Análisis de contenido: resumen e indización

Manela Juncà Campdepadrós

PID_00143964

Universitat Oberta de Catalunya
www.uoc.edu

Índice

Introducción

Este módulo os introduce en los procesos documentales de la segunda fase de la cadena documental, llamada **análisis de contenido,** formada por el resumen y la indización.

Itinerario de estudio

El módulo empieza con un capítulo dedicado al análisis de contenido, para situar al estudiante en las dos operaciones mencionadas, el resumen y la indización.

El apartado dedicado al **resumen** está diseñado para responder a las preguntas de qué es un resumen, quién lo redacta, qué utilidades tiene y cuántos tipos de resúmenes hay. Finalmente, se presentan los resúmenes automáticos, explicando su evolución y funcionamiento.

La **indización** es el grueso de esta asignatura y en este módulo tiene tres apartados. El primero de ellos trata de dar respuesta a las preguntas de qué es indizar, quién indiza, por qué hacen falta los lenguajes documentales y cómo se indiza. El apartado titulado "Lenguajes documentales" responde a las preguntas de qué son los lenguajes, cuántos hay, qué son los términos de indización, cómo han evolucionado, cuándo son necesarios y cuál es su uso en solitario o combinados. El último apartado, titulado "Tipología", trata de los diferentes criterios usados para clasificar los lenguajes.

Este es un módulo básico para el aprendizaje de la terminología que se usará en el resto de módulos.

Conceptos más importantes

Concepto	Ved
Resumen informativo Resumen indicativo Resumen selectivo Resumen automático	1. El resumen
Ambigüedad Lenguaje natural Exhaustividad Especificidad Traducción Univocidad	2. La indización

Concepto	Ved
Lenguaje documental Sistemas de clasificación Listados de encabezamientos de materia Listados de autoridades Tesauros Listados de descriptores libres Listados de palabras clave Notación Encabezamiento Descriptor Identificador o autoridad Palabra clave	4. Los lenguajes documentales
Codificado Natural Libre Controlado Precoordinado Postcoordinado Jerárquico Combinatorio Materias Conceptos Palabras clave	5. Tipología de los lenguajes documentales

Objetivos

Con el estudio de los materiales asociados a este módulo alcanzaréis los objetivos siguientes:

En cuanto al **resumen**:

1. Aprender a hacer resúmenes de manera intelectual: resúmenes informativos, indicativos y selectivos.

2. Aprender a hacer resúmenes con programas de resúmenes automáticos.

En cuanto a la **indización**:

1. Analizar los factores necesarios para que haya una buena comunicación documental: entender los problemas del lenguaje natural y la función de los lenguajes documentales dentro de esta comunicación.

2. Conocer los procesos de indización: examen del documento, selección y traducción.

En cuanto a los **lenguajes documentales**:

1. Conocer las características principales de los lenguajes documentales.

2. Conocer la evolución histórica de los lenguajes documentales.

3. Aprender a distinguir y saber utilizar la diferente tipología de los lenguajes documentales: sintéticos-analíticos, precoordinados-postcoordinados, controlados-libres, jerárquicos-combinatorios, materias-conceptos-palabras clave.

1. El análisis de contenido

El **análisis de contenido** se sitúa en la segunda fase de la cadena documental y reúne todo el conjunto de operaciones destinadas a representar la materia de los documentos para una posterior recuperación.

Son tareas de cariz intelectual en las que la formación y la habilidad del analista juegan un papel importante.

> "Representar la materia" o "describir el contenido" es responder a la pregunta: "¿cuál es el tema de un documento?".

Para representar el contenido de un documento el analista tiene que llevar a cabo dos operaciones:

1) El **resumen**, que condensa el contenido en un texto más breve y manejable.

2) La **indización**, que identifica los conceptos o temas principales. También se conoce como descripción característica.

Estas dos operaciones admiten una elaboración humana o automática. Por lo tanto, habrá resúmenes elaborados por documentalistas y resúmenes elaborados por programas, y también indizaciones hechas por analistas e indizaciones elaboradas por un software.

Operaciones humanas y automatizadas

	Humano	Automatizado
Resumen	Resumen informativo Resumen indicativo Resumen selectivo	Resumen automático
Indización	Sistemas de clasificación Listas de encabezamientos de materia Listados de autoridades Tesauros Listados de descriptores libres	Listado de palabras clave

Los dos sistemas tienen ventajas e inconvenientes. La calidad y coherencia que aporta un documentalista supera en estos momentos la que ofrecen los programas informáticos, pero en cambio los sistemas automáticos son instantáneos, baratos y capaces de asumir ingentes cantidades de documentos.

La rama científica que estudia cómo emular el conocimiento humano, en cuanto a la identificación de los conceptos y las frases con contenido relevante para el resumen y la indización, es el procesamiento en lenguaje natural.

El procesamiento en lenguaje natural (PLN[1]) es una rama de la inteligencia artificial y de la lingüística computacional que estudia los lenguajes que usan los humanos para interactuar con los ordenadores en contextos escritos y orales.

Ved también

Trataremos el procesamiento del lenguaje natural en el subapartado 2.4 y en el apartado 3.

(1)PLN es la sigla de *procesamiento en lenguaje natural*.

Lectura complementaria

I. Gil Leiva; J. V. Rodríguez Muñoz (1996). "El procesamiento del lenguaje natural aplicado al análisis del contenido de los documentos". Revista general de información y documentación (vol. 6, núm. 2, pág. 205-218).

A modo de conclusión

Para representar o describir el contenido de un documento el analista tiene que llevar a cabo dos operaciones:

- El resumen, que condensa el contenido en un texto más breve y manejable.
- La indización, que identifica los conceptos o temas principales. También se conoce como descripción característica.

Las dos operaciones se pueden llevar a cabo de manera humana o automática.

2. El resumen

Según la norma UNE 50-103-90 *Preparación de resúmenes*, un **resumen** es la presentación abreviada y precisa de un documento, sin interpretación ni crítica y sin mención expresa del autor del resumen.

Ved también

Encontraréis la norma UNO 50-103-90 en el espacio "Materiales y fuentes" de las aulas.

Cuando decimos documento nos estamos refiriendo a todo tipo de documento, sea cual sea su soporte material. Podemos resumir un texto, la imagen de una fotografía, un vídeo, audios, información en línea o hipertextos.

Los resúmenes, como la indización, pueden ser de elaboración humana o automática. En el primer caso hay cuatro tipos de personas que pueden redactar un resumen. En el caso de los resúmenes automáticos, se trata de un software.

1) Resumen humano:

a) El **autor** del documento. Los resúmenes elaborados por los propios autores son muy habituales en el mundo de las comunicaciones científicas y tecnológicas.

Resúmenes para revistas

Las revistas acostumbran a dar directrices a sus autores para la elaboración de resúmenes. Ved, por ejemplo, la revista *EPI* en su apartado "Instrucciones para los autores".

b) Un **especialista** en la materia de la que trata el documento.

c) La **editorial**. Son los resúmenes que aparecen en la contraportada de los libros impresos y que tienen una función claramente publicitaria.

d) Un **profesional de la documentación**. Aporta su conocimiento sobre la redacción de buenos resúmenes y los elabora pensando en las utilidades futuras.

2) Resumen automático: los programas se conocen como programas resumidores de textos o *Automatic Text Summarizer*.

Programas resumidores de textos

Un ejemplo de programas resumidores de textos es Swesum, que hace un análisis estadístico del texto y elabora el resumen con los fragmentos que contienen las palabras más ponderadas (más repetidas pero con significado).

La norma internacional ISO 214:1976, traducida por AENOR como norma UNE 50-103-90 *Preparación de resúmenes*, establece las directrices que se tienen que seguir para presentar los resúmenes en los documentos. Pone especial énfasis en la preparación de resúmenes por parte de los autores de los documentos primarios y en la misma publicación.

Redactar un resumen es fácil. Lo difícil es redactar un buen resumen. El punto de inflexión es la calidad del resumen, que lo hará más o menos útil en un sistema documental. Un resumen propagandístico no aportará muchos conceptos principales para indizar, aunque haya sido un buen reclamo para las ventas.

Ejemplo de resumen elaborado por la editorial con finalidad publicitaria

SAGAN, Carl. *Cosmos*. Traducció: Albert Santamaria i Martínez; pròleg: Ricard Guerrero. Barcelona: Publicacions i Edicions de la Universitat de Barcelona: Omnis Cellula, cop. 2006.

"He aquí una de las obras más destacadas de la literatura internacional de divulgación científica, publicada por primera vez en catalán. Una obra imprescindible de uno de los grandes maestros de la divulgación, que nos introduce en los grandes enigmas que la humanidad ha tratado de entender y explicar desde tiempos inmemoriales, y por los cuales ha nacido lo que llamamos ciencia.

Desde la infinitud del Universo hasta el mundo invisible de los átomos, desde el nacimiento de las estrellas hasta la aparición de la vida, Carl Sagan consigue transmitir los conocimientos de la ciencia actual de una manera clara y apasionante."

Para un analista sólo tendría utilidad el último párrafo, en qué aparecen términos como *universo*, *átomos*, *estrellas*, *vida*.

El resumen es útil en dos fases de la cadena, en los procesos de selección y adquisición que se da en la primera fase de la cadena y en la fase de salida, donde es un excelente instrumento de recuperación, ya que el resumen ofrece más datos que la simple referencia documental. La principal utilidad del resumen es la de difundir la información.

Difundir la información

Cada vez más bases de datos referenciales ofrecen el resumen de sus monografías y revistas, como por ejemplo Ebsco, Dialnet, Compludoc, CBUC, Eric database o *ISI current contents connect*. También lo hacen las bases de datos de novedades editoriales, por ejemplo la editorial Trea (recomendamos el acceso desde la biblioteca de la UOC).

En todos los casos es indudable el valor informativo que aporta el resumen para difundir el contenido del documento de la colección. Pero además, el resumen tiene otras utilidades, tal como dice la norma UNE 50-103-90:

a) Determinar la pertenencia: un resumen bien elaborado capacita a los lectores para identificar de forma rápida y precisa el contenido de un documento y decidir si hay que leerlo en su totalidad.

b) Evitar la lectura del texto completo en documentos de interés secundario. Un resumen bien elaborado proporciona suficiente información sobre temas que no sean de interés principal para el lector. Ahorra tiempo al usuario.

c) Ayudar en la búsqueda automatizada. Los resúmenes automatizados incorporados en los catálogos son muy útiles para:

- Extraer términos de indización de su texto, es decir, indizar a partir del resumen.

- Hacer búsquedas de palabras clave que no se encuentran en el título.

- Servir de control bibliométrico, al comparar los términos usados en una ecuación de búsqueda con los términos que aparecen en un resumen y así establecer la pertinencia de la recuperación.

- Ayudar a la difusión desde los servicios de alerta.

Según María Pinto (1992), las **características de un resumen** son las siguientes:

Lectura complementaria

Podéis ampliar la información sobre el resumen leyendo la obra siguiente:

M. Pinto Batanea (1992). *El resumen documental: principios y métodos*. Madrid: Pirámide/ Fundación Germán Sánchez Ruipérez (Biblioteca del Libro, Y).

- Brevedad. Se tienen que omitir datos preliminares o temas del conocimiento común.

- Pertinencia. El resumen se tiene que adecuar al mensaje principal del documento, sin obviar o interpretar los datos.

- Claridad y coherencia. Frases completas, dotadas de coherencia lineal y global.

- Profundidad. Varía en función del tipo de resumen o de los diferentes niveles de detalle que se persigan.

- Consistencia lingüística. Un resumen se tiene que adaptar a las pautas lingüísticas en uso y tiene que tener en cuenta las reglas morfológicas y sintácticas correspondientes.

- Proximidad cronológica entre las ediciones del documento original y el resumen. Es importante que el tiempo transcurrido entre la publicación del original y el resumen no sea excesivo, especialmente en ámbitos científicos y técnicos.

A modo de conclusión

- El resumen es la presentación abreviada y precisa de un documento, sin interpretación ni crítica y sin mención expresa del autor del resumen.

- El resumen puede ser redactado por el autor del documento, un especialista en la materia, la editorial, un documentalista o un programa informático.

- El resumen es útil en dos fases de la cadena: en los procesos de selección y adquisición que se da en la primera fase de la cadena y en la fase de salida, donde es un excelente instrumento de recuperación.

- La principal utilidad del resumen es la de difundir la información, pero además, el resumen tiene otras utilidades, como determinar la pertinencia, evitar la lectura del texto completo en documentos marginales y ayudar a la búsqueda automatizada.

- Los resúmenes automatizados incorporados en los catálogos son muy útiles para extraer términos de indización del texto, para hacer búsquedas de palabras clave que no se encuentran en el título, para servir de control bibliométrico y ayudar a la difusión a través de los servicios de alerta.

2.1. Tipos de resúmenes

Hay diversos tipos de resúmenes, según el tamaño, los usuarios y la profundización en el contenido. Los tipos más habituales son los resúmenes informativos, indicativos y selectivos.

1) Resumen informativo

Redactaremos el tema central, temas adicionales, naturaleza y objetivo del documento, metodología, resultados, conclusiones y anexos. La idea de fondo es que un resumen informativo puede sustituir en ocasiones la lectura del documento original. La norma UNE 50-103-90 recomienda que el esquema a seguir sea el de:

objetivo + metodología + resultados (o conclusiones)

Sin embargo, no hay que seguir forzosamente este orden, ya que hay entornos, como el técnico científico, donde se prefieren los resúmenes orientados a los resultados (para que la discriminación sea más rápida).

En cuanto al tamaño del resumen, la norma da pautas pero advirtiendo que el contenido del documento es más significativo que las pautas para determinar la extensión del resumen. De todas maneras la norma nos sugiere:

- Monografías, informes, tesis: 500 palabras.
- Artículos de revista, capítulos de monografías: 250 palabras.
- Comunicaciones breves: 100 palabras.

Ejemplo de resumen informativo

CONSUEGRA FERNÁNDEZ, Jesús: "El Ajedrez: evolución y claves de un juego milenario". En *Mundo antiguo*. Madrid: 2002. n° 3-4, año 1, p. 60-61.

"Artículo divulgativo sobre el juego del ajedrez, estructurado según sus orígenes, antigüedad, expansión, variantes y simbolismo.

El origen del ajedrez es hindú y el primer representante conocido es el Ghaturanga, aparecido entre el 3000 y el 2000 a.C. en Sri Lanka, aunque no aparece documentado hasta el siglo VII d.C.

Del Ghaturanga proceden en cascada las diferentes variantes del ajedrez: de la India viajó a Persia en el siglo VI d.C., donde pasó de los 4 jugadores originales a 2 en la versión persa Shatranj. Desde Persia se extendió hacia Occidente y hacia Oriente.

Hacia Occidente: paralela a la expansión árabe, el juego llega a la Península Ibérica durante la Alta Edad Media, y desde aquí se expande al resto de Europa y al resto del mundo en la época de las colonizaciones.

Hacia Oriente: en la China, en el s. VII d.C., el ajedrez toma la forma del ajedrez chino Xiang qi; en el Japón, el Shogi; en Indochina, el ajedrez birmano y tailandés. Tanto en Oriente como en Occidente, el ajedrez presenta innumerables variaciones locales.

El tablero y las fichas parecen poseer un significado simbólico. El tablero, con la alternancia de casillas blancas y negras, forma un mandala. El simbolismo de las fichas es menos esotérico y ha ido cambiando según los tiempos: obispos, elefantes, etc.

El autor concluye que el ajedrez, además de un juego, es una herramienta educativa de primer orden, casi una ciencia."

Como podéis comprobar, este resumen tiene 237 palabras.

2) Resumen indicativo

Redactaremos sólo las ideas centrales del documento. Su lectura no puede sustituir la lectura del original. Como su nombre sugiere, el resumen indicativo presenta de forma abreviada y muy sintética el contenido o la tipología del documento. Su extensión puede oscilar entre una frase o 4 líneas de texto.

Ejemplo de resumen indicativo

CONSUEGRA FERNÁNDEZ, Jesús: "El Ajedrez: evolución y claves de un juego milenario". En *Mundo antiguo*. Madrid: 2002. n° 3-4, año 1, p. 60-61.

"Artículo divulgativo sobre el juego del ajedrez, trata de su origen hindú, antigüedad, expansión histórica tanto en Oriente como en Occidente, variantes nacionales y simbolismo del tablero y las fichas."

3) Resumen selectivo

Redactaremos sólo una parte concreta del documento. El más habitual es el resumen de conclusiones, pero también hay otros tipos, como la reseña (*review*), que es un análisis del documento con elementos críticos. Este tipo de resumen se adapta muy bien a las necesidades de los usuarios, por ejemplo investigadores o técnicos que necesitan un dato muy concreto sobre el objetivo del documento o las conclusiones a las que llega.

Ejemplo de resumen selectivo

CONSUEGRA FERNÁNDEZ, Jesús: "El Ajedrez: evolución y claves de un juego milenario". En *Mundo antiguo*. Madrid: 2002. n° 3-4, año 1, p. 60-61.

"El ajedrez, además de un juego, es una herramienta educativa de primer orden, casi una ciencia."

A modo de conclusión

Los resúmenes más habituales son el resumen informativo, el indicativo y el selectivo:

- El **resumen informativo** consigna el tema central, temas adicionales, naturaleza y objetivo del documento, metodología, resultados, conclusiones y anexos. La idea de fondo es que un resumen informativo puede sustituir en ocasiones a la lectura del documento original.
- El **resumen indicativo** consigna sólo las ideas centrales del documento. Su lectura no puede sustituir a la lectura del original.
- El **resumen selectivo** consigna sólo una parte concreta del documento. El más habitual es el resumen de conclusiones, pero también hay otros tipos, como la reseña (*review*).

2.2. Resumen automático

Una de las necesidades más perentorias ante el aumento de información digital debido al crecimiento exponencial de Internet es manejar y filtrar el gran volumen de información. Una de las soluciones aportadas por el PLN han sido los programas de resumen automático, que actúan sobre textos, imágenes, webs y correo electrónico.

Hans Peter Luhn

Los primeros en trabajar en el campo de la automatización de los resúmenes fueron Hans Peter Luhn en el año 1958 y Edmundson en 1969, que aplicaron técnicas como la frecuencia de las palabras, o la posición de una frase dentro de un documento para redactar resúmenes sin intervención humana.

A partir de estas primeras investigaciones se han perfeccionado muchas técnicas diferentes basadas en conocimiento y recursos lingüísticos (como las de Lin y Hovy, 2002; Gotti *et al.*, 2007) o basadas en métodos estadísticos y de aprendizaje automático (Hirao *et al.*, 2002; Svore, 2007) (autores citados en Lloret *et al.*, 2008; y Mateo *et al.*, 2003).

Últimamente las investigaciones giran en torno al resumen multidocumento, es decir, resumir más de un documento (Goldstein *et al.*, 2000; Qiu, 2007; Huo y Chen, 2008) de contenidos afines o redundantes (autores citados en Lloret *et al.*, 2008; y Mateo *et al.*, 2003).

Los resúmenes automáticos se conocen también como *extracts*. La terminología anglosajona diferencia así los *extracts* y los *abstracts*. Los *extracts* son los resúmenes formados a partir de la extracción de algunas frases del texto previamente seleccionadas por un programa, mientras que los *abstracts* son los resúmenes elaborados por una persona.

La base de todas las técnicas de funcionamiento de un programa de resúmenes automático es el cómputo de la frecuencia de las palabras.

Hay diversas herramientas para hacer estos cálculos, por ejemplo WVTool. Se trata de contar cuántas veces sale una palabra no vacía en el texto.

Lecturas complementarias

Podéis consultar los resultados de las investigaciones de estos autores en los artículos siguientes:

E. Lloret; O. Ferrández; R. Muñoz; M. Palomar (2008). "Integración del reconocimiento de la impliación textual en tareas automáticas de resúmenes de textos". *Procesamiento del lenguaje natural*, n°. 41, pág. 183-190.

P. L. Mateo; J. C. González; J. Villena; J. L. Martínez (2003). Un sistema para resumen automático de textos en castellano.

Ved también

Encontraréis una explicación detallada sobre las palabras vacías en el módulo "Indización automática y descriptores libres".

Ejemplo de funcionamiento de un programa de resúmenes automático (extraído de Lloret *et al.*, 2008)

"Tropical storm Gilbert formed in the eastern Caribbean and strengthened into a hurricane Saturday night. There were no reports of casualties."

Oración 1:	Tropical (2) storm (6) Gilbert (7) formed (1) in (0) the (0) eastern (1) Caribbean (1) and (0) strengthened (1) into (0) a (0) hurricane (7) Saturday (4) night (2).
Oración 2:	There (0) were (0) no (0) reports (1) of (0) casualties (1).

Lo primero que vemos es que las palabras vacías, es decir, las palabras que no tienen significado (preposiciones, artículos, verbos) no se computan.

Al lado de cada palabra con significado vemos el número de veces que sale en todo el texto. Se suman los valores, de manera que la oración 1 tiene 3,2 puntos y la oración 2, 0,2. El programa seleccionará la frase 1 como más representativa para el resumen automático.

Este sistema de resumir a partir de las frases con las palabras más significativas en el texto parece simplista pero tiene cierta justificación. Según Kupiec *et al.* (1995) aproximadamente el 80% de las frases en resúmenes humanos están copiadas literalmente o con pequeñas modificaciones del texto original.

A partir de esta base estadística se incorporan otras técnicas para dotar al programa de más conocimiento y paliar la escasa coherencia del resultado, como puede ser, por ejemplo, la resolución de la anáfora o aplicar programas (por ejemplo, WordNet) que proporcionen relaciones como las de sinonimia o hiperonimia, o mecanismos para detectar y eliminar la redundancia.

Definimos brevemente qué son las anáforas y la hiperonimia:

a) Las **anáforas** son la relación de referencia entre un elemento lingüístico y otro anterior en el discurso.

b) Decimos que una palabra es **hiperónima** cuando tiene un campo significativo que incluye otro de menor extensión.

Los expertos consideran que la tecnología actual no tiene problemas para detectar las frases con más significado, pero sí para ordenarlas según su importancia.

Los programas funcionan a grandes rasgos de la siguiente manera: se copia el texto a resumir o bien se escribe la dirección del documento. Se escoge el tipo de documento (académico, periodístico, etc.) y el tanto por ciento de reducción del texto.

A continuación tenéis unos cuantos programas de los más conocidos:

Anáfora

"El Salón del Hobby ha tenido más de 60.000 visitantes este año. Este salón se ha convertido en la feria de ocio familiar más visitada".

En este ejemplo, la anáfora se da en "este salón", que hace referencia al Salón del Hobby, expresado en la frase anterior. Como se puede comprobar, si en el resumen automático aparece sólo la segunda frase, el lector no sabrá a qué salón hace referencia.

Hiperonimia

Color es un hiperónimo. Su contrario es hipónimo: *amarillo*, *naranja*, *verde* son hipónimos.

- Connexor
- Daedalus
- Extractor
- FociSum
- InTEXT (Dynamic Summarizing)
- Inxight Summarizer
- IslandInText
- K-Site de Daedalus
- Pertinence Summarizer
- Sinope Summarizer
- Summarizer
- SweSum[2]
- System Q
- TextAnalyst
- Trestle

[(2)]Podéis practicar con el programa Swesum, que es gratuito y traduce al español.

El programa K-Site de Daedalus

De entre los programas de resumen automático mencionados, veamos el funcionamiento del programa K-Site de Daedalus. Este programa tiene cinco módulos:

- **Módulo 1: Análisis morfosintáctico**. En este módulo se determina la categoría léxica de cada palabra: sustantivo, verbo, adjetivo, artículo, preposición, etc. También se determina el lema. Estas operaciones permiten distinguir las palabras con significado (sustantivos, adjetivos, verbos) de las vacías (artículos, preposiciones, pronombres, etc.). El lema permite agrupar todas las palabras que son flexiones de otra (info/informar/información/informador/informacional/etc.). El producto final es un listado con las palabras puntuadas y un listado de frases candidatas.

- **Módulo 2: Ponderación de frases**. Este módulo recibe las palabras etiquetadas por el módulo anterior, y su función es escoger entre todas las frases candidatas. Para hacerlo se ayuda de diversos submódulos que ponderan las frases según los parámetros siguientes: la frecuencia, la presencia de palabras indicativas (buscan palabras como *importante, esencial, conclusiones*, etc.), buscan frases que contengan palabras que aparezcan en el título, o que tengan nombres propios, o que la tipografía sea destacada (negritas, cursivas, tamaño superior, etc.) y seleccionan frases que aparezcan en posiciones destacadas en el texto (al principio de cada párrafo, al final a modo de conclusión).

- **Módulo 3: Detección de anáforas**. Una vez tiene las frases seleccionadas, puede ser que se dé el caso de anáforas mal resueltas (una frase contiene una anáfora que se encontraba en la frase previa y que no ha sido seleccionada). El programa busca las anáforas (especialmente los demostrativos pronominales o pronombres personales, por ejemplo *este, aquel, lo que*, *eso*) y su posición en la frase: al principio, entre las seis primeras palabras, en otras posiciones.

- **Módulo 4: Selección de frases**. Este módulo computa toda la información recogida en las fases anteriores: frases candidatas, puntuaciones, detección de anáforas. Selecciona las frases candidatas de puntuación más alta hasta llegar al tanto por ciento pedido por el usuario. Si entre estas frases hay alguna que contenga una anáfora, se selecciona la frase anterior (que contiene la palabra a la cual se está haciendo referencia) siempre y cuando forme parte de las frases candidatas y no sobrepase la longitud del resumen.

- **Módulo 5: Postprocesado del extracto**. Su función es detectar expresiones que conectan partes del texto, ya sea para mostrar causalidad, contraposición, etc. Son expresiones del tipo *por lo tanto, en contra*, etc. Como en el caso de las anáforas, si forman parte de una frase seleccionada, se procura incluir en el resumen la frase con la cual están relacionadas.

Por último, debemos recordar que algunos procesadores de textos, como Microsoft Word, también ofrecen esta opción (*Autosummarize* o Auto-resumen).

A modo de conclusión

- Los resúmenes automáticos (*extracts*) son una de las soluciones aportadas por el PLN para hacer frente al manejo de grandes volúmenes de información en línea.
- Los primeros en trabajar en el campo de la automatización de los resúmenes fueron Hans Peter Luhn en el año 1958 y Edmundson en 1969.
- Las técnicas han evolucionado de los primeros cómputos sobre la frecuencia de las palabras, o la posición de una frase dentro de un documento, a las técnicas basadas en conocimiento y recursos lingüísticos o en métodos estadísticos y de aprendizaje automático.
- La base de todas las técnicas es el cómputo de la frecuencia de las palabras. A partir de esta base estadística, se incorporan otras técnicas para dotar al programa de más conocimiento y paliar la escasa coherencia del resultado, por ejemplo la resolución de la anáfora o se aplican programas que proporcionen relaciones como las de sinonimia o hiperonimia o mecanismos para detectar y eliminar la redundancia.
- Los expertos consideran que la tecnología actual no tiene problemas para detectar las frases con más significado, pero sí para ordenarlas según su importancia.

3. La indización

"Indizar es la acción de describir o identificar un documento con relación a su contenido."

Norma UNE 50-121-91.

Indizar es el resultado de examinar el documento, seleccionar los conceptos y almacenarlos en una base de datos.

Esta definición implica tres acciones, de las cuales la más significativa es la selección de los conceptos y su traducción al lenguaje documental.

Al igual que se ha tratado en el resumen, la indización la puede realizar una persona o un programa.

Ved también

La indización se estudia en los módulos "Sistemas de clasificación documentales", "Listas de encabezamientos y listados de autoridades", "Los tesauros" y "Listado de descriptores libres y listado de palabras clave".

Si la indización es intelectual, es decir, la llevan a cabo personas, estas personas pueden ser:

- **Profesionales** (documentalistas), que llevan a cabo la tarea de indización de manera individual o en equipo. A su vez, los equipos pueden indizar de manera centralizada o coordinada.

- **Amateurs** (usuarios de Internet que indizan de manera social o *tagging* –por ejemplo, en Delicious–).

Ved también

La forma de indizar de los equipos se trata en el apartado 5 del módulo "La cadena documental" de esta asignatura.

El elemento humano permite un análisis más rico del documento, captando conceptos y matices que un programa no llegaría a detectar, pero tiene el inconveniente del tiempo que se tiene que dedicar y la coherencia entre indizadores.

La indización automática se realiza a través de un programa informático. Su funcionamiento es muy sencillo: extrae del título, resumen o texto completo las palabras más significativas. Es un método económico y muy rápido.

Ved también

La indización automática se estudia en el módulo "Listado de descriptores libres y listado de palabras clave".

3.1. Lenguaje natural y lenguaje documental

Para indizar necesitamos los lenguajes documentales. ¿Qué diferencia hay entre el lenguaje natural y el documental?

Por **lenguaje natural** entendemos el lenguaje que usamos de forma cotidiana: catalán, castellano, vasco, gallego, francés, etc.

Por **lenguaje documental** entendemos el listado o vocabulario de términos que usamos para indizar y que puede estar en formato libre o controlado.

¿Y por qué hay que controlar los términos del lenguaje natural? Porque el lenguaje natural es ambiguo, los conceptos se pueden representar de formas diversas, dando lugar a problemas de recuperación. El lenguaje natural es rico en terminología, en formas (plurales y singulares), tiempos verbales, acrónimos, sinónimos, polisemias, etc.

La riqueza del lenguaje natural

- Ejemplos de sinónimos del mismo concepto: Cosmos / Universo / Infinito / Firmamento / Cielo.
- Ejemplo del mismo concepto en formas diferentes, siglas o frases, y en idiomas diferentes: OTAN / NATO / Organització del Tractat de l'Atlàntic Nord / Organización del Tratado del Atlántico Norte / North Atlantic Treaty Organization.
- Ejemplo de polisemia: Banco / Planta / Carta / Sierra / Estrella / Lengua / Capital.

La principal diferencia entre el lenguaje natural y el documental controlado es precisamente el control terminológico, que permite representar los conceptos de forma unívoca, sin ambigüedades.

Para ser más concretos, las diferencias se dan en el número de términos del vocabulario, el control de las formas, el control del significado y las relaciones de significado entre términos.

Univocidad

La univocidad consiste en representar un concepto con un único término.

3.1.1. Número de términos

Los lenguajes documentales son entrópicos (Blanca Gil, 2004, pág. 20), es decir, tienden a la selección, a la restricción del vocabulario. Es el proceso contrario del lenguaje natural, que tiende a la abundancia, a la reiteración de conceptos, a la sinonimia en beneficio de una expresión más rica.

Los lenguajes documentales reducen considerablemente el número de términos del lenguaje natural, ya que sólo tienen en consideración los sustantivos y algunos sintagmas nominales, pero no adjetivos, preposiciones, conjunciones, adverbios, verbos, etc. Además, entre todos los sustantivos, escogen uno que representará al resto cuando el significado sea el mismo. Y entre diversas formas aceptadas por el mismo término, sólo una será la aceptada, como es el caso de las siglas.

Los lenguajes documentales son en esencia sencillos, su eficacia aumenta a medida que las reiteraciones y la redundancia son controladas en una única forma que reúne conceptos afines.

3.1.2. Control de las formas

Los lenguajes documentales controlan las formas plural/singular, el uso de acrónimos y siglas y la construcción de las frases, y de esta manera establecen unos modelos.

Ejemplo

Modelo	Ejemplo
Sustantivo	Pintura
Sustantivo + adjetivo	Pintura medieval
Sustantivo + preposición + sustantivo	Pintores de vitrales

Estas reglas gramaticales y sintácticas unifican las palabras seleccionadas y las frases.

Ved también

Los mejores ejemplos se ven en los módulos "Listas de encabezamientos y listados de autoridades" y "Los tesauros".

Ejemplos en las listas de encabezamientos de materia

- Se acostumbra a usar el singular para expresar conceptos abstractos. Así, por ejemplo, es *solidaridad* y no *solidaridades*.
- No se permite el uso de siglas; se prefiere la expresión entera del concepto y en la lengua del servicio de información y documentación (SID[3]). Por ejemplo, Organización del Tratado del Atlántico Norte.
- Es preferible la expresión natural del concepto compuesto, y no su forma inversa. Es correcto *Objetos de arte*, y no *Arte, objetos de*.

[(3)]SID es la sigla *de servicio de información y documentación.*

3.1.3. Control del significado

Los problemas más importantes en cuanto al significado son la sinonimia y la polisemia.

a) Sinonimia: decimos que las palabras son sinónimas cuando tienen el mismo significado. En un sistema documental, si no se controlan y se usan indiscriminadamente, comportan silencio documental. En el caso de "alimento, nutriente, comida, provisión", el usuario puede estar buscando por "alimento" y no recuperar documentos porque se encuentran indizados con otras formas, como "nutriente". La solución de los lenguajes controlados es recoger todos los términos sinónimos y seleccionar uno para representar a todo el conjunto de términos que tienen el mismo significado, porque dos sinónimos son sustituibles el uno por el otro en cualquier contexto.

Ejemplo

Una lista de encabezamientos de materia como la del Consejo Superior de Investigaciones Científicas (CSIC) recoge todos estos sinónimos:

- Hispanoamericanos.
- Iberoamericanos.
- Latinoamericanos.
- Sudamericanos.

Pero sólo da como término aceptado "Latinoamericanos". Si al SID[4] llegara un documento titulado "Los sudamericanos del siglo XX", el analista lo indizaría como **Latinoamericanos**, ya que es el término aceptado.

(4)A partir de ahora denotamos *servicios de información y documentación* con la sigla SID.

b) Polisemia: decimos que dos palabras son polisémicas cuando el mismo signo lingüístico, palabra o sonido tiene más de un significado. Habitualmente el contexto de la conversación o lectura donde está insertada la palabra deshace los problemas de ambigüedad, pero una palabra polisémica introducida en un sistema documental, sin el contexto, puede dar lugar a ruido documental.

Ejemplo

Un usuario puede estar buscando sobre columnas en arquitectura y recuperar datos sobre columnas tipográficas de diarios. Los lenguajes documentales controlan la polisemia diferenciando cada significado con paréntesis, usando el plural o el singular, adjetivando, etc.

Un tipo de polisemia es la homonimia. La diferencia entre ellas radica en la etimología de la palabra. Si la etimología de las dos palabras es la misma, hablamos de polisemia; si la etimología es diferente, hablamos de homonimia.

Ejemplos de polisemia y homonimia

Misma etimología = polisemia

La polisemia se da cuando una palabra tiene un único origen etimológico y acaba teniendo significados diferentes sin cambiar su categoría gramatical: por ejemplo, no pasa de sustantivo a verbo, como pasa en castellano entre el vino (bebida) y el vino (verbo venir). Es una palabra que con el tiempo ha ido adquiriendo diferentes significados, pero aun así, todos guardan entre sí una relación de significado; por ejemplo, en catalán y castellano fulla/hoja, que viene del latín *folia*, tiene diversos significados, como hoja de una planta, hoja de metal de una herramienta, página de un libro, cada una de las partes de una puerta doble o ventana, etc. Y en todos los significados lleva implícita la idea de una lámina.

Si queremos saber si una palabra es gramaticalmente polisémico, basta con consultar un diccionario etimológico y ver si proviene de un mismo origen. Encontraremos la palabra, un único origen y una lista de diferentes significados. En castellano podemos consultar el *Diccionario de la Real Academia*.

Más ejemplos de polisemia:

- *Servicio*, del latín *servitium*, que ha dado lugar a oficios religiosos, lavabos, misiones militares, cubiertos para comer y, en deportes, poner la pelota en juego. Y en todos ellos permanece la idea de ser útil.

- *Crucero*, del latín *crux*, significando 'cruz', intersección entre las dos naves de una iglesia, encargado de llevar la cruz a la cabeza de una procesión, viaje de placer por el mar, etc. En estos significados la idea es la de la forma de cruz, el cruzar como ir de un extremo a otro.

- *Columna*, del latín *columna*, que usamos para referirnos a los pilares arquitectónicos, las partes verticales de una página impresa de un diario, en física la forma que adoptan

algunos fluidos, como "columnas de humo", en el ámbito militar, la formación de barcos o soldados. Y la idea que permanece es la de verticalidad.

Diferente etimología = homonimia

La homonimia se da cuando dos conceptos han llegado a tener el mismo nombre, la misma forma, pero vienen de orígenes diferentes y, por lo tanto, tienen etimologías diferentes.

Por ejemplo, *metro* puede ser el transporte urbano, una unidad de medida o el utensilio para medir. Pero el origen etimológico entre el transporte y los otros dos significados es evidente: el primero es una abreviación de la palabra inglesa *metropolitan*, y en el segundo caso viene del griego μέτρον y significa medida.

Otro ejemplo: la palabra castellana *botín* puede venir del latín *bota* y significará 'calzado hasta el tobillo', o puede venir del alemán *bytin* y significará 'premio de una conquista'.

En castellano y catalán este fenómeno es menos frecuente que en otras lenguas, como el inglés o el francés, en las que abundan las palabras homónimas que dan mucho juego en los chistes.

Dentro de la homonimia podemos diferenciar las palabras que escribiéndose igual tienen significados diferentes, llamadas homógrafas, como las anteriores *metro* o *botín*, de las palabras que sonando igual también tienen significados diferentes, conocidas como palabras homófonas: *vell/bell* en catalán, o *tubo/tuvo* en castellano.

En resumidas cuentas, la sinonimia provoca silencio documental y la polisemia y variantes provocan ruido documental. El control terminológico del vocabulario garantiza el criterio de univocidad que tienen que tener los lenguajes documentales controlados, según el cual un concepto se representa con un término y un término sólo puede tener un significado.

3.1.4. Relaciones de significado de los términos

Por **relaciones de significado** entendemos la relación de genérico, específico o relacionado que puede tener un término con respecto a otro.

En el lenguaje natural estas relaciones son implícitas. Por ejemplo, cuando hablamos de manzanas todos entendemos que se trata de una fruta fresca y que las Fuji y las Golden son variedades concretas. Es decir, situamos el término "manzana" dentro de una jerarquía de términos conceptualmente más genéricos (fruta) y más específicos (Golden, Fuji). Incluso podemos relacionar por asociación de ideas la manzana con otras frutas, como la naranja o el plátano. Pero en un lenguaje documental hay que definir estas relaciones, agrupando y relacionando los términos afines.

La estructura que relaciona los términos es implícita en el lenguaje natural, pero en los lenguajes documentales hay que hacerla explícita. Eso se puede hacer de dos maneras:

a) En una secuencia jerárquica, donde la propia posición del concepto ya define sus términos genéricos y específicos. También deshace problemas de significado.

Ejemplo de la pesca

Ved el ejemplo de la pesca extraído de la *Clasificación Decimal Universal* (CDU). El concepto *pesca* puede ser la actividad económica o la pesca como deporte. Si nos fijamos en la cadena jerárquica vemos que cada uno cuelga de una clase diferente:

```
6 Ciencias aplicadas. Medicina. Tecnología
63 Agricultura y ciencias relacionadas
   639 Caza. Pesca

7 Bellas artes. Juegos. Deportes
79 Diversiones. Espectáculos. Juegos
   799 Caza deportiva. Pesca deportiva.
```

b) En una presentación alfabética donde cada término se acompaña de todos sus términos relacionados, ya sean equivalentes, genéricos, específicos o relacionados.

El tesauro del CSIC

En el tesauro de Psicología del CSIC, consultamos "Sueños" y encontramos:

Sueños

TG Dinámica de la personalidad

TE Contenido del sueño
TE Pesadilla

TR Déjà vu
TR Interpretación de los sueños
TR Sueño fisiológico
TR Sueño REM
TR Trastornos de conciencia

Las siglas nos informan del tipo de relación que establecen: TG significa término genérico (por encima de "Sueños" el tesauro tiene "Dinàmica de la personalidad"), TE son los términos específicos (son términos específicos de "Sueños": Contenido del sueño, Pesadilla) y los TR son los términos relacionados (se relacionan con "Sueño", "Déja vu", la "Interpretación de los sueños", el Sueño REM", etc.).

Finalmente, las principales ventajas e inconvenientes del lenguaje natural y el documental controlado son:

Ventajas e inconvenientes de los lenguajes documentales

	Ventajas	Inconvenientes
Lenguaje natural	Amigable Actualizado Económico	Dificulta la búsqueda Poco preciso
Lenguaje documental controlado	Unívoco Facilita la búsqueda	Caro Poco actualizado

A modo de conclusión

Indizar es la acción de describir o identificar un documento en relación con su contenido.

La indización la puede realizar una persona (de forma centralizada o de forma coordinada) o un programa.

Por lenguaje natural entendemos el lenguaje que usamos de forma cotidiana (catalán, castellano, vasco), y por lenguaje documental entendemos el listado o vocabulario de términos que usamos para indizar y que puede estar en formato libre o controlado. La principal diferencia entre el lenguaje natural y el documental controlado es el control terminológico:

- El control del número de términos del vocabulario: los lenguajes documentales son entrópicos, tienden a la selección, a la restricción del vocabulario.
- El control de las formas: los lenguajes controlados, controlan las formas plural/singular, el uso de acrónimos y siglas y la construcción de las frases.
- El control del significado: los lenguajes controlados controlan la sinonimia y la polisemia. Decimos que las palabras son sinónimas cuando tienen el mismo significado. Decimos que dos palabras son polisémicas cuando el mismo signo lingüístico tiene más de un significado. La sinonimia provoca silencio documental y la polisemia y variantes provocan ruido documental. El control terminológico del vocabulario garantiza el criterio de univocidad que tienen que tener los lenguajes documentales controlados, según el cual un concepto se representa con un término y un término sólo puede tener un significado.
- Las relaciones de significado entre los términos son las relaciones de genérico, específico o relacionado que puede tener un término con respecto a otro. En el lenguaje natural estas relaciones son implícitas pero en los lenguajes documentales hay que hacerlas explícitas a través de una secuencia jerárquica o una presentación alfabética.

3.2. ¿Cómo se indiza?

Ahora que ya hemos visto la necesidad de contar con lenguajes documentales para paliar la ambigüedad del lenguaje natural, estamos en condiciones de preguntarnos por el proceso de indización que lleva a cabo un analista.

A continuación presentamos las **fases** que proponen diversos autores antes de llegar a la que nos servirá como marco de referencia en este subapartado:

- Dos fases: análisis del texto y traducción (Chaumier, 1988; Fidel, 1994).
- Tres fases: análisis del texto, identificación de conceptos y traducción (Amat, 1989; Norma UNE 50-121-91).
- Cuatro fases: análisis del texto, identificación de conceptos, traducción y establecer enlaces sintácticos entre descriptores (Slype, 1991).
- Cinco fases: registro de datos, análisis del texto, identificación de conceptos, traducción y examen de la indización.

En este módulo seguiremos la **norma UNE 50-121-91** y sus tres etapas:

Norma UNE 50-121-91

UNE 50-121-91. *Métodos para el análisis de documentos, determinación de su contenido y selección de términos de indización.*

1) Examinar el documento para identificar su contenido.

2) Seleccionar los conceptos principales del contenido.

3) Traducir a un lenguaje documental.

Ejemplo

Examinamos un libro titulado *Mitos de antiguas civilizaciones*. Leemos el título, el resumen, el sumario, etc.

En una segunda etapa seleccionamos como conceptos principales: Mitos, Grecia, Roma, India, Japón, Indios norteamericanos.

En la tercera etapa indizamos. Si indizamos con un lenguaje libre podemos escribir el término como deseamos o como salga en el texto. Por ejemplo:

```
Mitología india americana.
```

En cambio, si indizamos con un lenguaje controlado tendremos que traducir estos conceptos a una forma controlada. Pongamos por ejemplo que pensábamos indizar Mitología india americana. Veamos cómo quedaría en tres lenguajes documentales diferentes:

```
CDU                 259.2
LEMAC               Mitologia ameríndia
LEM del CSIC        Indios de América - Religión y mitología
```

A continuación se detalla cada parte del proceso.

1) Examen del documento e identificación de los conceptos

El analista tiene que examinar con precisión el documento. La lectura completa es, a menudo, impracticable, pero sí que tiene que prestar atención al título, resumen, sumario, introducción, ilustraciones y palabras o frases destacadas en una tipografía diferente.

No se recomienda la indización sólo a partir del título, ya que hay títulos que llevan a error, y tampoco confiar en que el resumen sea un sustituto del texto, ya que no todos los resúmenes están bien elaborados.

Ejemplo de títulos y resúmenes que no aportan datos significativos para la indización

- CHESNEAUX, Jean. *¿Hacemos tabla rasa del pasado?* México: Siglo XXI Editores 1981. Su materia es *Historia, historiadores, historiografía*. En el catálogo de la Biblioteca Nacional de España (BNE[5]) lo encontramos indizado como Historia.
- MALLOL, Tomas. *Si la memòria no em falla*. Girona: CCG Ediciones 2005.
  Su materia es *Memorias, cine, coleccionismo*. En la Biblioteca de Catalunya (BC[6]) lo encontramos indizado como Cine amateur.

Si recordamos el resumen del libro de Carl Sagan, *Cosmos*, nos daremos cuenta de que no era suficiente para indizar el contenido de la obra. Por estos motivos se recomienda una lectura ágil del resto de partes significativas del documento.

[5]BNE es la sigla de *Biblioteca Nacional de España*.

[6]BC es la sigla de *Biblioteca de Catalunya*.

Ved también

Recordad que el ejemplo del resumen del libro de Carl Sagan, *Cosmos*, salía en el apartado 2 de este módulo.

2) Selección de los términos de indización

Tal como dice la norma UNE, el analista tiene que identificar las nociones que son elementos esenciales de la descripción del contenido. Si la indización es compartida, la institución que la patrocina tiene que establecer claramente los factores que considera importantes.

Para seleccionar los conceptos del documento, el analista tiene que ser consciente del número de conceptos (criterio de exhaustividad) y de la exactitud de los mismos (criterio de especificidad).

a) Exhaustividad

A medida que el analista va leyendo, tiene que ir tomando nota de los conceptos interesantes del documento.

Una buena praxis es la que identifica los conceptos relevantes sobre:

- El tema.
- Los nombres personales que puedan ser interesantes de indizar.
- Los nombres geográficos.
- Las fechas cronológicas.
- La forma en que se presenta el documento: artículo, estadística, formulario o divulgación, científico, etc.

La exhaustividad es un criterio relacionado con el número de conceptos que se tienen en cuenta para caracterizar el contenido entero de un documento. El principal criterio de selección es el valor potencial del concepto para los usuarios de su SID.

Podemos distinguir entre una exhaustividad baja, media y alta en función del número de descriptores. Es en este entorno donde la norma UNE 50-121-91 da sus recomendaciones en cuanto a la exhaustividad. Los criterios que el indizador tiene que tener en cuenta son:

- El tipo de SID y perfil de usuario. No es lo mismo indizar para una base de datos genérica que para una específica.
- El tipo de documento. No se indiza con el mismo número de descriptores una monografía que un artículo de revista, una tesis, etc.

Tal como recomienda la norma UNE, no es conveniente ser estrictos con el número de términos, no se tiene que limitar el número de forma arbitraria, tipo "para una monografía dos términos de indización", ya que puede conducir a una pérdida de objetividad y a una deformación de la información. Es preferible sugerir un baremo, entre tantos y tantos términos para cada tipo documental y SID y ser flexibles, ya que los criterios que tienen que regir son el propio contenido del documento y su posterior recuperación.

Ejemplo

Cuervo Herrero, C.; Fernández González, A.: "Objetos celestes erróneos". *Tribuna de Astronomía y Universo. Revista de Astronomía, Astrofísica y Ciencias del espacio*. 2000. II Época, n° 16 – octubre. p. 36-40.

A partir del siguiente resumen informativo, elaboraremos tres tipos de indizaciones sugiriendo un baremo (para esta asignatura y sus prácticas) y una finalidad:

"Anàlisis y descripción de los errores más frecuentes que cometen los profesionales y aficionados a la fotografía astronómica mientras intentan descubrir nuevos objetos celestes todavía no identificados.

Estos errores son debidos a cuatro causas: errores en el proceso de positivado de la copia como consecuencia de la presencia de partículas de polvo en los negativos o en las lentes del equipo de laboratorio; errores en el negativo debidos a defectos de lavado, deficiencias en la emulsión, rayas y rasguños o por el uso de películas de color destinadas a ser forzadas, y errores en las lentes de los objetivos, debidos a efectos de distorsión y a alteraciones en la refracción. Finalmente se describen otras causas: reflejos de la luz del sol sobre las antenas de satélites artificiales Iridum, retoques digitales o de fotocopiadoras y duplicadoras, uso de objetivos sencillos y poco potentes para captar imágenes de cielo profundo y, en último término, oscilaciones del condensador de luz del microscopio.

Todos estos errores pueden dar lugar a imágenes falseadas: objetos inéditos, diámetros erróneos, efectos de redondeo, alineaciones planetarias erróneas, etc. El artículo facilita imágenes de estos errores fotográficos.

Los autores concluyen que hace falta ser cauteloso y hacer las oportunas comprobaciones antes de dar a conocer el descubrimiento de un nuevo objeto celeste a las sociedades astronómicas."

Ejemplo de los tres grados de exhaustividad

Exhaustividad baja	**Exhaustividad media**	**Exhaustividad alta**
Baremo 1-3	**Baremo 4-6**	**Baremo 7 ...**
Ejemplo de uso: catálogo de una biblioteca pública	Ejemplo de uso: bases de datos de una biblioteca especializada en astronomía	Ejemplo de uso: bases de datos de una biblioteca especializada en astrofotografía
• Errores fotográficos • Fotografía astronómica	• Astrofotografía • Errores fotográficos • Descubrimientos • Identificación de objetos celestes • Objetos erróneos	• Alineaciones planetarias • Defectos de lavado • Deficiencias de la emulsión • Diámetros erróneos • Efectos de redondeo • Errores en el negativo • Errores en el positivado • Errores en las lentes • Objetos inéditos • Objetivos • Oscilaciones del microscopio • Partículas de polvo • Rayadas • Reflejos del sol • Retoques digitales

b) Especificidad

La especificidad está relacionada con la exactitud en que un concepto particular que aparece en un documento está representado por un término de indización.

Ejemplo

Si en el texto que estamos indizando aparece el concepto *Diplomacia*, y este término aparece en el lenguaje documental controlado, tenemos que indizar "Diplomacia". Si indizamos "Relaciones internacionales" o "Embajadores" no estaremos siendo específicos, como podéis ver en la tabla siguiente:

Ejemplo de especificidad

Materia	Correcto, y por lo tanto:	Incorrecto por:	
	Específico	Genérico	Demasiado específico
Diplomacia	**Diplomacia**	Relaciones internacionales	Embajadores

Los conceptos se tienen que identificar de la manera más específica posible, pero en determinados casos se pueden preferir nociones más genéricas:

- Cuando el indizador considere que un exceso de especificidad puede ser negativa en la recuperación; por ejemplo, puede decidir que un modelo muy específico de una máquina se indice con el nombre más genérico de este tipo de máquinas.

- Cuando la idea no esté plenamente desarrollada en el documento, o sólo se haga alusión a ella.

- Cuando se esté a la espera de validar el término más específico.

3) Traducción a un lenguaje documental controlado

Para traducir el concepto inicial escrito en lenguaje natural a un lenguaje documental, el indizador tiene que consultar las listas del lenguaje buscando la forma correcta de introducir el concepto.

Ejemplos

Concepto tal como sale en el texto	Traducción	Lenguaje documental utilizado
Tragicomèdia	791.221.28	Classificación Decimal Universal (CDU)
Eolític	Edat de la pedra	Lista de encabezamientos de materia en catalán
Matriz	Útero	Lista de encabezamientos del CSIC
Monarquía absoluta	Absolutismo	Tesauro de Historia contemporánea del CSIC

Cuando el analista procede a traducir el concepto del texto se puede encontrar en las siguientes situaciones:

a) Encuentra el concepto, solo o repartido por las tablas:

- Consulta el lenguaje y encuentra el concepto a la primera. Entonces indiza con este término de indización. Por ejemplo, buscaba "Eolític" y encuentra que tiene que indizar "Absolutismo".

- Consulta el lenguaje y encuentra el concepto o las partes del concepto repartidos por el lenguaje. Entonces tiene que conocer las reglas de combinación de las partes integrantes del término de indización. Ejemplos:
  – Una notación con CDU como 391.91(961.3) "Tatuajes de la isla de Samoa" está formada por 2 elementos, tatuajes + Samoa. Estos elementos van colocados en un orden determinado por las reglas de precoordinación de la CDU (primero la clase principal + auxiliar).

  – Un encabezamiento construido con la LEM del CSIC como Agua-Aspectos económicos está formado por dos partes: Agua + Aspectos económicos, que es un encabezamiento y un subencabezamiento respectivamente y van en este orden.

Con los lenguajes tesauros y listado de autoridades no hay una sintaxis de combinación.

b) No encuentra el concepto:

- Consulta el lenguaje y no encuentra el concepto. Entonces el indizador tiene que conocer las obras de referencia que su SID considera como autoridades reconocidas en la materia. Estas obras de referencia son diccionarios, enciclopedias, otros lenguajes documentales (especialmente los tesauros construidos de acuerdo con las normas ISO y UNE 50-106 y UNE 50-125), atlas, etc.

- Hay lenguajes, como tesauros, donde el indizador tiene que proponer el término nuevo como descriptor candidato y esperar a que la dirección del tesauro lo valide como descriptor. Mientras tanto indiza con un término más genérico.

3.3. Calidad y coherencia de la indización

La **calidad** y la **coherencia** de la indización dependen de factores como la competencia del indizador y la calidad de los instrumentos o lenguajes documentales. La coherencia es un factor importante en el comportamiento de un sistema de indización, especialmente cuando forma parte de una red de centros y la información se tiene que intercambiar entre ellos.

La coherencia se calcula de la siguiente manera: dos analistas indizan el mismo documento, con un lenguaje de descriptores como un tesauro. Se cuentan separadamente el número de descriptores idénticos entre los dos analistas sobre el total de descriptores.

Ejemplo

Como ejemplifica van Slype:

- 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 diferentes. Tasa de coherencia = 4/8 = 50% (van Slype, 1991, p. 123).

La consistencia en la indización suele oscilar entre el 20% de mínima y el 60% de máxima (Isidoro Gil, 2001).

A modo de conclusión

La norma UNE 50-121-91 *Métodos para el análisis de documentos, determinación de su contenido y selección de términos de indización* establece tres fases:

Lectures complementaries

Podéis ampliar la información sobre la coherencia en la indización leyendo las obras siguientes:

G. van Slype (1991). *Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales*. Madrid: Pirámide. Fundación Germán Sánchez Ruipérez. Biblioteca del Libro.

I. Gil Leiva (2001).

- Examinar el documento para identificar su contenido: el analista tiene que examinar con precisión el documento. La lectura completa es a menudo impracticable, pero sí que tiene que prestar atención al título, resumen, sumario, introducción, ilustraciones y palabras o frases destacadas en una tipografía diferente.

- Seleccionar los conceptos principales de los contenidos: el analista tiene que identificar las nociones que son elementos esenciales de la descripción del contenido, tiene que ser consciente del número de conceptos (criterio de exhaustividad) y la exactitud de los mismos (criterio de especificidad).

- Traducir a un lenguaje documental: para traducir el concepto inicial escrito en lenguaje natural a un lenguaje documental hay que consultar el listado del lenguaje buscando la forma aceptada.

4. Los lenguajes documentales

Un **lenguaje documental** es un vocabulario de términos en lenguaje natural o un sistema artificial de signos normalizados que facilitan la representación del contenido de los documentos.

Sus funciones principales son indizar el contenido de los documentos y permitir la recuperación a partir del campo materia.

Hay seis lenguajes documentales:

- Los sistemas de clasificación.
- Los listados de encabezamientos de materia.
- Los listados de autoridades.
- Los tesauros.
- Los listados de descriptores libres.
- Los listados de palabras clave.

En teoría todos los documentos se pueden indizar con cualquiera de estos seis lenguajes, pero en la práctica la tipología del SID (si es archivo, biblioteca o centro de documentación) y el tipo de usuario (general o especializado) condicionan que un SID indice y recupere con uno u otro lenguaje. En líneas generales:

- las bibliotecas indizan con sistemas de clasificación + listados de encabezamientos de materia + listados de autoridades;
- los centros de documentación indizan con tesauros + listados de palabras clave;
- los archivos, con sistemas de clasificación y/o tesauros.

Como podéis observar, los SID pueden trabajar con un solo lenguaje o con una combinación de lenguajes.

4.1. Los términos de indización

Llamamos **término de indización** a la representación de un concepto en lenguaje natural o un código de clasificación.

Los términos de indización pueden estar formados por una palabra o más de una.

Ejemplo

Ejemplos de términos de indización:

- De una palabra: "Bosques".
- De más de una palabra: "Fuente de información".

La parte más pequeña con significado de un término de indización se conoce como **unitérmino**.

La norma UNE 50-113-92/1 define unitérmino como:

"El elemento significativo más pequeño de un lenguaje documental utilizado para representar un concepto específico en un sistema de indización coordinado; no se debe confundir con palabra clave o descriptor".

UNE 50-113-92/1.

Ejemplo

El descriptor "Fuente de información" está formado por dos unitérminos: "Fuente" e "Información". La preposición "de" no se indiza.

Cada lenguaje documental da un nombre diferente a su término de indización. Esta es la terminología que usaremos en esta asignatura:

Términos de indización

Lenguaje documental	Su término de indización se conoce como
Sistemas de clasificación	Notación o símbolo de clase
Listados de encabezamientos de materia	Encabezamiento
Listados de autoridades	Autoridad, identificador o descriptor
Tesauros	Descriptor
Listados de descriptores libres	Descriptor
Listados de palabras clave	Palabra clave

La norma UNE 50-113-92/1 define estos conceptos de la siguiente manera:

- "Notación/Símbolo de clase: es la representación de una clase mediante la notación de un sistema de clasificación.
- Identificador: nombre utilizado como descriptor.
- Descriptor: términos de indización asignados por el analista fruto de alguna de las operaciones intelectuales que implica el proceso de indización.
- Palabra clave: una palabra o grupo de palabras seleccionadas de manera automática del título, resumen o texto de un documento del que representan su contenido y permiten la recuperación."

Norma UNE 50-113-92/1. *Documentación e información. Vocabulario. Parte 1. Conceptos fundamentales*.

Lectura recomendada

Para cuestiones de terminología recomendamos la consulta de la norma UNE 50-113-92/1. *Documentación e información. Vocabulario. Parte 1. Conceptos fundamentales*. En: *Documentación: Normas fundamentales*. Madrid: AENOR, 1994.

A modo de conclusión

Un lenguaje documental es un vocabulario de términos en lenguaje natural o un sistema artificial de signos normalizados que facilitan la representación del contenido de los documentos. Sus funciones principales son indizar el contenido de los documentos y permitir la recuperación a partir del campo materia.

Hay seis lenguajes documentales:

- Los sistemas de clasificación.
- Los listados de encabezamientos de materia.
- Los listados de autoridades.
- Los tesauros.
- Los listados de descriptores libres.

- Los listados de palabras clave.

Llamamos término de indización a la representación de un concepto en lenguaje natural o un código de clasificación. Los términos de indización pueden estar formados por una palabra o más de una.

4.2. Evolución histórica de los lenguajes documentales

Los primeros analistas mesopotámicos, egipcios o romanos leían el documento, copiaban las primeras líneas del texto o seleccionaban los conceptos que mejor representaban el contenido y los escribían en la tablilla, *pinake*, *cartela* o ficha correspondiente. Poco a poco estas materias fueron conformando un listado de temas. En la Edad Media sabemos de la existencia de catálogos de algunas grandes bibliotecas, como la de Lorsh en Alemania, que tenía 600 títulos clasificados en 63 materias.

Edad contemporánea

Ahora bien, para muchos autores la historia de los lenguajes documentales empieza en las bibliotecas del siglo XIX con los sistemas de clasificación, ya que fueron el primer intento serio de controlar las materias de los documentos.

Los sistemas de clasificación empezaron a ser considerados propiamente lenguajes en el siglo XIX con las **clasificaciones bibliográficas** de Brunet, Harris, Dewey, Cutter o la de la Library of Congress. Eran cuadros de clasificación jerárquicos, de cariz enciclopédico y sus clases se combinaban de una forma definida con anterioridad, es decir, precoordinada. Los conceptos se representaban con códigos, no palabras. Por ejemplo, el concepto "Fotografía" era el código 77 (ejemplo extraído de la CDU).

Library of Congress

El siguiente paso en la evolución de los lenguajes lo formuló Charles Ammi Cutter en 1876, creando un listado de materias escritas en lenguaje natural. Ya no se usaba un código, sino que se expresaba el concepto (como "Fotografía") con todas las letras. Estos listados, llamados **listas de encabezamientos de materia**, eran alfabéticos y se basaban en los principios de especificidad (hay que indizar con el término específico, no el genérico) y el de entrada directa (hay que respetar el orden natural de las expresiones y no optar por formas inversas del tipo "Electrónico, comercio").

Charles Ammi Cutter

Las colecciones bibliotecarias estaban cubiertas con estos dos lenguajes documentales: sistemas de clasificación + listados de encabezamientos de materia. Los listados de autoridades controlaban el resto de autoridades. Además, se combinaban en los registros bibliográficos con el fin de minimizar el inconveniente de la codificación, ya que no era de fácil comprensión para los usuarios. La indización era sintética, sumaria, dos o tres entradas por el campo materia, ya que tenemos que ser conscientes de que nacieron en sistemas no automatizados.

A medida que la producción científica iba generando cada vez más volumen de información, surgió la necesidad de indizar de una forma más analítica, con más conceptos. Se crean **centros de documentación** con una vocación más especializada que las bibliotecas. El uso de tecnología informática facilitaba el acceso a un documento a través de diversos puntos de acceso. Nacen los lenguajes especializados por excelencia, los **tesauros**. Se aplican a los centros de documentación y a algunos archivos históricos y administrativos.

Los tesauros recogen todo lo mejor de sus antecesores: la estructura arborescente de los sistemas de clasificación, que aplican a su presentación jerárquica, y la estructura combinatoria de las listas de encabezamientos de materia, que aplican a su presentación alfabética. Además incluyen nuevas estructuras de presentación, como la gráfica y la de índices permutados.

Los tesauros se automatizan y desde mediados de los años setenta el crecimiento de la industria de las bases de datos posibilita la consulta en línea de muchas publicaciones seriadas. Nace el último lenguaje documental, el **listado de palabras clave o indización automática**.

Internet

La última gran etapa la marca **Internet**. La globalización de la red a partir de la década de los años 1990 impulsa el acceso a la información, ya no hace falta que los SID dispongan en propiedad del documento, ya que la red permite acceder a la información alojada en cualquier otro centro de información. La cooperación impulsa a todos los lenguajes documentales a automatizarse y formar parte de proyectos colectivos (catálogos colectivos, consorcios, redes). En el mismo sentido se buscan pasarelas entre los diferentes lenguajes para solucionar problemas idiomáticos entre países.

Surge la necesidad de indizar la abundante producción de recursos electrónicos, como por ejemplo con el uso de metadatos para definir e intercambiar datos entre sistemas informáticos (etiquetas del tipo <subject>,<keywords>) y explotar la indización automática en los potentes robots de los buscadores. También los usuarios pueden indizar los recursos gracias a iniciativas de indización social o *tagging*.

Los expertos opinan que en la actualidad el problema principal no es tanto indizar o recuperar, sino presentar los resultados en algún orden significativo, lo que implica el uso de algoritmos que valoren los resultados.

A continuación reproducimos algunas de las fechas más significativas, extraídas de la cronología de Isidoro Gil (2008) sobre las listas de encabezamiento de materia, los tesauros y la indización automática.

Lectura complementaria

Podéis encontrar esta cronología en la obra siguiente:

I. Gil Leiva (2008). *Manual de indización. Teoría y práctica*. Gijón: Ediciones Trea (Biblioteconomía y Administración cultural, 193), pág. 110-114.

Cronología de la evolución de los lenguajes documentales

Fechas	Concepto	Breve explicación
30.000 a.C.	**Etiquetas de barro**	**Los antiguos escribas mesopotámicos guardaban las tablillas de barro (documentos) en cestas de mimbre. Por fuera, la cesta llevaba otra tablilla de barro con el contenido.**
Egipto	Las *cartelas* de Egipto	Los egipcios introducen el papiro como soporte documental. El papiro se enrollaba en torno a una varita de madera o metal. Para no desplegar completamente el rollo, ponían las primeras frases del documento en una etiqueta o *cartela* en un extremo.
1876	**Charles A. Cutter Rules for a dictionary catalog**	
1895	List of subject headings for use in dictionary catalogs	Publicada por la American Library Association (ALA) para bibliotecas medias y pequeñas, con fondos no especializados.
1909	**Library of Congress Subject Headings**	**Nace a partir de la lista de ALA y las reglas de Cutter. A partir de aquí esta lista se convierte en el referente de todas las listas de encabezamientos de materia del mundo.**
1923	List of subject headings for small libraries	Minnie Earl Sears es la autora de esta lista conocida como SEARS. Es una versión reducida de la LCSH para bibliotecas pequeñas.
1934	**Guía para los encabezamientos de materia**	**Juan Manrique Lara publica la primera lista de encabezamientos en castellano en México. Era una traducción de la Library of Congress Subject Headings (LCSH), el ALA y la SEARS.**
1946	Répertoire de vedettes-matière RVM	Primera lista de encabezamientos en francés (Universidad de Laval Canadà).
1951	**Descriptor**	**Calvin Mooers acuña el término.**
1952	Unitérmino	Mortimer Taube acuña el término.
1957	**Indización automática**	**Hans Meter Luhn empieza a trabajar en indización automática aplicando el método de la frecuencia.**
1960	Compatibilidad	En la década de los 60 se inician los primeros proyectos para hacer compatibles los diferentes lenguajes documentales mediante tablas de equivalencia.
1961	**Sistema SMART**	**Gerald Staton desarrolla el sistema SMART de análisis automático de textos.**
1967	Guidelines for the development of information retrieval thesauri	Directrices para elaborar tesauros confeccionadas por el US Federal Council for Science and Technology de Washington
1967	**Lista de encabezamientos de materia para bibliotecas**	**Lista compilada por Carmen Rovira y Jorge Aguayo en español para la Unión Panamericana.**
1974	Norma ISO 2788:1974 Guidelines for the establishment and development of monolingual thesauri	1ª edición de la norma ISO para la confección de tesauros monolingües.
1980	**Répertoire d'autorité-matière encyclopédique et alphabétique unifié RAMEAU**	**Primera lista de encabezamientos de materia de la Biblioteca Nacional de Francia. Se basaron en la RVM y la LCSH.**
1983	Bilindex	Lista de encabezamientos bilingüe en inglés y castellano. Es equivalente a la LCSH. En el año 2007 se editó la 15 ed.
1985	**Norma ISO 5963:1985 Methods for examinating documentos**	**Norma ISO que no sería traducida a norma UNE hasta 1991 con el número UNE 50-121-91.**

Fechas	Concepto	Breve explicación
1985	Norma ISO 5964:1985 Guidelines for the establishment and development of multilingual thesauri	1ª edición de la norma ISO para la confección de tesauros multilingües.
1986	**Abandono de los símbolos tradicionales de las listas de encabezamientos por los propios de los thesaurus**	**La LCSH, en su 10 edición, abandona los símbolos de x, see, xx, v, a por los propios de los tesauros Use, BT, NT, RT. Las demás listas mundiales también los adoptan.**
1986	Unified medical language system	El sistema unificado de lenguajes en medicina es un proyecto para integrar los diferentes vocabularios de ciencias de la salud. Es un proyecto de la Biblioteca Nacional de Medicina de EE.UU. (actualmente coordina el MESH, Medical Subject Headings).
1995	**Universalización de Internet**	**Internet ha difundido y popularizado conceptos, técnicas y prácticas propias de documentalistas.**
1995	Metadatos	Uso de metadatos para definir e intercambiar datos entre sistemas informáticos. Los lenguajes de marcaje tienen etiquetas para el resultado de la indización del tipo <subject>, <keywords>.
1997	**Proyecto MACS**	**Iniciativa de la Conference of European National Libraries CENL para hacer compatibles tres listas de encabezamientos de materia, la alemana SWD, la RAMEAU francesa y la LCSH usada en Gran Bretaña y Suiza.**

A modo de conclusión

Para muchos autores la historia de los lenguajes documentales empieza en las bibliotecas del siglo XIX con los sistemas de clasificación, ya que fueron el primer intento serio de controlar las materias de los documentos.

El siguiente paso en la evolución de los lenguajes lo formuló Charles Ammi Cutter en 1876, creando una lista de materias escritas en lenguaje natural.

A medida que la producción científica iba generando cada vez más volumen de información, surgió la necesidad de indizar de una forma más analítica, con más conceptos. Se crean centros de documentación con una vocación más especializada que las bibliotecas. Nacen los lenguajes especializados por excelencia, los tesauros.

Desde mediados de los años setenta el crecimiento de la industria de las bases de datos posibilita la consulta en línea de muchas publicaciones seriadas. Nace el último lenguaje documental, el listado de palabras clave o indización automática.

La última gran etapa la marca Internet. La globalización de la red a partir de los años 1990 impulsa el acceso a la información. La cooperación impulsa a todos los lenguajes documentales a automatizarse y formar parte de proyectos colectivos (catálogos colectivos, consorcios, redes). En el mismo sentido se buscan pasarelas entre los diferentes lenguajes para solucionar problemas idiomáticos entre países.

Surge la necesidad de indizar la abundante producción de recursos electrónicos, como por ejemplo el uso de metadatos para definir e intercambiar datos entre sistemas informáticos (etiquetas del tipo <subject>, <keywords>) y explotar la indización automática en los potentes robots de los buscadores. También los usuarios pueden indizar los recursos gracias a iniciativas de indización social o *tagging*.

4.3. ¿Cuándo son necesarios los lenguajes documentales?

Los lenguajes documentales son necesarios en dos momentos de la cadena documental:

- La fase de análisis y tratamiento > Análisis documental > Análisis de contenido > Indización.
- La fase de salida > Instrumentos de recuperación.

Tanto en la fase de indización como en la fase recuperación, el proceso de análisis-selección-traducción de conceptos es el mismo. En el momento de la indización el analista lee el documento, extrae conceptos y si hace falta los traduce a un lenguaje controlado para almacenarlos en el sistema. En el momento de la recuperación, el analista tiene que trabajar con la consulta del usuario, extraer los conceptos y traducirlos. Si se trata de un lenguaje postcoordinado, además tendrá que saber cómo convertir los descriptores a una ecuación de búsqueda.

Ejemplo de la fase de recuperación

- **Usuario**: "Necesito información sobre las instalaciones deportivas de hockey hierba que se construyeron en la ciudad de Terrassa con motivo de la celebración de los Juegos Olímpicos de 1992."

- **Analista**: selecciona los conceptos más relevantes para la búsqueda: instalaciones deportivas, hockey hierba, Terrassa, Juegos Olímpicos. El próximo paso es traducir los conceptos a un lenguaje documental, en el ejemplo, el "Tesaurus d'Història local de Catalunya". Como se puede apreciar entre la expresión en lenguaje natural del usuario y los descriptores aceptados del tesauro hay ciertas diferencias:

En la expresión del usuario:	Traducido al tesauro:
Instalaciones deportivas	Equipamientos deportivos (Equipaments esportius)
Hockey hierba	Hockey (Hoquei)
Terrassa	Terrassa
Olimpiadas	Juegos Olímpicos 1992 (Jocs Olímpics 1992)

Traducido a una ecuación de búsqueda: Equipamientos deportivos AND Hockey AND Terrassa AND Juegos Olímpicos 1992.

G. van Slype (1991, pág. 161) considera que los lenguajes documentales pueden intervenir, como máximo, hasta en seis momentos diferentes en la recuperación:

1) Selección de los sistemas documentales que se interrogarán: qué catálogos, qué bases de datos, etc.

2) Selección de los conceptos expresados por el usuario en su enunciado.

3) Traducción a un lenguaje documental controlado.

4) Formulación de la ecuación de búsqueda.

5) Extensión asistida por ordenador.

6) Evaluación final de la pertinencia de los resultados obtenidos.

Hay una tercera función dentro de la cadena documental, pero sólo afecta a un lenguaje documental concreto, que son los sistemas de clasificación:

- La fase de Análisis y Tratamiento > Procesamiento técnico > Ordenación.

Los códigos numéricos de los sistemas de clasificación jerárquicos, como la CDU, son la herramienta para ordenar los documentos en las estanterías de acuerdo a un orden secuencial de las materias (ordenación altamente significativa).

En teoría todo documento se podría indizar con cualquiera de los seis lenguajes. En la práctica cada tipología de SID tiende a utilizar un lenguaje o combinación de lenguajes concreta.

Lectura complementaria

Podéis ampliar la información sobre los lenguajes documentales en la obra siguiente:

G. van Slype (1991). *Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales*. Madrid: Pirámide. Fundación Germán Sánchez Ruipérez. Biblioteca del Libro.

Ejemplo: un documento y seis indizacions

Veamos cómo sería el resultado de indizar el mismo documento con cada uno de los seis lenguajes documentales:

El mercado del tabaco en España durante el siglo XVIII: fiscalidad y consumo / Santiago de Luxán Meléndez, Sergio Solbes Ferri, Juan José Laforet (ed.). Las Palmas de Gran Canaria: Universidad de Las Palmas de Gran Canaria, Servicio de Publicaciones, 2000.

Resumen:

"En este libro se ha querido poner el énfasis en un tema hasta ahora poco tratado como es el consumo de tabaco en España durante el siglo XVIII.

No obstante también se atienden otros aspectos como los fiscales. La obra se ha estructurado en tres partes: la primera se ocupa de la fiscalidad, la segunda atiende el área del monopolio y la tercera analiza los mercados regionales de Canarias y Navarra. El libro se cierra con un apartado dedicado al cultivo del tabaco."

Ejemplo de un único documento y seis indizaciones

Sistema de clasificación: CDU	**Listas de encabezamientos de materia: LEMAC**	**Listado de autoridades: Gran Enciclopèdia Catalana**
336.226(460)"17":663.97	Industria tabaquera- España- Historia - s. XVIII Tabaco - impuestos - España - Historia - s. XVIII	Canarias España Navarra
Tesauro: Tesaurus d'Història local de Catalunya (UAB)	**Listado de descriptores libres: Consultores de la asignatura**	**Listado de palabras clave: programa Swesum**
Tabaco Consumo Historia Impuesto de consumos Cultivos Monopolios Siglo XVIII	Canarias Cultivo Consumo España Fiscalidad Monopolio Navarra Siglo XVIII Tabaco	libro tabaco

Aunque en este momento el estudiante no conozca el funcionamiento de estos lenguajes, sí que está en disposición de observar algunos rasgos característicos de cada uno:

- El sistema de clasificación ha indizado un código, no son palabras. Es un código construido a base de números y símbolos, incomprensible a primera vista para un profano.

- La lista de encabezamientos de materia ha indizado dos términos en lenguaje natural, que están formados por diversas palabras separadas con guiones.

- La lista de autoridades ha indizado sólo nombres geográficos y ha prescindido del resto de conceptos. También ha usado el lenguaje natural.

- El tesauro ha indizado unos cuantos descriptores en lenguaje natural, poniendo un término bajo el otro.

- El listado de descriptores libres no se diferencia a simple vista de la indización con tesauro. En cambio, la diferencia es fundamental ya que el tesauro es controlado y los descriptores libres son libres.

- En el listado de palabras clave, la indización la ha realizado un programa informático, que ha seleccionado las palabras *libro* y *tabaco* porque salen dos veces en el texto, son las palabras más repetidas.

Ved también

Todos estos temas serán desarrollados en los módulos siguientes, dedicados a cada uno de los lenguajes documentales.

A modo de conclusión

Los lenguajes documentales son necesarios en dos momentos de la cadena documental:

- La fase de análisis y tratamiento > Análisis documental > Análisis de contenido > Indización.
- La fase de salida > Instrumentos de recuperación.

Los sistemas de clasificación también son útiles en:

- La fase de análisis y tratamiento > Procesamiento técnico > Ordenación.

4.4. Complementariedad de los lenguajes documentales

Indizar con más de un lenguaje documental al mismo tiempo es muy conveniente, porque así se suman las ventajas y minimizan los inconvenientes de los diferentes sistemas. Significa un esfuerzo añadido en el momento de la indización pero permite recuperar de manera más precisa. Es decir, combinamos lenguajes para recuperar mejor.

Algunas de las combinaciones posibles son las siguientes:

- Sistema de clasificación + listas de encabezamientos + listados de autoridades.
- Sistema de clasificación + listas de encabezamientos + listados de autoridades + palabras clave.
- Sistemas de clasificación + tesauro.
- Tesauro + listados de autoridades + palabras clave.

Ejemplo de combinación de lenguajes

Ejemplo de una captura de un registro del catálogo de la Biblioteca Nacional de España donde vemos un campo para la notación con CDU y otro para un encabezamiento de materia.

Historia de las ideas políticas [Texto impreso]
Touchard, Jean

CDU:	32(091)
Autor personal:	Touchard, Jean
Título uniforme:	[Histoire des idées politiques Español]
Título:	Historia de las ideas políticas [Texto impreso] / Jean Touchard ; con la colaboración de Louis Bodin ... [et al. ; traducción de J. Pradera]
Edición:	6ª ed.
Publicación:	Madrid : Tecnos, 2006
Descripción física:	659 p. ;24 cm
Serie:	(Colección de ciencias sociales. Serie de ciencia política)
Nota al tít.y menció:	Traducción de: Histoire des idées politiques
Encabez. materia:	Política -- Historia
N. depósito leg.:	M 7115-2006

A modo de conclusión

En teoría todos los documento se podrían indizar con cualquiera de los seis lenguajes. En la práctica cada tipología de SID tiende a utilizar un lenguaje o combinación de lenguajes concreta.

Indizar con más de un lenguaje documental al mismo tiempo es muy conveniente porque se suman las ventajas y minimizan los inconvenientes de los diferentes sistemas. Significa un esfuerzo añadido en el momento de la indización pero permite recuperar de manera más precisa. Es decir, se combinan lenguajes para recuperar mejor.

5. Tipología de los lenguajes documentales

Podemos clasificar los seis lenguajes documentales a partir de unas características o tipologías que los describen. Concretamente los lenguajes se tipifican según la naturaleza de sus términos, el nivel de control, el nivel de coordinación, la estructura y el nivel de análisis:

Tesauro

Un lenguaje es la suma de diversas características. Así, por ejemplo, un tesauro es natural, controlado, postcoordinado, jerárquico y combinatorio e indiza por conceptos.

Tipología de los lenguajes documentales

		Sistemas de clasificación	Listados de encabezamientos de materia	Listados de autoridades	Tesauro	Listado de descriptores libres	Listado de palabras clave
Según la naturaleza de los términos	**Codificado**	X					
	Natural		X	X	X	X	X
Según el nivel de control sobre los términos	**Libre**					X	X
	Controlado	X	X	X	X		
Según el nivel de coordinación de los términos	**Precoordinado**	X	X				
	Postcoordinado			X	X	X	X
Según la forma de agrupar los términos o estructura	**Jerárquico o Sistemático**	X			X		
	Combinatorio		X	X	X	X	X
Según el nivel de análisis	**Por materias**	X	X				
	Por conceptos			X	X	X	
	Por palabras clave						X

A continuación, vamos a ver estas características.

5.1. Naturaleza del término: codificado o natural

Los términos pueden expresarse en lenguajes codificados o naturales:

a) Lenguajes codificados. Entendemos por *codificado* el uso de un código artificial compuesto de números, letras y símbolos que traducen un concepto. Por ejemplo, el Sol, en un lenguaje como la CDU, sería 523.9.

Los lenguajes codificados son lenguajes sintéticos, muy usados en bibliotecas, ya que, además de clasificar el contenido del fondo documental, son operativos en cualquier idioma y permiten la ordenación de los fondos. Por otra parte, tienen el inconveniente de ser poco comprensibles por parte de los usuarios.

Sólo hay un tipo de lenguaje codificado: son los **sistemas de clasificación**.

Algunos lenguajes codificados

Son ejemplos de lenguajes codificados la Clasificación Decimal Universal (CDU), la Clasificación Dewey (DDC), la Clasificación de la Library of Congress (LCC) o la Clasificación Colon (CC).

b) Lenguajes naturales. Entendemos por *natural* el uso de palabras del lenguaje usual, habitual, no códigos. Es mucho más próximo al usuario, más amigable. Hay cinco lenguajes documentales naturales:

- Las listas de encabezamientos de materia.
- Los listados de autoridades.
- Los tesauros.
- Los listados de descriptores libres.
- Los listados de palabras clave.

Ved también

Los sistemas de clasificación se estudian con más profundidad en el módulo "Sistemas de clasificación documentales" de esta asignatura.

5.2. Nivel de control: libre o controlado

Hace referencia al control del vocabulario, es decir, si las palabras seleccionadas para indizar corresponden al lenguaje natural o a un lenguaje artificial construido para garantizar la indización y recuperación:

a) Lenguajes libres. Son listas de términos extraídos del lenguaje natural sin sufrir ningún tipo de control. Normalmente los lenguajes libres se utilizan en sistemas automatizados donde hay un fichero inverso o diccionario de la base de datos. Tienen muchas ventajas en la indización, como el gasto mínimo de construcción, la actualización inmediata, coherencia máxima y la riqueza terminológica. Pero presentan inconvenientes en la recuperación, ya que al trabajar con lenguaje natural, arrastran todos los problemas derivados de la ambigüedad (sinonimia, polisemia, homonimia).

Los lenguajes libres son dos:

- Los listados de descriptores libres.
- El listado de palabras clave.

b) Lenguajes controlados. Consideramos lenguajes controlados aquellos que están redactados previamente en forma de listas o listados de términos que se consideran aceptados y unívocos para la indización. Sólo los términos de la lista se pueden usar para indizar.

Son términos seleccionados tanto en su forma (plural, singular, sintagma nominal, adjetivado, siglas, etc.), como en su contenido (de todos los sinónimos se escoge uno, los homónimos se diferencian entre ellos, etc.) y como en sus relaciones de jerarquía y asociación (términos conceptualmente más genéricos o específicos y términos que se evocan mutuamente). Requieren unos gastos de construcción elevados, tanto en personal cualificado como en tiempo. Para muchos autores son los verdaderos lenguajes documentales. También se conocen por el nombre de **lenguajes artificiales**.

Su función documental es la de representar un concepto con un único término y que sólo haya un término por concepto, lo que se conoce como univocidad.

Los lenguajes controlados son cuatro:

- Los sistemas de clasificación.
- Las listas de encabezamientos.
- Los listados de autoridades.
- Los tesauros.

5.3. Nivel de coordinación: precoordinado o postcoordinado

a) Precoordinación. La precoordinación consiste en determinar a priori cómo se combinan los términos, tanto sea a la hora de construir el lenguaje como a la hora de indizar el documento o a la de recuperarlo.

Lenguajes precoordinados

Un ejemplo de construcción con un lenguaje precoordinado como la "Lista de encabezamientos de materia" como la del CSIC prevé que la materia *Construcción de viviendas* se represente como:

```
Viviendas - Construcción
```

Es decir, por este orden y separados con un guión.

Un ejemplo de indización con un lenguaje precoordinado, por ejemplo, de una materia compuesta por tres elementos como *Enciclopedia de los perros pastores europeos* se representa como:

```
Perros Pastores - Europa - Enciclopedias
```

El encabezamiento se hace en este orden concreto, y las reglas sintácticas del lenguaje evitan la posibilidad de otras combinaciones.

La precoordinación tiene dos grande ventajas:

- Agrupa en proximidad todos los documentos que tienen una temática afín, de manera que si consultamos el catálogo de una biblioteca *Viviendas*, también veremos otros documentos como:
  `Viviendas - Alumbrado`
  `Viviendas - Arrendamiento`
  `Viviendas - Calefacción y ventilación`

- Un solo término de indización reúne los elementos principales para la búsqueda.

La precoordinación era una auténtica necesidad en el entorno de las bibliotecas manuales, ya que no se podía buscar por una combinación de dos o más términos.

Ved también

El tema de la precoordinación se trata sobradamente en los módulos dedicados a los dos lenguajes precoordinados: "Sistemas de clasificación documentales" y "Listas de encabezamientos y listados de autoridades".

b) Postcoordinación. La postcoordinación consiste en combinar los términos de indización en el momento de la recuperación. Permite combinar múltiples términos de indización siguiendo la lógica de los operadores booleanos y de esta manera profundizar en el análisis de contenido. No tienen sintaxis en el momento de la indización. Cada término indizado es un punto de acceso al documento; cuantos más términos indizamos, más posibilidad tenemos de recuperarlo.

Lenguajes postcoordinados

Un lenguaje postcoordinado, como un tesauro, representaría el documento anterior sobre perros pastores como:

```
Perros pastores
Europa
Enciclopedia
```

que sería recuperado siguiendo la lógica de los operadores booleanos:

```
Perros Pastores AND Europa
```

Los lenguajes postcoordinados sólo tienen sentido en sistemas documentales automatizados que dispongan de un fichero inverso. El fichero inverso está donde se almacenan todos los descriptores que el analista va indizando, se sitúan uno detrás del otro de forma secuencial y asociados al documento al que hacen referencia.

Los lenguajes postcoordinados son cuatro:

- Listados de autoridades.
- Tesauro.
- Listados de descriptores libres.
- Listados de palabras clave.

Ejemplo de fichero inverso

Fichero inverso

Documento	Fichero inverso: concepto y n°. de documento
Documento 1	Alimentación (2) Enciclopedia (1,3) Entrenamiento (2) Europa (1) Perros Pastores (1,2) Química orgánica (3)
Perros pastores Europa Enciclopedia	
Documento 2	
Perros pastores Alimentación Entrenamiento	
Documento 3	
Química orgánica Enciclopedia	

5.4. Estructura: jerárquica o combinatoria

El vocabulario de los lenguajes documentales se organiza en dos estructuras básicas, en forma jerárquica o en forma combinatoria:

a) Jerárquica: en la estructura jerárquica o arborescente, el vocabulario se presenta en forma de cadena, con términos genéricos que agrupan términos más específicos. Todos los términos dependen de un término superior y de significado más genérico. Esta estructura permite agrupar los conceptos por temas y también situarlos en contexto, ya que la secuencia jerárquica nos informa de cuál es el campo temático al que está adscrito el concepto.

Ejemplo

Pongamos un ejemplo extraído de la CDU:

```
37 Educación
   371 Organización de la educación
   372 Contenido. Materias
   373 Tipo de escuelas
   374 Enseñanza extraescolar
   376 Escuelas especiales
   377 Formación profesional
   378 Universidades
```

Así, el concepto "Universidades" depende del concepto "37 Enseñanza", por lo tanto hace referencia a la educación que se imparte en la universidad y no a la arquitectura de las universidades (que estaría dentro de "72 Arquitectura").

Los lenguajes jerárquicos son dos:

- Los sistemas de clasificación.
- Los tesauros (en la parte de presentación sistemática o jerárquica).

b) Combinatoria: en la estructura combinatoria, los términos no forman cadena, están listados por orden alfabético. Este tipo de estructura surgió como reacción a la rigidez de la estructura jerárquica, que no era fácil de actualizar.

Ejemplo extraído de la Lista de encabezamientos del CSIC


Peajes
Pearcea
Pearl Harbor, Ataque a, 1941
Pecado
Pecado (Islam)
Pecado original
Pecados
Pecados capitales
Pecaris

La estructura combinatoria permite la inclusión de nuevos términos y la eliminación de los obsoletos sin afectar al resto de la estructura del lenguaje. La facilidad para actualizar el vocabulario los convierte en lenguajes adecuados para todo tipo de entornos: enciclopédicos, científicos y técnicos.

Los lenguajes de estructura combinatoria son cinco:

- Listas de encabezamientos de materia.
- Listados de autoridades.
- Tesauros.
- Listados de descriptores libres.
- Listados de palabras clave.

Como se puede observar, los tesauros participan de las dos estructuras: tienen una presentación sistemática en forma jerárquica y una presentación alfabética en forma combinatoria.

El descriptor "Còmic"

Veamos el descriptor "Còmic" tanto en una presentación como en la otra (extraído del Tesaurus d'història local de Catalunya).

Presentación jerárquica (izquierda) y alfabética (derecha)

```
Fuentes primarias                    Comics
UP  Documentos primarios             TG  Fuentes primarias
                                     TR  Dibujos
TG  Fuentes de información           TR  Ilustraciones
                                     TR  Láminas
TE  Almanaques
TE  Atlas
TE  Biografías
TE  Comics
TE  Cronologías
TE  Diccionarios
TE  Directorios
TE  Enciclopedias
```

5.5. Nivel de análisis: materias, conceptos, palabras clave

Los lenguajes pueden indizar más o menos conceptos, de manera que podemos establecer una última tipología según la cantidad de información que transmiten cada uno. En el punto más sintético, con uno o dos términos de indización, tenemos los lenguajes que indizan por materias; en el punto medio, los lenguajes de conceptos, también llamados de descriptores, y en el punto más analítico, los lenguajes de palabras clave.

Indizar por materias, conceptos y palabras clave está en relación directa con los dos paradigmas de búsqueda. La indización por materias es adecuada para sistemas de *browsing* (o de navegación o directorio). En cambio, las indizaciones por conceptos y palabras clave se adaptan mejor a los sistemas de interrogación en buscadores.

a) Por materias: responden a la pregunta:"¿cuál es el tema de este documento?". Los lenguajes que indizan por materias son dos:

- Los sistemas de clasificación.
- Las listas de encabezamientos de materia.

b) Por conceptos: indizar por conceptos significa indizar las ideas y nociones del texto sin reducirlo a un tema principal. Responden a la pregunta: "¿cuáles son los conceptos de este documento?". Van ligados necesariamente a sistemas automatizados, ya que no es factible elaborar tantas fichas de cartulina como conceptos se van a indizar.

Los lenguajes que indizan por conceptos son tres:

- Listados de autoridades.
- Tesauro.
- Listados de descriptores libres.

c) Por palabras clave: indizar por palabras clave significa indizar todas las palabras con significado del texto. Es el proceso más analítico que existe. No es una tarea de indización humana, sino automática. Los programas que indizan por palabras clave seleccionan sólo las palabras que tienen significado (preferentemente sustantivos).

Sólo hay un lenguaje por palabras clave, y es evidentemente el único lenguaje automático: el listado de palabras clave.

Ejemplo de indización con los tres niveles de análisis

Indizaremos con los tres niveles de análisis el siguiente resumen indicativo:

MUÑOZ CRUZ, Valle. El papel del gestor de la información en las organizaciones a las puertas del siglo XXI. A. *Los sistemas de información al servicio de la sociedad: actas de las jornadas*. Valencia: FESABID, 1998, vol. 2, p. 649-660.

"Artículo sobre el papel y funciones del gestor de la información, un nuevo profesional de la documentación, en las organizaciones del siglo XXI. Describe el panorama laboral español, analizando la Administración pública y la empresa privada. Propone desarrollar una política nacional de información y una formación adaptada a las necesidades organizativas de las instituciones."

Ejemplo de niveles de análisis

Por materias	Por conceptos	Por palabras clave	
Gestor de información	Gestor de información Documentación Administración pública Empresa privada Política de información	Adaptada Administración Artículo Documentación Empresa Español Formación Funciones Gestor Información Instituciones Laboral Nacional	Necesidades Nueve Organizaciones Organizativas Panorama Papel Política Privada Profesional Pública Siglo XXI

A modo de conclusión

Los lenguajes documentales se tipifican según:

- **La naturaleza de los términos**: los términos pueden expresarse en lenguaje codificado o natural. Entendemos por codificado el uso de un código artificial compuesto de números, letras y símbolos que traducen un concepto. Entendemos por natural el uso de palabras del lenguaje usual, habitual, no códigos.

- **El nivel de control del vocabulario**: los lenguajes pueden ser libres o controlados. Los lenguajes libres son listas de términos extraídos del lenguaje natural. Consideramos lenguajes controlados aquellos que están redactados previamente en forma de listas o listados de términos que se consideran aceptados y unívocos para la indización. Sólo los términos de la lista se pueden usar para indizar.

- **El nivel de coordinación**: precoordinado o postcoordinado. La precoordinación consiste en determinar a priori cómo se combinan los términos, ya sea a la hora de construir el lenguaje, a la hora de indizar el documento o a la hora de recuperarlo. La postcoordinación consiste en no establecer reglas a la hora de la indización y combinar los términos de indización en el momento de la recuperación siguiendo la lógica de los operadores booleanos.

- **La estructura**: el vocabulario de los lenguajes documentales se organiza en dos estructuras: jerárquica o combinatoria. En la estructura jerárquica o arborescente, el vocabulario se presenta en forma de cadena, con términos genéricos que agrupan términos más específicos. En la estructura combinatoria, los términos no forman cadena, están listados por orden alfabético.

- **El nivel de análisis**: materias, conceptos, palabras clave. Indizar por materias consiste en indizar la materia principal del documento. Indizar por conceptos significa indizar las ideas y nociones del texto. Indizar por palabras clave significa indizar todas las palabras con significado del texto. Es el proceso más analítico que existe. No es una tarea de indización humana, sino automática.

5.6. Conclusiones

El estudio de las tipologías de los lenguajes documentales permite elaborar la ficha descriptiva de cada uno.

Fichas descriptivas de cada lenguaje documental

Sistemas de clasificación	Listado de encabezamiento de materias	Listado de autoridades
• Sintético por materias • Símbolos de clase o notaciones • Humana • Codificado • Controlado • Precoordinado • Jerárquico	• Sintético por materias • Encabezamientos • Humana • Natural • Controlado • Precoordinado • Combinatorio	• Analítico por conceptos • Identificadores y descriptores • Humana • Natural • Controlado • Postcoordinado • Combinatorio
Tesauro	**Listado de descriptores libres**	**Listado de palabras clave**
• Analítico por conceptos • Descriptores • Humana • Natural • Controlado • Postcoordinado • Jerárquico • Combinatorio	• Analítico por conceptos • Descriptores • Humana • Natural • Libre • Postcoordinado • Combinatorio	• Analítico por palabras clave • Palabras clave • Automática • Natural • Libre • Postcoordinado • Combinatorio

Actividades

1. A partir del siguiente artículo elaborad un resumen informativo, uno indicativo, uno selectivo de conclusiones y uno automático que tenga una extensión parecida al informativo.

VALLEZ, M; PEDRAZA-JIMÉNEZ, R. "El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines" [en línia en http://www.hipertext.net/web/pag277.htm]. *Hipertext.net*, núm. 5, 2007. ISSN 1695-5498.

2. Indizad el mismo artículo con los tres niveles de exhaustividad. Argumentad en qué tipo de base de datos y SID podría ser útil cada uno.

3. Proponed dos títulos de documentos, reales o inventados, donde la materia se exprese a través de dos sinónimos.

4. Imaginad dos títulos más donde aparezcan dos polisémicos y proponed una manera de diferenciarlos. Buscad el origen etimológico de las palabras y decid si son polisémicas u homónimas.

5. Responded las siguientes afirmaciones justificando la solución:
a) ¿Todo lenguaje controlado es codificado?
b) ¿Todo lenguaje precoordinado es controlado?
c) ¿Todo lenguaje libre es natural?
d) ¿El lenguaje que tiene la tasa de coherencia más elevada es el Listado de palabras clave?

6. El siguiente texto es un compendio de errores y medias verdades. ¿Sabríais localizarlas y argumentar por qué no son correctas?

Usar lenguajes naturales en la indización y recuperación permite una buena comunicación documental. Los sistemas de clasificación representan la materia de los documentos a través de múltiples notaciones. Los lenguajes que indizan por materias son los tesauros y las listas de encabezamientos de materia. Para recuperar de manera precisa tenemos que utilizar sistemas de clasificación y listados de encabezamientos de materia. Los lenguajes controlados son muy amigables para el analista y el usuario. Los lenguajes precoordinados permiten ordenar los documentos en las estanterías.

Glosario

abstract *m* Terminología anglosajona para los resúmenes redactados por personas.

anáforas *f* Relación de referencia entre un elemento lingüístico y uno anterior en el discurso.

análisis de contenido *m* Operaciones de análisis que identifican y representan de manera precisa la materia de los documentos, con el objetivo de permitir la recuperación. Las operaciones son dos: el resumen y la indización. Esta parte del análisis documental establece el punto de acceso por materias.

análisis morfosintáctico *m* Análisis que determina la categoría léxica de cada palabra: sustantivo, verbo, adjetivo, artículo, preposición, etc. También determina el lema. Estas operaciones permiten distinguir las palabras con significado (sustantivos, adjetivos, verbos) de las vacías (artículos, preposiciones, pronombres, etc.). El lema permite agrupar todas las palabras que son flexiones de otra (info/informar/información/informador/informacional/etc.).

autoridad *f* Término de indización propio del lenguaje documental Listado de Autoridades. También se conocen con el nombre de identificadores y descriptores.

codificado *adj.* Tipología de lenguaje documental consistente en el uso de un código artificial compuesto de números, letras y símbolos que traducen un concepto. Sólo hay un tipo de lenguaje codificado, son los sistemas de clasificación.

combinatoria *f* Tipología de lenguaje documental consistente en estructurar los términos de indización por orden alfabético. La estructura combinatoria permite la inclusión de nuevos términos y la eliminación de los obsoletos sin afectar al resto de la estructura del lenguaje. Los lenguajes de estructura combinatoria son cinco: las listas de encabezamientos de materia, los listados de autoridades, los tesauros, el listado de descriptores libres y el listado de palabras clave.

controlado *adj.* Tipología de lenguaje documental consistente en listas de términos seleccionados tanto en su forma (plural, singular, sintagma nominal, adjetivado, siglas, etc.) como en su contenido (de todos los sinónimos se escoge uno, los homónimos se diferencian entre ellos, etc.) y como en sus relaciones de jerarquía y asociación (términos conceptualmente más genéricos o específicos y términos que se evocan mutuamente). Requieren unos gastos de construcción elevados, tanto en personal cualificado como en tiempo. Son los verdaderos lenguajes documentales. También se conocen por el nombre de lenguajes artificiales. Su función documental es la de representar un concepto con un único término y que sólo haya un término por concepto, lo que se conoce como univocidad. Los lenguajes controlados son cuatro: los sistemas de clasificación, las listas de encabezamientos, listados de autoridades y tesauros.

describir el contenido *loc. v.* *Ved* representar el contenido

descripción característica *f* *Ved* indización.

descriptor *m* Término de indización propio de tres lenguajes documentales: listado de autoridades, tesauros, listado de descriptores libres.

encabezamiento *m* Término de indización propio del lenguaje documental de las listas de encabezamientos de materia.

entropía *f* Calidad aplicable a los lenguajes documentales que tienden a la selección, a la restricción del vocabulario. Es el proceso contrario al lenguaje natural que tiende a la abundancia, a la reiteración de conceptos, a la sinonimia en beneficio de una expresión más rica.

especificidad *f* Criterio relacionado con la exactitud en que un concepto particular que aparece en un documento está representado por un término de indización.

estructura *f* Tipología de los lenguajes documentales que los clasifica en jerárquicos o combinatorios.

examen del documento *m* Primera fase del proceso de indización consistente en la lectura del título, resumen, sumario, introducción, ilustraciones y palabras o frases destacadas en una tipografía diferente.

exhaustividad *f* Criterio relacionado con el número de conceptos que se tienen en cuenta para caracterizar el contenido entero de un documento. El principal criterio de selección es el

valor potencial del concepto para los usuarios de su SID. Podemos distinguir entre exhaustividad baja, media y alta en función del número de descriptores.

extract *m* Terminología anglosajona para los resúmenes automáticos. Los *extracts* son los resúmenes formados a partir de la extracción de algunas frases del texto previamente seleccionadas por un programa.

fichero inverso *m* Fichero donde se almacenan todos los términos de indización. Estos se sitúan uno detrás del otro de forma secuencial y asociados al documento al que hacen referencia.

hiperónimo *adj.* Decimos que una palabra es hiperónima cuando tiene un campo significativo que incluye otro de menor extensión. Ejemplo: color es un hiperónimo con respecto a amarillo, naranja, verde...

hipónimo *adj.* Decimos que una palabra es hipónima cuando tiene un campo significativo que queda incluido en otro de mayor extensión. Ejemplo: amarillo, naranja, verde son hipónimos ya que pertenecen al término color.

homonimia *f* Tipo de polisemia. Se da cuando dos conceptos diferentes han llegado a tener el mismo nombre, la misma forma, pero vienen de orígenes diferentes y por lo tanto tienen etimologías diferentes.

identificador *m* Término de indización propio del lenguaje documental Listado de autoridades. También se conocen con el nombre de autoridad y descriptores.

indización *f* Acción de describir o identificar un documento en relación a su contenido. Norma UNE 50-121-91. Indizar es el resultado de examinar el documento, seleccionar los conceptos y almacenarlos en una base de datos. Esta definición implica tres acciones, de las cuales la más significativa es la selección de los conceptos y su traducción al lenguaje documental.

indización por conceptos *loc. v.* Indización de las ideas y nociones del texto, sin reducirlo a un tema principal. Responden a la pregunta "¿cuáles son los conceptos de este documento"?, van ligados necesariamente a sistemas automatizados. Los lenguajes que indizan por conceptos son tres: listados de autoridades, tesauros, listados de descriptores libres.

indización por materias *loc. v.* Indización sintética. Responden a la pregunta "¿cuál es el tema de este documento?". Los lenguajes que indizan por materias son dos, los sistemas de clasificación y las listas de encabezamientos de materia.

indización por palabras clave *loc. v.* Indización de todas las palabras con significado del texto. Es el proceso más analítico que existe. No es una tarea de indización humana, sino automática. Los programas que indizan por palabras clave seleccionan sólo las palabras que tienen significado (preferentemente sustantivos). Sólo hay un lenguaje por palabras clave, y es evidentemente el único lenguaje automático, el listado de palabras clave.

ISO 214: 1976 *f* Norma internacional, traducida por AENOR como norma UNE 50-103-90 *Preparación de resúmenes*.

jerárquica *adj.* Tipología de lenguaje documental consistente en estructurar los términos de indización de forma arborescente. El vocabulario se presenta en forma de cadena, con términos genéricos que agrupan términos más específicos. Todos los términos dependen de un término superior y de significado más genérico. Esta estructura permite agrupar los conceptos por temas.

lenguaje artificial *m* *Ved* controlado.

lenguaje documental *m* Vocabulario de términos en lenguaje natural o un sistema artificial de signos normalizados que facilitan la representación del contenido de los documentos. Sus funciones principales son indizar el contenido de los documentos y permitir la recuperación a partir del campo materia.

lenguaje natural *m* Lenguaje que usamos de forma cotidiana: catalán, castellano, vasco, gallego, francés, etc.

libre *adj.* Tipología de lenguaje documental consistente en listas de términos extraídos del lenguaje natural sin formar parte de ningún listado establecido a priori, ni haber pasado un proceso de control de su vocabulario. Los lenguajes libres son dos: los listados de descriptores libres y el listado de palabras clave.

listado de autoridades *m* Lenguaje documental. Analítico por conceptos, natural, controlado, postcoordinado y combinatorio. Su término de indización se conoce como identificador, autoridad o descriptor.

listado de descriptores libres *m* Lenguaje documental. Analítico por conceptos, natural, libre, postcoordinado y combinatorio. Su término de indización se conoce como descriptor.

listado de encabezamientos de materia *m* Lenguaje documental. Sintético por materias, natural, controlado, precoordinado y combinatorio. Su término de indización se conoce como encabezamiento.

listado de palabras clave *m* Lenguaje documental. Analítico por palabras clave, natural, libre, postcoordinado y combinatorio. Su término de indización se conoce como palabra clave.

natural *adj.* Tipología de lenguaje documental consistente en el uso de palabras del lenguaje usual, habitual, no códigos. Hay cinco lenguajes documentales naturales: las listas de encabezamientos de materia, los listados de autoridades, los tesauros, los listados de descriptores libres y los listados de palabras clave.

naturaleza de los lenguajes *f* Tipología de los lenguajes documentales que los clasifica en codificados o naturales.

nivel de análisis *m* Tipología de los lenguajes documentales que los clasifica en lenguajes de materias, conceptos y palabras clave.

nivel de control *m* Tipología de los lenguajes documentales que los clasifica en libres o controlados.

nivel de coordinación *m* Tipología de los lenguajes documentales que los clasifica en precoordinados o postcoordinados.

notación *f* Término de indización propio del lenguaje documental de los sistemas de clasificación.

palabra clave *f* Término de indización propio del lenguaje documental de las palabras clave o indización automática. Palabra o grupo de palabras seleccionadas de manera automática del título, resumen o texto de un documento que representan el contenido y permiten la recuperación.

palabra vacía *f* Palabra sin significado en las operaciones de indización y resumen. Son preposiciones, artículos, verbos, adverbios, etc.

polisemia *f* Propiedad de un signo lingüístico de tener más de un significado. Decimos que dos palabras son polisémicas cuando el mismo signo lingüístico, palabra o sonido, tiene más de un significado. La palabra tiene un único origen etimológico y acaba teniendo significados diferentes sin cambiar su categoría gramatical.

ponderación (de frases, de palabras) *f* Método que evalúa las frases y las palabras de un texto en función de parámetros como la frecuencia, la presencia de palabras indicativas (buscan palabras como *importante, esencial, conclusiones*, etc.), la aparición en lugares destacados, por ejemplo el título: al principio de cada párrafo, al final a modo de conclusiones, etc.

postcoordinación *f* Tipología de lenguaje documental consistente en combinar los términos de indización en el momento de la recuperación. Los lenguajes postcoordinados sólo tienen sentido en sistemas documentales automatizados que dispongan de un fichero inverso. Los lenguajes postcoordinados son cuatro: listados de autoridades, tesauros, listados de descriptores libres y listados de palabras clave.

precoordinación *f* Tipología de lenguaje documental consistente en determinar a priori cómo se combinan los términos, ya sea a la hora de construir el lenguaje, de indizar el documento, o de recuperarlo. Los dos lenguajes precoordinados son los sistemas de clasificación y las listas de encabezamientos de materia.

procesamiento en lenguaje natural (PLN) *m* Rama de la inteligencia artificial y de la lingüística computacional que estudia los lenguajes que usan los humanos para interactuar con los ordenadores en contextos escritos y orales. EL PLN estudia cómo emular el conocimiento humano, en cuanto a la identificación de los conceptos y frases con contenido relevante.

relación de significado *f* *Ved* relación semántica.

relación semántica *f* Relaciones de significado de las palabras. Las relaciones pueden ser de tipo genérico, específico o relacionado de un término con respecto a otro. En lenguaje natural estas relaciones son implícitas pero en un lenguaje documental hay que definir estas relaciones, agrupando y relacionando los términos afines.

representar el contenido *loc. v.* Expresión que significa describir el tema o los temas de un documento.

resumen *m* Presentación abreviada y precisa de un documento, sin interpretación ni crítica y sin mención expresa del autor del resumen. Norma UNE 50-103-90 *Preparación de resúmenes*.

resumen indicativo *m* Resumen que consigna sólo las ideas centrales del documento. Su lectura no puede sustituir la lectura del original.

resumen informativo *m* Resumen que consigna el tema central, temas adicionales, naturaleza y objetivo del documento, metodología, resultados, conclusiones y anexos. La idea de fondo es que un resumen informativo puede sustituir en ocasiones la lectura del documento original.

resumen selectivo *m* Resumen que consigna sólo una parte concreta del documento. El más habitual es el resumen de conclusiones, pero también hay otros tipos, como la reseña (*review*).

selección de los términos de indización *f* Segunda fase en el proceso de indización consistente en identificar las nociones que son elementos esenciales de la descripción del contenido. Los criterios de selección son el número de conceptos (criterio de exhaustividad) y la exactitud de los mismos (criterio de especificidad).

símbolo de clase *m* Ved Notación.

sinonimia *f* Palabras que tienen el mismo significado. Ejemplo: alimento, nutriente, comida, provisión. En un sistema documental, si no se controlan y se usan indiscriminadamente, comportan silencio documental.

sistema de clasificación *m* Lenguaje documental. Sintético por materias, codificado, controlado, precoordinado y jerárquico. Su término de indización se conoce como notación o símbolo de clase.

término de indización *m* Representación de un concepto en lenguaje natural o un código de clasificación. Los términos de indización pueden estar formados por una palabra o más de una.

tesauro *m* Lenguaje documental. Analítico por conceptos, natural, controlado, postcoordinado, jerárquico y combinatorio. Su término de indización se conoce como descriptor.

traducción a un lenguaje documental controlado *f* Buscar un concepto expresado en lenguaje natural en el listado de términos de un lenguaje documental controlado. Utilizar el término controlado para indizar y recuperar.

UNE 50-103-90 *Preparación de resúmenes* *f* Norma española que establece las directrices que se tienen que seguir para presentar los resúmenes en los documentos. Pone especial énfasis en la preparación de resúmenes por parte de los autores de los documentos primarios y en la misma publicación.

UNE 50-113-92/1 *f* Norma española titulada Documentación e información. Vocabulario. Parte 1. Conceptos fundamentales. En: *Documentación: Normas fundamentales*. Madrid: AENOR, 1994.

UNE 50-121-91 *f* Norma española titulada *Métodos para el análisis de documentos, determinación de su contenido y selección de términos de indización*. Basa el proceso de indización en tres fases: examinar el documento para identificar su contenido, seleccionar los conceptos principales de los contenidos y traducir a un lenguaje documental.

unitérmino *m* La parte más pequeña con significado de un término de indización. La norma UNE 50-113-92/1 define los unitérminos como el elemento significativo más pequeño de un lenguaje documental utilizado para representar un concepto específico en un sistema de indización coordinado; no se debe confundir con palabra clave o descriptor.

univocidad *f* Representar un concepto con un único término.

Bibliografía

Bibliografía sobre el resumen

AENOR (1990). *Documentación. Preparación de resúmenes. UNE 50 103 90.* Madrid: AENOR.

Climent, Salvador. "Sistemes de resum automàtic de documents". *Digit. Hum. Revista digital d'humanitats*. ISSN 1575-2275.

Lloret, E.; Ferrández, O.; Muñoz, R.; Palomar, M. (2008). "Integración del reconocimiento de la implicación textual en tareas automáticas de resúmenes de textos". *Procesamiento del lenguaje natural*, núm. 41, pág. 183-190.

Mateo, P. L.; González, J. C.; Villena, J; Martínez, J. L. (2003). Un sistema para resumen automático de textos en castellano.

Pinto Molina, M. (1992). *El resumen documental: principios y métodos*. Madrid: Pirámide/ Fundación Germán Sánchez Ruipérez (Biblioteca del Libro, Y).

Bibliografía sobre la indización

Abadal, E.; Codina, L. (2005). "Recuperación de Información". En: *Bases de Datos Documentales: Características, funciones y método* (cap. 2. p. 29-92). Madrid: Síntesis.

AENOR (1997). *Métodos para el análisis de los documentos, determinación de su contenido y selección de los términos de indización. Norma UNE 50-121-91*. Madrid: AENOR.

AENOR (1997). "Documentación e información. Vocabulario. Parte 6: lenguajes documentales". *Revista Española de Documentación Científica*, Norma UNE-50-113/6 (ISO 5127/6), vol. 20, núm. 4, pág. 417-436.

Cid, P.; Cuadrado, M.; Aguiriano, C. (1999). *Fonaments de llenguatges documentals*. [Document electrònic]. Barcelona: UOC.

Codina, L. (1994). "El papel del lenguaje natural en los sistemas multimedia: una reflexión sobre la tecno-simpleza y la ciber-ingenuidad". *Cuadernos de documentación multimedia*, núm. 3 (junio).

Gil Leiva, I. (2008). *Manual de indización. Teoría y práctica*. Gijón: Ediciones Trea (Biblioteconomía y Administración cultural, 193).

Gil, I.; Rodríguez Muñoz, J. V. (1996). "El Procesamiento del lenguaje natural aplicado al análisis del contenido de los documentos". *Revista general de información y documentación*, vol. 6, núm. 2, pág. 205-218.

Gil Urdiciain, B. (1992). "Función de los lenguajes documentales en el tratamiento de la información en las organizaciones". *Revista general de información y documentación*, vol. 2, Núm. 2, pág. 195-200.

Gil Urdiciain, B. (2004). *Manual de lenguajes documentales*. Gijón: Ediciones Trea (Biblioteconomía y Administración cultural, 106).

Norma UNE 50-113-92/1. *Documentación e información. Vocabulario. Parte 1. Conceptos fundamentales* (1994). En: *Documentación: Normas fundamentales*. Madrid: AENOR.

Slype, G. van (1991). *Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales*. Madrid: Pirámide. Fundación Germán Sánchez Ruipérez. Biblioteca del Libro.