Distr.
RESTRINGIDA

LC/DEM/R.79
Serie B, N° 70
24 de enero de 1990

ORIGINAL: ESPAÑOL

C E L A D E
Centro Latinoamericano de Demografía

GUIA PARA LA RECUPERACION DE INFORMACION EN UNIDADES DE INFORMACION SOBRE POBLACION: CONCEPTOS BASICOS

Esta Guía fue preparada por la Organización Internacional para las Migraciones (OIM) y el Centro Latinoamericano de Demografía (CELADE), y forma parte del conjunto de materiales de instrucción preparados para la Red de Información sobre Población para América Latina y el Caribe, Red IPALCA.

PRESENTACION

El Centro Latinoamericano de Demografía (CELADE) y la Organización Internacional para las Migraciones (OIM), han sido, hasta 1989, las instituciones encargadas de mantener la base de datos bibliográficos sobre población que con carácter de regional (América Latina y El Caribe) se organizó en 1976.

Con la institucionalización de la Red de Información en Población, Red IPALCA, coordinada conjuntamente por el CELADE y el Programa Latinoamericano de Actividades en Población (PROLAP), y la existencia de nuevas tecnologías, se espera que sean las unidades de información existentes a los países, las que uniendo sus esfuerzos amplíen la cobertura de la base y sus servicios.

Para esos centros, los participantes en la Red IPALCA, el CELADE/DOCPAL y la OIM/CIMAL han preparado un conjunto de guías e instructivos, con el propósito que su aplicación acelere la difusión de información a los usuarios de la región.

La presente guía comprende los conceptos básicos y su aplicación práctica a todas las tareas relacionadas con la función de recuperación de información en un centro especializado en población y materias conexas.

II. DEFINICIONES BASICAS

Búsqueda o recuperación de información

La búsqueda o recuperación de información es el proceso que permite identificar -del total de documentos existentes en una unidad de información (biblioteca, centro de documentación, etc.)- sólo aquellos relevantes para responder a una determinada necesidad de información.

Este proceso implica una labor intelectual para identificar los términos por los cuales se efectuará la búsqueda y una manipulación de las herramientas de almacenamiento en las cuales se registró dicha información.

El contenido temático del documento ha sido registrado a través de la indización [2], en tanto que la identificación bibliográfica -que permitirá recuperar documentos relevantes a una determinada consulta (por autor, título, nombre de conferencia, etc.) - ha sido registrada mediante el proceso de descripción bibliográfica [3].

De lo anterior se desprende que una adecuada recuperación de información está supeditada a procesos previamente realizados, como son la descripción bibliográfica y del contenido temático de un documento (indización).

Como se sabe, la indización se realiza mediante la identificación de los conceptos que describen el contenido del documento, los que posteriormente se vierten a palabras o conjuntos de palabras (descriptores). A la inversa, al efectuar la recuperación de información, se analiza el contenido de la pregunta, asignándole la palabra o conjunto de palabras que mejor describe la consulta del usuario, las que permiten identificar los documentos que cumplen los requisitos por él estipulados. Es así como el mismo vocabulario utilizado para la indización del contenido temático del documento servirá para indizar la consulta, actuando este vocabulario como enlace entre el lenguaje de los documentos y el de la consulta.

Indización

La indización consiste en la descripción del contenido temático de un documento. Mediante un proceso de análisis efectuado a través de un examen detenido del documento, se identifican los conceptos que se vierten a palabras o conjuntos de palabras.

La indización puede efectuarse utilizando el lenguaje natural (palabras utilizadas por el autor, la interpretación que el indizador hace de estas palabras, etc.). Sin embargo, esta simplificación en el proceso de la indización puede traer complicaciones en la recuperación de información. De allí que para los procesos documentarios se prefiera la utilización de un lenguaje normalizado, el cual recibe el nombre genérico de vocabulario controlado. Por ejemplo, listas de encabezamientos de materias, listas de descriptores, tesauros, etc.

TABLA DE CONTENIDO

I. INTRODUCCION 1

II. DEFINICIONES BASICAS 2

III. FORMULACION DE ESTRATEGIAS DE BUSQUEDA DE INFORMACION 5

A. Estrategia de búsqueda de un aspecto 5
B. Estrategia de suma lógica 6
C. Estrategia de producto lógico 8
D. Estrategia de producto lógico y suma lógica 9
E. Estrategia de diferencia lógica 10

IV. CARACTERISTICAS DE LA RECUPERACION DE INFORMACION 11

V. VOCABULARIOS CONTROLADOS DISPONIBLES PARA LA RECUPERACION DE INFORMACION EN EL CAMPO DE LA POBLACION 12

I. INTRODUCCION

Todas las operaciones (selección, adquisición, indización de documentos) que se efectúan en una unidad de información (biblioteca, centro de documentación, etc.), están orientados a proporcionar a los usuarios potenciales todos aquellos documentos en la colección que, en el momento de la solicitud de información, cumplen con lo solicitado por el usuario y a excluir los documentos que no son de interés.

Por esta razón, el contenido temático de los documentos ingresados a la colección se registra mediante el proceso de indización, para posteriormente registrarla (almacenarla) en un determinado medio (tarjeta de catálogo, tarjeta Unitérmino, diskette, disco duro, etc.) para que el sistema, enfrentado a la necesidad de satisfacer los pedidos de información de los usuarios, pueda realizarlo mediante el proceso de búsqueda o recuperación de información.

Este manual recoge la experiencia del Sistema de Información sobre Población en América Latina del Centro Latinoamericano de Demografía (CELADE/DOCPAL) y el Centro de Información sobre Migraciones en América Latina de la Organización Internacional para las Migraciones (OIM/CIMAL) en la recuperación de información a partir de los documentos ingresados a la colección. Tiene como propósito el servir de guía a las unidades de información que han adoptado técnicas manuales o automatizadas para el tratamiento de la documentación sobre población. Incluye definiciones de la terminología propia de la recuperación de información, para presentar, a continuación, una sección sobre formulación de estrategias de búsqueda y características de la recuperación de información. El manual contiene, asimismo, una reseña de los vocabularios controlados disponibles en el campo de la población, con especial enfásis en el Tesauro Multilingüe de Población [1/].

Tesauro

Un tesauro puede definirse de acuerdo a su función o de acuerdo a su estructura.

Desde el punto de vista de su función, un tesauro es una herramiento de control terminológico utilizada para convertir a un lenguaje más restringido (lenguaje documentario, lenguaje de información) el lenguaje natural presente en los documentos y utilizados por los indizadores o analistas de información y los usuarios.

Desde el punto de vista de su estructura, es un vocabulario controlado y dinámico de términos que guardan relación entre sí. Las relaciones pueden ser de tipo semántico, es decir, referentes al significado de las palabras, o bien relaciones de tipo jerárquico que expresan superioridad o subordinación [4/].

A diferencia de un diccionario que entrega definiciones de palabras o de términos, un tesauro entrega las palabras o descriptores, expresando relaciones entre ellas y sin incluir, obligatoriamente, explicaciones o definiciones.

Descriptor

Es un término o símbolo autorizado y normalizado que figura en un tesauro y que se usa para representar sin ambigüedad los conceptos contenidos en los documentos y en los pedidos de recuperación de información.

Un descriptor puede comprender una o varias palabras, de preferencia una sola.

Las relaciones entre términos (descriptores cuando se trata de un tesauro) son básicamente tres y en todos los casos de caracter recíproco, es decir, la relación opera en ambos sentidos. Estas relaciones son las de equivalencia, jerárquicas y de asociación.

La relación de equivalencia controla la sinonimia o cuasi-sinonimia. Se denominan sinónimos aquellos términos que tienen el mismo sentido en una disciplina específica, en tanto que se entiende que son cuasi-sinónimos los términos cuyo significado puede ser diferente en otro vocabulario especializado, pero que se consideran como sinónimos para las necesidades de un determinado sistema de información.

La relación se expresa por los términos USE (Utílicese) y a la inversa, UF ("Used For": usado por).

Ejemplo:

DISTRIBUCION ESPACIAL
USE: DISTRIBUCION GEOGRAFICA

DISTRIBUCION GEOGRAFICA/GEOGRAPHIC
DISTRIBUTION/REPARITTION GEOGRAPHIQUE
[04.01.03]
UF: DISTRIBUCION ESPACIAL

La relación jerárquica controla las relaciones de superioridad y subordinación entre descriptores, utilizándose:

BT ("Broader Term") en español, "término genérico o dominante" para indicar el descriptor que representa un concepto del cual forma parte otro descriptor, y

NT ("Narrower Term") en español, "término específico o subordinado", para indicar términos subordinados.

Ejemplo:

DISTRIBUCION GEOGRAFICA/GEOGRAPHIC
DISTRIBUTION/REPARITTION GEOGRAPHIQUE
[04.01.03]

BT: DISTRIBUCION DE LA POBLACION [04.01.03]

Como estas relaciones son de caracter recíproco, bajo DISTRIBUCION DE LA POBLACION se encuentra:

DISTRIBUCION DE LA POBLACION/POPULATION DISTRIBUTION/
REPARITTION DE LA POPULATION
[04.01.03]

NT: DISTRIBUCION GEOGRAFICA [04.01.03]

La relación de asociación se utiliza para indicar relaciones entre conceptos que presentan afinidad temática. Se representan por medio del término RT ("Related Term"), en español, "término relacionado o afín".

Ejemplo:

DISTRIBUCION GEOGRAFICA/GEOGRAPHIC
DISTRIBUTION/REPARITTION GEOGRAPHIQUE
[04.01.03]

RT: ESPACIO [04.01.02]
GEOGRAFIA [04.01.01]
GEOGRAFIA DE LA POBLACION [04.01.01]
HABITAT [04.01.03]
LOCALIZACION [04.01.02]
MOVILIDAD GEOGRAFICA [15.02.01]

III. FORMULACION DE ESTRATEGIAS DE BUSQUEDA DE INFORMACION

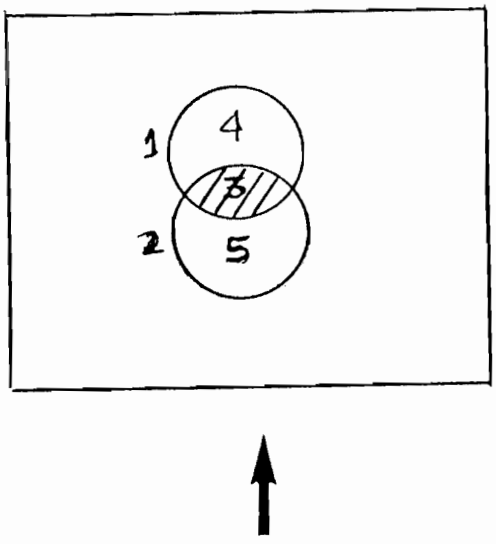
Para efectuar búsquedas en una unidad de información, es necesario formular una estrategia de búsqueda que permita -ante una determinada consulta- recuperar todos los documentos pertinentes ingresados al sistema y excluir los no pertinentes.

Si se graficara lo expresado tendríamos:

1. Círculo que representa los documentos pertenecientes a la consulta.

2. Círculo que representa todos los documentos recuperados

3. Area que representa los documentos de interés recuperados en la búsqueda.



4. Area que representa los documentos de interés en el Sistema que no fueron recuperados en la búsqueda.

5. Area que representa los documentos recuperados que no eran de interés.

Todos los documentos
en la base de datos

Así, el propósito de la estrategia de búsqueda es minimizar el número de documentos de interés no recuperados en la búsqueda (área 4) y el número de documentos recuperados que no son de interés (área 5) (ver sección IV).

La elección de la estrategia de búsqueda es muy importante para una correcta recuperación de información. El método más comúnmente usado para formularlas se basa en los principios del álgebra booleana.

Aplicando la lógica de Boole al campo de la información, se pueden presentar los casos que se detallan a continuación.

A. ESTRATEGIA DE BUSQUEDA DE UN ASPECTO

Este tipo de búsqueda, el más sencillo, se utiliza cuando interesa recuperar del sistema de información los documentos que tienen un aspecto en común.

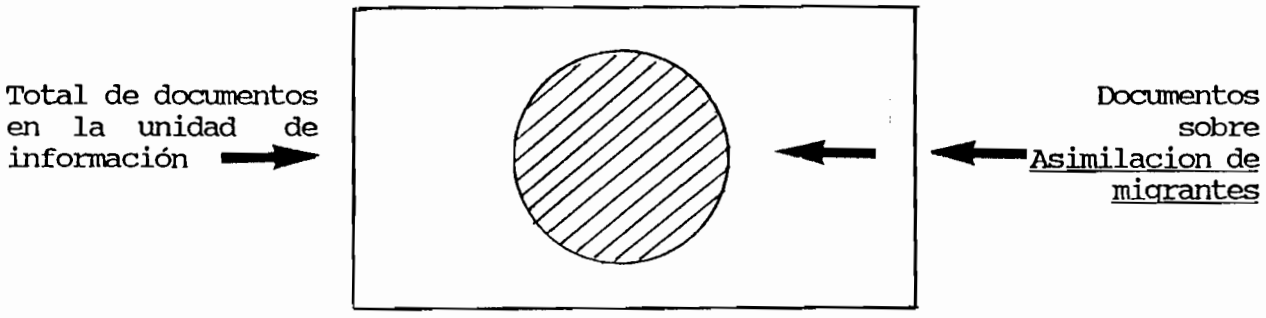
Ejemplo: Un usuario puede solicitar todos los documentos en el sistema cuyo autor haya sido Juan Ckackiel; en este caso, debería buscarse por el nombre del autor. Lo mismo podría ocurrir con una solicitud sobre un determinado tema: todos los documentos que existen en el sistema sobre Asimilación de migrantes.

Este tipo de búsqueda es el que normalmente se efectúa en sistemas tradicionales. En el caso de los ejemplos ya mencionados, bastaría ir al catálogo por autor o por temas y allí, en estricto orden alfabético, se encontrarían todos los trabajos que se tienen de ese autor o sobre ese tema.

Esta estrategia de búsqueda se expresaría como:

[ASIMILACION DE MIGRANTES] [5]

Forma gráfica de representación



B. ESTRATEGIA DE SUMA LOGICA ("OR" lógico, inclusivo))

Este tipo de estrategia se utiliza cuando se indican dos o más aspectos de interés en una búsqueda. Por ejemplo, dados dos conceptos, A y B, se puede pedir los documentos que tienen:

[A] o [B] [6]

Esta búsqueda incluirá documentos que tienen sólo A, B solamente o ambos conceptos (es decir, "o inclusivo").

A mayor número de aspectos incluídos, normalmente mayor número de documentos recuperados.

Ejemplo 1: Esta estrategia podría usarse para recuperar todos los documentos sobre un tema dado (es decir, todos los documentos de un término genérico con sus correspondientes términos específicos).

Por ejemplo, interesa el concepto Migrantes, pero se busca también bajo los términos más específicos: Emigrantes, Inmigrantes, Personas desplazadas, Refugiados.

Esta estrategia de búsqueda se expresaría como:

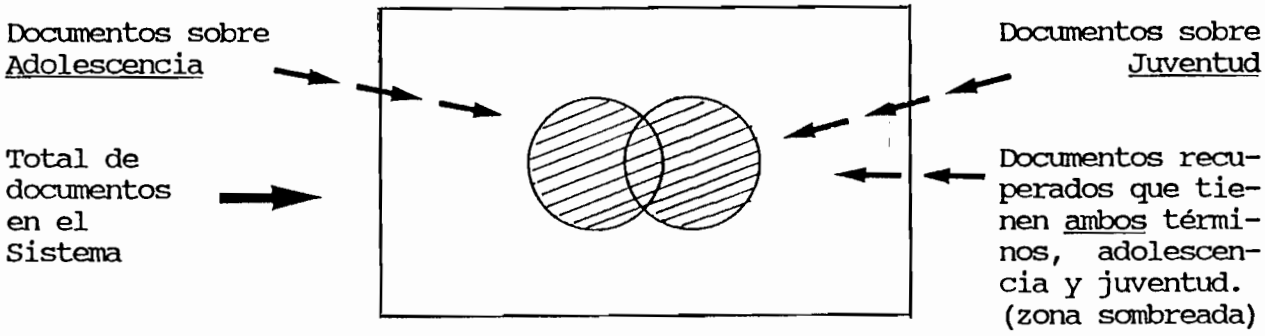
[MIGRANTES] o [EMIGRANTES] o [INMIGRANTES]
o [PERSONAS DESPLAZADAS] o [REFUGIADOS]

Ejemplo 2: A la inversa, se podría estar buscando información sobre un concepto específico, pero porque hay pocos documentos que traten este concepto se pide también un término más genérico, con el propósito de encontrar más documentos. Por ejemplo, un usuario podría necesitar documentos que traten la Adolescencia, pero debido a que se sabe que hay pocos documentos sobre el tema incluye también el término Juventud.

Esta estrategia de búsqueda se expresaría como:

[ADOLESCENCIA] o [JUVENTUD]

Forma gráfica de representación



Todo documento que tenga uno o ambos términos será recuperado.

En un sistema tradicional, se debería ir al catálogo y buscar bajo ambos términos, Adolescencia y Juventud.

La búsqueda se responderá con los documentos que se encuentren bajo cualquiera de estos términos.

C. ESTRATEGIA DE PRODUCTO LOGICO ("AND" lógico)

Este tipo de estrategia se utiliza cuando dos o más condiciones deben estar presentes en cada documento para satisfacer las necesidades del usuario.

Por ejemplo, dados los conceptos A y B, se pueden pedir los documentos que tienen:

[A] y [B]

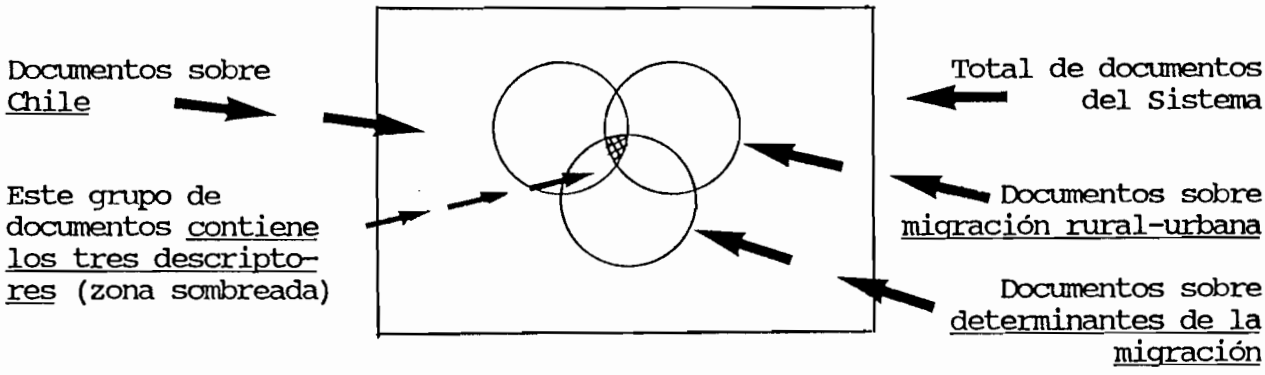
A mayor número de condiciones establecidas, normalmente menor número de documentos recuperados.

Ejemplo: Un usuario puede necesitar todos los documentos ingresados a la colección sobre "Determinantes de la migración rural-urbana en Chile". En este caso, para recuperar un documento es necesario que los términos Migración rural-urbana y Determinantes de la migración y Chile, se encuentren presentes en la indización para que ese documento sea seleccionado.

Esta estrategia de búsqueda se expresaría como:

[MIGRACION RURAL-URBANA] y [DETERMINANTES DE LA MIGRACION]
y [CHILE]

Forma gráfica de representación



En un sistema tradicional, se debería buscar en el catálogo bajo los términos Migración rural-urbana, Determinantes de la migración y Chile.

Los documentos que tienen estos tres términos (descriptores) constituyen la respuesta a la consulta.

D. ESTRATEGIA CON PRODUCTO LOGICO ("AND") Y SUMA LOGICA ("OR")

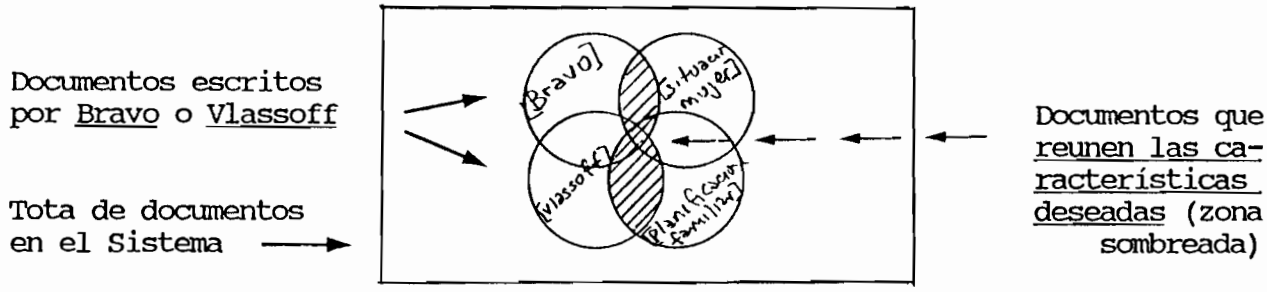
Se utiliza esta estrategia para recuperar documentos que tienen uno o más grupos de términos en común.

Ejemplo: Un usuario desea documentos escritos por Bravo o Vlassoff sobre "situación de la mujer" o "planificación familiar".

Esta estrategia de búsqueda se expresaría como:

([BRAVO] o [VLASSOFF]) y
([SITUACION DE LA MUJER] o [PLANIFICACION FAMILIAR])

Forma gráfica de representación



En un sistema tradicional, debe buscarse en el catálogo bajo los autores Bravo y Vlassoff. Si no se encuentran documentos bajo cualquiera de estos autores, la búsqueda se da por terminada. Si se encuentran documentos por uno o ambos autores, debe buscarse por Situación de la mujer y planificación familiar. Si no se encuentran bajo estos términos, la búsqueda se da por terminada.

La respuesta a la consulta la constituyen los documentos que se encuentren bajo cualquiera de estos términos y cuyos autores sean Bravo o Vlassoff.

E. ESTRATEGIA DE DIFERENCIA LOGICA ("BUT NOT")

En este caso, para que un documento sea recuperado, no debe tener presente uno o más descriptores que se especifican. Es decir, recuperar todos los documentos con A que no tienen B, lo que se puede escribir:

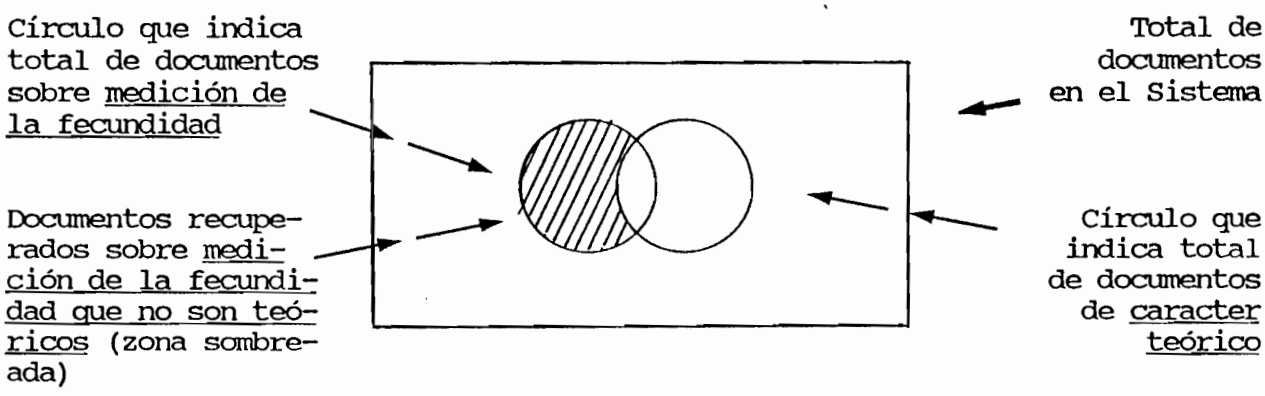
[A] y no [B]

Ejemplo: Se necesitan los documentos sobre "medición de la fecundidad" con excepción de aquellos que son puramente teóricos, sin datos. En este caso se recuperan todos los documentos sobre medición de la fecundidad aplicados a cualquier país de América Latina, pero se eliminan los que no tienen datos empíricos (es decir, aquellos con código de país igual a "ZZ", donde ZZ = documentos teóricos, sin datos empíricos)

Esta estrategia de búsqueda se expresaría como:

[MEDICION DE LA FECUNDIDAD] y no [METODOLOGIA Y/O TEORIA]

Forma gráfica de representación



En un sistema tradicional, debería buscarse en el catálogo bajo Medición de la fecundidad. Si no se encuentran documentos bajo este término, la búsqueda se da por terminada. Si se encuentran documentos bajo este término (descriptor) se debería buscar luego bajo Metodología y/o teoría para descartar los documentos que tienen ambos términos.

La consulta se responderá con todos los documentos que aparezcan bajo el término Medición de la fecundidad pero que no aparezcan bajo Metodología y/o teoría.

IV. CARACTERISTICAS DE LA RECUPERACION DE INFORMACION

Según se señalaba anteriormente, la recuperación de información es una de las tareas básicas que debe realizar una unidad de información. Esto significa que, en respuesta a una solicitud de información específica, el sistema debe estar capacitado para seleccionar, del total de documentos ingresados a la colección, sólo aquellos pertinentes a la consulta.

Factores de importancia -previos a la recuperación de información- para lograr este objetivo son la formulación de una correcta estrategia de búsqueda (véase sección III), la exhaustividad en la indización [7], y la especificidad del lenguaje de indización que permitan describir adecuadamente los conceptos tratados en el documento.

La eficiencia de un sistema de información para cumplir con esta tarea se mide básicamente en términos de la exhaustividad y precisión del resultado de la búsqueda.

Exhaustividad

La exhautividad se entiende como la relación del número de documentos relevantes recuperados en relación con el número total de documentos relevantes existentes en la colección.

Precisión

La precisión se refiere a la relación del número de documentos relevantes recuperados en relación con el número total de documentos recuperados (tanto relevantes como no relevantes). Así, la precisión es una medición de la capacidad del sistema para eliminar los documentos no-relevantes.

Existe una relación inversa entre los factores de exhaustividad y precisión: a mayor precisión, menos exhaustividad. Por consiguiente, a mayor número de condiciones en la formulación de la consulta, en busca de una mayor precisión, hay menos exhaustividad porque hay más posibilidades de pérdida de documentos relevantes.

Para asegurar que no se pierda ningún documento relevante, se podrían entregar todos los documentos del sistema. En este caso la exhaustividad sería perfecta (100% de los documentos relevantes están entregados). Pero esta solución no sería muy útil porque el usuario tendría que revisar todos los documentos para separar los relevantes de los no-relevantes. Es decir, la precisión de la respuesta al pedido sería muy baja.

V. VOCABULARIOS CONTROLADOS DISPONIBLES PARA LA RECUPERACION DE INFORMACION EN EL CAMPO DE LA POBLACION

La literatura mundial hace permanente hincapié en que la calidad de la indización del contenido de los documentos, y por ende, la recuperación de información depende en mucho del vocabulario utilizado para realizar estos procesos. En CELADE/DOCPAL y OIM/CIMAL se utiliza el Tesauro Multilingüe sobre Población, para el tratamiento de documentos de caracter demográfico o estudios de población. Este vocabulario es producto de una real y efectiva cooperación internacional en la que participaron el Comité Internacional de Cooperación de Investigaciones Nacionales sobre Demografía (CICRED), el Centro Latinoamericano de Demografía (CELADE), el Fondo de las Naciones Unidas en materia de Población (FNUAP), en el marco de la acción de la Red Informativa sobre Población (POPIN).

Para 1990 se prevee la publicación en español del International Thesaurus of Refugee Terminology [8]. Este vocabulario será un valioso aporte y será utilizado para complementar la terminología sobre refugiados incluída en el Tesauro Multilingüe sobre Población. La preparación de este tesauro, publicado bajo los auspicios del la Red Internacional de Documentación sobre Refugiados, ha estado a cargo de un Grupo de Trabajo internacional y fue coordinado por el Centro de Documentación sobre Refugiados del Alto Comisionado de las Naciones Unidas para los Refugiados (ACNUR/CDR). OIM/CIMAL participó en este Grupo de Trabajo y estuvo a cargo de la versión en español.

Dado que los estudios poblacionales se ubican en un contexto más amplio del desarrollo económico y social, cuya terminología no siempre está presente en estos Tesauros, se utiliza también el Macrothesaurus de la OECD [9].

Tesauro Multilingüe de Población

El Tesauro Multilingüe sobre Población se encuentra disponible en 5 versiones (española, francesa, inglesa, portuguesa y árabe). Todas incluyen la misma terminología y se ciñen al mismo formato de impresión. Esta característica permite que los países de la región se comuniquen entre ellos y posibilita el intercambio de información de la región con el resto del mundo.

En este tesauro -que consta de tres partes que se interrelacionan- cada descriptor aparece identificado por un número de 6 dígitos que representa el lugar (faceta) en que se inserta el término en los campos de interés (campos semánticos).

1. Tesauro alfabético

En esta sección los descriptores aparecen de acuerdo con sus equivalentes linguísticos, con su número del tema al que se refieren, con su eventual nota de alcance y con sus diferentes relaciones en el conjunto del vocabulario. Los no-descriptores remiten explicitamente a los descriptores a ser utilizados.

2. Presentación temática de los descriptores

En esta segunda sección, los términos se presentan ordenados según la faceta a la que pertenecen, según sus afinidades semánticas.

3. Indice permutado

En este índice los descriptores aparecen ordenados alfabéticamente de acuerdo a las palabras significativas que lo forman, apareciendo tantas veces como palabras tenga el descriptor. En este índice se incluyen los sinónimos e indica, para cada término incluído, su número de faceta o sub-faceta.

NOTAS

1. Tesauro de POPIN. Tesauro Multilingüe sobre Población. Segunda edición. POPIN, CICRED, FNUAP, 1985.

2. Guía para completar la Hoja de Análisis de Contenido HAC en una Unidad de Información sobre Población.

3. Véase Sistema de Información Bibliográfica: Uso de Hojas de Trabajo (HDB y HAC y Tarjeta de Registro Bibliográfico). Santiago. CEPAL, 1984. 169 p. (Manual de procedimiento - Sistema de Información Bibliográfica de la CEPAL No. 1). E/CEPAL/G.1224.

4. UNESCO. UNISIST. Guidelines for the establishment and development of monolingual thesauri. Paris, UNESCO, 1973. (SC/WS/JJJ).

5. El paréntesis cuadrado indica todos los documentos que tratan el concepto "Asimilación de migrantes".

6. El paréntesis cuadrado indica todos los documentos que poseen un determinado concepto. En este caso, A señala todos los documentos que tienen el concepto A.

7. op. cit.

8. International Thesaurus of Refugee Terminology, prepared by Jean Aitchison. Dordrecht, Martinus Nijhoff Publishers, 1989.

9. Macrothesaurus para el tratamiento de la información relativa al desarrollo económico y social. Nueva edición española. Paris, OECD, 1979.