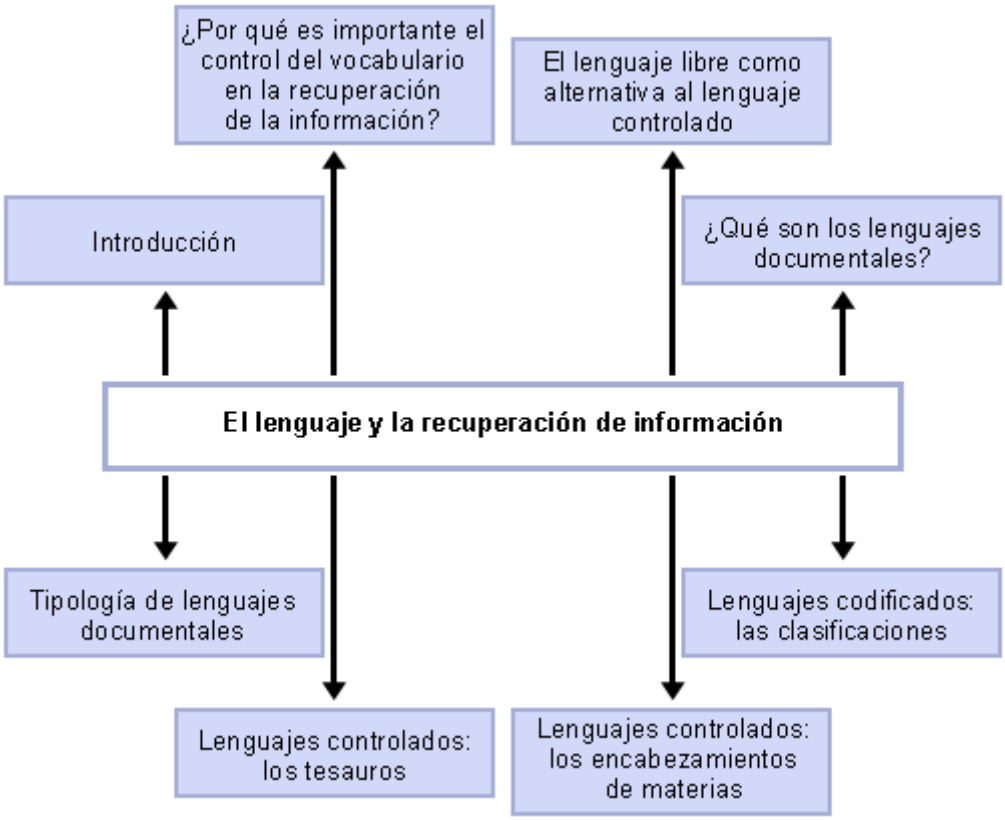
El lenguaje y la recuperación de información

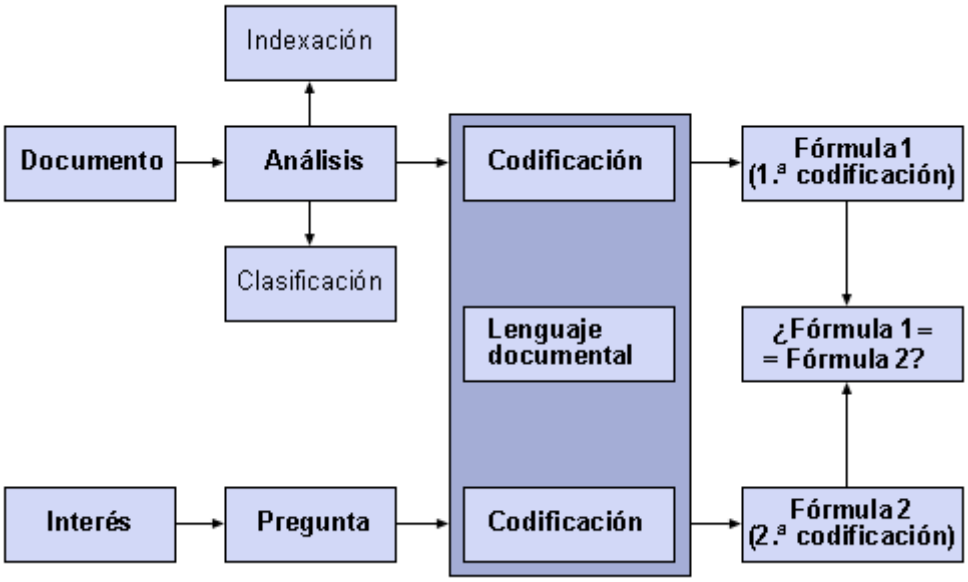


Introducción

En la búsqueda de información, el éxito o el fracaso del resultado obtenido dependerá, en gran medida, de la correcta elección de los términos por los cuales el contenido de los documentos ha sido identificado.

> **Un lenguaje de indización puede definirse como los términos o códigos que deben ser utilizados como puntos de acceso a un índice.**
>
> **Un lenguaje de recuperación puede ser definido como los términos o códigos que utiliza la persona que efectúa la búsqueda cuando especifica un argumento de búsqueda.**

Jacques Maniez denomina la comunicación entre lo que pide y lo que suministra la información como "mediación documental", que se realiza mediante un complejo proceso con "doble codificación", y que esquematiza de este modo:



Fuente: J. Maniez (1993)

El proceso de doble codificación es, por lo tanto, una comparación entre el resultado de los dos códigos: el del analista/indizador (fórmula 1) y el del documentalista o usuario final (fórmula 2). Por lo tanto, para que estas operaciones tengan éxito es necesario que la fórmula 1 y la fórmula 2 tengan una perfecta identidad formal.

Aplicamos este proceso de doble codificación en el ejemplo siguiente:

Un usuario quiere localizar artículos que traten el tema de los derechos de autor. Hace una búsqueda en una base de datos concreta dentro del campo de *descriptor* y no localiza ningún artículo relacionado con el tema. Analizamos el porqué de este resultado y vemos que los artículos que tratan este tema han sido indizados por el término *propiedad intelectual*.

En este caso, podemos decir que la comunicación documental ha fracasado, porque una materia no ha sido codificada del mismo modo por la persona que la ha indizado (indizador) y por la persona que hace la búsqueda (documentalista o usuario final).

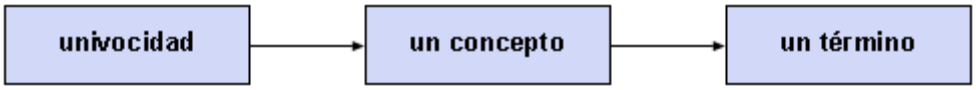
¿Por qué es importante el control del vocabulario en la recuperación de la información?

En el contexto anterior, podemos afirmar que, para que la mediación documental tenga éxito, es necesario que elaboremos nuestra búsqueda mediante un **lenguaje controlado** que nos asegure que:

- Un único término denomina un único objeto, de modo que:
  - Un término no puede definir más de un objeto, y
  - Un objeto no puede ser definido por más de un término.

Se precisa, pues, un lenguaje unívoco en el que:

1. Los términos estén relacionados desde el punto de vista semántico en los siguientes ámbitos:
   - Sinónimos y cuasisinónimos: palabras con el mismo significado o que quieren decir prácticamente lo mismo.
   - Antónimos: palabras que significan lo contrario.
2. Se establezca algún tipo de diferenciación entre los términos homónimos: por ejemplo, la misma palabra pero con significados diferentes.
3. Se establezcan también relaciones de tipo jerárquico entre los términos de diferentes grados: más amplios, más específicos y relacionados.



Dentro del proceso de recuperación, el control del vocabulario es importante, principalmente, en dos fases:

1. La **fase de entrada**: implica el análisis conceptual y su traducción a un lenguaje determinado en el momento de inidizar el documento. En la mayoría de los sistemas hay un vocabulario controlado; es decir, un conjunto limitado de términos que se debe utilizar para representar las materias de los documentos.

2. La **fase de salida**: los usuarios realizan diferentes peticiones en el centro de documentación y los documentalistas preparan las estrategias de búsqueda para estas peticiones. En estas estrategias de búsqueda también hay que considerar las fases del análisis conceptual y de traducción. La primera fase implica un análisis de la petición con el fin de determinar lo que realmente busca el usuario, y la segunda consiste en la traducción del análisis conceptual al vocabulario propio del sistema utilizado.

El hecho de que una base de datos tenga un campo de "descriptor" no quiere decir que el lenguaje esté controlado; se puede tratar de un campo que se llena sin ninguna herramienta de indización como base.

En el ámbito de la documentación, los lenguajes controlados que utilizamos se denominan *lenguajes documentales*.

El lenguaje libre como alternativa al lenguaje controlado

El lenguaje libre consiste en el uso de los términos que aparecen en el mismo documento, y en la misma forma en que lo hacen. Estos términos pueden ser palabras significativas del título, del resumen o del texto.

Un directivo de empresa, responsable de un equipo de trabajo, está muy interesado en localizar artículos sobre un nuevo concepto aplicado a la dirección de personal, conocido como "inteligencia emocional", ya que piensa que su aplicación puede hacer mejorar el bienestar y rendimiento de sus trabajadores. Busca información en bases de datos y catálogos de bibliotecas utilizando este término, pero resulta que los documentos que le pueden interesar no han sido indizados con este término, ya que se trata de un concepto de aparición muy reciente. Estos documentos han sido indizados por *dirección de personal* y *estrategias de dirección*.

El lenguaje libre es muy utilizado en búsquedas efectuadas en Internet y en bases de datos a texto completo. Es recomendable utilizar el lenguaje libre cuando no conocemos el sistema en el que haremos la búsqueda, y también cuando buscamos temas muy actuales.

El lenguaje libre o natural puede ejercer un papel muy importante en la búsqueda de la información y puede suplir algunos inconvenientes de los lenguajes documentales, ya que:

- Es el lenguaje de comunicación de la persona que busca.
- Es muy adaptable a los cambios tanto de nuevas palabras como de frases que pueden aparecer en un vocabulario especializado o en una disciplina.
- Con el uso del lenguaje natural se evita a la persona que hace la búsqueda la necesidad de entender cómo funciona, por ejemplo, un tesauro.
- Permite realizar la búsqueda por más términos.
- Evita los gastos de creación y mantenimiento de un lenguaje controlado.

A pesar de ello, y como ya hemos comentado, el lenguaje libre presenta grandes inconvenientes, ya que:

- Es necesario que la persona que busca realice el esfuerzo de pensar en todas las posibles variantes del término que busca: sinónimos, cuasisinónimos, etc.
- Es muy posible que el resultado de la búsqueda contenga muchos documentos no pertinentes.

¿Qué son los lenguajes documentales?

B. Gil define el lenguaje controlado o documental como:

"... todo sistema artificial de signos normalizados, que facilitan la representación formalizada del contenido de los documentos para permitir la recuperación, manual o automática, de información solicitada por los usuarios".

B. Gil Urdiciaín (1996, pág. 324-353)

Por otro lado, A. Large lo define como:

"... un lenguaje artificial que ha sido creado para un sistema de recuperación de la información en particular o un grupo de sistemas relacionados. Al igual que otros lenguajes, un vocabulario controlado tiene su propio vocabulario y sintaxis y debe asegurar la consistencia en la representación de los conceptos, tanto en el momento de crear las bases de datos como cuando se hacen búsquedas".

A. Large y otros (1999)

Y por último Van Slype lo define como:

"... todo sistema de signos que permite representar el contenido de los documentos con la finalidad de recuperar los documentos pertinentes en respuesta a consultas que tratan sobre este contenido".

G. Van Slype (1991)

Según Van Slype, la definición del lenguaje documental como lenguaje artificial —expresada en las definiciones anteriores— es cierta en el caso de los sistemas de clasificación, las listas de materias y los tesauros; sin embargo, no lo es para los listados de palabras claves extraídas de los títulos, resúmenes y textos de los documentos, términos cada vez más importantes en el momento de buscar en bases de datos a texto completo y en Internet, y que corresponden a lo que él considera "sistemas menos controlados".

Conclusión

El lenguaje documental es un lenguaje intermediario en la medida en que sirve de puente entre las informaciones contenidas en los documentos y las informaciones solicitadas por los usuarios. Así pues, el lenguaje documental identifica materias y no se refiere a otros criterios utilizados en la búsqueda documental: autor del documento, lengua del texto, fecha de publicación, etc.

Tipología de lenguajes documentales

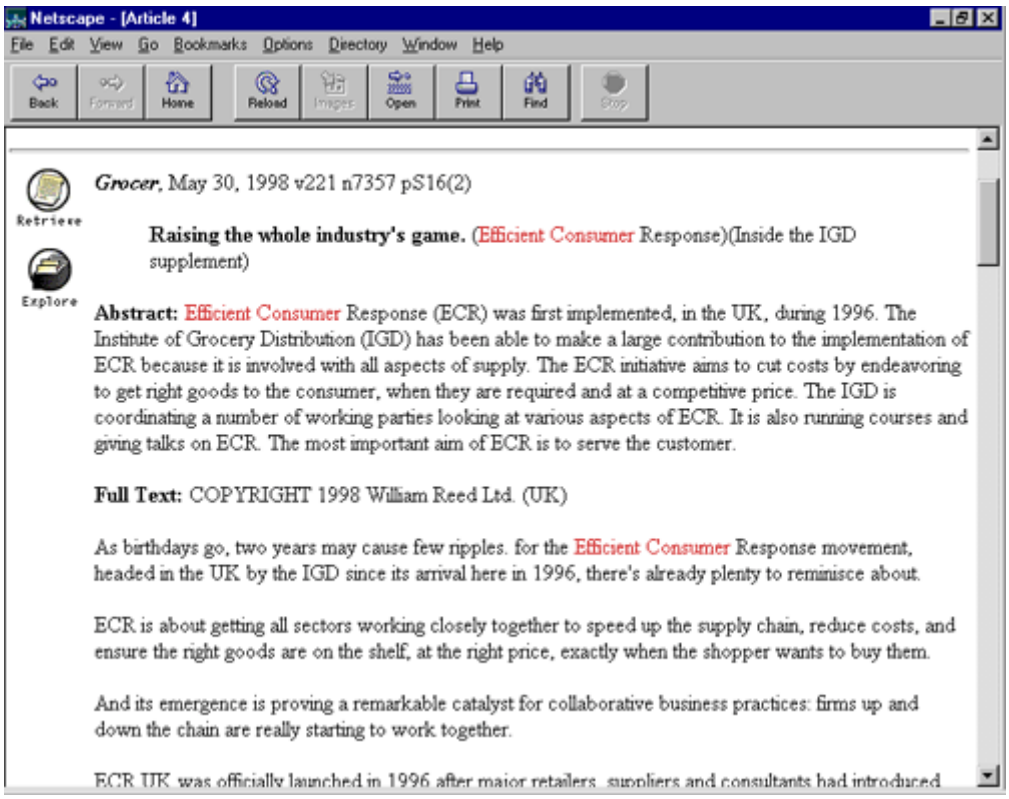
La mayoría de los autores coincide en dividir los lenguajes documentales según las tipologías siguientes:

1. **Lenguajes libres**, constituidos *a posteriori*, sobre la base de la indización en lenguaje libre o natural de los documentos.

   Los lenguajes libres pueden presentarse bajo dos tipologías:

   - **Palabras clave**

Corresponde al tipo de lenguaje libre más habitual. Esta tipología consiste en una lista de palabras significativas ordenadas alfabéticamente. Cuando hablamos de palabras significativas, nos referimos al hecho que son palabras "no vacías" (es decir, todas las palabras que no son artículos, conjunciones, pronombres, preposiciones, numerales y algunos verbos y adverbios). Estas palabras se extraen de forma automática por ordenador, a partir del título, del resumen y, cada vez más, del texto completo de los documentos registrados.



Artículo de un texto completo en el que podemos ver destacados los términos por los que se ha efectuado la búsqueda en lenguaje libre.

- **Lista de descriptores libres**

  Una lista de descriptores es un listado (ordenado alfabéticamente) de conceptos destacados, por un proceso intelectual, a partir de los documentos registrados dentro de un sistema documental determinado. Estos conceptos se expresan mediante palabras o expresiones extraídas de los documentos, o bien los proponen los documentalistas, sin verificar si existen previamente en una lista establecida *a priori*.

  Vemos que las características son idénticas a las que presentan las palabras clave. La única diferencia consiste en que en los descriptores libres se han excluido las palabras vacías y también los casos de polisemia o sinonimia más evidentes, de manera que el vocabulario está depurado a un nivel elemental, lo cual facilita la tarea de recuperación de información.

2. **Lenguajes controlados**, construidos *a priori* antes de indexar los documentos de una colección.

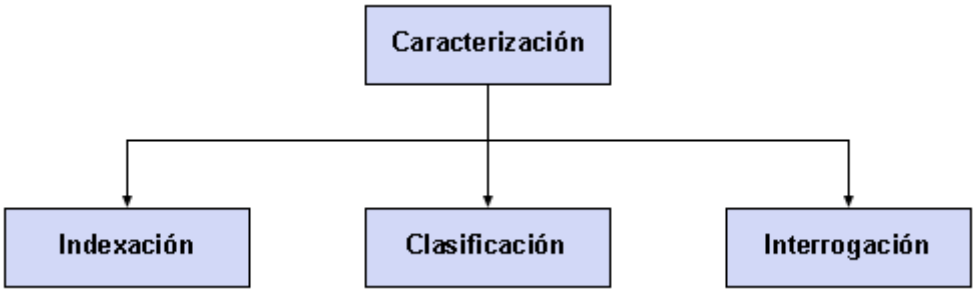
   Existen dos tipos principales de lenguajes controlados:

   - Los encabezamientos de materias
   - Los tesauros de descriptores

3. **Lenguajes codificados** representados, principalmente, por las clasificaciones.

   Teniendo en cuenta su importancia, estos tres últimos tipos de lenguajes controlados serán tratados más adelante en este mismo bloque.

Así pues, vemos que muchos autores utilizan el mismo término para referirse tanto a los lenguajes de indización como a los lenguajes de clasificación. No obstante, J. Maniez establece una distinción de las diferentes tipologías de los lenguajes documentales que se desvía de la mayoría de los autores y que pensamos que es interesante:



Fuente: J. Maniez (1993).

Según el mismo autor:

> "... la clasificación se distingue claramente de la indización porque el análisis del documento se orienta hacia la búsqueda de la materia "dominada" y no hacia los conceptos claves que lo caracterizan, como haría un indizador."
>
> **J. Maniez** (1993)

En este sentido, cuando clasificamos estamos haciendo una formulación sintética (una única formulación por documento), mientras que la indización nos permite representar el contenido de un documento o de una consulta de manera analítica (es decir, enumerando los conceptos y combinándolos entre sí en su posterior recuperación). No establece, por lo tanto, un límite tan restrictivo; el indizador puede introducir tantos términos como considere oportuno.

Por otra parte, utilizamos los lenguajes de interrogación como un instrumento para elaborar nuestra estrategia de búsqueda con los términos de indización: álgebra de Boole, lógica aritmética, etc.

Esta distinción, establecida como hemos visto por J. Maniez, es muy interesante en la búsqueda documental, ya que la mayoría de los sistemas de recuperación de información nos permite interrogarlos por medio de lenguajes de indización (básicamente tesauros y listas de encabezamientos de materias). La clasificación, por otra parte, se utiliza más para localizar documentos mediante las técnicas de "hojear" o *browsing*, que también estudiaremos más adelante.

¿Qué problemas podemos encontrarnos si utilizamos un lenguaje controlado o documental en una búsqueda?

Hasta ahora hemos visto las ventajas que nos proporciona el uso de un vocabulario controlado en el momento de hacer búsquedas. Hay que tener en cuenta, sin embargo, que también puede presentar una serie de inconvenientes. Según A. Large, estos inconvenientes son los siguientes:

1. La restricción del vocabulario a uno o unos cuantos términos relacionados provoca que se pierda la especificidad que permite el lenguaje libre. En la práctica, los conceptos son representados por el término más adecuado del lenguaje controlado en cuestión. Como consecuencia de ello, algunos ítems son indizados por un término que no es exactamente el concepto.

2. El número de términos por los cuales se indiza un ítem en un lenguaje controlado, aunque sean 10 términos por ítem, es drásticamente menor que en el caso del lenguaje libre, en el cual puede haber centenares o millares.

3. El lenguaje controlado es un lenguaje artificial creado por especialistas de la información, y no puede reflejar completamente la terminología más actual en un campo. Un lenguaje controlado no puede estar continuamente actualizado, y sólo de forma casual puede ser modificado.

4. Es difícil determinar la materia de un documento y asignarle una serie de términos controlados que representen esta materia. En la práctica, los indizadores no siempre tienen la misma consistencia en la asignación del mismo término controlado para representar el mismo concepto.

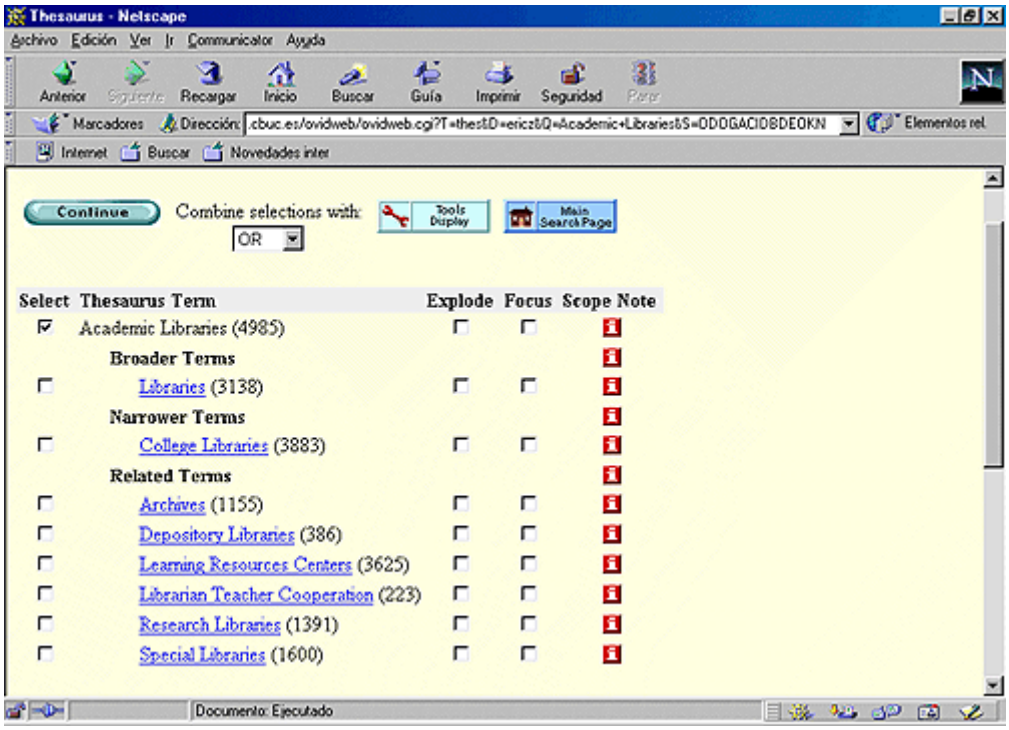
5. La creación y el mantenimiento de un lenguaje controlado, sobre todo de los tesauros, es muy caro.

Lenguajes controlados: los tesauros

Podemos decir también que:

Los tesauros son el elemento del lenguaje controlado más típico dentro de los contextos especializados. Normalmente organizan las materias de un ámbito del conocimiento concreto.

- Un tesauro es una lista estructurada de conceptos, destinados a representar de manera unívoca el contenido de los documentos y de las consultas dentro de un sistema documental determinado.



Extracto del tesauro especializado en educación que incorpora la base de datos ERIC.

Lectura obligatoria

Gil Urdiciaín, B. (1996). "Los lenguajes documentales II". En: J. López Yepes. *Manual de información y documentación* (pág. 354-374). Madrid: Pirámide.

- Los conceptos están extraídos de una lista finita, establecida *a priori*.
- Sólo los términos que figuran en esta lista pueden ser utilizados para indexar los documentos y las consultas.
- La ayuda al usuario la proporciona la estructura semántica del tesauro, fundamentalmente, las relaciones de equivalencia, de jerarquía y de asociación.

Lenguajes controlados: los encabezamientos de materias

Como ya hemos comentado, los encabezamientos de materias son una tipología de lenguaje de indización que nos permite realizar un control del vocabulario por medio de un listado alfabético de términos.

Tradicionalmente, las listas de encabezamientos de materias se han utilizado para consultar por materias los documentos de una biblioteca mediante su catálogo.

Una lista de encabezamientos tiene las características siguientes:

- Está constituida por una colección de conceptos ordenados alfabéticamente y destinados a representar de manera unívoca el contenido de los documentos y de las consultas dentro de un sistema documental.
- Estos conceptos se expresan mediante palabras o expresiones extraídas de una lista finita, establecida *a priori*.
- Sólo los términos que figuran en esta lista pueden ser utilizados para indexar los documentos y las consultas.
- En un mismo encabezamiento pueden combinarse diferentes conceptos, mediante la aplicación de subdivisiones que concretan el alcance del término.

Documentos	Escoja línea
18	Economía
2	Economía -- Aspectos ambientales
1	Economía -- Aspectos morales
3	Economía -- Aspectos políticos
1	Economía -- Aspectos políticos -- Congresos
1	Economía -- Aspectos sociales -- Recopilación de escritos
1	Economía -- Aspectos sociológicos
2	Economía -- Bases de datos
1	Economía -- Cataluña
1	Economía -- Cataluña -- 1642/1862
2	Economía -- Cataluña -- 1955/1995
1	Economía -- Cataluña -- 1986/1993
1	Economía -- Comunidad Europea, Países de la
15	Economía de empresa

Extracto del listado de encabezamientos de materias del catálogo de la Biblioteca de la UOC.

Lectura obligatoria

Gil Urdiciaín, B. (1996). "Los lenguajes documentales I. Listas de encabezamientos de materias". En: J. López Yepes (coord.). *Manual de información y documentación* (pág. 335-341). Madrid: Pirámide.

Lenguajes codificados: las clasificaciones

Una clasificación es un conjunto ordenado de conceptos que se presentan distribuidos sistemáticamente en clases, subclases y apartados, y conforman una estructura.

> **Los lenguajes de clasificación representan el contenido mediante códigos numéricos o alfanuméricos que corresponden a las categorías o clases en las que previamente se ha dividido el ámbito del conocimiento.**

Durante los últimos años, antes de la aparición de Internet, el uso de clasificaciones en las bibliotecas se había limitado a la función de topográfico.

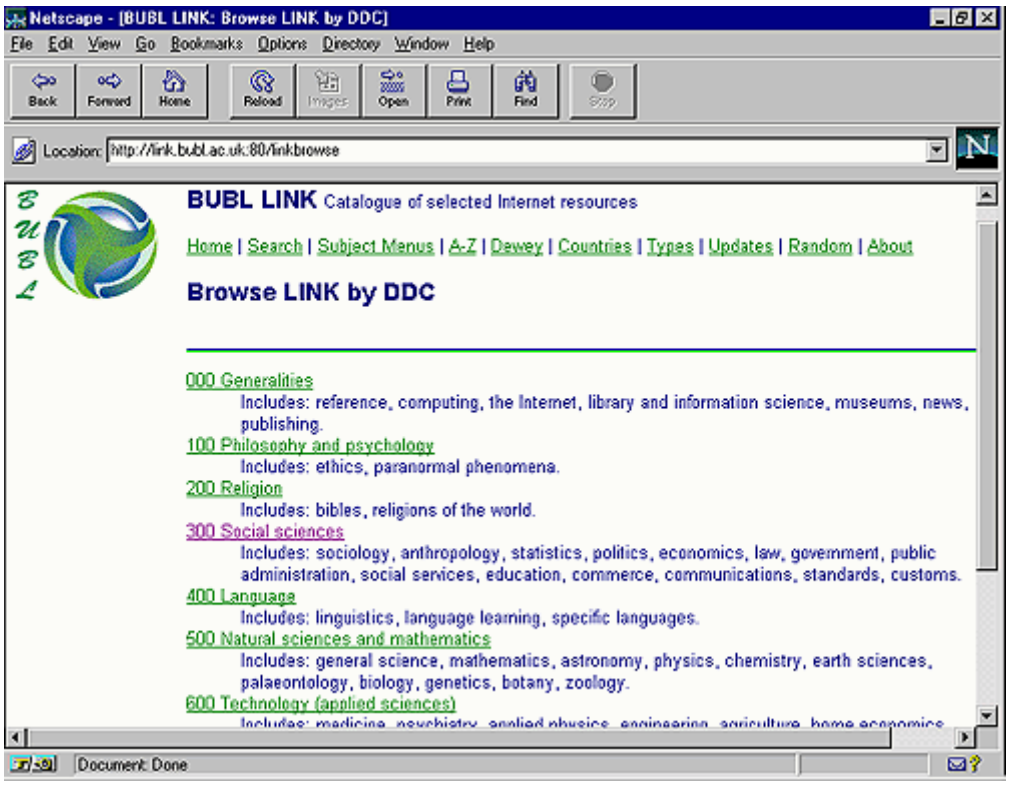
De hecho, y tal y como dice Carme Caro en su artículo "Sistemas de clasificación y organización de la información en Internet", hasta hace dos décadas las clasificaciones no tenían un futuro muy claro en el mundo de las bibliotecas, ya que su función se había quedado limitada a la ordenación física de los documentos, y para la indización se utilizaban los vocabularios controlados del tipo tesauro o listados de encabezamientos.

No obstante, el desarrollo de las tecnologías de la información ha ampliado y cambiado la naturaleza y aplicación de las clasificaciones, de manera que cada vez son más utilizadas en los sistemas de recuperación de la información, sobre todo en la búsqueda en Internet.

"Cómo organizar y recuperar los recursos electrónicos accesibles en Internet es uno de los retos que se han planteado desde sus orígenes. Durante el primer periodo del desarrollo de los servicios de información en red muchos especialistas, sobre todo los relacionados con la comunidad informática, cuestionaron el valor de los métodos de descripción de contenido utilizados tradicionalmente en las bibliotecas insistiendo en las ventajas de los sistemas automatizados de indización y recuperación que utilizaban el texto completo de los documentos. Sin embargo, esta percepción ha variado con el incremento del uso de Internet y especialmente de la World Wide Web para el almacenamiento y recuperación de la información. Han emergido dos vías diferentes para localizar recursos en la Red. Una aproximación consiste en el desarrollo de motores de búsqueda que utilizan palabras clave para recuperar la información. Son extremadamente útiles, aunque tienden a recuperar grandes cantidades de información irrelevante. La otra posibilidad consiste en crear directorios organizados temáticamente que ayuden a los usuarios a visualizar los recursos y navegar por la WWW. La categorización de los recursos condujo a la adopción de esquemas de clasificación que proporcionaban la necesaria estructura semántica. Motores de búsqueda tan populares como *Yahoo*! diseñaron su propio esquema de clasificación para dar una estructura jerárquica a los recursos de Internet que se habían indizado, mientras que servicios más especializados, que sólo facilitaban el acceso a una selección de documentos, también entendieron que una estructura que permitiera la visualización, basada en un sistema de clasificación, era un buen complemento a sus servicios de búsqueda.

La idea de utilizar en estos servicios los sistemas de clasificación tradicionalmente empleados por las bibliotecas —Clasificación Decimal Universal (CDU), Dewey Decimal Classification (DDC) o Library of Congress Classification (LCC)— surgió en el verano de 1994 en una serie de listas de discusión y noticias. Entre los argumentos que se utilizaron destacó la defensa del papel profesional e institucional que históricamente han jugado bibliotecas y bibliotecarios en la organización de la información: aplicar en la biblioteca virtual las técnicas y herramientas utilizadas por los bibliotecarios debía suponer un claro beneficio para los usuarios. Con este debate sin concluir, lo cierto es que cada vez son más los servicios que, además de seleccionar los recursos, utilizan el potencial de las clasificaciones para organizar y recuperar los recursos electrónicos disponibles en Internet."

C. Caro Castro (1998)



Extracto de la clasificación DCC (clasificación decimal de Dewey) utilizada por el buscador especializado BUBL para organizar los recursos.

Lectura obligatoria

Gil Urdiciaín, B. (1996). "Los lenguajes documentales I. Clasificaciones". En: J. López Yepes (coord.). *Manual de información y documentación*. Madrid: Pirámide.

Ejercicios de autoevaluación

1. Explicad en qué consiste el proceso de *doble codificación*.
2. Comentad en qué dos fases del proceso documental es importante el control del vocabulario.
3. Elaborad un cuadro en el que se reflejen las principales ventajas y los principales inconvenientes del lenguaje controlado. Comparadlo con el lenguaje libre.
4. ¿Cuándo y en qué fuentes de información es muy recomendable utilizar el lenguaje libre como lenguaje de búsqueda?
5. Explicad brevemente las diferentes tipologías de lenguajes documentales.
6. En las listas de encabezamientos de materias, ¿qué clases de encabezamientos encontraremos? Explicadlas brevemente.
7. Comentad la nueva función de las clasificaciones como lenguaje de recuperación de la información.
8. ¿Cuántas clases tiene la CDU (clasificación decimal universal)? Enumeradlas.
9. Explicad los conceptos de sinonimia y polisemia como accidentes lingüísticos y comentad qué recursos utilizan los tesauros para controlarlos.
10. Determinad qué relaciones jerárquicas pueden establecerse entre los descriptores de un tesauro.

Bibliografía

Caro Castro, C. (1998). "Sistemas de clasificación y organización de la información en Internet". En: *VI Jornadas Españolas de Documentación* "Fesabid 98". Valencia.

Gascón García, J.; Abadal Falgueras, E. (1997). "Documentació". En: *Documentació i arxivística* (pág. 30-70). Barcelona: UOC.

Lancaster, F.W. (1995). *El control del vocabulario en la recuperación de información*. Valencia: Universitat de València.

Large, A.; Tedd, A.L.; Hartley, R.J. (1999). *Information seeking in the online age: principles and practice*. London: Bowker-Saur, cop.

López Yepes, J. (coord.) (1996). *Manual de información y documentación*. Madrid: Pirámide.

Maniez, J. (1993). *Los lenguajes documentales y de clasificación: concepción, construcción y utilización en los sistemas documentales*. Salamanca: Fundación Sánchez Ruizpérez; Madrid: Pirámide.

Slype, G.V. (1991). *Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales*. Salamanca: Fundación Sánchez Ruizpérez; Madrid: Pirámide.