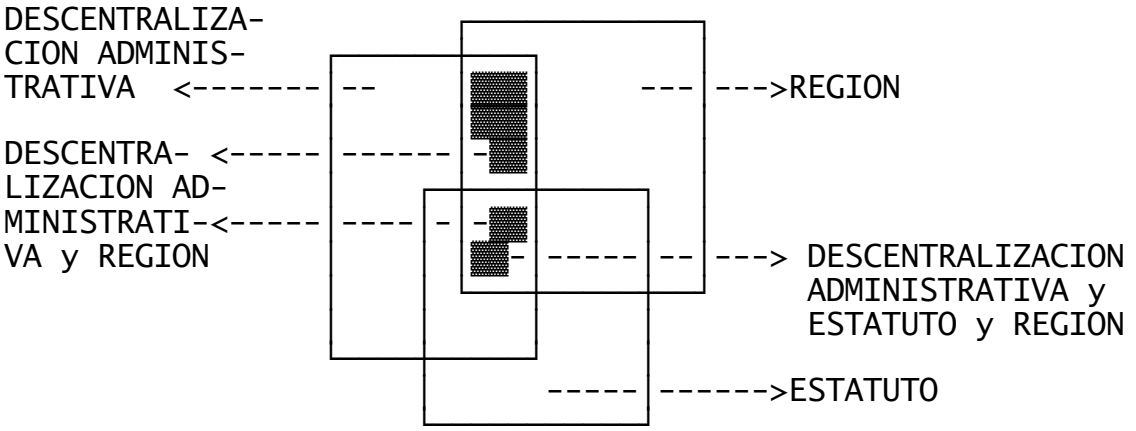
-al añadir Y disminuye el número de respuestas.

Ejemplo:

DESCENTRALIZACION ADMINISTRATIVA y ESTATUTO y REGION

recuperará, en principio, menos documentos que:

DESCENTRALIZACION ADMINISTRATIVA y REGION



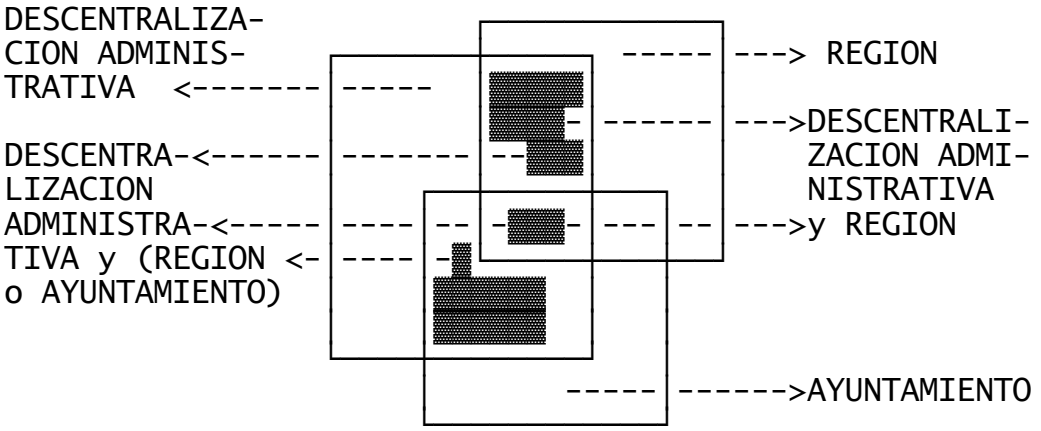

-al añadir O aumenta el número de respuestas.

Ejemplo:

DESCENTRALIZACION ADMINISTRATIVA y (REGION o AYUNTAMIENTO)

recuperará, en principio, más documentos que:

DESCENTRALIZACION ADMINISTRATIVA y REGION




En los primeros sistemas documentales automatizados, en los que la búsqueda se realizaba por lotes (tratamiento simultáneo de un lote económico de 10 a 50 consultas), esta estrategia de búsqueda era explotada «a priori» por los documentalistas: en efecto, si, tras haberse ejecutado la búsqueda, parecía que una ecuación estaba mal formulada y debía reformularse, antes de que la nueva ecuación pudiera ser introducida en máquina pasaban de varias horas a varios días, hasta que se hubiera reunido un nuevo lote económico de consultas.

Para prevenirse contra ese retraso, frecuentemente insoportable, los documentalistas introducían, de forma sistemática para cada consulta, tres ecuaciones de búsqueda:

- -una ecuación «normal» que parecía corresponder a la necesidad expresada por el peticionario;
- -una ecuación más amplia, que contenía menos Y y más O;
- -una ecuación más ceñida, que contenía más Y y menos O.

Tras la interrogación, el documentalista valoraba, a partir de las tres listas de referencias obtenidas en respuesta a la consulta, cuál era la bibliografía que más se adecuaba a la petición y suministraba esa bibliografía al peticionario.

En la actualidad todos los sistemas de búsqueda documental automatizados son interactivos, y la estrategia de búsqueda se construye a lo largo de toda la búsqueda:

- -el usuario introduce en su terminal uno o varios términos de indización, o directamente una ecuación de búsqueda.
- -a continuación el sistema le responde indicándole:
  - +la frecuencia de indización de cada uno de los términos
  - +el número de documentos que responden a la ecuación de búsqueda
- -si esta frecuencia, o este número, no le parecen adecuados, el usuario corrige su ecuación:
  - +introduciendo nuevos términos de indización

+suprimiendo términos de indización
+suprimiendo o añadiendo Y u O

hasta que el número de referencias correspondientes a la consulta le parezca adecuado.

-el usuario puede entonces visualizar en su terminal cada una de las referencias así encontradas:

+si le son útiles, detendrá aquí la búsqueda y solicitará que se imprima su bibliografía
+si no le resultan útiles, modificará de nuevo su ecuación de búsqueda.

.4 Características cualitativas

Al igual que en la indización de documentos, en la formulación de las consultas se han de tomar en cuenta dos factores:

-la exhaustividad: todos los conceptos presentes en la consulta y ligados por los «Y», o sólo parte de ellos, son traducidos por términos dentro de la ecuación de búsqueda.

Es necesario saber que la exhaustividad en la formulación lleva a:

+disminuir la llamada, es decir, la proporción de referencias pertinentes que serán extraídas del sistema en respuesta a la consulta
+aumentar la precisión, es decir, la proporción de referencias extraídas del sistema en respuesta a la consulta y que son pertinentes.

-la especificidad: los conceptos presentes en la consulta pueden ser traducidos por términos situados al mismo nivel de especificidad o a un nivel diferente. Una elevada especificidad en la formulación de la consulta tiene el mismo efecto que una elevada exhaustividad (al contrario de lo que sucede cuando se indizan los documentos, momento en que esos dos factores producen efectos divergentes):

+disminución de la llamada
+aumento de la precisión.

.2 Método del truncamiento

Este método permite buscar las raíces de las palabras no teniendo en cuenta:

-o bien los sufijos: truncamiento a la derecha (es el caso más frecuente)
-o bien los prefijos: truncamiento a la izquierda (caso más raro, al menos en la primera fase de una búsqueda)
-o bien cadenas de caracteres en el interior de una palabra: enmascaramiento (frecuente).

Permite acceder (en el fichero inverso) a todos los términos de una misma raíz. Los términos así obtenidos son enlazados entre sí por un operador implícito de unión.

Lo más frecuente es utilizar este método en los sistemas con indización en lenguaje libre, para evitar que el usuario tenga que introducir manualmente dentro de su ecuación todas las flexiones gramaticales, y a veces todas las formas derivadas de una misma raíz.

Ejemplos:

Una consulta con truncamiento a la derecha (señalado aquí simbólicamente por tres puntos) sobre ADMINISTR... hará que aparezcan todas las referencias que contengan las palabras: administrar, administra, administrador, administrado, administración, administraciones...

Una consulta con enmascaramiento (indicada aquí por el mismo símbolo: tres puntos) sobre PALABRA...CLAVE hará que aparezcan todas las referencias que contengan: palabra clave o palabra-clave.

.3 enlaces de proximidad

La búsqueda booleana simple, con el operador de intersección, permite exigir la presencia de dos o más términos en una noticia para que ésta sea reconocida como pertinente; no se impone ninguna restricción sobre la localización relativa de tales términos dentro de la noticia.

El método de los enlaces de proximidad permite, por el contrario, definir tal restricción: consiste en enunciar todos los términos de una pregunta y enlazarlos entre sí, de forma no explícita, por un operador de intersección, y de forma explícita, por un operador que precisa la localización relativa que deben tener los términos correspondientes dentro de las noticias (adyacente uno con otro; separados uno de otro por menos de x palabras; dentro de la misma frase o dentro del mismo párrafo), para que tales noticias sean consideradas pertinentes.

Este método se aplica:

-o bien a los sistemas en que los descriptores han sido registrados en el momento de la indización de los documentos siguiendo un orden significativo; por ejemplo: descriptores integrados dentro de un resumen;
-o bien, más generalmente, a los sistemas con indización en lenguaje libre, en los que se registra el título y el resumen (a veces el texto completo) del documento, y la búsqueda consiste entonces en descubrir la presencia de palabras simples y/o la coocurrencia de palabras que forman conceptos expresados por expresiones que se componen de dos o más palabras.

Los enlaces de proximidad pueden ser:

+*la adyacencia:* para que un documento sea recuperado, es necesario que los términos que figuran en la pregunta

aparezcan uno junto a otro dentro del documento, sin ninguna palabra que les separe.

Ejemplo:

AUTONOMIA (adyacente) REGIONAL

-permitirá recuperar un documento en el cual figura la frase:

La *autonomía regional* exige una transferencia de competencias a las autoridades locales,

-no permitirá que se recupere:

La *autonomía* de esta población se manifiesta en la producción de una especialidad *regional*.

*+la presencia a menos de x palabras:* para que un documento sea recuperado es necesario que los términos que figuran en la pregunta aparezcan dentro del documento separados uno de otro por un número de palabras inferior al que se ha precisado dentro de la consulta.

Ejemplo:

DESCENTRALIZACION (menos de 4 palabras) ADMINISTRATIVA

-permitirá recuperar un documento en el cual figura:

La *descentralización* de la autoridad *administrativa* ha permitido...

-no permitirá que se recupere:

La *descentralización* ha reforzado considerablemente la función *administrativa*.

*+la presencia dentro de la misma frase:* para que un documento sea recuperado es necesario que los términos que figuran en la pregunta aparezcan dentro de la misma frase del documento.

Ejemplo:

PODER (misma frase) DECISION

-permitirá recuperar un documento en el cual figura:

El *poder* del presidente regional ha sido ampliado, ya que en lo sucesivo será responsabilidad suya la *decisión* sobre...

-no permitirá que se recupere:

El *poder* central ha sido siempre muy fuerte hasta hace poco. Se ha tomado esta *decisión* para reducirlo en provecho de las regiones.

*Nota:*

Evidentemente, hemos escogido unos ejemplos adecuados para:

-seleccionar documentos pertinentes,
-eliminar documentos no pertinentes

Pero, como la búsqueda por proximidad opera fundamentalmente sobre palabras clave en lenguaje libre, consideradas dentro de su contexto topológico, pero separadas de su contexto semántico, y el lenguaje libre es ambiguo, existe una gran posibilidad de seleccionar así mismo documentos no pertinentes, o de eliminar documentos pertinentes.

Ejemplos:

1) documentos seleccionados, aunque indudablemente no pertinentes

-en el caso de la adyacencia:

La falta de *autonomía regional* ha producido...

-en el caso de la presencia a menos de x palabras:

La *descentralización* de la tramitación *administrativa* permite...

-en el caso de la presencia dentro de la misma frase:

El *poder* ha considerado finalmente la posibilidad de tomar la decisión de...

2)documentos eliminados, aunque indudablemente pertinentes

-en el caso de la adyacencia:

Una *autonomía* verdaderamente *regional* exige una transferencia de competencias a las autoridades locales

-en el caso de la presencia a menos de x palabras:

La *descentralización* y una importante ampliación, recientemente aprobadas, de la autoridad *administrativa* permite...

-en el caso de la presencia dentro de la misma frase:

El *poder* del presidente regional ha sido ampliado. Puede, en efecto, tomar una *decisión* sobre...

+*la proximidad y el orden de palabras:* los enlaces de proximidad pueden así mismo hacer intervenir la noción de

orden de los términos de búsqueda: se puede, según la consulta, solicitar:

-que se respete el orden de los términos: para que un documento sea recuperado es necesario que los términos que figuren en la pregunta aparezcan dentro del documento:

+respetando, por una parte, la restricción de proximidad impuesta en la consulta,
+y, por otra parte, dentro del mismo orden que en la consulta.

Ejemplo:

PODER (menos de 2 palabras; mismo orden) DECISION
->permitirá recuperar un documento en el que figure:

El *poder* de *decisión* sobre...

->no permitirá que se recupere:

La *decisión* de *poder*...

-que aparezcan en cualquier orden: sólo se establece la restricción de proximidad; los términos pueden estar en cualquier orden

Ejemplo:

DESCENTRALIZACION (misma frase; cualquier orden) REGIONAL

->permitirá recuperar:

La *descentralización regional* de...

->y también:

Esta característica *regional* de la *descentralización*...

*Nota:* la búsqueda sobre las palabras clave de los títulos, resúmenes y/o textos en lenguaje natural utiliza muy a menudo la combinación del truncamiento y de la proximidad.

Ejemplo:

DESCENTRALI... (misma frase) REGION...
permite recuperar documentos que contienen frases como:

-la *descentralización regional* ha permitido...
-la *descentralización* de *regiones* ha conducido a...
-las *regiones descentralizadas*.

.4 Comparación

Este método permite realizar búsquedas sobre criterios cuantitativos (fechas, valores críticos) y establecer

condiciones de igualdad o de no igualdad (entre números), de similitud o de divergencia (entre cadenas de caracteres), de presencia o ausencia (de valores). Raramente se lleva a cabo sobre los descriptores o las palabras clave.

Ejemplo:

-Documentos fechados en 1976.
-Documentos que contengan más de 20 referencias bibliográficas.
-Documentos en los que aparezca la cadena de caracteres «revue des deux mondes» dentro de la zona título (no invertido).
-Documentos en los que la zona «fecha de muerte de la publicación periódica» está vacía.

.5 Búsqueda asistida por el ordenador

El ordenador puede ayudar al usuario en la formulación de su pregunta y/o en la elección de las respuestas:

-o bien en modo fundamentalmente algorítmico;

+por el método de extender los descriptores de la ecuación a los descriptores que están ligados a ellos dentro del thesaurus por una relación semántica;
+por el método de la ponderación;
+por el método denominado bucle de pertinencia, es decir, a base de tener en cuenta los primeros resultados de la búsqueda para mejorar la formulación de la ecuación;
+y/o con tratamiento lingüístico y/o estadístico de la consulta;

-o bien en modo fundamentalmente heurístico, por medio de un sistema experto.

.1 Método de extensión

Es el método que permite explotar las relaciones semánticas del thesaurus para añadir, de forma automática, todos los descriptores ligados al descriptor o descriptores especificado(s) dentro de la ecuación de búsqueda por medio de enlaces:

-o bien de equivalencia (en los sistemas de lenguaje post-controlado; ejemplo: STAIRS-TLS),
-o bien de jerarquía (uso relativamente frecuente en los sistemas de lenguaje precontrolado; ejemplo: MISTRAL, GOLEM),
-o bien de asociación (raro; ejemplo: MISTRAL, GOLEM).

Los descriptores así añadidos son ligados entre sí, y con el descriptor especificado al principio, por un operador de unión (o).

Ejemplo:

Una consulta sobre FRUTA + TS hará que aparezcan todas las referencias indizadas por: fruta, manzana, pera, naranja, etc., sin que el usuario deba escribir estos descriptores específicos dentro de la ecuación de búsqueda, que se genera automáticamente de la siguiente forma: FRUTA o MANZANA o PERA o NARANJA...

.2 Método de la ponderación

Uno de los mayores inconvenientes de la búsqueda booleana es que realiza en cada documento de una colección una apreciación binaria sobre su pertinencia con respecto a una consulta: un documento o es pertinente o es no pertinente. Sería deseable poder matizar esta apreciación, definiendo para cada documento un grado de pertinencia probable con respecto a una consulta, que variaría

-de 1: para un documento de pertinencia máxima,
-a 0: para un documento de pertinencia nula.

La elección, por parte del usuario, de un umbral de pertinencia (por ejemplo: 0,6) permitirá al sistema conservar, dentro de la bibliografía suministrada como respuesta, sólo los documentos cuyo grado de pertinencia sea superior a ese umbral.

Una ventaja considerable de este método es permitir que las referencias suministradas se puedan disponer por orden de pertinencia: los documentos de pertinencia elevada al principio de la bibliografía, los documentos con una pertinencia equivalente al umbral, al final de la lista.

Otro de los grandes inconvenientes del método booleano se encuentra en que el usuario está obligado a introducir los operadores (Y, O, NO) entre cada criterio de búsqueda. Ahora bien, el significado de estos operadores, y el de su combinación dentro de diferentes niveles de paréntesis, no es algo evidente para el usuario ocasional o neófito. En este caso también, emerge, muy rápidamente, el interés por unos métodos más sencillos, que se basen en una simple yuxtaposición de los términos de la pregunta, sin intervención de la lógica de Boole.

Desde las primeras aplicaciones de la automatización documental, una serie de sistemas operaron utilizando algoritmos muy sencillos, basados en una ponderación establecida por el documentalista.

Este método consiste en comparar todos los términos (descriptores o palabras clave), no enlazados por operadores, de una pregunta, con los términos de los documentos, considerando como pertinentes los documentos que tengan en común con la pregunta un determinado número, o un porcentaje, fijado de antemano, de términos. Así es posible, además, organizar los documentos suministrados como respuesta por el sistema dentro de un orden decreciente de pertinencia probable: los documentos que tengan más términos en común con la consulta se colocan al principio.

Ejemplo de pregunta:

DESCENTRALIZACION ADMINISTRATIVA, AUTONOMIA REGIONAL, REGION, PROVINCIA
ponderación mínima: 3
documentos pertinentes: todos los que tienen al menos 3 términos de la lista anterior.

Como cada uno de los dos grupos ficticios (aquí):

-DESCENTRALIZACION ADMINISTRATIVA, AUTONOMIA REGIONAL
-REGION, PROVINCIA

contiene dos términos, las respuestas contendrán al menos un término de cada uno de esos grupos y serán, por tanto, pertinentes, como si se hubiera utilizado la lógica de Boole.

Una variante de este método permite asignar un coeficiente de ponderación a cada término de la consulta, mejorando así la organización de las respuestas por orden decreciente de pertinencia.

Ejemplo:

una consulta:DESCENTRALIZACION ADMINISTRATIVA (5), AUTONOMIA REGIONAL (1), COLECTIVIDAD DESCENTRALIZADA (3), REGION (2), PROVINCIA (2), AYUNTAMIENTO (2)

ponderación mínima: 8

irá por delante la mayor parte de las combinaciones de términos en los que intervengan DESCENTRALIZACION ADMINISTRATIVA y COLECTIVIDAD DESCENTRALIZADA.

Una segunda variante de este método consiste en que, además de asignar pesos diferentes a cada término de la consulta, se explota también la ponderación asignada a los términos de indización de los documentos (cf. § 1.4.4.2).

Estos métodos, puramente intelectuales, encuentran sus límites dentro de su mismo empirismo: un mismo término, en un mismo documento o en una misma pregunta, recibirá dos pesos diferentes, según sea asignado por:

-dos indizadores diferentes;
-un mismo indizador, en dos momentos distintos.

Por tanto, se prefiere utilizar métodos estadísticos basados en:

-calcular la frecuencia de los términos dentro de la colección y dentro de los documentos;
-calcular un grado de proximidad, o de similitud entre documentos y preguntas.

Estos métodos se denominan también «estadísticos», si se basan en la aplicación de cálculos estadísticos, y/o «probabilísticos», si se basan en el cálculo de probabilidades.

Muchos investigadores se han destacado en la búsqueda de algoritmos para el cálculo de la ponderación y de la similitud; citemos especialmente a Salton (1983), Bookstein (1985), Robertson (1982) y Van Rijsbergen (1979).

Con la excepción de los trabajos de Andreewski y Fluhr, que han obtenido como resultado un producto industrial (SPIRIT), todas las investigaciones, si bien algunas se remontan a hace más de 20 años, han desembocado sólo en prototipos universitarios, que se han probado con colecciones de varios cientos o, excepcionalmente, varios miles de documentos, y algunas decenas o cientos de consultas.

Citemos así mismo el programa STAIRS, de IBM, que incluye tres algoritmos de ponderación (a elección del usuario); estos algoritmos pueden intervenir, tras una búsqueda booleana clásica, para organizar por orden decreciente de pertinencia probable las referencias obtenidas: todos los términos de la consulta, ya estén enlazados por Y o por O, son tomados en cuenta por el algoritmo que mide la coocurrencia con los términos de indización de los documentos.

No es posible describir todos los algoritmos presentados por los investigadores anteriormente citados: todos los años aparecen otros nuevos, más complejos que los anteriores. En ODDY et al. (1981) se encontrará la descripción de una serie de ellos, y especialmente en el texto de NOREAULT (T) et al., allí incluido, se presenta una lista de:

- -37 algoritmos de cálculo de ponderación de términos en los documentos y/o las preguntas;
- -64 algoritmos de cálculo de similitud.

A título de ejemplo, expondremos aquí el modelo vectorial de SALTON.

Se dispone de:

- -una colección de documentos D, representado cada uno por un conjunto de términos (descriptores y/o palabras clave extraídas del título y del resumen) $d_1$, $d_2 \ldots d_t$;
- -una serie de preguntas Q, representada cada una por un conjunto de términos $q_1$, $q_2 \ldots q_t$.

Cada término se representa por un vector cuya longitud es proporcional al peso del término dentro de un espacio multivectorial.

La similitud entre un documento dado, $D_i$, y una pregunta $Q_i$, se mide por la similitud global de los dos espacios vectoriales, la cual es medida por la función coseno:

$$S(Di, Qi) = \cos(Di, Qi) = \frac{\sum_{k=1}^{t} dik \times qjk}{\left(\sum_{k=1}^{t} (dik)^2 \times \sum_{K=1}^{t} (qik)^2\right)^{1/2}}$$

donde:

-$S(D_i, Q_i)$ = medida de la similitud entre el documento i y la pregunta j; puede variar de 0 (cuando los dos vectores no tienen ningún término en común) a 1 (cuando los dos vectores tienen todos sus términos en común);
-dik = peso del término representativo k del documento; este peso puede variar entre un máximo de 1 y un mínimo de 0;
-qjk = peso del término representativo k de la pregunta; este peso puede variar entre un máximo de 1 y un mínimo de 0;
-t = número de términos representativos, ya sea del documento i, o de la pregunta j.

El peso de los términos se calcula, por ejemplo, con la siguiente fórmula:

-*término de un documento* (es decir, del título y del resumen de ese documento)

$dik = fik/B_k$

donde:

fik = frecuencia = número de apariciones del término k dentro del documento i

$B_k$ = número de documentos dentro de la colección a los que el término k ha sido asignado.

El peso de un término dentro de un documento (es decir, su valor discriminante) es tanto más elevado cuanto ese término tiene:

+una frecuencia más elevada dentro de ese documento,
+una frecuencia más reducida dentro de la colección.

Este peso varía de 1 (el término sólo está presente en el documento examinado) a 0 (el término no está presente dentro del documento)

-*término de una pregunta* (es decir, del enunciado de la pregunta)

$qjk = fjk/B_k$

.3 Bucle de pertinencia

La formulación de una ecuación de búsqueda es una operación difícil, que exige a menudo mucho cuidado y mucho tiempo: desde varios minutos, en los casos sencillos, hasta una hora o más en los casos complejos, con una media habitual de 5 a 20 minutos aproximadamente.

Esta operación es completamente manual en casi todos los sistemas en funcionamiento; la ayuda que aporta el ordenador (informando sobre el thesaurus, indicando el número de referencias pertinentes, extendiendo la ecuación según la cadena jerárquica de los descriptores) permite mejorar la

calidad de la búsqueda, pero no hace ganar tiempo, o lo hace poco.

La aproximación mediante el método del bucle de pertinencia (en inglés: relevance feed-back) es de todo punto diferente; y permite automatizar, al menos en parte, la formulación de la ecuación de búsqueda.

Robson y Longman (1976), por ejemplo, parten de comprobar que las referencias que responden a una ecuación de búsqueda presentan, por definición, una serie de características comunes: descriptores, autores...

De ahí la idea de reemplazar la operación de construir la ecuación de búsqueda por la de introducir, en ordenador, la identificación de una serie de referencias conocidas en principio por el usuario y consideradas por él pertinentes.

Los criterios de búsqueda de esas referencias (descriptores, autores...) son automáticamente extraídos por el ordenador, que los utiliza para construir, siempre automáticamente, una ecuación de búsqueda, aplicando el método de la ponderación.

Se calculan automáticamente los pesos para cada criterio de búsqueda de la siguiente manera:

- -se calcula el número de referencias señaladas al principio por el usuario como pertinentes que contienen ese criterio (Fr);
- -se busca (dentro del fichero inverso del fondo) el número total de noticias del fondo que contienen ese mismo criterio (Fn);

- -se calcula la ratio $Fr^2/Fn$, denominada especificidad del criterio.

Esta ratio es la que constituye el factor de ponderación asignado al criterio correspondiente. En efecto, cuanto más específico es un término, es decir, cuanto más frecuentemente interviene dentro de la muestra de documentos pertinentes suministrados por el usuario, y menos interviene dentro del conjunto de la colección, mayor es su poder discriminante, y su peso debe ser más elevado dentro de la ecuación de búsqueda.

El ordenador confronta la ecuación de búsqueda así formada con el fichero de búsqueda, y aparece en la pantalla del terminal una lista de referencias que responden a la ecuación y que son, por tanto, susceptibles de ser pertinentes.

El usuario puede, a la vista de estas referencias, señalar de entre ellas las que son más pertinentes, y se desencadena un proceso de reajuste automático de la ecuación, siguiendo el método que a continuación se describe, pero a partir de una muestra de referencias mayor. El proceso se puede volver a realizar varias veces, en sucesivos ciclos, que mejoran cada vez el resultado de la búsqueda.

El trabajo del operador se limita, pues, a:

- -introducir una primera vez varias referencias que considera pertinentes, lo que lleva mucho menos tiempo que formular una ecuación de búsqueda;

-valorar la pertinencia de las referencias que le son suministradas, lo cual ha de realizarse en cualquier caso.

Un método, más flexible todavía, que, según se presenta (Oddy, 1977), permite organizar la búsqueda a base de ir navegando por un diálogo entre el hombre y la máquina. El primero introduce en el terminal una o varias palabras que caracterizan el objeto general de su búsqueda; no debe aportar ninguna precisión sobre este objeto (ya que, por otra parte, a estas alturas de la búsqueda es incapaz de hacerlo: no puede describir lo que todavía no conoce) y no debe utilizar ningún operador lógico. Esta o estas palabras forman una primera «imagen» de la necesidad del usuario. El programa reacciona ante esta imagen seleccionando y mostrando en el terminal la referencia más «implicada»: la implicación de un elemento se define como el resultado de dividir el número de elementos asociados incluidos dentro de la imagen entre el número total de elementos asociados conocidos.

La noticia mostrada en el terminal contiene el título, la fuente, los autores y los descriptores de la referencia así seleccionada por el sistema.

El usuario es invitado inmediatamente a indicar si el título, cada uno de los autores y cada uno de los descriptores son o no pertinentes; puede así mismo introducir nuevos nombres y nuevas palabras que describan su consulta. El programa construye, a partir de esas referencias, una nueva imagen de la necesidad del usuario: los elementos (autores y descriptores) elegidos y las referencias a los que éstos están ligados son añadidos a la imagen, y se retiran los elementos rechazados. El programa elige una nueva referencia y la muestra en el terminal; el usuario indica si acepta o rechaza el título, cada uno de los autores y cada uno de los descriptores.

El diálogo prosigue así de forma muy natural:

-el usuario sólo debe indicar sus preferencias; jamás debe escribir la ecuación de búsqueda; puede modificar sus reacciones a la vista de las referencias sucesivas (confirmar, por ejemplo, que un descriptor que, al principio del diálogo, le parecía que debía ser conservado, no tiene en realidad interés para él);

-el ordenador construye, por aproximaciones sucesivas, una imagen cada vez más precisa de la necesidad del usuario y, a partir de esta imagen, le propone elecciones que le permitan abordar cada vez mejor su consulta.

Señalemos un tercer método, que tiene la ventaja de que lo usa de hecho un importante distribuidor, ESA-IRS (Martin - 1983): el usuario introduce una ecuación de búsqueda sencilla, relativamente tosca, seguida del comando ZOOM; el sistema muestra, como respuesta, la lista de:

-todos los descriptores (en lenguaje controlado o no),
-y, si el usuario lo desea, todas las palabras del título y del resumen

que caracterizan el conjunto de los documentos que responden a la ecuación de búsqueda.

Se encuentran, en esta lista, no sólo los términos de la ecuación de búsqueda, sino también todos los demás criterios de búsqueda presentes dentro de los documentos encontrados por medio de la ecuación de búsqueda. Estos términos son organizados por orden decreciente de frecuencia de aparición dentro de los documentos recuperados; se muestra la frecuencia de cada término.

Al usuario le basta entonces con examinar esta lista y escoger de ella los términos más adecuados para enriquecer su ecuación. Se introduce la nueva ecuación, de nuevo con el comando ZOOM, lo que tiene como resultado un enriquecimiento de la lista inicial. La experiencia muestra que, con dos o tres ciclos de este tipo, el usuario termina por recuperar todos los criterios de búsqueda pertinentes que están presentes dentro de la base de datos.

.4 Método lingüístico

El programa informático ALEXDOC, ya descrito anteriormente por su aplicación de indización asistida (cf. § 3.6.3), ha sido así mismo concebido para permitir la formulación asistida de consultas.

Contiene los mismos automatismos que se han descrito en el § 3.6.3; a los que se añaden:

-la corrección ortográfica, que permite corregir omisiones de letras, dislexias, letras repetidas y cambiadas. Esta corrección no se realiza en la fase de indización de los documentos;
-y sobre todo la construcción de una ecuación booleana a partir de la consulta del usuario.

El proceso de búsqueda es el siguiente:

-el usuario introduce su pregunta, expresada en lenguaje natural, con el teclado de su terminal;
-ALEXDOC consulta los conjuntos de datos «thesaurus» y «diccionario» y aplica las reglas de:

+corrección ortográfica,
+derivación morfológica,
+desambiguación,
+reestructuración,
+transformación semántica,

para establecer la lista de descriptores que corresponden a los términos de la consulta;

-Alexdoc construye una ecuación booleana insertando o bien el operador Y o bien el operador O entre los descriptores conservados. En algunos casos el operador booleano corresponde a la conjunción de coordinación Y u O, tal y como figura dentro de la consulta del usuario.

Ejemplos:

manzana o pera

ordenador y documentación.

En el caso de una pregunta que contenga una enumeración, las sucesivas comas son reemplazadas por el operador booleano correspondiente a la conjunción que aparece delante del último término de la enumeración.

Ejemplos:

manzana, pera o cereza
se convierten en
MANZANA O PERA O CEREZA

automatización, cultivo y fruta
se convierten en
AUTOMATIZACION Y CULTIVO Y FRUTA.

En otros casos, estas reglas de desestructuración se utilizan no sólo para pasar de la formulación del usuario a la ecuación booleana, sino también para elegir el operador booleano adecuado.

Ejemplo:

automatización de la producción
se convierten en
AUTOMATIZACION Y PRODUCCION.

Incluso en algunos casos el sistema llega a reemplazar la conjunción Y por el operador booleano O.

Ejemplo:

consumo de fruta y legumbres
se convierten en
CONSUMO Y (FRUTA O LEGUMBRES).

-ALEXDOC consulta inmediatamente el conjunto de datos bibliográficos y muestra:

+la fórmula booleana, indicando el número de referencias que responden a la consulta;
+un block de notas, que contiene cada uno de los descriptores de la ecuación, acompañado de su frecuencia de indización en la base de datos y de un número de orden dentro del block de notas;

-el peticionario puede entonces:

+navegar a través del thesaurus, añadiendo en el block de notas los descriptores que tengan, con los que ya están en el block de notas, una o más categorías de relaciones semánticas (TG, TS, TA...);
+modificar la ecuación de búsqueda;

->o bien introduciendo en ella nuevos descriptores que aparecen en el block de notas, o

suprimiendo los descriptores que le parezca que no son adecuados;
->o bien añadiendo nuevos términos en lenguaje natural, que el sistema tratará como acabamos de describir;

+pedir que se visualicen una o más noticias;
+a la vista de las noticias, navegar a través del conjunto de datos bibliográficos pidiendo ver, por ejemplo, las noticias que

->tengan el mismo autor que la que acaban de ver;
->estén ligadas a la que acaban de ver (es el ejemplo de un caso de jurisprudencia dictada en aplicación de una ley).

.5 Método lingüístico-estadístico

El programa informático SPIRIT, ya descrito anteriormente por su aplicación de indización automática (cf. § 3.4), puede así mismo ser utilizado para ayudar al peticionario a formular su consulta.

El usuario enuncia su pregunta de manera natural, sin que deba aprender a utilizar:

-un lenguaje de comandos;
-la lógica de búsqueda: operadores booleanos, enlaces de proximidad...;
-un thesaurus de descriptores.

La búsqueda se realiza en siete etapas:

-se mecanografía la consulta, en lenguaje libre: palabras significativas y vacías, sin ninguna sintaxis entre los términos;
-se indiza automáticamente la consulta, siguiendo el mismo proceso que en la indización de un documento (cf. § 3.4.2): se obtiene una lista de palabras (o expresiones) normalizadas y ponderadas que caracterizan la consulta;
-se hace la búsqueda en el fichero inverso, siguiendo una variante del método de la ponderación:

+se unen todos los descriptores de la pregunta por medio del operador O; se extraen los números de los documentos que tengan uno, dos o más descriptores en común;
+se calcula una ponderación para cada número de documento así extraído; esta ponderación, denominada «cálculo de proximidad» está en función de:

->el número de términos comunes a la pregunta y al documento;
->la ponderación de los términos comunes dentro de la pregunta y del documento;
->el carácter de los términos comunes: peso más elevado para las palabras compuestas y menos elevado para las simples;

-se organizan los números de los documentos que responden a la consulta por orden decreciente del peso así calculado;
-se editan los números de los documentos que responden a la consulta, acompañados de las palabras simples y compuestas que han permitido su selección;
-se eligen manualmente los documentos que se van a visualizar;
-se visualizan los documentos conservados, en orden decreciente de ponderación.

*Notas:*

1) Para las bases de datos textuales, en las que cada documento puede incluir decenas o cientos de páginas, SPIRIT permite acceder directamente sólo a aquellas páginas en las que aparezca la información que responda a la consulta.

Esto se realiza aplicando a cada una de las páginas de los documentos obtenidos, durante una primera fase de la búsqueda, un proceso de comparación ponderado, análogo al que ordena, en una primera fase, los documentos pertinentes.

2) SPIRIT puede producir automáticamente las ecuaciones de búsqueda, no sólo por el método de la ponderación (con la intervención de operadores de unión O entre los términos de la búsqueda), sino también según la lógica booleana (interviniendo además operadores de intersección Y).

Esta técnica ha sido puesta a punto para la interrogación de bases de datos documentales fuertemente estructuradas.

Supongamos, por ejemplo, una estructura de noticia que contenga los campos siguientes:

-autor(es) personal(es) de un documento;
-persona(s) en cuanto tema(s) de un documento;
-fecha del documento;
-tipo de documento;
-texto del documento.

Tal base de datos podrá ser interrogada de forma estructurada, pero, no obstante, aún muy libre, pues aparece en la pantalla una «cuadrícula de pantalla multicriterio» que invita al peticionario a introducir su consulta reemplazando las casillas adecuadas de la cuadrícula.

Esta cuadrícula tendrá, por ejemplo, la forma siguiente:

Autor(es) de los documentos buscados
Autor(es) tratado(s) en los documentos buscados
Fecha de los documentos buscados
Tipo de documento
Tema.

Supongamos que el peticionario introduce la pregunta siguiente: ¿existe una obra de Breton sobre Fourier, que trate sobre la analogía universal y esté publicada en 1947?

En vez de introducir su pregunta en la forma anterior, el peticionario completará la cuadrícula registrando:

-en la primera línea: BRETON

-en la segunda línea: FOURIER
-en la tercera línea: 1947
-en la cuarta línea: LIBRO
-en la quinta línea: ANALOGIA UNIVERSAL.

Los cuatro primeros criterios de búsqueda, de tipo factual, serán asociados por un operador de intersección dentro de la ecuación de búsqueda generada automáticamente por el sistema.
El tema del documento, de tipo textual, es:

-asociado por un operador de intersección con los cuatro criterios factuales;
-tratado, además, por el sistema como una consulta en lenguaje libre, es decir, creando una ecuación de búsqueda con ponderación y con una operación de unión entre los dos términos normalizados ANALOGIA, UNIVERSAL.

3) SPIRIT puede así mismo guiar una búsqueda en modo «bucle de pertinencia» (cf. § 6.5.3), a partir de la referencia de un documento pertinente, conocido por el usuario y presente en la base, o de un documento pertinente ya obtenido como respuesta a una consulta.
El proceso es el siguiente:

-se marca ocasionalmente un umbral para el cálculo de proximidad (en forma de número mínimo de palabras comunes al documento consulta y a cada documento respuesta; el valor por defecto es el resultado de dividir por dos el número de palabras comunes a la consulta y al documento que tengan la mayor cantidad de palabras comunes con el documento consulta);
-la indización del documento conocido es considerada como expresión de la consulta en palabras y expresiones normalizadas y ponderadas; así es conducida, por último, la segunda fase de la búsqueda a partir de una consulta.

## .6 Búsqueda documental por medio de un sistema experto

Un sistema experto es un sistema informático concebido para resolver problemas de una forma similar a la que utiliza un experto humano dentro de su dominio de competencia.
Un sistema experto se asienta sobre dos elementos fundamentales:

-una base de conocimientos que almacena:

+por una parte, todos los datos que constituyen el saber de uno o varios expertos dentro de su especialidad; por ejemplo, en documentación el descriptor MANZANA es un específico del descriptor FRUTA;
+por otra parte, un conjunto de reglas que el experto aplica para explotar sus conocimientos; por ejemplo, en documentación,

si se me hace una consulta sobre un término genérico y no encuentro suficientes respuestas, entonces puede tener interés extender la consulta a los específicos de ese término;

-un motor de inferencia, que sirve para detectar las condiciones iniciales, para seleccionar la regla o reglas que hay que aplicar en función del estado de esas condiciones y de los elementos de saber que hay que explotar, y para inferir una modificación de las condiciones; por ejemplo:

+se introduce una pregunta sobre el concepto de fruta;
+el sistema documental encuentra 10 documentos indizados por «FRUTA» y comprueba que el descriptor FRUTA es un término genérico;
+el sistema experto asociado toma en cuenta las condiciones iniciales:

10 respuestas
FRUTA = término genérico
+selecciona
->una primera regla:
-si se obtienen menos de 20 respuestas (por ejemplo), la pregunta debe ser ampliada, si es posible;
->una segunda regla:
-una pregunta puede ser ampliada extendiéndola a los específicos de un genérico;
+selecciona
->una serie de datos:
-MANZANA es específico de FRUTA
-PERA es específico de FRUTA
-etc.
+infiere una nueva pregunta, formulada:
FRUTA o MANZANA o PERA...

La masa de datos y de reglas que maneja un experto humano dentro de su dominio de especialización es enorme; y generalmente no está formalizada en su totalidad; por tanto, es difícil registrarla dentro de un sistema informático.

Este es el motivo por el que los sistemas expertos sólo están comenzando a ser operativos en algunos pocos dominios en los que ha sido posible enumerar y enunciar los conocimientos utilizados y las reglas seguidas para resolver problemas. Es el caso sobre todo del diagnóstico médico, de la prospección geológica y de la detección de causas de averías, y siempre dentro de dominios muy limitados todavía.

En lo que se refiere a la búsqueda documental, comienzan a aparecer sistemas expertos en una serie de sectores:

-en Francia: el prototipo «DIALECT» de sistema experto para la búsqueda documental esencialmente en lenguaje libre y controlado, de Bassano (1985);
-en el Reino Unido: el sistema «CANSEARCH» de búsqueda documental dentro del dominio de la terapéutica del cáncer, basado en la explotación de un thesaurus, de Pollitt (1984);

-en Estados Unidos:

+CITE (Current Information Transfer in English), sistema utilizado por la National Library of Medicine y que se basa en el thesaurus médico de la NLM (Doszkocs - 1983);
+CONIT y EXPERT, sistema experto para la búsqueda documental multidistribuidor y multibase fundado en el lenguaje libre y en el lenguaje controlado (Marcus - 1983);
+IIADA (Individualised Instruction Aids for Data Access), sistema de formación y de ayuda a la búsqueda documental multibase para el distribuidor DIALOG (Meadows - 1982);
+RITA (The Rule Intelligent Terminal Agent), sistema experto de diálogo para el acceso a la base de datos del New York Times (Waterman - 1979).

El objetivo de los productores de sistemas expertos para la búsqueda documental es permitir a un usuario final, que no conozca nada sobre las técnicas documentales, acceder a los documentos que necesite y que se encuentren dentro de las bases de datos bibliográficas, siendo guiado por el sistema experto y sin recurrir a un documentalista.

Este objetivo es muy ambicioso; no es seguro que se pueda alcanzar alguna vez, teniendo en cuenta

-la extrema complejidad de los problemas que se han de resolver para seguir la cadena de operaciones descritas al principio del § 6;
-el carácter muy evolutivo de los sistemas documentales;
-la reducida prioridad concedida a los trabajos de investigación y desarrollo relacionados con los sistemas de almacenamiento y recuperación documentales, y, por tanto, la limitación de recursos disponibles para construir las bases de conocimientos.

Los sistemas expertos que existen hoy día, dentro del dominio de la búsqueda documental, tienen generalmente objetivos muy limitados y tratan de resolver, sin que siempre lo consigan, una parte de los problemas que acarrea la búsqueda documental.

Vamos a describir a continuación las grandes orientaciones de tres de sus sistemas:

-el primero, basado en un thesaurus: CITE;
-los otros dos, basados a la vez en el lenguaje libre y en el lenguaje controlado: CONIT/EXPERT y DIALECT.

.1 Sistema experto basado en un thesaurus

CITE está concebido para hacer corresponder las preguntas en lenguaje natural con la indización en lenguaje controlado de los documentos de la National Library of Medicine; CITE está operativo y accesible al público.

El usuario introduce su consulta en lenguaje natural.

Ejemplo: The use of biofeedback for the treatment of tension headache.

El sistema responde mostrando una lista de términos extraídos del fichero inverso (palabras del texto y descriptores del thesaurus) que:

-o bien contienen todos o parte de los términos de la pregunta;
-o bien tienen una relación semántica, dentro del thesaurus, con un término de la consulta.

Los términos son mostrados en un orden de pertinencia decreciente, en función del número de palabras comunes entre los términos de la pregunta y los términos del fichero.

El sistema indica para cada término mostrado si es una palabra del texto del documento o si es un descriptor del thesaurus:

Ejemplo:

```
 1 BIOFEEDBACK (texto)
 2   FEEDBACK (descriptor)
 3   BIOFEEDBACK (PSYCHOLOGY) (descriptor)
 4 HEADACHE (texto)
 5   HEADACHE (descriptor)
 6   HEMORRHAGIC FEVERS, VIRAL (descriptor)
 7   HEADACHE (texto)
 8 TENSION (texto)
 9   TRANQUILIZING AGENTS, MINOR (descriptor)
10   TENSILE STRENGTH (descriptor)
11   TENSIONS (texto)
12 TREATMENT (texto)
13   THERAPY (descriptor)
14   THERAPEUTICS (descriptor)
```

Para poder recuperar esos términos el sistema utiliza:

-un algoritmo de detección y corrección de errores ortográficos;
-un análisis morfológico de la pregunta, para establecer el radical de las palabras a partir de una lista de afijos (formada ésta al principio a partir de un análisis del contenido del fichero inverso);
-la búsqueda de todas las palabras del fichero inverso que contengan esos radicales;
-la búsqueda de los descriptores correspondientes dentro del fichero thesaurus, utilizando, si es el caso, las relaciones semánticas del thesaurus.

A la vista de esta lista, se invita al usuario a que indique los términos que conservará para su búsqueda.

El sistema efectúa la búsqueda, indica el número de documentos encontrados y muestra el título y la indización de los documentos que responden total o parcialmente a la consulta.

El usuario designa los documentos más pertinentes, y el sistema procede entonces a un bucle de pertinencia, basado en

las palabras y descriptores de los documentos conservados por el usuario.

## .2 Sistemas expertos basados en el lenguaje libre y en el lenguaje controlado

*+CONIT/EXPERT*

Este sistema está en su quinta versión; hemos pensado que es interesante presentar las tres últimas versiones, con el fin de ilustrar la manera en que un productor hace evolucionar su sistema.

*Descripción de CONIT 3*

Un interface permite interrogar tres distribuidores y varios cientos de bases de datos por medio de comandos normalizados (PICK: selección y conexión a una base; FIND: búsqueda acerca de un criterio; COMBINE: búsqueda acerca de varios criterios; SHOW: visualización; EXPLAIN: curso programado).

Sólo son activados los comandos fundamentales de los distribuidores, que son así mismo los más usuales; a los usuarios principiantes se les enseña un número más limitado todavía de comandos.

El curso programado contiene una serie de menús, con una estructura jerárquica de explicaciones muy detalladas (de 10 a 20 minutos de clase para el curso completo) y algoritmos de decisión sensibles al contexto que permiten adaptar las instrucciones y las explicaciones a cada situación particular.

El lenguaje documental utilizado es a la vez el lenguaje libre y el lenguaje controlado: en el momento de la búsqueda se comparan las raíces de las palabras de la pregunta con las palabras clave libres (extraídas de los títulos y resúmenes) y con los descriptores controlados (procedentes de la indización); se cuenta un acierto por cada coincidencia entre una raíz de palabra introducida por el usuario y una cadena de caracteres de una palabra o una expresión de los documentos buscados; se combinan en un grupo por medio de una unión (O booleano) todos los términos (palabras o expresiones) encontrados así a partir de una raíz); se combinan por una intersección (Y booleano) los grupos correspondientes a las diferentes raíces de la consulta.

*Mejoras añadidas a CONIT 4*

El usuario ya no ha de introducir las raíces de las palabras; puede introducir o bien las palabras o bien expresiones (ejemplo: transplantation rejection); un algoritmo para la búsqueda de radicales va a extraer los radicales de esas palabras (ejemplo: transplant- y reject-); el sistema realiza inmediatamente una búsqueda con truncamiento en todos los términos del fichero inverso dentro de la base a la que el usuario está conectado; los conjuntos así recuperados se combinan por medio de una intersección (Y booleano); el sistema muestra los resultados de cada conjunto y de su combinación; si un radical no aporta ningún

resultado, el sistema propone mostrar los términos alfabéticamente vecinos; si, por el contrario, un término truncado suministra un número demasiado grande de referencias, el sistema reemplaza el término truncado (el radical) por la palabra completa, para disminuir el número de referencias.

Cuando dentro de una base los descriptores compuestos por dos o más palabras figuren en el fichero inverso sin haber sido descompuestos en sus palabras constituyentes, el sistema efectúa una búsqueda de la expresión completa introducida por el usuario, además de la búsqueda de la raíz de cada palabra constituyente (ejemplo: transplantation rejection, además de transplant- y reject-).

CONIT 4 ha sido objeto de una evaluación comparativa muy detallada, sobre búsquedas documentales gestionadas por:

-usuarios finales con ayuda del sistema experto;
-expertos documentalistas en modo tradicional.

Síntesis de los resultados:

-los expertos utilizan tres veces menos tiempo de conexión y dos veces menos para la búsqueda propiamente dicha;
-la utilización de CONIT aumenta el coste de la búsqueda (distribuidores + telecomunicaciones) de un 10 a un 30 %;
-los usuarios finales recuperan un 65 % más de documentos pertinentes que los documentalistas durante la sesión, y un 10 % más en total (durante la sesión + impresión en diferido).

Comentario: CONIT 4 se presenta más como un interface entre el usuario y un conjunto de distribuidores y de bases de datos que como un verdadero sistema experto, en el sentido de que obliga al usuario a manipular, y por tanto a aprender, un lenguaje de comandos comunes.

*Descripción de CONIT EXPERT* (o EXPERT)

EXPERT funciona preparándole al usuario las consultas y no tiene, por tanto, necesidad de un lenguaje de comandos: muestra unos menús y el usuario señala las respuestas exactas o reemplaza los blancos.

Una búsqueda comienza por la explicación, suministrada por el sistema, del principio de la búsqueda documental. El sistema solicita al usuario que precise el número de referencias que espera recibir y que defina cada uno de los conceptos sobre los que desea realizar la búsqueda.

El usuario introduce el número de referencias esperadas y el enunciado de su pregunta (ejemplo: LINGUISTIC THEORY y LANGUAGE LEARNING); el sistema deduce una intersección de conceptos, siendo cada concepto expresado por una unión de términos. Ejemplo: (LINGUISTIC THEORY or TRANSFORMATIONAL GRAMMAR or GENERATIVE GRAMMAR) and (LANGUAGE LEARNING or LANGUAGE ACQUISITION).

El sistema pide a continuación al usuario que seleccione, dentro de una clasificación por temas presentada

en un menú, una clase bastante amplia (ejemplo: SOCIAL SCIENCE, EDUCATION AND HUMANITIES); después, en un menú más específico, una o más clases específicas; a partir de estas elecciones, el sistema dispone las bases de datos en orden probable de pertinencia (ejemplo: ERIC, Psychology Abstracts...) y el usuario escoge una base. El sistema se conecta a esa base y traduce los términos de la pregunta en una serie de palabras clave/radicales truncados (buscando en todos los campos invertidos) y combinados por una intersección de las palabras/radicales que constituyen descriptores compuestos (ejemplo: LINGUISTIC THEORY se traduce por LINGU- and THEOR-); a continuación se combinan todos los términos de cada concepto por medio de un O, y todos los conceptos por un Y. Ejemplo: ((LINGU- and THEOR-) or (GENER- and GRAMMAR-) or (TRANSFORM- and GRAMMAR-)) and ((LANGU- and LEARN-) or (LANGU- and ACQUISIT-)). Se suministran al usuario las frecuencias de cada término. El sistema compara el número total de referencias obtenidas con el número de referencias esperadas por el usuario y sugiere, si es necesario, cómo ampliar o restringir la búsqueda.

Para permitir al usuario que reformule su consulta, si es necesario, el sistema desencadena un bucle de pertinencia: muestra los títulos y autores de una serie de referencias; el usuario selecciona las diez referencias que le parecen más pertinentes, y el sistema le muestra los títulos, autores, resúmenes, fuentes y descriptores controlados. El usuario es entonces invitado a escoger dentro de cada documento las palabras del título o de los resumenes y/o de los descriptores que considera buenos términos adicionales de búsqueda. Cuando el usuario piensa que ha visto suficientes documentos, el sistema reformula la consulta escogiendo una serie de opciones y explicando por qué lo ha hecho: el usuario recibe primero la opción de añadir cada término que acaba de escoger como término de búsqueda suplementario dentro de la lista de términos propios a uno o más conceptos; recibe a continuación la opción de reemplazar, añadir o suprimir términos. Este proceso es denominado por su autor «computer directed relevance feed-back».

Cuando la lista de términos ha sido así actualizada, el sistema la procesa de nuevo dentro de la base de datos, pero:

- -sin tener que repetir las búsquedas sobre los términos que figuran ya dentro de la primera formulación (sin prolongar, por tanto, el tiempo de búsqueda para esos términos);
- -realizando sólo, para los descriptores controlados, la comparación con las expresiones completas y no con los radicales compuestos.

Los procesos de retroacción y de reformulación son reiterados hasta que el usuario esté satisfecho con los resultados.

*+DIALECT*

DIALECT es un prototipo de sistema experto para la búsqueda documental, construido sobre el SGBD (Sistema de Gestión de Bases de Datos) ADABAS.

Se basa en:

-el análisis de preguntas, enunciadas en lenguaje natural por el usuario, para descomponer una consulta formada por estructuras elementales, que integran relaciones gramaticales y que son susceptibles de ser ligadas por relaciones booleanas;
-la búsqueda en el fondo documental (noticias y/o texto íntegro) de frases que tengan estructuras elementales próximas a la pregunta y respondiendo a la lógica booleana;
-el enriquecimiento de la consulta por medio de informaciones extraídas de las primeras frases pertinentes recuperadas;
-la búsqueda de noticias bibliográficas que contengan estructuras elementales conservadas tras el enriquecimiento.

El sistema puede funcionar en tres modos:

-usuario final (no experimentado en documentación);
-documentalista (no experto en los dominios cubiertos);
-experto (experimentado en documentación y experto en los dominios cubiertos).

El principal interés de DIALECT reside en su capacidad de trabajar sobre el texto libre (texto de resúmenes, indización libre, incluso texto íntegro). La indización por descriptores que formen parte de un thesaurus no es, pues, indispensable para el funcionamiento de DIALECT. Sin embargo, esta indización no va a ser rechazada. DIALECT utiliza los descriptores, si es el caso, para ampliar una pregunta, en función de las reglas estratégicas ad hoc que son desencadenadas, por ejemplo, después de 3 a 5 iteraciones sobre texto libre, cuando la pregunta en lenguaje natural ha sido suficientemente ampliada.

La búsqueda documental con DIALECT se realiza en dos etapas:

-inicialización y enriquecimiento de la consulta;
-extracción de los documentos que responden a la consulta.

++*Inicialización y enriquecimiento de la consulta*

-*Introducción de la pregunta del usuario en lenguaje natural*

Ejemplo: «modelo de evaluación de sistemas documentales».

-*Análisis morfo-léxico*

+Se detectan las palabras (cadenas de caracteres entre signos de puntuación o espacios en blanco).
+Se buscan las palabras dentro del diccionario.

+Para las palabras reconocidas, se extraen los datos gramaticales: género, número, forma normalizada (enunciado del nombre o del adjetivo en masculino singular, del verbo en infinitivo...) y categoría gramatical (o CG; ejemplos: NX = sustantivo, AX = adjetivo, DC = pronombre preverbial, DX = artículo, LZ = punto, VX = verbo en infinitivo, VP = verbo en participio pasado, PX = preposición..).

Ejemplo:

```
MODELO    NX
MODELOAX
MODELOVX
DEPX
EVALUACIONNX
DEPX
SISTEMANX
DOCUMENTALAX
```

+Para las palabras no reconocidas:

->en modo «usuario final»: la palabra se trata «por defecto», es decir, a partir de reglas que se basan en las categorías vecinas y en la forma (terminación) de la palabra;
->en los otros dos modos: a elección del usuario, la palabra es o bien tratada por defecto, o bien añadida al diccionario con sus datos gramaticales, suministrados por el usuario.

*-Resolución de ambigüedades gramaticales*

Una ambigüedad gramatical existe cuando una palabra es homógrafa, es decir, cuando está caracterizada por más de una categoría gramatical.

Ejemplo: modelo es verbo, nombre o adjetivo.

La resolución de la ambigüedad se realiza utilizando una regla que se basa en examinar las categorías gramaticales de los términos que preceden y que siguen al término ambiguo.

Ejemplo: la ambigüedad de «modelo»

-podrá ser solucionada en las expresiones

+un modelo de sistema de evaluación (NX)
+un sistema modelo de evaluación (AX)
+modelo un sistema de evaluación (VX)

-no podrá ser solucionada en la expresión

+el sistema parte del modelo

(porque parte es así mismo ambigua y la frase es realmente ambigua).
Las elecciones efectuadas tienen repercusiones antes y después.

Ejemplo: en la expresión «el sistema parte del modelo de evaluación», la desambiguación de «modelo», gracias a la presencia de la expresión «modelo de evaluación», permite la desambiguación de «parte».

Para las ambigüedades no resueltas:

+en modo «usuario experto»: el sistema presenta las diferentes soluciones al usuario, que escoge la categoría gramatical correcta;
+en los otros dos modos: el sistema conserva y trata todas las soluciones; en algunos casos, el sistema da preferencia sistemáticamente a una de las soluciones.

Ejemplo:

MODELO NX
DEPX
EVALUACIONNX
DEPX
SISTEMANX
DOCUMENTALAX.

*-Análisis sintáctico*

El análisis sintáctico reposa en una centena de reglas de análisis destinadas a detectar los conceptos del documento y a insertarlos dentro de los «enunciados».
Un enunciado, denominado también «tema», está formado por una terna: dos términos y una relación, así como por una serie de informaciones complementarias; tiene la siguiente forma:
(término D(, R, $C_1$, $C_2$), término A) (Lib R, Lib $C_1$) P1
donde:

-término D = forma normalizada del término que designa el objeto fuente (de partida);
-término A = forma normalizada del término que designa el objeto destino (llegada);
-relación:

->R =categoría gramatical que precisa la presencia (VX, VI..) o la ausencia (VZ) de un verbo entre D y A;
->$C_1$ =categoría gramatical que precisa la presencia (PX, PR..) o la ausencia (WW) de una preposición entre D y A;
->$C_2$ =categoría gramatical de A (NX, AX..)

-informaciones complementarias:

->Lib R =forma normalizada del verbo R;

->Lib $C_1$ =forma normalizada de la preposición $C_1$;
->P1 =plausibilidad del enunciado, calculado por el sistema en función de la naturaleza del vínculo sintáctico R, $C_1$, $C_2$, entre D y A; varía de 1,0 (vínculo fuerte) a 0,5 (vínculo débil).

Ejemplo:

(MODELO(, VZ, PX, NX), EVALUACION) ( ,DE ) 0,80

Una regla de análisis se presenta bajo la forma CONDICION/CODIGO DE PROCEDIMIENTO,
donde:

CONDICION es el orden de las categorías gramaticales de dos términos que se siguen (de manera contigua o no) dentro de la pregunta.
CODIGO DE PROCEDIMIENTO designa el procedimiento que se ha de aplicar, es decir, acciones del tipo:

+crear un enunciado D, R, $C_1$, $C_2$, A;
+suprimir un enunciado;
+modificar un enunciado;
+transformar la cadena de entrada: desplazamiento, inserción.

Los principales tipos de enunciados que el sistema puede detectar son:

+A es complemento directo, indirecto o circunstancial de D;
+A es complemento del nombre o adjetivo calificativo de D;
+A es atributo de D;
+A está coordinado con D (por una conjunción Y u O o por una coma).

Los tipos de enunciados no están, sin embargo, codificados dentro del sistema tras la colocación de los enunciados.
Se pueden presentar dos casos de ambigüedad:

+en una frase que contenga un nombre, seguido de un verbo, al que a su vez siguen varios sustantivos, sólo el primero de los sustantivos que siguen al verbo será ligado al verbo por un vínculo fuerte, mientras que los enunciados formados por el verbo y cada uno de los siguientes sustantivos serán considerados enunciados ambiguos: serán conservados, pero con un vínculo más débil;
+cuando varios sustantivos y preposiciones se suceden dentro de la misma frase, el sistema establece un enunciado de tipo ambiguo (= plausibilidad débil) entre cada sustantivo de la cadena y, por una parte, su predecesor inmediato y, por otra, el primero de la cadena.

Ejemplo:

```
1 (MODELO, VZ, PX, NX, EVALUACION)  ( ,DE )   0,080
2 (MODELO, VZ, PX, NX, SISTEMA)     ( ,DE )   0,60
3 (EVALUACION, VZ, PX, NX, SISTEMA)( ,DE )    0,70
4 (SISTEMA, VZ, WW, AX, DOCUMENTAL)( ,   )    0,80
```

El analizador conserva los enunciados ambiguos en el orden de enunciados; a medida que avanza el análisis, el sistema o bien podrá resolver la ambigüedad, o bien, si no puede, la registrará definitivamente en el orden de enunciados, pero con una plausibilidad débil.

Al final del análisis, se presenta al usuario (en modo documentalista o experto) el orden de enunciados, para que pueda decidir suprimir algunos enunciados.

Ejemplo: supresión del enunciado nº 2.

Al término de este proceso, la pregunta es transformada en una consulta, formada por la secuencia de los enunciados conservados.

Ejemplo:

```
1 (MODELO, VZ, PX, NX, EVALUACION)( ,DE )0,080
2 (EVALUACION, VZ, PX, NX, SISTEMA)( ,DE )    0,70
3 (SISTEMA, VZ, WW, AX, DOCUMENTAL)( ,   )    0,80
```

++*Extracción de los documentos que responden a la consulta*

Esta extracción se realiza a lo largo de un proceso complejo, durante el cual la consulta se desarrolla y enriquece, con el fin de introducir en ella los elementos no explícitamente contenidos dentro de la pregunta.

*-Selección inicial*

El sistema construye automáticamente una ecuación booleana, en la que:

+para cada línea de la secuencia de enunciados:

->todos los términos semánticamente equivalentes al término del tipo D, buscados dentro de una lista de sinónimos y de cuasi-sinónimos introducida a priori, forman un primer grupo de criterios de búsqueda ligados por operadores de unión O.

Ejemplo:

MODELO O SIMULACION

->todos los términos semánticamente equivalentes al término del tipo A forman un segundo grupo de criterios ligados por medio de O.

Ejemplo:

EVALUACION O MEDIDA O ESTUDIO

->los dos grupos, procedentes de una misma línea, son ligados por un operador de intersección Y.

+las parejas de grupos procedentes de diferentes líneas son, en principio, ligadas por medio de Y.

Ejemplo:

((MODELO O SIMULACION) Y (EVALUACION O MEDIDA O ESTUDIO)) Y ((MODELO O SIMULACION) Y SISTEMA) Y ((EVALUACION O MEDIDA O ESTUDIO) Y SISTEMA) Y (SISTEMA Y DOCUMENTAL).

La pregunta podrá ser modificada a continuación transformando algunos «Y» en «O».

Ejemplo:

(MODELO O EVALUACION O SISTEMA) Y (EVALUACION O SISTEMA O DOCUMENTAL).

Se confronta esta ecuación con el fichero inverso y luego con el fichero bibliográfico; el sistema va a extraer todas las frases de los documentos que respondan a la ecuación; en esta fase, el sistema extrae, pues, frases consideradas pertinentes, pero no todavía los resúmenes completos, ni los otros datos bibliográficos de los documentos que contengan esas frases.

Ejemplo de frases extraídas:

«se suministra por medio de un modelo analítico de evaluación de los costes y de la eficacia de los sistemas documentales on-line: los ejemplos de simulación...»

*-Análisis morfo-léxico*

Todas las palabras de las frases extraídas son objeto de un análisis morfo-léxico, que es completamente similar al que se realiza con las palabras de la pregunta.

*-Resolución de las ambigüedades gramaticales*

Igual que con las palabras de la pregunta.

*-Construcción de estructuras elementales*

En una primera etapa, el sistema construye la secuencia de los enunciados extraídos de las frases pertinentes apoyándose en las mismas reglas de análisis sintáctico que las utilizadas en el análisis de la pregunta.

Ejemplo:

```
1 (MODELO, VZ, WW, AX, ANALITICO)   ( ,   )   0,80
2 (MODELO, VZ, PX, NX, EVALUACION)  ( ,DE )   0,70
3 (EVALUACION, VZ, PX, NX, COSTE)   ( ,DE )   0,70
4 (EFICACIA, VZ, PX, NX, SISTEMA)   ( ,DE )   0,70
5 (SISTEMA, VZ, WW, AX, DOCUMENTAL)( ,   )    0,80
6 (SISTEMA, VZ, WW, AX, ON-LINE)    ( ,   )   0,80
7 (SISTEMA, VZ, PX, NX, SIMULACION)( ,DE )    0,70
```

En una segunda etapa, el sistema va a reconocer las dependencias entre enunciados: hay dependencia entre dos enunciados (sucesivos o no) cuando esos enunciados contienen dos términos idénticos; para cada enunciado que contiene dos términos, existen cuatro tipos de dependencias; las dos primeras son:

+dependencia de tipo D: los dos objetos de partida son idénticos.

Ejemplo: los enunciados nº 1 y 2 anteriores (objeto común: MODELO)

+dependencia de tipo A: el objeto de llegada del primer enunciado es idéntico al objeto de partida del segundo.

Ejemplo: los enunciados nº 2 y 3 anteriores (objeto común: EVALUACION)

Por último, en una tercera etapa, el sistema va a establecer «estructuras elementales» entre los enunciados dependientes.

Ejemplo:

+*Estructura elemental 1:*

```
1 (MODELO, VZ, WW, AX, ANALITICO)   ( ,   )   0,80 D
2 (MODELO, VZ, PX, NX, EVALUACION)  ( ,DE )   0,70 D
```

+*Estructura elemental 2:*

```
2 (MODELO, VZ, PX, NX, EVALUACION)  ( ,DE )   0,70 A
3 (EVALUACION, VZ, PX, NX, COSTE)   ( ,DE )   0,70 A
```

+*Estructura elemental 3:*

```
4 (EFICACIA, VZ, PX, NX, SISTEMA)   ( ,DE )   0,70 A
5 (SISTEMA, VZ, WW, AX, DOCUMENTAL)( ,   )    0,80 A
```

+*Estructura elemental 4:*

```
4 (EFICACIA, VZ, PX, NX, SISTEMA)   ( ,DE )   0,70 A
6 (SISTEMA, VZ, WW, AX, ON-LINE)    ( ,   )   0,80 A
```

+*Estructura elemental 5:*

5 (SISTEMA, VZ, WW, AX, DOCUMENTAL)( , ) 0,80 D
6 (SISTEMA, VZ, WW, AX, ON-LINE) ( , ) 0,80 D

Las estructuras elementales así construidas, denominadas «posibles estructuras», representan lo fundamental de la información de las frases que verifican la consulta.

*-Desarrollo de la consulta*

El sistema va a comparar los enunciados de la consulta con las estructuras elementales de las frases que verifican esta consulta, con el fin de añadir a la secuencia de enunciados de la consulta nuevos enunciados similares a los enunciados originales. Este enriquecimiento se realiza aplicando una de las veinte reglas disponibles.

Ejemplo:
1 (MODELO, VZ, PX, NX, EVALUACION) ( , ) 0,80

va a permitir seleccionar la primera y segunda estructuras elementales anteriores y añadir a la lista de los enunciados de la consulta los enunciados complementarios de las dos estructuras elementales:

1 (MODELO, VZ, WW, AX, ANALITICO) ( , ) 0,80
3 (EVALUACION, VZ, PX, NX, COSTE) ( ,DE ) 0,70

La segunda línea de la consulta:

2 (EVALUACION, VZ, PX, NX, SISTEMA)( ,DE ) 0,70

no permite seleccionar estructuras elementales.
La tercera línea de la consulta:

3 (SISTEMA, VZ, WW, AX, DOCUMENTAL)( , ) 0,80

va a permitir seleccionar la tercera y quinta estructuras elementales anteriores y añadir a la lista de los enunciados de la consulta los enunciados complementarios:

4 (EFICACIA, VZ, PX, NX, SISTEMA) ( ,DE ) 0,70
6 (SISTEMA, VZ, WW, AX, ON-LINE) ( , ) 0,80

Como resultado de esta etapa, la secuencia de los enunciados de la consulta, así enriquecida, se presenta como sigue:

1 (MODELO, VZ, PX, NX, EVALUACION)
2 (EVALUACION, VZ, PX, NX, SISTEMA)
3 (SISTEMA, VZ, WW, AX, DOCUMENTAL)
4 (MODELO, VZ, WW, AX, ANALITICO)
5 (EVALUACION, VZ, PX, NX, COSTE)
6 (EFICACIA, VZ, PX, NX, SISTEMA)
7 (SISTEMA, VZ, WW, AX, ON-LINE).

*-Fusión de las clases semánticas*

Durante esta etapa, los enunciados de la consulta desarrollada van a ser comparados entre sí (y ya no más con los enunciados procedentes de las frases que respondan a la consulta), utilizando el mismo análisis de dependencia que anteriormente, con el fin de fusionar los enunciados.

Ejemplo:

+fusión de los enunciados 2 y 5 (factor común: EVALUACION)
EVALUACION, VZ, PX, NX, SISTEMA, COSTE;
+fusión de los enunciados 3 y 7 (factor común: SISTEMA)
SISTEMA, VZ, WW, AX, DOCUMENTAL, ON-LINE;
+por el contrario, los enunciados 1 y 4 (factor común: MODELO) no serán fusionados, ya que los datos R, $C_1$ y $C_2$ no son los mismos en los dos enunciados.

*-Iteraciones*

El sistema busca nuevas frases pertinentes a partir de la consulta así desarrollada, y se realiza un nuevo ciclo de enriquecimiento de la consulta.

Este ciclo se repite varias veces, hasta que se produce una condición de parada, cuando el desarrollo resulta ya imposible. Estas condiciones de parada se producen por utilización de reglas estratégicas que controlan las iteraciones; estas reglas permiten controlar el desarrollo global de la búsqueda y la elección de las reglas de reformulación. El sistema gana así en eficacia y rapidez durante estas fases de iteración.

*-Presentación de la consulta y de las respuestas*

El sistema muestra, a petición del usuario:

+el estado final de la consulta.

Ejemplo:

1((EVALUACION, MODELO, EFICACIA), VZ, PX, NX, (SISTEMA, ADQUISICION, DIFUSION, COSTE))
2((SISTEMA, ADQUISICION, DIFUSION, COSTE), VZ, PX, NX, (DOCUMENTACION, DOCUMENTO, BUSQUEDA))
3((DOCUMENTACION, DOCUMENTO, BUSQUEDA), VZ, WW, AX, (AUTOMATICO))
4((EVALUACION, MODELO, EFICACIA), VZ, WW, AX, (GLOBAL, ANALITICO))
5((SISTEMA, ADQUISICION, DIFUSION, COSTE), VZ, WW, AX, (DOCUMENTAL, ON-LINE))

+así como los resúmenes y referencias de una docena de documentos, de forma que se permita al usuario valorar la pertinencia de las respuestas.

En modo «usuario final» es posible lanzar una búsqueda complementaria, de cara a obtener todos los documentos pertinentes.
En modo «documentalista» es posible modificar la consulta antes de lanzar la edición de las respuestas.

.6 Apreciación final de la pertinencia

Algunos servicios de documentación organizan una evaluación, continua o por sondeos, de la pertinencia de las referencias obtenidas en respuesta a las consultas.
Esta evaluación se realiza a dos niveles:

-la macro-evaluación;
-la micro-evaluación.

.1 Macro-evaluación

La calidad de las prestaciones de un sistema documental se puede apreciar siguiendo diferentes criterios; los más importantes, desde el punto de vista del lenguaje documental utilizado, son:

-la llamada
-la precisión

Consideremos una colección de documentos y un conjunto de documentos que son extraídos de ella en respuesta a una pregunta o a un conjunto de preguntas.
Denominemos:

a:el número de documentos pertinentes (en función de determinados criterios que serán definidos más adelante) y efectivamente extraídos;
b:el número de documentos extraídos pero que al examinarlos parecen no pertinentes: es el «ruido» propio de todo sistema de información;
c:el número de documentos pertinentes, pero no extraídos: es el «silencio», igualmente normal en la mayoría de los sistemas documentales;
d:el número de documentos no pertinentes y no extraídos.

| | Pertinentes | No pertinentes |
|---|---|---|
| Extraídos | a | b |
| No extraídos | c | d |

La tasa de llamada es igual a:

a / (a+c)

donde

a=número de documentos pertinentes suministrados en respuesta a una pregunta (o a un conjunto de preguntas)
a+c=número total de documentos pertinentes dentro del fondo, suministrados *y* no suministrados en respuesta a la pregunta (o al conjunto de preguntas).

La tasa de llamada mide un conjunto compuesto de elementos, con referencia a la calidad de la indización de los documentos y de la formulación de las consultas. Se sitúa, en general, entre el 0,6 y el 0,8, y a veces es menor.

Ejemplo:

-Sistema Medlars (Lancaster, 1968) : 0,58;
-Sistema Badadug (Balcer, 1979) : 0,41.

La determinación del valor de *c* es relativamente complicada: para conocer los documentos pertinentes dentro de un fondo dado que no han sido suministrados en respuesta a una pregunta, es necesario explorar ese fondo:

-ya sea sistemáticamente, examinando las referencias una a una (lo que tiene el riesgo de durar mucho);
-o bien interrogando de nuevo este fondo con una serie de ecuaciones muy amplias (con pocos o ningún Y, con muchos O), incluso basándose en la clasificación.

La tasa de precisión es igual a:

a / (a+b)

donde

a=número de documentos pertinentes suministrados en respuesta a una pregunta
a+b=número total de documentos, pertinentes *y* no pertinentes, suministrados en respuesta a la misma pregunta.

Se distingue:

-la precisión sistema/documentalista, tal y como es medida por el documentalista a la vista de:

+las referencias extraídas más o menos automáticamente del fichero de búsqueda en función de criterios relativamente pobres: los descriptores que han servido para indizarlos;
+las referencias que quedan tras el filtro manual que realiza el documentalista, sobre la base de una representación más rica del contenido de los documentos: sus títulos y/o su resumen.

Esta precisión se sitúa, según los sistemas, entre 0,2 y 0,8.

Ejemplo:

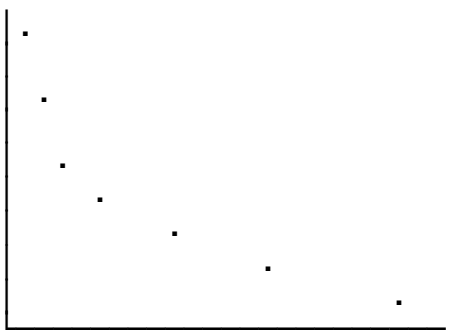
-Sistema Medlars (Lancaster, 1968) : 0,50;
-Sistema Badadug (Balcer, 1979) : 0,58;

-la precisión sistema/usuario, que es casi siempre más reducida que la precisión sistema/documentalista, ya que el usuario, en función de su necesidad real de información, procede a una segunda eliminación de los documentos que no le interesan.

La determinación de la tasa de precisión es más cómoda que la de la tasa de llamada: basta con examinar las referencias extraídas «habitualmente» del sistema y clasificarlas en dos lotes: pertinentes y no pertinentes.

*Nota:*

Dentro de un sistema documental dado, la precisión varía siempre en función inversa de la llamada según el carácter más o menos amplio o restringido de las ecuaciones de búsqueda: se mejora uno de estos criterios siempre en detrimento del otro (ley de Cleverdon). Para mejorar simultáneamente los dos criterios se necesita poner en funcionamiento nuevos medios (cualificación de los documentalistas, calidad del lenguaje documental, sofisticación del programa informático de búsqueda), generalmente costosos.



.2 Micro-evaluación

La macro-evaluación de la calidad de las respuestas a las consultas permite apreciar globalmente la llamada y la precisión de un sistema documental, pero no basta para determinar sus puntos débiles. En efecto, si se desea mejorar sus prestaciones, es necesario identificar las causas de la disfuncionalidad.

Este es el objeto de la micro-evaluación, que consiste en analizar los resultados de una muestra de varios cientos de búsquedas documentales.

Lancaster (1968) ha sido el primero en efectuar este tipo de evaluación sobre el sistema MEDLARS de la NLM - National Library of Medicine, en Washington, estudiando, consulta por consulta y documento por documento:

-la formulación de las preguntas;
-la indización de los documentos recuperados, pero no pertinentes (ruidos);
-la indización de los documentos no recuperados, aunque pertinentes (silencios).

Este examen permite:

-realizar un diagnóstico sobre los motivos de las respuestas inadecuadas:

+indización inadecuada del documento, debida a:
->una inadecuada comprensión de su contenido;
->una inadecuada traducción de sus conceptos en descriptores del thesaurus;
->lagunas en el thesaurus;
+formulación inadecuada de la pregunta, debida a:
->una inadecuada expresión, por parte del usuario, de su necesidad;
->una interpretación inadecuada, por parte del documentalista, de la petición;
->una inadecuada traducción de los conceptos de la consulta en descriptores del thesaurus;
->lagunas en el thesaurus;

-mejorar la formación de los documentalistas, especialmente intentando mejorar la coherencia de la indización, a través de una serie de ejercicios de indización en grupo. La experiencia muestra que al principio la coherencia es bastante reducida (del 20 % al 40 %) y que, tras dos o tres jornadas de trabajo y debate en común, la coherencia aumenta considerablemente (del 60 % al 80 %);
-asegurar un mantenimiento del thesaurus basado en un análisis de las disfuncionalidades y no en valoraciones teóricas. Nosotros mismos hemos puesto a punto un método de análisis de este tipo, desgraciadamente bastante pesado al llevarlo a cabo. Se trata de la determinación del grado de confianza que se puede asignar a un descriptor, medida por el grado de coherencia de su utilización en el momento de la indización. Este grado de confianza se mide de la siguiente manera:

+dos (o dos grupos de) documentalistas indizan un conjunto considerable (varios miles) de documentos por medio del mismo thesaurus, trabajando independientemente uno de otro;
+se cuenta, para cada descriptor, separadamente:
->por una parte: el número de documentos a los que los dos documentalistas han asignado ese descriptor;
->por otra parte: la frecuencia total de indización de ese descriptor, sea cual sea el documentalista que lo ha asignado;
->el grado de confianza de ese descriptor es la ratio entre estos dos números.

*Ejemplo:* consideremos los tres descriptores D1, D2 y D3, utilizados por los dos documentalistas A y B para indizar los diez documentos numerados de 1 a 10:

| Descriptores | Documentalistas | Documentos nº 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| D1 | A | | X | | X | X | | | X | | |
| | B | | X | X | X | | | | X | | X |
| D2 | A | X | | X | | | | | X | | |
| | B | | X | | | | X | | | | |
| D3 | A | | | X | | | | X | | X | |
| | B | X | | X | | | | X | | X | |

Fiabilidad:

Descriptor D1 : 3/6 = 50 %;
Descriptor D2 : 0/5 = 0 %;
Descriptor D3 : 3/4 = 75 %.

## 7. **Bibliografía**

### .1 Indización


ANDREEWSKY (A.)
*Apprentissage, analyse automatique du langage, application à la documentation* - Paris, Dunod, 1973, 275 p., FR ISSN: 0085-4786.

BROWN (A.G.) et al. -
*An introduction to subject indexing* - London: Clive Bingley, 1982, 180 p., ISBN: 0-85157-331-2.

CAMPBELL (D.J.) -
*Survey of British practice in co-ordinate indexing in information/library units* - London: Aslib, 1975, 76 p., ISBN: 85142-006-4.

CHAN (L.M.) -
*Cataloguing and classification: an introduction* - London: Mac Graw Hill, 1981, 397 p., ISBN: 0-07-010498-0.

CHAUMIER (J.) -
*Analyse et langages documentaires: le traitement linguistique de l'information documentaire* - Paris: Enterprise Moderne d'Edition, 1982, 186 p.

CLEVELAND (D. B.) and CLEVELAND (A.D.) -
*Introduction to indexing and abstracting* - London: Libraries Unlimited, 1983, 348 p., ISBN: 0-87287-346-3.

DAHLBERG (I.) -
*Major developments in classification. - in: Advances in librarianship* - vol. 7, 1977, New York: Academic Press, 1977, 348 p., ISBN: 0-12-785007-4.

DEBILI (F.) -

Analyse syntaxico-sémantique fondée sur une acquisition automatique de relations lexiales-sémantiques - Orsay: Université de Parix XI, thése de doctetur es sciences informatiques, 1982, 290 p.

DEWEZE (A.) -
Réseaux sémantiques: essai de modélisation ; aplication à l'indexation et à la recherche documentaire - Lyon: Université Claude Bernard, thèse de docteur en sciences mathématiques, 1981, 435 p.

DUCROT (J.M.) -
TITUS IV: système de traduction, automatique et simultanée en quatre langues - Commutication présentée à EURIM 5, mai 1982, 9 p.

DUPRAT (A.) -
*Spirit, version 2.1., manuel d'utilisation* - Paris: CISI Télématique, 1985, 83 p.

FEINBERG (H.) -
*ed., Indexing specialised formats and subjects* - Metuchen, N.J.: Scarecrow Press, 1983, 288 p., ISBN: 0-8108-1608-3.

FLUHR (C.) -
Algorithmes à apprentissage et traitement automatique des langues - Orsay: Université de Paris Sud, thèse de docteur es sciences, 1977, 274 p.

GRANDJEAN (E.) -
*Projet PIAF: application à la documentation automatique ; définition et utilisation du produit - prototype PIAFDOC* - Grenoble: IMAG, 1978, 40 p.

ISO -
*Documentation - Methods for examining documents, determining their subjects, and selecting indexing terms* - Genève: ISO, 1985, 5463-1985.

KNIGHT (G.N.) -
*Indexing the art of: a guide to indexing of books and periodicals* - London: Allen & Unwix, 1979, 218 p., ISBN: 0-04-029002-6.

LANGRIDGE (D.W.) and MILLS (J.) -
*An introduction to subject indexing* - London: Bingley, 1982, 180 p., ISBN: 0-85157-331-2.

MARON (M.E.) -
*On indexing, retrieval and the meaning of about, Journal of ASIS* - vol. 28, nº 1, Januyary-February 1977, pp. 38-43.

PRESCHEL (B.M.) -
*Indexer consistency in perception of concepts and in choice of terminology* - New York: Columbia University, School of Library Science, June 1972, 278 p. (ED - 063 942).

RAJAN (T.N.) -
*Indexing systems: concepts, models and techniques* - Calcutta: IASLIC, 1981, 270 p.

RICHTER (N.) -
*Grammaire de l'indexation alphabétique* - Le Mans: Bibliothèque de l'Université du Maine, 1985, 155 p., ISBN: 2-904037-04-7.

ROWLEY (J.E.) -
*Abstracting and indexing* - London: Bingley, 1982, 155 p., ISBN: 0-85157-336-3.

SPARCK JONES (K.) and BATES (R.G.)

*Research on automatic indexing* - Cambridge: University of Cambridge, Computer Laboratory, 1977,vol. 1,121 p., vol. 2,210 p.
VICKERY (B.C.) -
*Classification and indexing in science* - London: Butterworths, 1975, 228 p., ISBN: 0-408-70662-7.
NF Z 47-102 1978-08 -
*Documentation - Principes généraux pour l'indexation des documents* - Paris: Afnor, 1978.
UNISIST -
*Principes d'indexation* - Paris: UNESCO, septembre 1975, 13 p., rapport SC/75/WS/58.
LE LOARER (P.) -
*Apports de la «compréhension» du langage naturel, outils de dialogue homme/machine* - Charenton: ERLI, 1985, 1986, 21 p., document interne.

.2 Boletines de índices

CAMPEY (L.H.) -
*Generating and printing indexes by computer* - London: Aslib, Aslib occasional publication, nº 11, 1972, 101 p., ISBN: 85412-047-8.
HALL (A.M.) -
*Case studies of the use of subject indexes* - London: Institution of Electrical Engineers, 1972, Inspec report, nº R 72/8.
KEEN (E.M.) -
On the generation and searching of entries in printed subject indexes - *Journal of Documentation*, vol. 33, nº 1, March 1977, pp. 15-45.
RICHMOND (P.A.) -
*Introduction to PRECIS for North American usage* - London: Library Unlimited, 1981, 340 p., ISBN: 0-87287-240-8.
ROWLEY (J.E.) -
*A future for printed indexes.* - Aslib Proceedings - vol. 35, nº5, May 1983, pp. 234-238.
WEATHLEY (A.) -
*Manual on printed subject index* - London: British Library, 1981, 426 p., BLR & D report 5680.

.3 Búsqueda documental

BALCER (M.) et GONING (J.P.) -
Réactions de l'usager face à l'utilisation du système de repérage en mode dialogue Badadug - *Documentaliste*, vol. 16, nº2, mars-avril 1979, pp. 55-61.
BASSANO (J.C.) -
*Un système convivial pour la recherche documentaire* - RIAO, 1985, PP. 28-48.
BASSANO (J.C.) -
DIALECT, un système experte pour la recherche documentaire - Orsay: Universite de Paris-Sud, 1986, 322 p., thèse d'état, nº ordre 3127.
BOOKSTEIN (A.) -
*Implication of boolean structures for probabilistic retrieval: research and development in information*

*retrieval* - Montréal: Association for computing machinery, 1985, pp. 11-17, ISBN: 0-89791-159-8.
CHAUMIER (J.) -
*L'accès automatisé à l'information* - Paris: Enterprise Moderne d'Edition, 1982, 147 p., ISBN: 2-7101-0397-4.
CHEN (C.C.) and SCHWEIZER (S.) -
*Online bibliographic searching: a learning manual* - New York: Neal-Schumen, 1981,227 p., ISBN: 0-918212-59.
CLEVERDON (C.W.) -
*Report on the testing and analysis of an investigation into the comparative efficiency of indexig systems* - Crandfield: The College of Aeronautics, 1962, 305 p.
CORDOLIANI (H.F.A) -
*Les techniques modernes de la recherche documentaire dans les sciences biomédicales* - Rueil-Malmaison: Sandoz Editions, 1982, 254 p.
DEWEZE (A.) -
*L'accès en ligne aux bases documentaires* - Paris: Masson, 1983, 165 p., ISBN . 2-225-0037-5.
DOSZKOCS (T.E.) -
*Automatic vocabulary mapping in online searching* - International classification, vol. 10 nº 2, 1983, pp. 78-83.
DUPRAT (A.) -
*Spirit, version 2.1., manuel d'utilisation* - Paris: CISI, Télématique, 1985, 83 p.
FLUHR (G.) -
Spirit: un système syntaxique et probabiliste d'indexation et de recherche d'informations textuelles -, in: ADBS - IDT81 - Paris: ADBS, 1981, pp. 113-116.
GILCHRIST (A.) -
*The thesaurus in retrieval* - London: Aslib, 1971, 18 p., ISBN: 0-85142-036-2.
HARTNER (E.P.) -
*An introduction to automated literature searching* - New York: Dekker, 1981, 145 p., ISBN: 0-8274-1293-5.
HOOVER (R.E.) ed. -
*On line search strategies* - White Plains, NY: Knowledge Industry, 1982, 345 p., ISBN: 0-86729-004-8.
INRIA -
*Informatique et information scientifique et technique* Le Chesnay: Institut National de Recherche en Informatique et en Automatique, 1983, 396 p., ISBN: 2-7261-0318-9.
JONES (K.P.) ed. -
*Intelligent information retrieval* - London: ASLIB, 1984, 149 p., ISBN: 0-85142-187-3.
KEEN (E.M.) -
In the processing of printed subject entries during searching -, *Journal of Documentation,* vol.33, nº 4, December 1977, pp. 266-276.
LANCASTER (F.W.) -
*Evaluation of the Medlars demand search service* - Washington DC: National Library of Medicine, 1968.
LANCASTER (F.W.) -
*Information retrieval system: characteristics, testing and evaluation* - New York: John Wiley, 1979, 381 p., ISBN: 0-417-04673-6.
MANIEZ (J.) -

*Le rôle de la syntaxe dans les systèmes de recherche documentaire* - Dijon: Institut Universitaire de Technologie, 1976-1977, 2 vol.

MARCUS (R.S.) -
An experimental comparison of the effectiveness of computers and humans as search intermediaries -, *Journal of the American Society for Information Science,* vol. 34, nº 6, November 1983, pp. 381-404.

MARTIN (W.A.) -
Methods for evaluating the number of relevant documents -, *Journal of the American Society for Information Science,* vol. 34 nº 6, November 1983 pp. 173-177.

MEADOWS (C.T.) et al. -
A computer intermediary for interactive database searching ; part 1: design, part II: evaluation - *Journal of the American Society for Information Science,* vol. 33, nº 5, pp. 325-332; vol. 33. nº 6, pp. 357-364, 1982.

ODDY (R.N.) et al. -
Information retrieval through man-machine dialogue -, *Journal of Documentation,* vol. 33, nº 1, March 1977, pp. 1-14.

ODDY (R.N.) et al. -
Information retrieval research - London: Butterworths, 1981, 389 p., ISBN: 0-408-10775-8.

POLLITT (A.S.) -
A «front-end» system: an expert system as an online search intermediary -, Aslib Proceedings, vol. 36, nº 5, May 1984, pp. 229-234.

ROBERTSON (S.E.) -
Theories and models in information retrieval - *Journal of Documentation,* - vol. 33, nº 2, June 1977, pp. 126-148.

ROBERTSON (S.E.) -
The probability ranking principle in IR -, *Journal of Documentation,* vol. 33. nº 4, December 1977, pp. 294-304.

ROBERTSON (S.E.) et al. -
Probability of relevance: a unification of two competing models for documental retrieval -, *Information technology research and development,* vol. 1, nº 1 - 1982, pp. 1-21.

ROBSON (A.) and LONGMAN (J.S.) -
Automatic aids to profiles construction - *Journal of ASIS,* vol. 27, nº 4, July-August 1976, pp. 213-223.

SALTON (G.) AND MCGILL (M.J.) -
*Introduction to modern information retrieval* - New York: McGraw-Hill, 1983, 448 p., ISBN: 0-07-054488-0.

SPARCK JONES (K.) -
*Information retrieval experiment* - London: Butterworths, 1981, 352 p., ISBN: 0-408-10648-4.

STANDERA (O.) -
On-line retrieval systems: some observation on the user/system interface -. In: *Proceedings of the ASIS Annual Meeting,* vol. 12: *Information revolution,* Washington: ASIS, 1975, pp. 38-40.

TURNER (C.) -
*The basics of organizing information* - London: Clive Bingley, 1985, 160 p., ISBN: 0-85157-379-7.

VAN RIJSBERGEN (C.J.) -

*Information retrieval* - London: Butterworths, 1979, 218 p., ISBN: 0-408-70929-4.

WATERMAN (D.A.) -
User oriented systems for capturing expertise: a rule base approach -, in: Michie (D.), ed. - *Expert systems in the micro-electronic age* - Edinburgh: Edinburgh University Press, 1979, pp. 26-34.