«Thesaurus»: la seva importància en l'emmagatzematge i recuperació de la informació

En tot sistema de recuperació de la informació, existeixen les següents etapes:
1) Anàlisi i indexació dels documents pertinents al sistema.
2) Emmagatzematge de les representacions d'aquests documents sobre un suport adient, obtingudes per l'anàlisi anterior.
3) Recuperació de la informació corresponent a una qüestió donada.
Prescindint de la complexitat i de la grandària del sistema, les operacions bàsiques que cal realitzar són les mateixes i totes elles depenen de la primera, això és de la indexació dels documents, consistent en essència en l'assignació a cadascun dels documents d'aquells mots-clau, descriptors o termes d'indexació que permetin expressar de la forma més completa possible el seu contingut.

El problema central de tot sistema documental, consisteix justament en l'assignació adient dels termes d'indexació i l'èxit o el fracàs del sistema dependrà de la idoneïtat d'aquesta assignació.

Per a solucionar aquests tipus de problemes s'han dissenyat els *thesaurus*, que constitueixen una eina de treball relativament recent, quan es consideren per a recuperació de la informació, ja que el nom *thesaurus* fou encunyat per en C. Roget, el 1852, en publicar per primera vegada l'obra ara coneguda per «Roget's Thesaurus», que es troba en la seva 90ª edició, un llibre clàssic dins la literatura anglesa i que consta de dues parts principals: la primera una llista de mots ordenats alfabèticament dins dels següents apartats principals: abstraccions, espai, matèria, intel·ligència, volició i afecció, i la segona formada per una relació alfabètica de tots els mots amb indicació dels apartats o sub-apartats on han aparegut dins de la primera part. Un equivalent, aproximat, en castellà, del Roget seria el «Diccionario ideológico» d'en J. Casares.

Els *thesaurus*, en el sentit considerat en aquest treball, neixen durant la dècada dels cinquanta, quan les necessitats d'informació i documentació fan convenient la disposició d'una eina flexible que permeti, més que el tractament de grans quantitats de documents, el tractament adequat d'un nou tipus de documents que no fan més que reflectir les necessitats canviants de les comandes de la societat nova; aquelles que no són pas ara de tipus monodisciplinari, sinó justament al contrari, típicament interdisciplinari. La societat d'avui s'afronta a reptes tals com la recerca de noves fonts d'energia, la lluita

contra la contaminació, la lluita contra el càncer o bé la detectació i prevenció dels nous riscs professionals, que no es poden resoldre mitjançant el coneixement d'una sola disciplina, sinó ben al contrari, que necessiten l'aportament de moltes ciències i tècniques per aconseguir prendre una decisió adequada.

Aquells mètodes d'arxiu de documents o de les seves representacions, basats en esquemes classificatoris dissenyats fonamentalment per a col·locar els documents en prestatges, del tipus de la CDD (Classificació Decimal Dewey), base de la CDU, la classificació de la Biblioteca del Congrés dels EE.UU., etc., no són suficients, encara que s'hagin fet successives adaptacions per a resoldre els problemes creats pel tractament de la informació multidisciplinària pròpia del món d'avui.

Ja en els anys 50 s'inicià l'anàlisi dels documents mitjançant l'assignació de termes d'indexació independents de qualsevol esquema classificatori previ, com pot ésser l'exemple dels Uniterm d'en Motimer Taube, antecessors directes dels mots-clau i descriptors emprats ara. També en aquells moments comencen a usar-se les classificacions facetades, fonamentades en les idees teòriques d'en Ranganathan i aplicades per en Vickery.

Si bé tots els mètodes moderns d'indexació s'aboquen a l'assignació de mots-clau, per a poder obtenir una bona recuperació final, sembla convenient d'emprar un vocabulari controlat i, encara més, per aconseguir un rendiment òptim en la recuperació de la informació emmagatzemada, és necessari que existeixi una interrelació dels termes d'indexació, bé mitjançant un sistema classificatori per facetes, bé emprant un mitjà més flexible de relacions terminològiques, tant conceptuals com sintàctiques, allò que realment constitueix un thesaurus-facetat o simplement un thesaurus.

Abans de donar una definició de thesaurus, potser val la pena que en donem de diferents:
1) Definició de la UNESCO: Vocabulari controlat i dinàmic de termes relacionats semànticament i genèrica, que cobreix de forma exhaustiva un camp específic del coneixement. Aquest vocabulari és una col·lecció sistemàtica i alfabètica de descriptors i no descriptors, així i com d'indicadors de les seves interrelacions.
2) Definició d'en Moisés Asín (Cubà): Dispositiu de control terminològic per al tractament i la recerca d'informació que s'empra en la traducció dels

conceptes, expressats en llenguatge natural en els documents, pels indexadors o els usuaris, a un llenguatge artificial que manca d'ambigüitat. Estructuralment és un diccionari d'unitats lèxiques, basat en el llenguatge natural on aquestes unitats lèxiques tenen relacions jeràrquiques i semàntiques entre elles i cobreixen un camp específic del coneixement.

3) Definició d'A.I. Mijáilov i R.S. Guiliarevskii: Diccionari de referència destinat a ajudar l'usuari de la informació a definir les seves necessitats d'informació en els termes del llenguatge de descriptors i facilitar una indexació en detall dels documents i comandes d'informació, mitjançant aquests termes. Ha de contenir tots els descriptors emprats pel llenguatge d'un sistema donat i posant en evidència les seves relacions conceptuals, així i com els mots-clau dins del sistema que es consideren com a sinònims d'aquests descriptors.

4) Definicions d'en Soergel, que en considera dues, segons es consideri en forma restringida o bé àmplia: En el primer cas el defineix com un sistema estructurat de conceptes amb indicació de les relacions jeràrquiques i associatives entre ells, de manera que per a cada concepte tots els termes que el desitgin (sinònims) es troben junts en el control terminològic, i que les relacions conceptuals (paradigmàtiques) són de 2 tipus, relacions entre conceptes (p. ex., relacions jeràrquiques) o bé relacions entre termes sinònims.

En el segon cas, i dins del camp de l'emmagatzament i recuperació de la informació, és una llista de termes i/o altres signes (o símbols) amb indicació de les interrelacions d'aquests elements segons els següents criteris, la llista conté una proporció significativa de termes no preferents i preferents no emprats com a descriptors i amb un intent de control terminològic.

5) Definició de la norma ISO 2788 (1974): Segons la seva funció, és un sistema de control de la terminologia emprada en la translació a un llenguatge més estricte (Llenguatge documental), el llenguatge natural emprat en els documents i pels catalogadors i usuaris. Segons la seva estructura, és un vocabulari controlat i dinàmic de termes que es troben interrelacionats semànticament i genèrica i que s'aplica a un camp donat del coneixement.

Del conjunt d'aquestes definicions se'n pot donar una que les resumeixi i que crec que és prou adient:

Un *thesaurus* és una eina emprada per al tractament de la informació continguda en els documents (emmagatzemament i recuperació), que

conté una relació alfabètica de mots de frases o de símbols que representen els conceptes significatius d'aquests documents, amb indicació expressa de les interrelacions que existeixen entre aquells, tant conceptuals (paradigmàtiques) com sintàctiques (sintagmàtiques).

Relacions entre els termes d'un *thesaurus*

Les principals relacions existents són dels següents tipus: d'equivalència (sinònims i quasi-sinònims), jeràrquiques i associatives.
Aquestes relacions s'indiquen dins el *thesaurus* mitjançant la terminologia següent:
Empreu (E), traducció del terme anglès «use» (U)
Empreu per (EP), traducció de l'anglès, «use for» (UF)
Terme ampli (TA), traducció de l'anglès «broader» (BT)
Terme específic (TE), traducció de l'anglès «narrower term» (NT)
Terme relacionat (TR), traducció de l'anglès «related term (RT)
Les expressions empreu i empreu per, E i EP, s'apliquen fonamentalment, per a no dir exclusivament a les relacions d'equivalència o sigui entre sinònims i quasi-sinònims.

En canvi, les expressions de terme ampli i terme específic són sempre de tipus jeràrquic, corresponent les associatives als termes relacionats.

Dins de les relacions jeràrquiques podem distingir les d'inclusió de classe, o sigui classe/nombre d'una classe, o classe/subclasse (jerarquia en el sentit lògic). En aquest cas, sempre es pot dir que els conceptes més específics tenen totes les característiques del concepte més ampli i, a més, una característica pròpia com a mínim.

Exemples de classe/membre d'una classe poden ésser:

Neumoconiosi	Silicosi
	Asbestosi
	Estannosi
Alcohols	Etanol
	Metanol
	Alcohol amílic
Adhesius	Cola de peix
	Adhesius sintètics

Exemples de classe/subclasse:

Felí	Gat
	Tigre
	Pantera
Amines	Amines alifàtiques
	Amines aromàtiques

En el cas de la inclusió tòpica, els termes inclosos ho són perquè es troben dins de la temàtica del terme



principal, per exemple:

Llibre	Fulls, capítol
Avió	Ales, motor
Cèl·lula	Protoplasma, nucli, cromosomes

Les altres connexions entre els termes d'un *thesaurus* són les que coneixem per termes relacionats (TR), que són fonamentalment de dos tipus: causa-efecte i associatives.

Com a exemple de causa efecte podem indicar: soroll-nens (encara que a vegades els grans fan més soroll que els nens), pols-neumoconiosi, TV-avorriment, diode-corba característica.

Dins de les relacions associatives podem distingir-ne, principalment, de dos tipus, semblança de significat i contigüitat contextual:

Com a exemples de semblança de significat podem indicar els següents:
Adaptació-Encaix-Acoblament
Sorció-Adsorció-Absorció
Dermatosi-Dermiti-Dermatopatia

La contigüitat contextual es pot produir segons:
Definició i en són exemples, instruments de mesura-calibració-error
Coneixement empíric, p.ex., alcohol, dissolvent, anticongelant
Freqüència de combinació, com en poden ésser exemple, dret-legislació, administració-govern, malalties, metges, salut.

Mots homògrafs i polisèmics

Quan es disenya un thesaurus té molta importància el fet de definir acuradament el significat de cadascun dels termes acceptats, en general nomenats descriptors, el que fa que en certs casos sigui necessari l'acotar el camp semàntic del terme emprat mitjançant una nota explicativa (en anglès una «scope note»). És per això que no és innecessari dir quelcom respecte als homògrafs i mots polisèmics.

Es designen per mots homògrafs aquells que designen conceptes diferents mitjançant la mateixa paraula i tenen diferències semàntiques, mentre es denominen molt polisèmics aquells que si bé designen amb la mateixa paraula conceptes diferents, aquests comparteixen característiques semàntiques. Moltes vegades no hi ha una diferència clara entre homògrafs i polisèmics. Podem acceptar com a exemple d'homògrafs o homònims (en el cas de la pronúncia i no de l'escriptura) els següents:

Gat (Felí), Gat (Aparell per aixecar pesos)
Cargues (Electricitat), Cargues (Explosives), Cargues (Pes), Cargues (Finances)
Són exemples de mots polisèmics:
Integració (Matemàtiques), Integració (Economia)
Càlcul (Medicina), Càlcul (Aritmètica)
Camp (Esport), Camp (Agricultura), Camp (Temàtic), Camp (Informàtica).

Diferències entre les interrelacions jeràrquiques en un thesaurus de les tradicionals

La diferència essencial consisteix en el fet que en un thesaurus s'empra la polijerarquia, la qual quan ha estat dissenyada adequadament no és res més que un dispositiu per tal d'ajudar a la indexació dels documents i a la realització de recerques. Les interrelacions jeràrquiques són pragmàtiques, no conceptuals i es troben orientades vers la recerca i la recuperació de documents.

Segons aquestes idees el concepte A és més ampli que el concepte B, quan en qualsevol recerca que inclogui A, tots els ítems de B es troben. Inversament B és més específic que A. Com exemple, transport per terra és més ampli que transport per ferrocarril i en una recerca referent a transport terrestre, tots els documents que facin referència a transport per ferrocarril s'hauran de trobar.

La classificació jeràrquica tradicional, en canvi, consistia a dividir un conjunt de conceptes en grups mútuament exclusius i cadascún d'ells era també exclusiu.

La nova concepció jeràrquica és completament diferent, per a cada parell de conceptes A i B, s'analitzen per a veure si compleixen les condicions dites abans sobre si un és o no un concepte més ampli o bé específic. Si compleixen la condició esmentada s'estableix una interrelació jeràrquica i si no fos així s'estableix aquesta relació.

Descomposició en factors semàntics

En tot thesaurus resulta molt important la descomposició en factors semàntics, però, cal tenir en compte que aquesta descomposició no es refereix a factors lingüístics, sinó als significats. Així, una expressió relativament complexa pot indicar un concepte elemental i, al contrari, un sol mot pot representar un concepte

complex. Per exemple, metall de les terres rares és un concepte simple i vaixell (vehicle per a transport marítim) o neumoconiosi (malaltia produïda per la pols en els pulmons) són conceptes complexos.

Sempre que es vulgui fer una descomposició semàntica d'un concepte compost és convenient seguir un camí pragmàtic i aplicar el sentit comú. Com a guia purament indicativa, val la pena de tenir en compte les següents qüestions:
1) De quina classe és el concepte, un membre d'aquesta o una subclasse?
2) De què és part?
3) De què és fet?
4) Quines són les seves propietats distintives?
5) De què és capaç?
6) És o no és acompanyat de quelcom més?
7) És un estat, condició o circumstància específica?
8) És un medi ambient específic?
9) Determina, causa, influencia, produeix o actua sobre quelcom més? És determinat o causat per altres coses?
10) És una teoria de quelcom o un aspecte de considerar quelcom?
11) Té un propòsit específic, és un mitjà o instrument per aconseguir quelcom més? És una meta o fi per a ell mateix?

Quan s'empra la divisió semàntica cal tenir en compte com a regla d'or la de no descompondre un terme de moltes paraules si els mots corresponents considerats per ells mateixos tenen un significat diferent del seu significat en la combinació. Per exemple, Casa Blanca o Casa Rosada, seus respectives dels governs dels EUA i Argentina, no tenen res a veure, excepte el color de les seves parets amb altres cases blanques o rosades. Un altre exemple ambigu seria Documentació-Computadores, ja que no se sap si es tracta de documentació sobre computadores o de tractament de la documentació mitjançant computadores.

Com a criteris per a decidir quins conceptes es poden emprar com a descriptors precombinats indiquem els següents:
a) Freqüència de la indexació i/o recerca: Emprar un concepte precombinat quan s'usa sovint en la indexació i/o recerca.
b) Freqüència dels components: Emprar un concepte compost quan s'usa sovint en la indexació.
c) Ajuntar conceptes específics en un de més ampli.
Per altra banda, és convenient tenir en compte quins conceptes compostos es poden representar millor per una combinació de descriptors senzills, pels

quals, en general, es poden aplicar les següents regles:
1) Objecte-matèria del que és fet: un descriptor per l'objecte i un altre pel material, p. ex.: Tubs-acer; caixes-fusta.
2) Compostos químics: Sodi-Acetat; calci-sulfat.
És molt important saber emprar conceptes de tipus general per a delimitar perfectament el significat d'altres, p. ex.:
Pressió sang-Mesura; Raigs X-Llindar; Contaminant-Anàlisi.

Com a exemples de conceptes de tipus general, es poden citar entre altres: capacitat, govern, administració, vulnerabilitat, control, anàlisi, mesura.

Parts d'un thesaurus i thesaurus adjunts

En tot thesaurus es poden distingir les següents parts:
1) Introducció: En aquesta part es dóna el camp temàtic a què correspon i totes les indicacions precises per a la seva utilització, així com les fonts i les relacions amb altres thesaurus.
2) Índex de grans apartats classificat.
3) El thesaurus de termes per ordre alfabètic i amb totes les interrelacions de E, EP, SIN, NE, TA, TE i TR.
4) Una relació alfabètica de tots els mots o bé d'altres expressions contingudes en el thesaurus amb la referència corresponent de la part principal o thesaurus de termes on es troba cadascun dels mots.

A vegades, per a simplificar un thesaurus determinat o bé per aconseguir una relació amb d'altres thesaurus o classificacions ja existents i amb certa afinitat amb el camp temàtic d'un thesaurus, es poden emprar com a complements aquests altres thesaurus o classificacions ja preexistents, que es coneixen com a thesaurus adjunts.

Per exemple en el camp dels riscs professionals i la seva prevenció es pot considerar com a thesaurus adjunt la Classificació Internacional de Malalties Professionals de l'OMS.

Domènec Turuguet i Mayol

Bibliografia general sobre thesaurus

1. Aitchison, J. i Gilchrist, A. *Thesaurus construction. A practical manual.* London, ASLIB, 1972.
2. American National Standard Institut (ANSI) *Project de l'American National Standard guidlines for Thesaurus Structure, Constructions and Use.*

ANSI 239.9, 1973.
3. DIN. *Principis directrius per l'establiment i el desenvolupament de tesauros monolingües.* DIN 1463, 1972.
4. Gilchrist, A. *The thesaurus in retrieval.* London, ASLIB, 1971.
5. International Standard Organisation. *Documentació. Principis bàsics per la redacció i desenvolupament de thesaurus monolingües.* ISO 2788, 1974.
6. Lancaster, F.W. *Vocabulary control for information retrieval.* London, ASLIB, 1971.
7. Soergel, D. *Indexing Languages and Thesauri: Construction and Maintenance.* Los Angeles, Melville Publishing Co., 1974.
8. UNESCO. *Principles directeurs pour l'etablissement et le développement de thesaurus scientifiques et techniques monolingues destinés à la recherche documentaire.* Paris, UNESCO, 1971.

Thesaurus dipositats a la Biblioteca del C.N.I.D.

1. *Thesaurus of structured terms in occupational safety and health.* Rockville, NIOSH, 1971.
2. *UNESCO Thesaurus.* Compilat per Aitchison, J. Old Working, Union Brothers, 1977, 2 vols. ISBN 42-3-101469-2.
3. O.I.T. *Thesauro O.I.T.* (Tesauro de los descriptores empleados para el tratamiento de la información en la biblioteca del O.I.T. Ginebra, O.I.T., 1976. ISBN 92-2-001264-2.
4. Centre International d'Informations de Securité et d'Hygiène du Travail (CIS) *Thesaurus CIS. Sécurité et Higiène du Travail.* Gèneve, CIS, O.I.T., 1976, ISBN 92-9096-201-5.
5. International Atomic Energy Agency. INIS: Thesaurus. IAEA-INIS-13 (Rev. 6). Vienna, I.A.E.A., 1974.