

MPN 7216  
# Cat. 1253

# XII COLOQUIO SOBRE INVESTIGACIONES BIBLIOTECOLOGICAS

CENTRO UNIVERSITARIO  
DE INVESTIGACIONES BIBLIOTECOLOGICAS

UNAM

7-8 julio 1994

BIBLIOTECA



CENTRO UNIVERSITARIO  
DE INVESTIGACIONES  
BIBLIOTECOLOGICAS

**Mesa Redonda**  
**Lenguajes de Indización**  
**Libres o Controlados**

"Lenguajes de indización libres o controlados:  
la experiencia de la Biblioteca  
de El Colegio de México"

Pilar María Moreno Jiménez  
Biblioteca Daniel Cosío Villegas  
El Colegio de México

INFOBILA

## Lenguajes de indización libres o controlados: la experiencia de la Biblioteca Daniel Cosío Villegas de El Colegio de México

### LENGUAJE LIBRE VERSUS LENGUAJE CONTROLADO

#### Introducción

La cuestión de qué tipo de lenguaje de indización adoptar presenta mayor interés si tenemos en cuenta que, como han demostrado numerosas encuestas en los Estados Unidos, al menos la mitad de las búsquedas en los catálogos en línea son temáticas<sup>1</sup>.

En mi opinión, el principal beneficio que puede reportar el uso de un lenguaje de indización controlado es la eficacia en la recuperación. De otra parte, el elemento disuasorio de mayor peso lo constituiría el costo.

#### Control de calidad y control de autoridad

Independientemente del lenguaje usado, debe existir un control de calidad al introducir los datos al sistema, puesto que los datos capturados con escasa calidad afectan en forma negativa la eficiencia del mismo, mermando la precisión, integridad y/o consistencia de la base de datos.

La calidad de las bases de datos se mide, entre otras cosas, por la ausencia de errores de diversos tipos. En dos estudios sobre los encabezamientos de materia de la base de datos de OCLC llevados a cabo por O'Neill y Aluri<sup>2</sup>, y O'Neill y Vizine-Goetz<sup>3</sup> respectivamente, los autores categorizaron los errores que se presentaron en los encabezamientos de la siguiente manera:

1. Inconsistencias en el uso de espacios, signos de puntuación y mayúsculas
2. Errores tipográficos y ortográficos. Estos a su vez divididos en cuatro tipos:
  - omisión de caracteres (educación)
  - adición de caracteres (edducación)
  - substitución de caracteres (educasión)
  - trasposición de caracteres (educacino)

---

<sup>1</sup>Simone Klugman, "Failures in subject retrieval", en: Holley, Robert P., ed. Subject control in online catalogs. New York, Haworth Press, 1989, p.10

<sup>2</sup>Edward T. O'Neill y Rao Aluri. A method for correcting typographical errors in subject headings in OCLC records. Dublin, OCLC Office of Research, 1980.

<sup>3</sup>Edward T. O'Neill y Diane Vizine-Goetz, "Computer generation of a subject authority file", Proceedings of the ASIS Annual Meeting, v.19, 1982, p.220-223

3. Formas no admitidas de encabezamientos
4. Etiquetas de campos o subcampos MARC incorrectas (errores de codificación)
5. Otros

Macaferri señala que "el control de autoridad es más importante en un contexto automatizado que en un medio manual, dada la literalidad con la que las computadoras almacenan y recuperan la información"<sup>4</sup>. Para una computadora la cadena de caracteres: "Educación de adultos" es diferente a la cadena: "Educación de adultos". Los errores de espacios, puntuación y ortográficos, por lo general, alejan considerablemente los encabezamientos o términos en los índices y, en el caso de los errores ortográficos, imposibilitan también su recuperación por palabra. Si busco por la palabra "Cultura" y en un registro esta palabra aparece mal escrita como "Cutura", éste no aparecerá en el resultado de mi búsqueda. Por tanto, es necesario algún tipo de control en este sentido. Hay que destacar, sin embargo, que algunos sistemas almacenan los encabezamientos en los índices con mayúsculas y sin acentos ni diacríticos, con lo cual, aunque existan diferencias de este tipo entre encabezamientos, al quedar indizados juntos, éstas no afectan la recuperación. Por ejemplo, al buscar "BIOLOGIA" el sistema podría recuperar registros en los que aparecen como temas: "Biología", "biología" o "biologia".

Por otro lado, las inconsistencias en el lenguaje de indización pueden dar lugar a la dispersión de documentos sobre un mismo asunto. Si los documentos que tratan, por ejemplo, sobre taquigrafía tienen asignados indistintamente los temas "Taquigrafía" y "Estenografía" al buscar por la palabra "Taquigrafía" recuperaré únicamente los documentos que tenían asignado ese tema y no los que tenían como tema "Estenografía". McDonald afirma que "a medida que los catálogos en línea proliferan la importancia y relevancia del control de autoridad aumenta...[ya que]...: es una herramienta que puede usarse para mantener la consistencia e integridad del catálogo en línea...[y]... establece una estructuración conocida en el catálogo que facilita la recuperación de información"<sup>5</sup>.

Los errores de etiquetado y codificación pueden impedir asimismo la recuperación. Por ejemplo, si el tema de un documento queda codificado en un campo diferente, como pudiera ser un campo para coautor, y la búsqueda que realizamos es por tema, no aparecerá ese registro en el resultado de la búsqueda, ya que el campo para coautor lo más seguro es que no se encuentre incluido en nuestra

---

<sup>4</sup>James Tilio Maccaferri, "Managing authority control in a retrospective conversion project", Cataloging and Classification Quarterly, v.14, nos.3-4, 1992, p.145

<sup>5</sup>David R. McDonald, "Data dictionaries, authority control, and online catalogs: a new perspective", Journal of Academic Librarianship, v.11, September 1985, p. 221

definición para una búsqueda temática.

Por último, los errores conceptuales, como puede ser la asignación de temas inadecuados, además de afectar la recuperación, afectan la imagen profesional de los bibliotecarios frente a la comunidad.

### **Esfuerzo del usuario**

Lancaster señala que "aunque el vocabulario que se use en un sistema de recuperación no tiene efectos en la cobertura del sistema, el esfuerzo que el usuario debe realizar para explotarlo aumentará y, consecuentemente, también aumentará el tiempo de respuesta del mismo"<sup>6</sup>. Es importante que nuestra base de datos contenga mucha información muy relevante pero también que ésta pueda ser explotada con el menor esfuerzo y el menor tiempo por parte del usuario.

### **Auxiliares en las búsquedas temáticas**

Los catálogos automatizados, ya sea en línea o en discos compactos, presentan un mayor número de puntos de acceso que los catálogos tradicionales de tarjetas. En búsquedas temáticas el usuario puede generalmente acceder a los temas por campo, subcampo, o por palabras, dentro de éstos o en forma libre usando operadores booleanos y otros operadores de búsqueda, así como limitar su búsqueda temática por lugar geográfico o período de tiempo a través de los códigos en campos fijos asignados con ese propósito. La búsqueda por palabras en el título también facilita el acceso temático en algunos casos. Las búsquedas por números de clasificación, especialmente de la Clasificación Decimal, han demostrado ser igualmente útiles.

Muchos sistemas, además, proveen a los usuarios ayudas en línea o en manuales impresos para auxiliarlos en las búsquedas. Por ejemplo, para orientarles sobre qué otras estrategias usar cuando el resultado de una búsqueda es cero o permitiendo la consulta del thesaurus o vocabulario empleado para la indización.

Se ha dicho que en el medio automatizado el concepto de arreglo alfabético reviste menor importancia que en los catálogos manuales. Es frecuente, sin embargo, el uso de búsquedas mediante la revisión en los índices, posibilidad que la mayoría de los sistemas ofrece. La búsqueda en los índices presenta algunas ventajas, como son: se puede llegar a los términos por aproximación, sin necesidad de conocerlos en forma exacta previamente, y se pueden encontrar términos relacionados próximos en el índice, aunque se pierden los términos relacionados que no están próximos. Por ejemplo, si no estoy seguro si el término adecuado es "Biblioteconomía" o

---

<sup>6</sup>F.W. Lancaster, "Vocabulary control in information retrieval systems", en Advances in Librarianship, ed. por Melvin J. Voight y Michael H. Harris. New York, Academic Press, 1977, p.1-40

"Bibliotecología" y descubro en el índice la palabra "Bibliotecología" ahí mismo puedo encontrar fácilmente también "Bibliotecarias", "Bibliotecarios", "Bibliotecas", "Bibliotecas infantiles", "Bibliotecas y estado", pero me quedará más lejos "Bibliología" y aún más lejos "Educación bibliotecológica".

### Características del sistema

El éxito en la recuperación de los documentos depende también de las técnicas de recuperación, es decir, en buena medida, de las características del sistema. La mayoría de los sistemas de gestión documental permiten la utilización de los operadores booleanos "y" "o" "no", otros incluyen además operadores de proximidad como "cerca de" "adyacente a" "a no más de dos, tres, ... palabras más allá de" "en la misma frase que" "en la misma ocurrencia que", y otras técnicas como truncamientos, sustitución de caracteres, etc. que facilitan aún más las búsquedas y generalmente acortan el tiempo de respuesta y presentan una mayor precisión en el resultado.

### Referencias cruzadas

Aunque, en principio, pudiera parecer que la potencia en la capacidad de búsqueda que poseen algunos de los actuales sistemas para bibliotecas hace innecesario el esfuerzo de desarrollar estrategias de control de autoridad, lo cierto es que, como algunos estudios han demostrado, sin una estructura adecuada de referencias cruzadas la recuperación de la información en las bases de datos no es tan eficiente. Incluso en un catálogo en línea con capacidad para realizar búsquedas por palabras clave, búsquedas booleanas y truncamientos, es necesaria una estructura de referencias cruzadas<sup>7</sup>.

Johnston asegura que "a la computadora le falta la capacidad intelectual para reconocer asociaciones entre términos que no estén explícitamente ligados"<sup>8</sup>.

El catálogo automatizado debe dar al usuario la mayor cantidad posible de asociaciones que le permitan encontrar lo que busca, sin importar si el lenguaje de sus preguntas está o no ortodoxamente formalizado en términos bibliotecarios.

Las referencias de "véase" y "véase también" sirven al usuario para recuperar información que aparece en diferentes formas y para relacionar unos términos con otros.

Si bien es cierto que las referencias cruzadas siguen siendo

---

<sup>7</sup>Alexis J. Jamieson, Elizabeth Dolan y Luc Declerck, "Keyword searching vs. authority control in an online catalog", Journal of Academic Librarianship, v.12, no.5, November 1986, p.277-283

<sup>8</sup>Sarah Hager Johnston, "Current offerings in automated authority control: a survey of vendors", Information Technology and Libraries, September 1989, p.236-237

necesarias aun con sistemas capaces de recuperar por palabras clave y truncamientos a la derecha, también es cierto que con tales sistemas no es necesario generar tantas referencias como en los catálogos manuales<sup>9</sup>. Por ejemplo, tanto "Educación de niños" como "Niños-Educación" pueden recuperarse en la búsqueda por palabras clave: "Niños" y "Educación". Por tanto, si alguna de las dos formas es la establecida en el catálogo en línea, por ejemplo: "Educación de niños", no es necesario elaborar una referencia de la otra forma: "Niños-Educación" véase "Educación de niños". De la misma manera "Mujer" y "Mujeres" puede recuperarse con la palabra truncada a la derecha "Mujer\*". Sin embargo, en casos como "Química biológica" véase "Bioquímica", sí será necesaria la referencia, ya que ni con truncamiento a la derecha ni buscando por palabra podríamos encontrar el encabezamiento sin conocerlo de antemano.

Algunos sistemas pueden, durante el proceso de catalogación, cambiar automáticamente en los registros bibliográficos un término no autorizado, pero ligado a otro por un "véase" en un registro de autoridad, por el término correcto. Asimismo, el sistema puede llevar directamente al usuario que consulta el catálogo en línea hacia los registros bibliográficos correspondientes a un término no establecido, siempre que éste se encuentre ligado al término aceptado por una referencia de "véase". En estos casos, se discute sobre la conveniencia o no de que estos procesos sean "visibles" para el usuario. Tal vez el usuario de la base de datos no tiene por qué estar demasiado interesado en cómo recuperó el sistema la información que buscaba, aunque en los casos en que el usuario desconoce totalmente la relación entre los términos podría pensar que el sistema no está respondiendo adecuadamente.

En este contexto, puede decirse también que la elección del término "correcto" reviste menor importancia, siempre que estén consignados todos los sinónimos y establecida la relación entre ellos.

En lugar de archivos de autoridad algunos sistemas poseen únicamente "diccionarios" o índices de encabezamientos. Ello ahorra espacio en el sistema pero a su vez reviste serias limitantes, ya que no es posible generar referencias cruzadas a partir de los índices, pues éstos no admiten que se añada ni este ni otro tipo de información sobre ellos.

#### **Tiempo de respuesta**

Algunos sistemas almacenan el texto de los encabezamientos

---

<sup>9</sup>Véanse: Catherine M. Thomas, "Authority control in manual versus online catalogs: an examination of 'see' references", Information Technology and Libraries, v.3, December 1984, p. 393-398; y

M. R. Watson y Arlene G. Taylor, "Implications of current reference structure for authority work in online environments", Information Technology and Libraries, v.6, 1987, p.10-19

únicamente en los registros de autoridad, con el fin de ahorrar espacio. Ello puede representar la desventaja de que el tiempo requerido para el despliegue de los registros sea mayor, ya que el sistema tiene que recuperar los encabezamientos del archivo de autoridad previamente al despliegue de los registros bibliográficos.

### **Personal especializado**

El uso de un lenguaje de indización controlado requiere de personal especializado en análisis documental y con conocimientos básicos en el área científica de que se trate. Cuando se opta por automatizar los mecanismos de control de autoridad, generalmente también se requiere de personal especializado en cómputo, además de máquinas y sistemas adecuados capaces de realizar operaciones complejas para el manejo de la información bibliográfica y documental.

### **Tamaño de los archivos**

Una de las limitantes más serias puede ser el tamaño de los archivos de autoridad, que pueden llegar a ocupar un espacio similar al que ocupa el catálogo bibliográfico dentro del sistema.

### **Costos**

Finalmente, la creación y mantenimiento de catálogos de autoridad automatizados supone una inversión en tiempo, equipo y personal especializado que no todas las instituciones pueden costear.

### **EL PROYECTO DE CONTROL DE AUTORIDAD DE LA BDCV**

El proyecto de automatización de la Biblioteca Daniel Cosío Villegas se inicia en el año 1982. En 1990 este proyecto adquiere un nuevo impulso con la adquisición de un manejador de bases de datos denominado STAR, que se instaló en una computadora SUN.

En 1991 comienza a caminar el proyecto de control de autoridad, bajo la coordinación de la que suscribe, con el objetivo de dotar a la biblioteca de un sistema de control de autoridad completamente automatizado con el manejador STAR.

STAR es un manejador de bases de datos con prototipos desarrollados para bibliotecas que requiere de cierto nivel de programación.

STAR cuenta con los siguientes operadores de búsqueda:

#### **Booleanos**

- Intersección: AND
- Unión : OR
- Exclusión : NOT

## Otros

- Proximidad : NEAR (hasta cinco palabras antes y después)
  - : W/n (Words/número de palabras antes y después)
  - : W/P (palabras en el mismo párrafo)
  - : W/O (palabras en la misma ocurrencia)
  - : W/F (palabras en el mismo campo)
  - : W/S (palabras en el mismo subcampo o frase)
- Adyacencia : ADJ (palabras juntas en el orden establecido)
  - : W/O (palabras juntas sin importar el orden)
- Truncamiento: \* (ilimitado, en cualquier posición)
  - ? (limitado, en cualquier posición)
- Intervalo : ":" (permite la unión de varias búsquedas consecutivas)
- Prefijación : [ ] (datos presentes o ausentes, en cualquier posición)
- Nulo : "" (ausencia de datos en un campo o subcampo)
- Numéricos : = (igual a)
  - : < (mayor que)
  - : <= (mayor o igual)
  - : > (menor que)
  - : >= (menor o igual)

Todas las búsquedas se realizan en los índices, donde las palabras aparecen en mayúsculas, sin diacríticos ni acentos.

Con STAR pueden construirse índices por campos y subcampos así como por las palabras contenidas en ellos.

Además, STAR permite la definición de los llamados "pseudoíndices" para recuperar información presumiblemente contenida en varios campos y/o subcampos. Los principales pseudoíndices definidos en la base de datos bibliográficos de la BDCV, denominada MONOGRAFIAS, son: AUT (autores), TIT (títulos), CONF (conferencias), SER (series) y TEM (temas).

Los archivos de autoridad se denominan en STAR archivos de "lookup". Los archivos de autoridad sirven como referencia para validar o rechazar los encabezamientos y emitir mensajes de ayuda al catalogador en el momento de la captura. También sirven para orientar las búsquedas en el catálogo, dirigiéndolas de términos no aceptados a términos aceptados.

Los procesos de catalogación y clasificación se han venido desarrollando conforme a la 1a. edición de las Reglas de Catalogación Angloamericanas, la lista de encabezamientos de I. M. Kidder y la clasificación decimal.

La adición de nuevos temas a la lista establecida se hace conforme a criterios y etapas bien definidos dentro de las políticas del Departamento de Control Bibliográfico.

La biblioteca mantenía un sistema de referencias cruzadas en los catálogos de autores y materias mediante tarjetas de envíos, así como un catálogo manual de equivalencias de temas traducidos de las Listas de Encabezamientos de Materia de la Biblioteca del Congreso de los Estados Unidos. Todo este material está sirviendo para alimentar los catálogos de autoridad automatizados. Hasta el momento, se han capturado únicamente 1500 registros, la mayoría de materias pero también de algunos autores y esperamos alcanzar la cifra de 5000 registros al finalizar el año.

La creación y mantenimiento de los catálogos automatizados de autoridad de temas y autores se vio primero como una acción preventiva de errores para los procesos de catalogación, pero resulta claro que de ello se derivan también beneficios para los usuarios.

A continuación les voy a mostrar una serie de ejemplos reales para ilustrar las diferentes ventajas que supone el funcionamiento de estos archivos tanto para el proceso de asignación de temas como para la recuperación temática en nuestro caso particular.

Ejemplos para la demostración del funcionamiento de archivos de autoridad durante la catalogación y recuperación con STAR

Aspecto	Ejemplo	Escribe/Encuentra
Sinónimos	Planificación familiar	Natalidad, Limitación de la
Términos en inglés	Real property	Bienes raíces Subc: Compra, venta, etc. Impuestos Valoración
Subdiv.	Biografía	Subc: No subdivida geog. Use también como subd. No bajo autores
	Investigación agrícola	Agricultura Subc: Investigación
	Capuchinos	Subc: Para subd. geog. use Capuchinos en...
Refs. de v.t.	Medicina folklórica	V.t.: Botánica médica Medicina mágica, mística y espagírica
Refs. complejas de v.t.	Bebidas	V.t.: tipos de bebidas como -Bebidas alcohólicas- y nombres de bebidas como -Pulque-
Notas de alcance	Trabajadores agrícolas	Nota: Aquí se asientan las obras sobre los peones y asalariados del campo Para los pequeños propietarios de tierra use -Campesinos-
Homónimos	Papas	V.t.: los nombres de Papas bajo: --Iglesia Católica-- Nota: Para referirse a los tubérculos use: --Patatas--
Mayúsculas, acentos y diacríticos	educacion cívica	Educación cívica
	bilingüismo	Bilingüismo
Lugares geográf.	República de China	Taiwan

Lugares  
geográf.

Malawi

V.t.: Rodesia y Nyasalandia  
Nyasalandia  
(Protectorado)

Nota: Nombre oficial del  
antiguo protectorado  
británico de Nyasalandia  
a partir de 1964. Para  
obras sobre este  
territorio entre 1907  
y 1952 use -Nyasalandia  
(Protectorado)- Para  
obras sobre el período  
entre 1953 y 1963 use  
-Rodesia y Niasalandia-

Países nórdicos

V.t.: Suecia, Islandia,  
Finlandia, Dinamarca,  
Noruega

Seudónimos

Concolorcorvo

Carrió de la Vandera, Alonso  
Subc: 1715-1783  
El lazarrillo de ciegos  
caminantes

Lepot, Francois

Oliva, Enrique  
Subc: 1923-

Autores con  
dificultad  
para deter-  
minar el  
apellido

José Pascual Buxó

Pascual Buxó, José  
Subc: 1931-

José Agustín

Ramírez, José Agustín  
Subc: 1944-

Autores con  
nombres  
iguales o  
muy parecidos

Cardenal Richelieu

Richelieu, Armand Jean du  
Plessis  
Subc: cardenal, duque de,  
1582-1642  
Nota: No confundir con el  
mariscal Richelieu,  
asentado bajo:  
-Richelieu, Louis  
Francois Armand de  
Vignerot du Plessis,  
duque de, mariscal de  
Francia, 1696-1788-

Dumas, Alejandro

Dumas, Alexandre  
Subc: 1802-1870. Los tres  
mosqueteros  
V.t.: Dumas, Thomas Alexandre,  
1762-1806  
Dumas, Alexandre, hijo,  
1824-1895. La dama de

Nombres iguales o muy parecidos Unamuno, Miguel

Nombres difíciles de escribir y/o muy largos Chimalpahin

Siglas organismos IFLA

Autores corporativos en otros idiomas Partido Comunista Chino

las camelias

Subc: 1932-

Nota: Autor uruguayo.  
Diferente al filósofo español asentado bajo:  
-Unamuno y Jugo, Miguel de-

Chimalpahin Quauhtlehuanitzin,  
Domingo Francisco de San Antón Muñón

Subc: 1579-1629

International Federation for  
Library Associations and  
Institutions

Chung-Kuo Kung Ch'an Tang