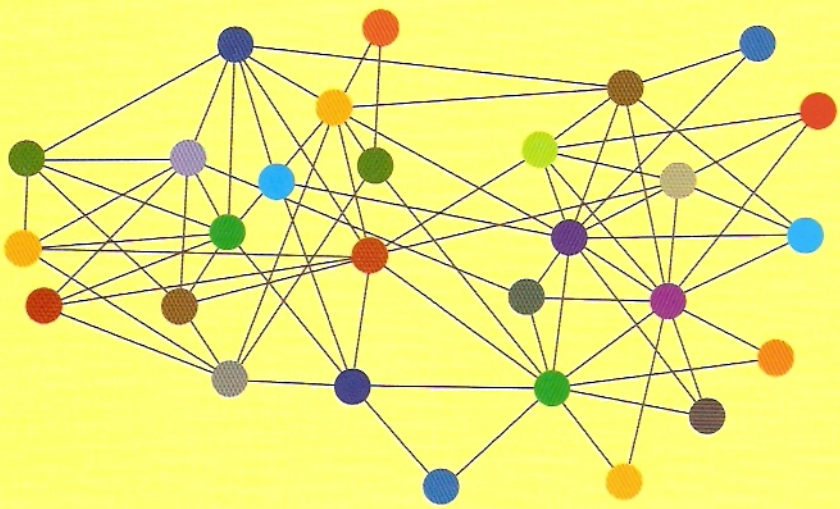
TREA

ISIDORO GIL LEIVA

Manual de indización

Teoría y práctica



MANUAL DE INDIZACIÓN
Teoría y práctica

Isidoro Gil Leiva

EDICIONES TREA, S. L

BIBLIOTECONOMíA Y ADMINISTRACIÓN CULTURAL – 193

Polígono Industrial de Somonte
María González, la Pondala, 98, nave D
33393 Somonte, Cenero. Gijón (Asturias)
Tel.: 985 303 801. Fax: 985 303 712
trea@trea.es
www.trea.es

Dirección editorial: Álvaro Díaz Huici
Coordinación editorial: Pablo García Guerrero
Producción: José Antonio Martín
Maquetación: María Álvarez Menéndez
Cubiertas: Impreso Estudio (Oviedo)
Impresión: Gráficas Apel, S. L. (Gijón)
Encuadernación: Encuadernaciones Cimadevilla, S. L. (Gijón)

Depósito legal: As. 2271-2008
ISBN: 978-84-9704-367-0

Impreso en España – *Printed in Spain*

PRÓLOGO

Con la publicación de este nuevo libro, el profesor Isidoro Gil Leiva amplía el ámbito del conocimiento sobre la indización iniciado en su anterior libro, *La automatización de la indización de documentos* (Trea, 1999), aportando, gracias a su experiencia académica, una visión didáctica y científica desde la génesis del proceso intelectual de la indización hasta la evaluación de sus resultados.

Mi contacto académico con el profesorGil Leiva se inició, justamente, gracias a su libro sobre la automatización de la indización, ya que lo he utilizado como texto de referencia para impartir clases de indización en biblioteconomía. Posteriormente, nuestra colaboración científica se concretó durante su estancia en noviembre del 2007 como investigador visitante en mi universidad, concretamente en el Departamento de Ciencia de la Información de la Facultad de Filosofía y Ciencia de la UNESP, campus de Marília, por medio de una ayuda otorgada por la FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) (Proc. 2006/61516-0).[1] Desde entonces, venimos compartiendo conocimientos de una forma continua y provechosa. En este sentido, hacer el prólogo de esta publicación es para mí una tarea de doble importancia. En el plano académico y de investigación, porque el área carece de libros con suficiente fundamento teórico y metodológico sobre la indización; y en el plano pedagógico, porque la enseñanza de la indización en las aulas necesita de libros cuyo contenido teórico y metodológico tenga claridad y consistencia para garantizar la formación del indizador.

Con esa finalidad, los dos primeros capítulos son conceptuales y dedicados a una contextualización de la indización en lo que respecta a la actividad intelectual. El primer capítulo del libro destaca por la importancia atribuida al proceso cognitivo que confiere a la indización la categoría de actividad intelectual compleja y, de esta manera, la torna singular. De esta forma, el libro se distingue de los demás en la medida en que aborda la indización como proceso cognitivo que busca la comprensión para alcanzar su ob-

[1] Informe científico aprobado por la Dirección científica de la FAPESP el 22/04/2008.

jetivo. En este contexto, este capítulo inicial pone el énfasis en el indizador que realiza un proceso cognitivo orientado a la indización y avanza desde la variable texto, en cuanto elemento de comunicación, hasta el lector, con su conocimiento previo necesario para el proceso de comprensión.

Es conveniente resaltar que los estudios cognitivos vienen ofreciendo importantes hallazgos respecto a la mente humana y sus capacidades, entre ellas la comprensión lectora. La concepción de la comprensión lectora se amplió, considerablemente, en las últimas décadas en lo que respecta a la participación del lector. La actitud del lector frente al texto, anteriormente vista como recepción pasiva de mensajes, pasó a considerar el procesamiento mental de información de la comprensión y evolucionó hacia una perspectiva de interacción entre el lector y el texto.

El conocimiento previo para la comprensión depende del conocimiento existente en la memoria a largo plazo, un repositorio de conocimientos con un tiempo y una capacidad de almacenamiento permanente e ilimitado y que posee una estructura de conocimiento basada en una red semántica de informaciones que conecta sus «nos» mediante asociaciones significativas entre conceptos, hechos, acciones, etcétera, allí representados. Para realizarse el proceso de comprensión, es necesario que la memoria a largo plazo tenga los llamados *esquemas* o representaciones generalizadas de ambientes, situaciones familiares e informaciones para que se hagan asociaciones con todo aquello que se está viendo, escuchando y leyendo. Para el indizador, el dominio de las tipologías documentales y de las estructuras textuales son dos tipos de conocimientos previos que podrán aumentar su comprensión durante el proceso de búsqueda de la temática textual para la identificación y selección de conceptos. Entonces, cuando hablamos de lectura para la indización, podemos decir que el indizador necesita comprender el texto para identificar y seleccionar conceptos, pues solamente lo realizará satisfactoriamente cuando hay comprensión. La lectura documental corresponde a la primera fase del abordaje del lector-indizador con el texto durante el análisis del asunto. La finalidad, en ese primer momento, es la identificación de conceptos que caracterizan el asunto tratado en el documento y, en un segundo momento, la selección de los conceptos, teniendo en cuenta el uso de esos conceptos.

El capítulo referente a la indización se abre con la conceptualización del proceso, donde se lleva a cabo una necesaria distinción entre la elaboración de índices y la indexación y el proceso para la representación del contenido documental mediante identificación y selección de conceptos. La construcción de índices es una práctica bastante antigua en el tratamiento de los documentos. Basta recordar que en las bibliotecas de la Antigüedad ya existían listas de documentos almacenados de ese modo. Entretanto, a partir del momento en que la ordenación de esas listas necesitó de una organización por asunto, se llevaron a cabo profundos cambios en el abordaje del proceso mecánico

de construir índices, es decir, se introdujo un proceso de análisis del contenido de los documentos con la finalidad de representación documental. En el resto del capítulo, además de la normalización de la indización y de la relación de esta con la recuperación, se abordadan, con una buena ejemplificación práctica, las cualidades de la indización —exhaustividad, especificidad, corrección y consistencia—, así como un interesante y oportuno epígrafe sobre la indización en Internet.

Los capítulos siguientes están organizados según una secuencia lógica en cuanto a la realización del proceso y su evaluación, lo que permite la comprensión natural tanto por aprendices de la indización como por profesionales, en la medida en que se dedican al uso de las herramientas en la indización, la práctica de la indización, la indización automática y la evaluación de la indización.

Teniendo en cuenta la importancia de los lenguajes de indización como herramientas de mediación de la comunicación del contenido del documento, el capítulo las identifica como lista de palabras clave, lista de descriptores, códigos de categoría temática, así como las más utilizadas: lista de encabezamientos de materia y tesauros. En el ítem dedicado a los tesauros, el libro ofrece un contenido dedicado a los *softwares* de gestión de tesauros para dominios de asunto que necesitan de controles de vocabularios más específicos, lo que se torna más conveniente para unidades de información, archivos, bibliotecas y centros de información, cada vez más especializados.

Esas herramientas de control del vocabulario, conocidas en la literatura como *lenguajes documentales* o *lenguajes de indización,* son un conjunto controlado de términos dotados de reglas sintácticas y semánticas cuyo objetivo es la representación de los conceptos significativos de asuntos de los documentos durante la indización, en la fase de traducción, y representación del asunto de interés del usuario durante la búsqueda.

La práctica de la indización es vista desde distintas perspectivas: desde el proceso realizado con el uso, tanto del lenguaje natural como de los vocabularios controlados (tesauro o listas de encabezamientos de materia); desde la tipología de documentos audiovisuales, sonoros, gráficos o textuales, y desde las políticas en grandes bases de datos documentales como en Agricola e INSPEC, entre otras. Cabe resaltar la importancia de abordar con claridad la existencia de esas políticas de indización en sistemas de información que producen bases de datos con el fin de legitimar la consistencia y la uniformidad en la actuación del indizador.

En el capítulo quinto, sobre la indización automática, se aborda la complejidad del proceso a partir del conocimiento teórico o metodológico de áreas que contribuyen a la creación interdisciplinar de un conjunto de herramientas. En el epígrafe dedicado a los prototipos para la indización automática, se presenta el Sistema de Indización Semiautomático (Sisa), un *software* diseñado por el autor que es objeto de análisis y evaluación. Durante la estancia en Brasil del profesor Gil Leiva, comentada anterior-

mente, llevó a cabo la presentación de los principales marcos teóricos de la indización automática, la evaluación mediante índices de consistencia de catálogos y bases de datos bibliográficas y, especialmente, del *software* Sisa, por el que los alumnos mostraron mucho interés, tanto acerca de su funcionamiento como de su evaluación. En un contexto de aprendizaje, el *software* Sisa es una herramienta que ofrece la posibilidad de la comprensión teórica y metodológica del proceso de indización con una doble ventaja: la identificación automática de términos y la selección manual compatible con un lenguaje documental para el control del vocabulario y de criterios cualitativos de indización.

El último capítulo aborda la evaluación de la indización en sus aspectos intrínseco y extrínseco. Esa distinción se refiere, por un lado, a la evaluación intrínseca, cualitativa o cuantitativa, como los resultados de la indización, los descriptores, encabezamientos o identificadores, y, por otro lado, a la evaluación extrínseca, cuando se usan los resultados de la indización en estudios comparados con diferentes catálogos o herramientas de recuperación de la información. De modo muy didáctico y, también, innovador, el autor expone las fórmulas de evaluación intrínseca y extrínseca acompañadas de ejemplos que esclarecen la aplicabilidad de sus resultados. Es absolutamente imprescindible la evaluación del proceso de indización por parte del indizador, aunque, en la práctica, no se priorizan en los sistemas de información. Entretanto, este *Manual de indización* ofrece la posibilidad de diversos esclarecimientos con relación a la práctica continua de los métodos de evaluación.

Como reflexión final, cabe señalar que el mérito de este obra es conciliar la teoría y la práctica de la indización, una tarea aparentemente simple cuando se piensa en la identificación de palabras clave de un texto, pero innovadora, porque entendemos que la actuación del indizador no está aislada, sino inmersa en una política de indización.

Mariângela Spotti Lopes Fujita
Departamento de Ciencias de la Información
de la Universidad Estadual Paulista (UNESP)
(Marília, São Paulo)

ÍNDICE

1. El proceso cognitivo y la indización 15
1.1. Organización de la comunicación 16
1.1.1. Discurso textual 16
1.1.1.1. Concepto de texto 16
1.1.1.2. Criterios de textualidad 18
1.1.1.3. Estructura del texto 19
1.1.1.4. Tipos de texto 22
1.2. Percepción sensorial de la información 28
1.3. Activación de la memoria 28
1.3.1. Memoria sensorial 30
1.3.2. Memoria a corto y memoria a largo plazo 30
1.4. Comprensión 32
1.4.1. Estrategias y procesos en la comprensión 32
1.4.2. Elementos para la comprensión 36
1.4.2.1. Cohesión discursiva 37
1.4.2.2. Coherencia discursiva 40
1.4.2.2.1. Tema oracional [42]. 1.4.2.2.2. Tema textual [47].

2. La indización 52
2.1. Concepto de indización 52
2.1.1. Índice e indexación *versus* indización 61
2.2. Cualidades de la indización 67
2.2.1. Exhaustividad 67
2.2.2. Especificidad 68
2.2.3. Corrección 69
2.2.4. Consistencia 69
2.3. Indizaciones de un documento 73
2.4. Zonas de extracción de conceptos y tiempo dedicado 79
2.5. Normas sobre indización 80
2.6. Relación entre indización y recuperación 81
2.7. Indización en Internet 90
2.8. Cronología de la indización 107

3. Herramientas para la indización ... 113
3.1. Lenguaje natural *versus* lenguaje controlado ... 113
3.2. Listas de palabras clave ... 115
3.3. Listas de descriptores ... 116
3.4. Códigos de categoría temática ... 119
3.5. Listas de encabezamientos de materia ... 122
3.4.1. Definición ... 122
3.4.2. Aportaciones para su configuración ... 123
3.4.3. Principios y reglas ... 129
3.4.4. Relaciones semánticas ... 141
3.6. Tesauros ... 146
3.6.1. Definición y uso ... 146
3.6.2. Composición ... 148
3.6.3. Normas y directrices ... 151
3.6.3.1. La norma ISO 2788-1986: Tesauros monolingües ... 153
3.6.4. Construcción de tesauros ... 187
3.6.4.1. Software de gestión de tesauros ... 202
3.6.5. Mantenimiento y actualización ... 208
3.6.6. Evaluación ... 213
3.6.6.1. Evaluación intrínseca ... 213
3.6.6.2. Evaluación extrínseca ... 215
3.6.7. Lenguajes de marcado para tesauros ... 217
3.6.7.1. Skos-Core ... 218
3.6.7.2. Zthes ... 220
3.6.8. Tesauros *versus* ontologías ... 224
3.7. Interoperabilidad entre vocabularios controlados ... 233

4. Práctica de la indización ... 245
4.1. Proceso de la indización ... 245
4.1.1. Indización con lenguaje natural ... 247
4.1.2. Indización con vocabulario controlado ... 251
4.1.2.1. Indización con tesauro ... 252
4.1.2.2. Indización con listas de encabezamientos de materia ... 259
4.2. Indización de documentos ... 261
4.2.1. Documentos audiovisuales ... 261
4.2.2. Documentos sonoros ... 269
4.2.3. Documentos gráficos ... 276
4.2.4. Documentos textuales ... 288
4.3. Políticas de indización ... 298
4.3.1. Bases de datos documentales ... 300
4.3.1.1. La indización en AGRÍCOLA ... 300
4.3.1.2. La indización en INSPEC ... 304
4.3.1.3. La indización en CURRENT CONTENTS ... 308
4.3.1.4. La indización en ERIC ... 309
4.3.1.5. La indización en MEDLINE ... 314

5. Indización automática 319
5.1. Concepto 319
5.2. Interdisciplinariedad en la indización automática 322
5.2.1. Lingüística 324
5.2.2. Terminología 325
5.2.3. Informática 326
5.2.4. Lingüística computacional 327
5.2.5. Estadística 328
5.3. Herramientas para la indización automática 329
5.3.1. Listas de palabras vacías 330
5.3.2. Ponderación de términos 333
5.3.2.1. Ley de Zipf 333
5.3.2.2. Frecuencia del término 334
5.3.2.3. *Inverse document frequency* 336
5.3.2.4. Valor de discriminación del término 337
5.3.3. Analizadores lingüísticos 338
5.3.3.1. Analizador morfológico 339
5.3.3.2. Analizador sintáctico 345
5.3.3.3. Analizador semántico 349
5.3.4. Algoritmos 361
5.3.5. Vocabularios controlados y ontologías 363
5.3.6. Reconocedores de nombres propios y siglas 364
5.3.7. Heurísticas 365
5.4. Prototipos para la indización automática 366
5.4.1. SISA 368

6. Evaluación de la indización 385
6.1. Evaluación intrínseca 385
6.1.1. Evaluación intrínseca cualitativa 385
6.1.2. Evaluación intrínseca cuantitativa 386
6.2. Evaluación extrínseca 388
6.2.1. Evaluación extrínseca mediante la interconsistencia 388
6.2.2. Evaluación extrínseca mediante la recuperación 392

Anexo 1: Recomendaciones para un buen posicionamiento web 401

Anexo 2: Lenguajes de encabezamientos de materia en bibliotecas nacionales 403

Anexo 3: Ejemplo de metadatos usando el esquema de tesauros RDF/XML 407

Bibliografía 411

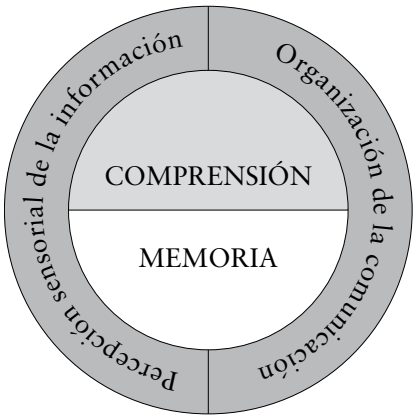
Capítulo 1

EL PROCESO COGNITIVO Y LA INDIZACIÓN

A las operaciones mentales llevadas a cabo por los seres racionales para la recepción selectiva de información, para su codificación simbólica y su almacenamiento y recuperación, se las denomina *proceso cognitivo*. La psicología cognitiva es la disciplina que estudia procesos cognitivos como la percepción sensorial de la información, el aprendizaje (lenguaje, lectura y escritura), la memoria o la capacidad de razonamiento.

Para producir palabras clave, términos de indización o los encabezamientos de materia para un documento, durante la indización, se desencadena una sucesión interactiva y simultánea de procesos mentales que tienen que ver precisamente con la percepción, la manera en la que se organizan la información, la memoria y la comprensión. Para explicar ello nos vamos a acercar a disciplinas como la lingüística textual, la psicología cognitiva o la comunicación de masas.

Si bien casi todas las actividades mentales del proceso cognitivo están interconectadas y son concurrentes durante la ejecución de la indización, aquí las presentamos de forma secuencial para conseguir una mayor claridad expositiva.



Proceso cognitivo en la indización