

INFOBILA

BIBLIOTECA



CENTRO UNIVERSITARIO
DE INVESTIGACIONES
BIBLIOTECOLÓGICAS

MARIA GRACIELA CAÑETE

CONSTRUCCION DE TESAUROS

RESUMEN DE LA UNIDAD III DEL CURSO

SOBRE LENGUAJES DE INDIZACION

(UNESCO-CAICYT)

TRABAJO PARA SER PRESENTADO A LA

ESCUELA DE BIBLIOTECARIOS DE LA UNIVERSIDAD

NACIONAL DE CORDOBA - 1982

Previamente a la consideración de los lineamientos a seguir para la construcción de un tesauro es necesario especificar: Qué es un TESAURO.

Según UNESCO, 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 herramienta de control terminológico utilizada para traducir a un lenguaje más restringido (lenguaje documentado, lenguaje de información) el lenguaje natural usado en los documentos y utilizados por los indizadores y los usuarios.

Desde el punto de vista de su estructura, es un vocabulario controlado y dinámico de términos que guardan entre sí relaciones semánticas y genéricas, que se aplica a un campo específico del conocimiento.

Los Tesauros son lenguajes de indización que se utilizan para sistemas post-coordinados de recuperación de la información, ésto significa sistemas que permiten combinar los conceptos en el momento de la búsqueda mediante operaciones lógicas de union, intersección o complementación.

A través de un tesauro se pueden expresar las múltiples relaciones implícitas entre los conceptos de un documento.

El vocablo, tesauro, se utilizó antiguamente como nombre de ciertos diccionarios, dados por sus autores.

Antecedentes de tesauros,,se pueden encontrar en dos propuestas presentadas en Estados Unidos, a fines de la década del 40 y principio de los años 50; propuestas para recuperar información.

- 1) El uso de vocabularios muy reducidos formados por términos de significado muy amplio.
- 2) La indización por medio de palabras únicas tomadas del texto analizado, sin ningún control del vocabulario (sistema unitérmino)

Debido a los problemas producidos por falsas coordinaciones de las palabras

únicas se llegó a sustituirlas por conceptos formados por una o varias palabras. De "Unit~~er~~terminos" se pasó a "Uniconceptos". Como resultado de esta evolución surgieron los "Tesauros", que son vocabularios totalmente controlados.

La idea principal antes de construir un tesauro, es analizar el sistema de información al que servirá y considerar las decisiones alternativas que se presentan.

Se debe considerar:

- a) alcance
- b) tipos de documentos
- c) volumen de la literatura
- d) usuarios
- e) tipos de consulta
- f) nivel de la indización
- g) implementación física

Es necesario averiguar si existen tesauros en el campo temático de interés. Si existiera uno, aceptable como marco general del que se quiere construir, se lo puede adaptar ampliando o restringiendo algunas clases.

Por lo tanto se puede: Adoptar, Adaptar, Traducir, o hacer un microtesauro compatible.

Existen dos formas básicas para construir un tesauro: Aposteriori de la indización (método empírico) y Apriori de la misma. Es conveniente usar una combinación de ambos métodos, aunque, Lancaster recomienda comenzar por el empírico. Un tesauro debe ser capaz de describir conceptos que se encuentran en la literatura sobre ese campo y también ser capaz de interrogar sobre dichos conceptos al sistema, por ello se procede a la recolección de términos. Es fundamental extraer términos de la literatura primaria. Incluir uno nuevo se justifica si es utilizado por los usuarios.

El registro de los términos que se van recolectando deben incluir toda la información que abarcará el tesauro.

Los términos seleccionados se van registrando en una ficha o en un formulario. Es importante indicar la fuente de donde se tomó: diccionarios, otros tesauros, etc. La ficha puede hacerse por duplicado para permitir el ordenamiento sistemático de un juego.

Los conceptos deben sistematizarse, por facetas y por jerarquías.

El análisis por facetas, aunque se hace en forma empírica, debe tener como marco de referencia ciertas categorías fundamentales que se determinan previamente de acuerdo con el conocimiento del campo temático que se está cubriendo. Se entiende por categoría: clase de alta generalidad que agrupa conceptos más específicos, o una faceta general que se aplica a todo el campo temático. Las subfacetas se organizan sobre la base de un solo principio de división.

La organización de los términos registrados empieza por su agrupamiento en pilas, de acuerdo con el reconocimiento de analogías y diferencias. Cada pila formada en forma empírica representará una faceta diferente. Faceta es un aspecto o punto de vista desde el cual puede ser tratado un concepto, clase, categoría o campo temático.

Aunque Lancaster en la organización de los términos recomienda hacer esta tarea en forma empírica, se debe tener como marco de referencia ciertas categorías fundamentales como ya se mencionó anteriormente.

Una vez que los términos están agrupados en facetas se empieza a reconocer las subfacetas -separación de conceptos, etc. dentro de una faceta por medio de una sola característica- y a establecer relaciones jerárquicas, es decir relaciones semánticas entre dos conceptos donde uno se encuentra subordinado al otro. En un tesoro se indican con siglas: TG (término genérico) y TE (término específico). Cada subfaceta se determina sobre la base de un solo principio de división. Las diferentes facetas pueden tener una longitud variable. No es necesario que todas las facetas tengan la misma cantidad de términos. Depende de las necesidades de los usuarios y de las características de la colección. Una faceta puede consistir

en un solo término.

En el proceso de revisión de los términos que es el paso siguiente se decidirá cuál es el término más apropiado para cada concepto y se revisarán exhaustivamente los términos recopilados, antes de tomar una decisión.

Para lograr la consistencia del tesoro se deben revisar cuidadosamente los términos registrados para verificar si se ha aplicado siempre el mismo criterio.

Es necesario tener en cuenta que la especificidad de un tesoro depende de la habilidad de sus términos para expresar la materia con la profundidad y con el detalle que requieren los usuarios a los cuales se dirige.

El nivel de precoordinación de un lenguaje de recuperación depende de que los conceptos se expresen por una sola palabra o por varias. Un sistema en el que predominan los utérminos tiene un nivel bajo, uno que usa conceptos expresados por dos o tres palabras tiene un nivel medio y uno que tiene una gran proporción de conceptos expresados por dos, tres o más palabras tiene un alto nivel de precordinación.

El problema de los términos compuestos está relacionado estrechamente con la especificidad del vocabulario.

Cuanto más complejos son los términos de indización, más específico y más preciso es el vocabulario y a su vez más extenso el número total de términos.

Uno de los items principales en el tesoro es la nota de alcance, que debe precisar el significado de un término dentro del sistema al que el tesoro servirá.

El método empírico recomendado para la conexión del mismo parte de la recolección de términos justificados por la literatura y por el usuario; en una segunda etapa, agrupa dichos términos en clases -conjunto de conceptos con una o varias características comunes- de acuerdo con sus afinidades semánticas; luego, los organiza en jerarquías y reconoce relaciones entre esas jerarquías, finalmente, vuelca todo el esquema armado en una organización alfabética.

Para establecer las relaciones jerárquicas debe trabajarse con las jerarquías inmediatamente superiores e inferiores.

La relación jerárquica entre los términos, es decir TG-TE, es una relación género-especie o cosa-classes de.

La relación todo-parte no es relación género-especie; aunque las directrices de la UNESCO la permiten, no es recomendable excepto para relaciones geográficas y partes del cuerpo.

La mayor dificultad de las relaciones TG-TE reside en que muchos términos pueden pertenecer legítimamente a varias jerarquías y, por lo tanto, hay que elegir una de ellas.

Los TR están vinculados por una relación semántica que no es jerárquica ni de sinonimia.

Las referencias USE dirigen de un término que no puede ser usado en la indización y búsqueda a un término que puede ser usado.

La referencia USE puede remitir desde sinónimos, abreviaturas, casi sinónimos y otras palabras alternativas; en esos casos los dos términos (referido y referido a) son esencialmente equivalentes.

La referencia USE es empleada también para remitir desde términos específicos no usados en la indización o búsqueda a los términos genéricos más apropiados.

Las notas de alcance pueden ser instrucciones para usar el descriptor en la indización y recuperación o definiciones tipo diccionario (especialmente para términos que no son familiares y cuyo significado puede ser ambiguo).

Antes de presentar el tesauro es recomendable hacer pruebas prácticas mediante la indización de un número suficiente de documentos.

Hay que probar el tesauro con preguntas que puedan señalar defectos y mostrar la necesidad de introducir nuevos descriptores.

Un vocabulario controlado para la indización y la búsqueda no puede permanecer estático: debe crecer.

La extensión del vocabulario depende no sólo del campo temático sino también de la especificidad de los términos y del tipo de términos usados.

El crecimiento continuado del vocabulario depende del crecimiento del sistema al

que sirve.

La actualización del tesoro puede implicar la cancelación de un término. Las estadísticas del uso de un término en la indexación son indicadores muy valiosos en las operaciones de actualización.

La adición de nuevos términos, aunque sea sobre una base provisional, debe ser permanente; la cancelación de términos exige el análisis de estadísticas.

Para considerar terminada la estructura de un tesoro es aconsejable indexar 1000 documentos por lo menos. A partir de entonces seguirá creciendo mediante la continua indexación, pero se mantendrá la misma estructura.

Las pruebas mostrarán si:

- a) la estructura propuesta toma en cuenta las necesidades más recientes.
- b) Los descriptores elegidos son útiles para la indexación y la búsqueda.
- c) Las relaciones son correctas y suficientes.
- d) es capaz de responder a las preguntas que se formulan.

Por supuesto, un vocabulario desarrollado a través de las operaciones de indexación crecerá muy rápidamente al principio, pero alcanzará una cierta estabilidad a partir de la que aumentará muy poco.

Una vez que el vocabulario llega a un cierto nivel puede producirse un crecimiento gradual para aumentar la especificidad y adecuarse a los nuevos tópicos, pero no debe esperarse ningún crecimiento brusco (a menos que el Centro se proponga abarcar áreas completamente nuevas). El crecimiento continuado del vocabulario provendrá del mismo sistema. Se fundamentará en la justificación literaria. Los términos nuevos surgirán de las operaciones de indexación y búsqueda.

En algunos sistemas es costumbre introducir un nuevo descriptor por un período de ensayo, observar su uso y retenerlo o descartarlo después de un período de 19 meses por ejemplo (Experiencia en el MEDLARS).