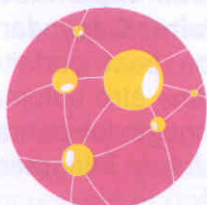


# - DES NOUVELLES d' ASCODOCPY -

JUIN 2003

numéro spécial



**ascodocpsy**

## Un outil d'aide à l'analyse documentaire : le thesaurus

Nous profitons du vaste chantier que conduit ascodocpsy en 2003 pour essayer de vous rendre familier cet étrange mot qu'est "thesaurus" : à quel moment du traitement de l'information les documentalistes l'utilisent-ils ? Quel rapport avec l'indexation ? Pourquoi utiliser un langage documentaire ? Et enfin, cœur de notre propos, qu'est-ce qu'un thesaurus ?

### La fonction traitement

Qu'il s'agisse d'un organisme privé ou public, le système documentaire peut être schématisé sous l' image souvent utilisée de la chaîne documentaire : fonction entrée (collecte des documents), fonction traitement (indexation et catalogage) et fonction sortie (produits documentaires et diffusion). Le traitement de l'information documentaire a deux volets : l'analyse documentaire et la recherche documentaire.

La fonction traitement est l'ensemble des opérations effectuées pour la transformation ou mise en forme, la mise en mémoire et la restitution selon les besoins, des informations contenues dans les documents collectés. C'est au long de la chaîne documentaire que l'on va dépouiller les revues, sélectionner les articles, les analyser et les répertorier dans des bases documentaires.

Pour faire coïncider analyse documentaire et recherche documentaire, les documentalistes ont recours à des outils documentaires : les langages documentaires.

Cette opération s'effectue aussi bien sur les ouvrages que sur tout autre type de document. La fonction traitement n'a d'intérêt qu'en fonction de l'opération suivante, à savoir la recherche documentaire.

L'indexation est une méthode d'analyse documentaire parmi d'autres. C'est l'opération centrale de tout système documentaire.

**Selon l'Afnor, l'indexation est un procédé destiné à représenter par les éléments d'un langage documentaire ou naturel des données résultant de l'analyse du contenu d'un document ou d'une question.**

L'indexation se décompose en trois étapes :

- \* la prise de connaissance du contenu
- \* le choix des concepts à représenter
- \* la traduction des concepts en descripteurs à l'aide du langage documentaire

## Les langages documentaires

Les langages documentaires ont un double usage

- ils servent principalement à l'**indexation** lors de la description de contenu,
- et encore à la **recherche** de l'information traitée.

Selon l'Afnor, un langage documentaire est un langage artificiel constitué de représentations de notions et de relations entre ces notions, et destiné, dans un système documentaire, à formaliser les données contenues dans les documents et dans les demandes des utilisateurs.

Les langages documentaires se répartissent en deux grandes catégories :

- les langages à structure hiérarchique : les classifications
- les langages analytiques ou langages à structure combinatoire : les thesaurus.

Historiquement, les classifications sont les premiers types de langage à apparaître. Encyclopédiques comme la Dewey ou la Classification décimale universelle (CDU), ou très spécialisées pour correspondre à un fonds bien délimité, elles fonctionnent toutes sur le même principe : **Aller du générique au spécifique.**

Une sous-classe est entièrement englobée dans la classe qui la précède, et englobe celle qui la suit. Logique semblable pour les répertoires (ou annuaires) de type "Yahoo" du web ! Il s'agit essentiellement ici d'un langage **pré-coordonné** où les combinaisons possibles entre les divers éléments sémantiques servant à décrire les concepts sont déterminés à l'avance.

Ainsi, les descripteurs sont souvent des mots composés qui recouvrent la totalité d'une notion.

Exemple : culture irriguée des céréales, pont routier métallique à haubans

## Le thesaurus

Le thesaurus est apparu plus tardivement, quand on a, à la fois, perçu les limites des classifications et les possibilités offertes par l'ordinateur pour le traitement du langage, l'informatique documentaire facilitant les manipulations combinatoires du langage. **On peut donc considérer que la finalité dernière d'un thesaurus est son exploitation par un logiciel documentaire.** Cependant, un centre de documentation peut parfaitement utiliser un thesaurus sans être automatisé. L'adoption d'un tel langage lui permet d'être compatible avec une future intégration dans un système automatisé.

Selon l'Afnor, un thesaurus est un langage documentaire fondé sur une structuration hiérarchisée d'un ou plusieurs domaines de la connaissance et dans lequel les notions sont représentées par des termes d'une ou plusieurs langues naturelles et les relations entre notions par des signes conventionnels.

### 1- Principes de base

C'est l'outil le mieux adapté à l'indexation par ses trois principes de base.

#### 1.1 un concept = un descripteur, et un seul !

L'organisation du langage s'effectue au niveau de chaque concept élémentaire. Les notions sont réduites à leurs éléments constitutifs les plus simples. (On ne cherche plus à mettre dans un tiroir, mais à décomposer en autant d'éléments signifiants pour la recherche). On recherche en fait une "traduction" en langage artificiel dépourvu d'ambiguïté car les deux principaux troubles sont la polysémie et la synonymie. Il est alors impératif de choisir (de façon la plus justifiée possible) un terme accepté et utilisé systématiquement : il sera LE descripteur ; les autres mots devenant des synonymes.

Les concepts synonymes doivent être représentés par **un seul et unique terme**.

Les unités lexicales (ou entrées) d'un thésaurus se composent donc de :

- **descripteurs**,
- de **mots outils** : descripteurs sans signification précise quand ils sont employés seuls (ex. : *méthode, comparaison, activité...*)
- et de **termes équivalents** (ou synonymes).

Les descripteurs sont les termes autorisés, à l'exception de tout autre, pour l'indexation des documents et l'interrogation du corpus.

## 1.2 un langage univoque et post-coordonné

A l'inverse des classifications, tout concept complexe doit être décomposé en expressions simples que l'on coordonne au moment de l'interrogation : c'est là que l'aspect combinatoire intervient.

Ainsi, pour un document ou une recherche sur la commercialisation des produits issus de l'agriculture biologique, on croisera les descripteurs :

- PRODUIT AGRICOLE
- DISTRIBUTION PAR PRODUIT
- AGRICULTURE BIOLOGIQUE

C'est l'usage de cette post-coordination que l'on observe avec les opérateurs booléens (ET / OU / SAUF).

### Un thésaurus n'est pas :

- un dictionnaire : recueil de mots rangés par ordre alphabétique et suivis de leurs définition
- un lexique : dictionnaire spécialisé regroupant les termes utilisés dans une science
- un glossaire : dictionnaire qui donne l'explication de mots anciens, spéciaux ou mal connus
- un index : table alphabétique accompagnée de références

## 1.3 des relations entre les termes

Le thesaurus induit plusieurs types de relations qui visent à faciliter l'utilisation de cet outil linguistique.

**Les relations d'équivalence ou de synonymie** renvoient les divers synonymes d'un concept vers le descripteur choisi, ce qui permet de limiter le nombre de descripteurs.

Elles sont indiquées par les expressions Employer (EM) et Employé pour (EP).

Exemple :

Derme EM Peau ;  
Peau EP Derme.

### Sa présentation

Il comprend trois entrées principales :

- la liste alphabétique des descripteurs, les synonymes renvoyant aux descripteurs et les différentes notes
- la présentation des champs sémantiques (où l'arborescence des descripteurs est représentée soit par décalage vers la droite, soit de façon graphique (schéma fléché (cf.p.5), arbre, terminogramme).
- l'index permuté des termes qui permet de regrouper les termes des noms composés qui se trouvent dispersés dans la liste alphabétique.

Une introduction présente les objectifs poursuivis, les domaines couverts, les nombres de termes et les normes adoptées. L'organisation et la présentation sont précisées ainsi que le mode d'utilisation et les sources employées.

Les thésaurus existent sous format papier, et de plus en plus sous formats électroniques.

Deux exemples sont accessibles en ligne :

Le MESH (Medical subject Headings), thesaurus de Medline traduit en français par l'Inserm : <http://dicdoc.kb.inserm.fr:2010/>

La banque de données en Santé Publique, élaboré par l'ENSP : <http://www.bdsp.tm.fr/TSP3/Default.asp>

Les relations d'équivalence sont de différents types :

- quasi synonymie : Panthère et Jaguar
- évolutions de concepts : Calculateur électronique et Ordinateur
- noms de marque : Réfrigérateur et Frigidaire
- appellations courantes : Hifi et Haute Fidélité
- origines linguistiques différentes : Géomagnétisme et magnétisme terrestre

**Les relations hiérarchiques (ou de filiation ou encore génériques)** qui expriment les rapports de subordination entre les termes et constituent l'ossature du thesaurus. La position des termes dans une arborescence est représentée, pour chaque terme, par le lien avec son Terme générique (TG) et ses Termes spécifiques (TS).

Exemples :

A partir d'un thesaurus concernant l'informatique ordinateur

- . unité centrale
- . périphérique
  - . . écran (TS de périphérique)
  - . . imprimante (TS de périphérique)
  - . . lecteur de CD Rom (TS de périphérique)
- . programmation

A partir d'un thesaurus concernant l'automobile automobile (TG)

- . catégories (TS1)
  - . . berline (TS2)
  - . . break (TS2)
  - . . coupé (TS2)
- . parties (TS1)
  - . . carrosserie (TS2)
  - . . moteur (TS2)
  - . . portière (TS2)

On nomme champ sémantique les regroupements effectués autour d'un concept. **Un descripteur ne doit être rattaché qu'à un seul champ sémantique.**

**Les relations d'association ou de voisinage** indiquent des proximités de sens entre des termes situés dans des hiérarchies différentes. Ce sont des passerelles entre les différents champs sémantiques du domaine.

Elles sont exprimées par les mentions **VA** (voir aussi) ou **TA** (Terme associé).

Exemple :

Physiologie végétale TA Botanique  
Maladie de carence VA Malnutrition

Enfin, ne pas oublier la **relation de définition**, introduite par la mention NA (Note d'usage ou Note d'application) qui définit l'emploi sémantique d'un terme.

Exemple : ALTO

NA définit une tessiture de violon

## 2- Sa construction

Trois méthodes peuvent être utilisées pour collecter le vocabulaire :

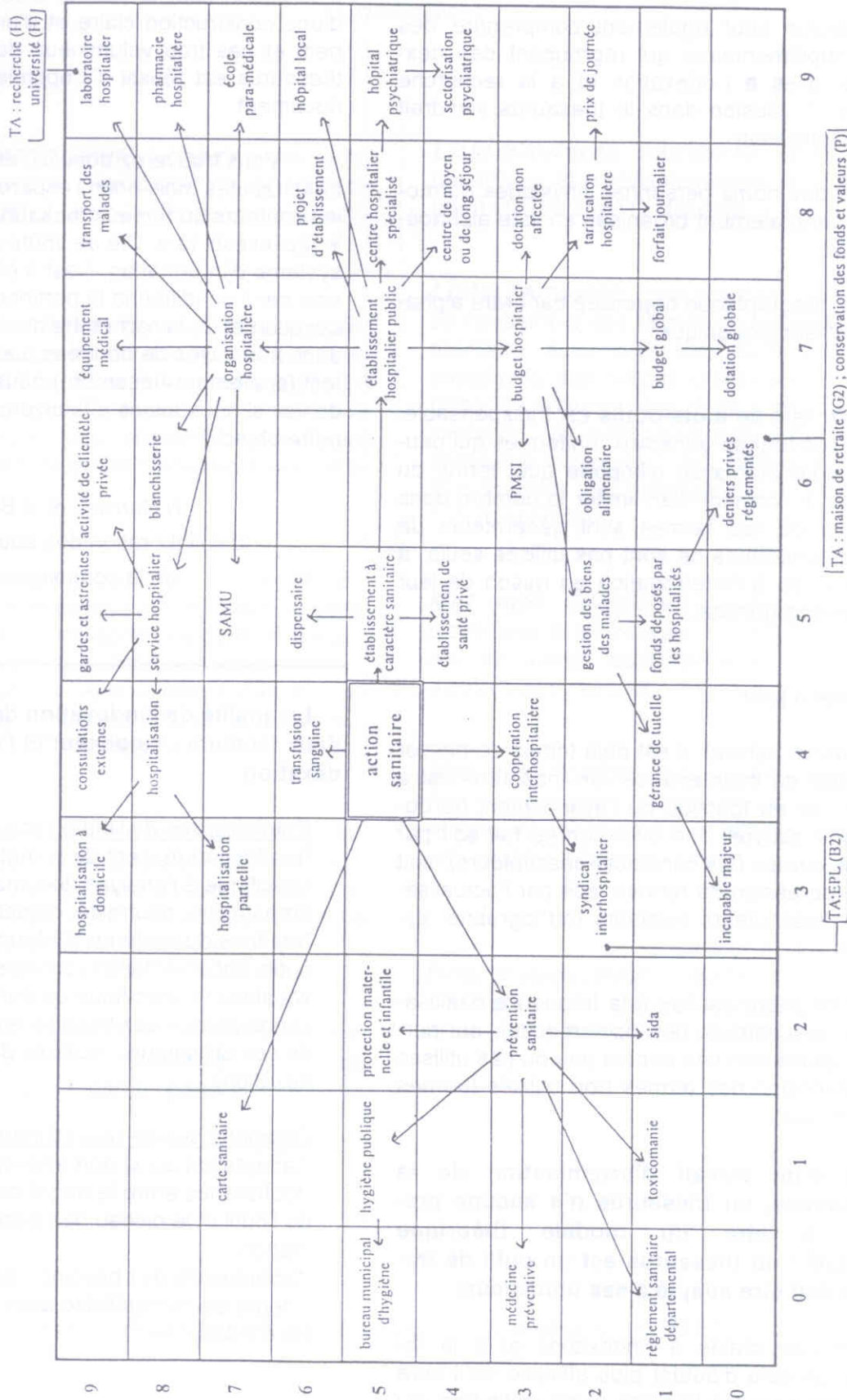
- **Méthode déductive** : extraction des concepts des documents et des questions des utilisateurs, à partir des index d'une base de données
- **Méthode inductive** : collecte des mots à partir de sources de références (dictionnaires, nomenclatures, thesaurus voisins...)
- **Méthode mixte** : combinaison des deux méthodes précédentes, elle permet de bien refléter le vocabulaire d'origine, et de vérifier l'exhaustivité de la collecte. **C'est cette troisième méthode qu'a privilégié la commission vocabulaire.**

Les termes retenus sont ensuite ventilés et hiérarchisés à l'intérieur de champs sémantiques. Cette étape doit impérativement être travaillée en collaboration avec des spécialistes des domaines concernés. Ceux-ci vérifient l'adéquation avec leur domaine d'étude, donnent leur aval sur l'emploi de tel ou tel terme ou sur la définition d'un mot.

Les premières hiérarchisations ne sont que les ébauches du thesaurus. Pour parvenir à un équilibre de l'édifice, les thèmes travaillés doivent être repris en comparaison avec les autres thèmes.

Un exemple de schéma fléché : le thesaurus des juridictions financières

G1 : ACTION SANITAIRE ET SOCIALE : ACTION SANITAIRE



## Des listes annexes

Le thesaurus peut également comprendre des listes complémentaires qui regroupent des descripteurs utiles à l'indexation et à la recherche mais dont l'inclusion dans le thesaurus viendrait alourdir l'utilisation :

- Liste des noms personnes physiques ou morales généralement organisée en liste alphabétique
- Liste géographique organisée par ordre alphabétique et hiérarchique

Enfin, une liste de **mots-outils** est indispensable. Ce sont des termes généraux, unitermes qui peuvent se combiner avec n'importe quel terme du thesaurus. Il convient d'en limiter le nombre dans la mesure où ces termes sont générateurs de bruit. Ces unitermes ne sont pas utilisés seuls à l'indexation ou à l'interrogation en raison de leur caractère trop général.

## 3- Sa mise à jour

Un thesaurus achevé, il est déjà temps de passer à sa phase de maintenance. Un thesaurus est à faire évoluer en fonction de l'avancement du domaine qu'il couvre. Son évolution se fait soit par l'ajout de termes (les candidats-descripteurs), soit par la suppression de termes, soit par l'actualisation des descripteurs existants (orthographe, synonymes ...).

Cette mise à jour est liée à la fréquence d'utilisation des descripteurs (les occurrences), qui permet la suppression des termes peu ou pas utilisés et la redéfinition des termes trop utilisés (termes trop généraux).

**Produit d'un travail d'organisation de la connaissance, un thesaurus n'a aucune prétention à être un modèle théorique intellectuel : un thesaurus est un outil de travail, qui doit être adapté à ses utilisateurs.**

C'est un outil d'aide à l'indexation et à la recherche qui sera d'autant plus efficace qu'il aura été élaboré pour les besoins d'une institution (ou d'un réseau).

C'est également un outil qui doit être facile d'utilisation pour le lecteur d'où la nécessité d'une construction claire et d'un choix pertinent et pas trop volumineux de termes. Un thesaurus est réussi s'il optimise l'accès au document.

Vous trouverez donc ici, en quelques traits rapides mais nous l'espérons précis, les contours du fameux thesaurus sur lequel le gip investit tant. Clé de voûte de notre système d'information, il est à élaborer avec soin car il conditionne la pertinence de l'indexation et de la recherche documentaire dans nos bases de données. Les six mois de test (septembre-décembre) nous permettront de voir si nous avons effectivement atteint notre objectif.

N.Berriau et V.Beltrame avec  
la collaboration des autres membres  
de la commission vocabulaire

### **La qualité de l'indexation dépend de deux facteurs : l'indexeur et l'outil d'indexation.**

#### Compétences de l'indexeur :

- \*maîtrise et respect de la méthodologie spécifique à l'analyse documentaire (impartialité, neutralité, objectivité),
- \*maîtrise du système d'information et des outils documentaires (connaissance du vocabulaire spécifique du domaine, connaissance des besoins en information de ses utilisateurs, maîtrise de l'outil d'indexation).

#### Compétences de l'outil d'indexation :

- \*adaptation au(x) domaine(s) couvert(s),
- \*cohérence entre le degré de spécificité de l'outil et le niveau des besoins en information,
- \*adaptabilité aux besoins nouveaux,
- \*degré de compatibilité avec d'autres outils d'indexation.

## Glossaire

**Candidat descripteur** : terme nouveau, créé lors de l'indexation d'un document et proposé pour son insertion éventuelle dans un thesaurus.

**Champ sémantique** : dans un thesaurus, ensemble thématique de notions qui ont des liens de proximité sémantique entre elles et qui sont regroupées et structurées autour d'une notion clé. L'ensemble des champs sémantiques constitue la présentation thématique d'un thesaurus.

**Descripteur** : terme normalisé et contrôlé, retenu dans un thesaurus, pour représenter, sans ambiguïté, une notion contenue dans un document ou dans une demande de recherche documentaire. Ce peut être un nom commun ou un nom propre, un uniterme ou un terme composé, un sigle ou un acronyme.

**Liste alphabétique structurée** : présentation d'un thesaurus listant, par ordre alphabétique, l'ensemble des descripteurs et non descripteurs ainsi que leurs relations sémantiques et les notes d'application associées.

**Liste annexe** : partie d'un thesaurus complétant la partie principale, présentation minimale obligatoire. Elle concerne généralement les noms propres.

**Liste d'autorité** : liste des termes normalisés qui doivent être obligatoirement et nécessairement utilisés dans l'indexation

**Liste permutée** : présentation alphabétique de l'ensemble des termes d'un langage documentaire ainsi que des mots significatifs constitutifs des termes composés de ce langage, à l'exclusion des mots vides. Cette présentation est particulièrement utile pour les thesaurus comportant des termes composés. Elle contient les descripteurs et parfois les non-descripteurs.

**Mot outil** : Descripteur qui n'est jamais employé seul, ni à l'indexation ni à la recherche. Il n'a de valeur documentaire qu'associé à d'autres descripteurs.

**Non-descripteur** : dans un thesaurus, terme non retenu pour représenter une notion. Il renvoie au descripteur à utiliser à sa place, par une relation d'équivalence.

**Note d'application** : courte note explicative précisant l'acception ou les conditions particulières d'utilisation d'un descripteur.

(Définitions issues du Thésauroglossaire des langages documentaires/ DEGEZ D/ MENILLET D/ Paris : ADBS, 2001)

## Bibliographie

- AFNOR / Règles d'établissement des thésaurus monolingues (NF Z 47-100), Paris : Afnor, décembre 1981
- ACCART JP/RETHY MP : Le métier de documentaliste.- Paris : Editions du Cercle de la Librairie, 1999
- COLLECTIF : Dictionnaire encyclopédique de l'information et de la documentation.- Paris : Nathan, 1997
- DEGEZ D/ MENILLET D : Construire un thesaurus,- Paris : ADBS, 2001
- DEGEZ D/ MENILLET D : Thésauriglossaire des langages documentaires : un outil de contrôle sémantique.- Paris : ADBS, 2001
- GUILLOT M : Pourquoi un service de documentation d'entreprise.- Paris : Management France, juillet 1973, p. 39-42
- GUINCHAT C/ MENU M/ BLANQUET MF : Sciences et techniques de l'information et de la documentation.- Paris : UNESCO, 1990
- GUINCHAT C/ SKOURI Y : Guide pratique des techniques documentaires, vol 2. – Paris : EDICEF/ AUPELF, 1996



Bulletin édité par le GIP ascodocpsy  
Directeur de la publication: M. Paul MONOT  
Comité de rédaction: Commission vocabulaire, Nathalie BERRIAU  
Mise en page: Nathalie DEREMAUX  
Impression : C.H St Jean de Dieu Lyon

Pour toute information complémentaire, contacter Nathalie BERRIAU, Coordinatrice ascodocpsy  
CH Saint-Jean-de-Dieu, 290 route de Vienne 69373, Lyon cedex 08  
Tél. 04 37 90 13 07 - Fax. 04 37 90 13 37 - Mobile : 06 82 44 18 24  
mél : [nberriau@ch-st-jean-de-dieu-lyon.fr](mailto:nberriau@ch-st-jean-de-dieu-lyon.fr)