UNIVERSITE
DES SCIENCES SOCIALES
DE GRENOBLE

UNIVERSITE
CLAUDE BERNARD
LYON

D.E.S.S. INFORMATION SPECIALISEE

LES LANGAGES DOCUMENTAIRES
ET
LE THESAURUS EN PARTICULIER

Mémoire présenté
et soutenu par

DANG VINH THIEN
sous la direction de M. ANDRE DEWEZE

LYON, 1984

UNIVERSITE
DES SCIENCES SOCIALES
DE GRENOBLE

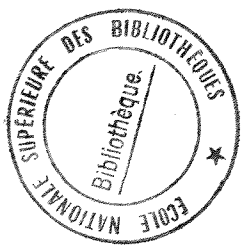
UNIVERSITE
CLAUDE BERNARD
LYON

D.E.S.S. INFORMATION SPECIALISEE

LES LANGAGES DOCUMENTAIRES
ET
LE THESAURUS EN PARTICULIER

Mémoire présenté
et soutenu par

DANG VINH THIEN
sous la direction de M. ANDRE DEWEZE



LYON, 1984

TABLE DES MATIERES

INTRODUCTION

Les langages documentaires ont un rôle essentiel à jouer dans la chaîne documentaire. Ils en constituent, en effet, un outil indispensable dans les maillons centraux depuis l'analyse du contenu, l'indexation jusqu'à la recherche des informations.

L'efficacité des services d'une unité d'information et de documentation scientifiques et techniques dépend largement de l'utilisation d'un langage approprié qui devrait être suffisamment sophistiqué pour permettre une description précise du contenu des documents et, à la fois, assez concis et simple pour faciliter leur recherche tout en assurant à l'utilisateur la possibilité de formuler correctement sa demande d'information.

Étant donné le développement accéléré des sciences et des techniques de nos jours, un langage documentaire doit également être élaboré de telle facon qu'il soit souple et adaptable aux besoins éventuels dus au surgissement des notions, voire des disciplines, nouvelles de la connaissance et du génie humains.

Ces contraintes ont été résolues de manière plus ou moins heureuse dans les pays développés où l'information et la documentation scientifiques et techniques sont devenues une science véritable et une activité établie qui, soutenue par des moyens informatiques, est susceptible d'apporter une contribution considérable à des efforts de R & D. Il n'en est pas de même, cependant, dans les pays en développement, surtout dans ceux où la langue nationale est utilisée en tant que moyen de communication dans la communauté scientifique et technologique comme c'est, par exemple, le cas du Vietnam.

En dépit de l'existence depuis une vingtaine d'années d'une terminologie scientifique et technique en langue vietnamienne dans presque tous les domaines, qui est ré-

gulièrement mise à jour pour se tenir au courant (bien qu'avec un certain décalage) de l'évolution du langage scientifique et technique du monde, et de l'achèvement d'une version vietnamienne de la CDU (actuellement sous presse) les documentalistes du pays se heurtent souvent à des difficultés lors des opérations de traitement intellectuel des documents et de recherche de l'information, faute d'un langage documentaire adéquat. Pour combler cette lacune, la solution ne consiste pas simplement dans la traduction ou l'adaptation des langages en usage général de par le monde, mais plutôt dans l'élaboration de ses propres outils d'indexation et de recherche de l'information (au moins pour ce qui concerne les travaux de production interne), tout en se servant des progrès déjà acquis ailleurs en la matière et en tenant compte des conditions du pays et surtout des particularités de la langue nationale.

C'est dans cet esprit que ce mémoire est écrit. Il représentera l'effort de l'auteur, d'une part, de réfléter ce qu'il a pu apprendre dans le cadre du présent CPU et dans l'étude des ouvrages portant sur ces problèmes et, d'autre part, de contribuer sur le plan méthodologique à l'élaboration d'un langage documentaire, ou plus spécifiquement, d'un thesaurus - spécialisé ou de caractère géneral - en langue vietnamienne.

Ce ne sera, de toute façon, qu'une approche tout à fait générale qui se prête bien à une présentation plus ou moins formelle que l'auteur a adoptée pour son travail. En outre, dans la dernière partie la méthode comparative sera utilisée là où une analogie (ou même anti-analogie) est observable entre le vietnamien et une langue européenne, notamment le français ou l'anglais.

CHAPITRE 1. LA RECHERCHE DOCUMENTAIRE

1.1. Généralités sur la chaîne documentaire.

Une unité d'information et de documentation fonctionne sur un ensemble d'opérations successives et logiquement liées l'une à l'autre que l'on appelle la "chaîne documentaire".

A l'entrée de la chaîne se trouve l'opération de collecte des documents primaires (monographies, périodiques, documents non-publiés etc.) pour constituer le fonds documentaire. L'opération suivante est le contrôle et l'enregistrement des documents admis dans le fonds documentaire, opération qui consiste à faire l'inventaire du stock au fur et à mesure du flux d'entrée et qui aboutit au rangement et à la conservation des objets documentaires. Ces deux opérations constituent le "traitement matériel" des documents qui se poursuit par des opérations de "traitement intellectuel" /1/: la description bibliographique et l'analyse du contenu. La première a pour but d'établir une carte d'identité du document, portant sur ses caractéristiques formelles (auteur, titre, source, date, langue etc.). La suivante consiste à extraire des informations essentielles contenues dans le document, les exprimer en langue naturelle éventuellement sous une forme canonique (rédaction du résumé) et les interpréter en termes d'un langage artificiel (indexation). Les produits de ces deux opérations sont ensuite mis en mémoire sous forme de différentes sortes de fichiers qui seront utilisés ultérieurement pour les dernières, mais non les moindres, opérations de la chaîne : la recherche documentaire et la diffusion. De fait, se sont ces deux fonctions qui constituent "le fondement des services offerts aux utilisateurs, la raison d'être de l'unité" /1/.

1.2. Les systèmes de recherche de l'information

La recherche de l'information est un ensemble d'opérations qui a pour but de fournir à l'utilisateur les renseignements répondant à ses questions occasionnelles ou permanentes /1/.

Ces opérations s'effectuent au sein des systèmes de recherche de l'information (SRI) qui, en fonction de ce qu'ils ont à fournir à la clientèle, peuvent être classés en deux catégories principales :

- Les SRI factuelles, ou banques de données, qui diffusent des données directement utilisables aux utilisateurs, et
- Les SRI documentaires, ou bases de données, qui donnent des références aux documents contenant des informations demandées et, éventuellement, ces documents eux-même. Dans le cas où le système ne s'occupe que de la fourniture des références ils sont appelés "systèmes de recherche d'informations bibliographiques" ou, plus récemment, "référothèques" /2/.

En dépit de cette catégorisation, il n'y a pas de différences de principe entre ces deux types de SRI. En fait, ils ont une caractéristique commune, c'est de ne donner que des informations qui y ont été mémorisées, d'où le terme "recherche" ("retrieval" en anglais). De toute façon, la distinction étant plutôt au niveau de l'accès aux informations stockées, ce qui fait l'objet de cette étude n'aura trait qu'au deuxième type de SRI que nous appelons, pour abréger, SRD (système de recherche documentaire).

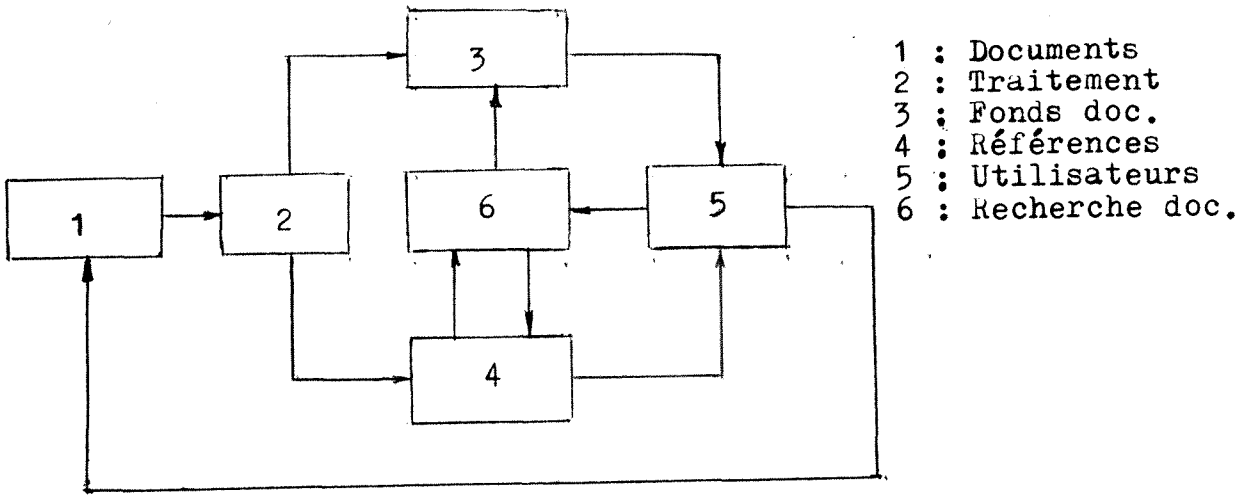


Fig. 1. Schéma d'un SRD

Le caractère occasionnel ou permanent des questions posées par les utilisateurs donne lieu, respectivement, à deux modes de recherche :

- La recherche rétrospective qui consiste à retrouver tout l'ensemble ou une partie des documents enregistrés au préalable et se rapportant à une question ponctuelle, et
- La recherche courante, c'est-à-dire, l'identification des documents répondant à une question donnée en permanence qui viennent d'être répertoriés au cours d'un certain laps de temps (semaine, dizaine, mois...)/1/. Ce mode de recherche est souvent connu en association avec son corollaire - la diffusion sélective de l'information (DSI).

La procédure de recherche documentaire, qui est sensiblement la même pour les deux modes, peut être décrite en termes d'un ensemble D des documents disponibles dans le corpus documentaire ou accessible au système,

d'un ensemble Q des questions posées par les utilisateurs et d'une relation R par laquelle on fait correspondre à une question $q \in Q$ un sous-ensemble $d \subset D$ qu'on appelle réponse à q.

Evidemment, la méthode la plus fiable serait de parcourir le contenu de tout élément $x \in D$ et de tenter d'en vérifier la relation avec Q. Cette démarche s'avère pourtant trop onéreuse dans la pratique pour être applicable, étant donné le nombre souvent assez grand des documents à trier de telle façon. Pour cette raison, au lieu d'un dépouillement dans l'ensemble D lui-même, on en fait un tri sur l'ensemble D' des images y de $x \in D$ par une application I telle que :

$$D' = I(D)$$
$$\forall y \in D', \exists x \in D \ : y = I(x)$$
$$\forall x \in D , \forall x' \in D, x \neq x' \longrightarrow I(x) \neq I(x')$$

Dans la pratique documentaire, I représente l'opération de traitement intellectuel et D' - un fichier des références qui sont des "modèles réduits" des documents correspondants, d'où une réduction notable du caractère onéreux de l'opération de tri. D'autre part, l'ensemble D' est systématisé et mémorisé sur des supports papiers ou magnétiques, ce qui facilite encore plus la recherche.

A son tour, $y = I(x)$ est défini par :

$$y = \{a_x, b_x, m_x, r_x\}, \ a_x < b_x < m_x < r_x$$

où

a_x désigne l'adresse de $x \in D$,

b_x = l'ensemble des caractères formelles de x,

m_x = l'ensemble des caractèristiques du contenu (formant le "modèle de recherche") de x.

r_x = le résumé de x.

Quand q porte sur un (ou des) élément(s) de b_x, ce qui n'est pas souvent le cas, la recherche consiste simplement à en comparer les valeurs respectives à celles indiquées dans q.

Il en est pas de meme quand it s'agit d'une q portant sur m_x. En fait, l'établissement de la relation de correspomdance entre le contenu de x et celui de q présuppose une interprétation de ce-dernier sous forme d'équation de recherche en des termes compatibles avec les éléments de m_x.

L'outil indispensable qui sert de "dénominateur commun" à cet effet c'est le langage documentaire.

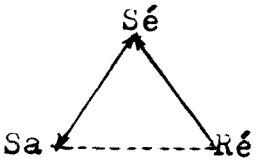
x x
x

CHAPITRE 2. LES LANGAGES DOCUMENTAIRES

2.1. Quelques aspects linguistiques.

Un langage est un système de signes d'une certaine nature physique, qui assure la fonction cognitive et communicative dans des activités humaines /3/.

Un tel signe peut être représenté par un diagramme en forme de triangle proposé par Ogden et Richards /2/ :



où :

Sa = signe = ce qui signifie (signifiant),
Sé = concepte = ce qui est signifié (signifié)
Ré = référent = ce à quoi le signe se réfère (la chose nommée)

Les relations signifiant-signifié (Sa-Sé) et signifiant-signifiant (Sa-Sa) sont établies soit de façon spontanée, naturelle, soit de façon artificielle, d'où le langage naturel et le langage artificiel.

Dans le langage naturel qui sert de véhicule de pensée et de communication de l'homme dans la vie courante et qui ne fonctionne pas selon des conventions et des règles strictement établies et inchangées, il ne saurait y avoir une correspondance bi-univoque entre les mots (signifiants) et leurs sens (signifiés) ni une manière unique d'associer des mots pour exprimer une même idée entière (un ensemble de signifiés), ce qui résulte de l'existence au sein du langage lui-même des homonymies, des synonymies, des polysémies et d'une variété de nuances d'expression, des contextes différents, sans compter des facteurs psychologiques subjectifs de l'utilisateur qui varient selon les cirscontances.

Pour ces raisons, le langage naturel ne se prete pas à une description sans équivoque du contenu essentiel des documents et des questions dans un SRD (+).

Crée par l'homme dans un but défini, tout langage artificiel, en l'occurence le langage documentaire, ne pourrait exister sans se baser sur le langage naturel, mais en même temps, devrait s'en débarrasser de toute ambiguité possible.

Un langage documentaire L_D peut etre décrit en termes d'un lexique L_x et d'une syntaxe S :

$$L_D = \{L_x, S\}$$

où:

L_x = l'ensemble de signifiants (mots-clés, descripteurs ou indices de classification),

S = l'ensemble de règles d'association des signifiants.

Des contraintes primordiales sur la base des relations linguistiques formulées par Morris /2/ s'imposent à L_D en tant qu'outil d'indexation et de formulation des équations de recherche :

1. Dans L_x, la relation Sa-Sé (relation sémantique) doit être débarrassées de toute équivoque :

   Soit Σé = l'ensemble des Sé, on a :

   $\forall Sé \in \Sigma é, \exists Sa \in L_x : Sé = R(Sa)$

   $\forall Sa \in L_x, \forall Sa' \in L_x, Sa \neq Sa' \longrightarrow R(Sa) \neq R(Sa')$

(+) A moins d'être contrôlé de manière ou d'autre, mais dans ce cas il ne restera plus "naturel". DEWEZE /2/ a montré que si l'utilisation du langage naturel (à ce propos) n'est pas impossible elle est pour le moins hasardeuse et très couteuse et de ce fait moins recommendable.

R est donc une relation de correspondance bi-univoque.

2. Au niveau de la relation Sa-Sa (relation syntaxique) toute règle $\rho \in S$ définie dans $L_x \times L_x$ ou dans $p \subset P(L_x \times L_x)$ qui en est munie, doit être univoque :

$$\forall \rho \in S,\ \forall \rho' \in S,$$
$$\forall Sa \in L_x,\ \forall Sa' \in L,\ Sa \neq Sa'$$
$$\rho \neq \rho' \longleftrightarrow Sa\rho Sa' \neq Sa\rho' Sa'.$$

3. La relation Sa-Ut (Ut = utilisateur (+)), ou relation pragmatique, doit etre eliminée.

Soit Fu = l'ensemble des facteurs subjectifs de l'utilisateur,

Us = l'Univers conceptuel du système,

On doit avoir :

$$L_D \times (Fu - Us) = \phi \text{ , donc}$$
$$Fu - Us = \phi$$

Celà veut dire que tout élément personnel subjectif qui n'est pas absorbé dans l'univers du système doit être exclu.

Comme un langage documentaire est exclusivement un langage écrit utilisant des symboles graphiques (numériques ou alphabétique) pour représenter des signifiants, on peut se passer de la phonétique pour aborder, au premier abord, la morphologie qui représente l'ensemble de règles et de moyens pour la formation et la transformation des mots.

L'unité de base à ce propos est le morphème qui est la plus petite structure linguistique en tant que signifiant, par opposition à son équivalent sémantique qu'est le sémème qui est, au contraire, décomposable en une plus petite unité, c'est le sème - l'élément de la signification.

(+) Il s'agit de l'utilisateur du langage, donc y compris l'indexeur.

On distingue des morphèmes lexicaux (ou lexèmes) qui constituent les bases lexicales des mots et des morphèmes grammaticaux (ou grammèmes) qui sont des affixes servant à modifier l'aspect grammatical du mot. Pour des raisons pratiques, les termes d'un langage documentaire sont de préférence des substantifs singuliers ou des phrases nominales /4/ on peut se passer de l'aspect grammatical au niveau des grammèmes pour ne s'intéresser qu'aux lexèmes. Un lexème peut être représenté soit par un seul symbole d'un alphabet comme, par exemple, le premier chiffre d'un indice dans la table principale de la CDU, soit par une suite de caractères dans un langage dont le lexique est composé des termes extraits du langage naturel.

Dans les langages de type classification hiérarchique l'unité structurale élémentaire est le lexème et elle le reste toujours quelque soit la longeur de l'indice (voir chapitre 3), tandis que dans d'autres elle est le mot qui est composé d'un ou de plusieurs lexèmes. Mais dans les deux cas, faisant partie du lexique, elle est souvent connue comme un terme.

Sur le plan syntaxique, la formation des phrases par l'association des mots pour exprimer des notions plus compliquées ou des idées entières se fait au moyen des outils lexicaux spéciaux comme, par exemple, les signes +, /, :, =, ( ) etc. dans la CDU. Dans les langages dont les termes sont empruntés au langage naturel elle s'effectue soit par simple association des mots avec ou sans indicateurs de rôle, soit à l'aide des opérations logiques (ET, OU, SAUF) lors de l'interrogation.

En ce qui concerne la sémantique l'élimination des ambiguïtés des termes et leur systématisation donne lieu à des considérations plus détaillées qui feront l'objet du chapitre consacré aux langages post-coordonnés.

2.2. Typologie des langages documentaires.

C'est selon la formule "menu" ou "à la carte" lors de l'indexation et de l'interrogation que l'on peut distinguer deux types principaux de langages documentaires :

- Langages pré-coordonnés
  - + Classifications thématiques
  - + Classifications à facettes
  - + Vedettes-matière
- Langages post-coordonnés
  - + Liste de descripteurs
  - + Thesaurus.

Dans le premier type des groupes de termes ou des phrases exprimant des notions plus compliquées sont établis au préalable sous les entrées des termes constituants plus génériques. De cette façon, lors de l'indexation ou de la recherche documentaire le contenu essentiel du document ou de la question établi, il faut chercher parmi des "phrases types" disponibles dans le langage celle qui en est la plus proche pour la représenter comme modèle de recherche du document ou comme l'équation de recherche de la question. L'acheminement vers la construction-cible est guidé par un système d'aiguillage qu'est la classification.

En ce qui concerne le deuxième type, la formation des modèles de recherche ou des équations de recherhce ne s'effectue qu'au cours de l'indexation ou de l'interrogation de telle manière qu'avec un jeu fini de termes donnés on puisse élaborer, en principe, un nombre pratiquement infini de leurs combinaisons.

Les principes généraux de ces deux types de langages documentaires seront analysés dans les chapitres suivants.

x x

x

CHAPITRE 3. LES LANGAGES PRÉ-COORDONNÉS

Les différents langages documentaires de ce type sont connus sous l'appellation commune de "classification" qui en constitue la base de structuration. Ainsi s'agit-il de la CDU ou de la classification à facettes et d'autres leur principe de systématisation en général reste le même et se fonde sur les relations hiérarchiques entre les termes, soit dans leur ensemble, soit au moins au niveau des différentes classes et sous-classes.

3.1. Principe général de classification.

La classification d'un ensemble d'objets consiste à en définir des classes d'équivalence selon une certaine caractérisque ou relation.

Soit T l'ensemble des termes d'un langage documentaire, une relation d'équivalence va définir dams T des classes d'équivalence T_1, T_2, T_3... T_n, telles que :

- $T_1 \cup T_2 \cup T_3 \cup \ldots T_n = T$
- $T_1 \cap T_2 \cap T_3 \cap \ldots T_n = \phi$
- $T_1 \cap T_2 = \phi$, $T_2 \cap T_3 = \phi$, $T_1 \cap T_3 = \phi$, etc.

Au niveau d'une classe, T_1 par exemple, la même relation R définit une partition semblable :

$$T_{11} \cup T_{12} \cup T_{13} \cup \ldots T_{1n} = T_1$$

...

...

et ainsi de suite.

3.2. Les classifications thématiques.

En définissant la relation d'équivalence d'une manière explicite : R = "avoir pour thème T_i" et en donnant à T_1, T_2, T_3... T_n leurs valeurs sémantiques correspondantes, nous aurons une classification thématique dans laquelle, au premier niveau de partition, les divisions principales sont les classes représentées respectivement par les termes thématiques T_1, T_2, T_3... T_n. Le même principe s'applique

thématiques T_1, T_2, T_3... T_n. Le même principe s'applique au niveau de partition suivant, ce qui nous donne des subdivisions T_{11}, T_{12}, T_{13}... T_{1n}; T_{21}, T_{22}, T_{23}... T_{2n}; etc. La hiérarchisation peut se continuer ainsi jusqu'à n'importe quel niveau. Une représentation graphique en démontre une structure arborescente :

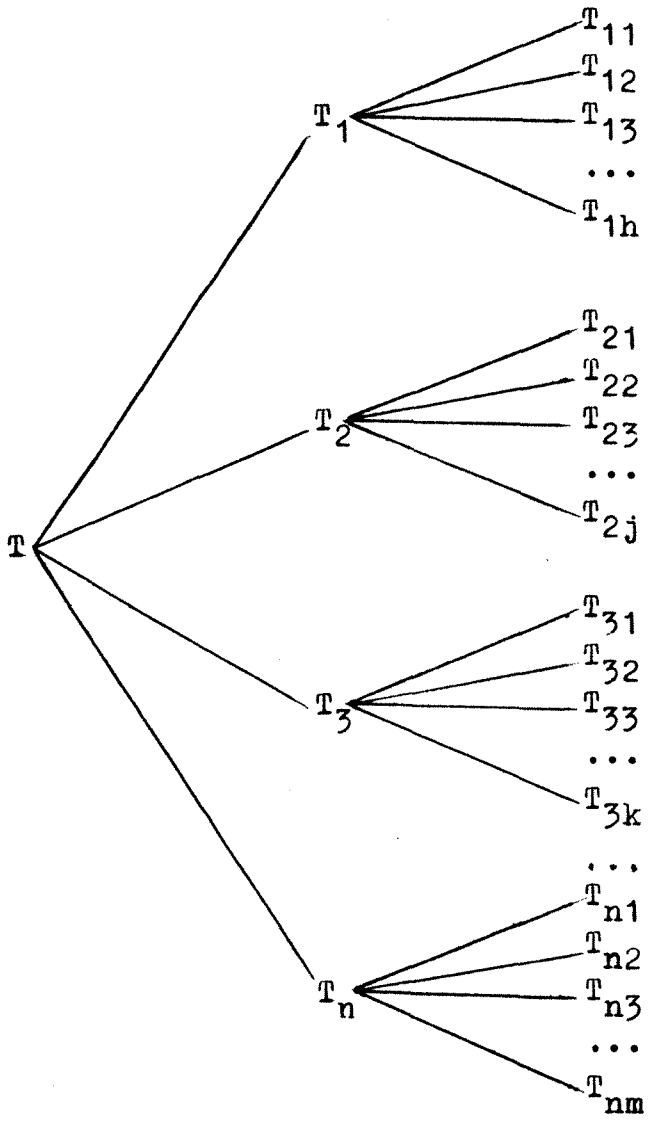


Fig. 2. Structure arborescente d'une classification

Les classes ainsi définies dans T par la relation R formeent un nouvel ensemble appelé ensemble quotient de T par R et qu'on note T/R. Il est à noter que nous pouvons avoir :

$$\text{Card. } T/R \neq \text{Card. } T_1/R \neq \text{Card. } T_2/R \ldots$$

et que le cardinal de chacun de ces ensembles est d'un nombre fini bien que le nombre des niveaux en soit, en principe, infini.

La classification typique de ce genre est la CDU dans laquelle on a toujours :

$$\text{Card. } T/R = \text{Card. } T_1/R = \text{Card. } T_2/R \ldots = 10$$

et ce à tout niveau.

Comme il est montré dans fig.2, chaque terme y occupe une position bien définie représentée par un indice (numérique, alphanumérique ou alphabétique selon le système) qui en indique l'appartenance à une certaine classe et ce à tel ou tel niveau. L'indice se substitue au terme dans les différentes opérations documentaires car il est plus court et plus maniable. La conversion réciproque de l'un à l'autre, c'est-à-dire la traduction du langage documentaire en langage naturel (contrôlé, bien sûr) et inversement, se fait au moyen des tables systématiques.

L'indice en tant que signifiant dans les langages de ce type n'est donc qu'un lexème dans le sens qu'il ne se laisse décomposer sans perdre sa capacité significative. En fait, dans la suite de symboles qui en font partie seul le premier a une signification à part, celle de chacun des autres étant déterminée par tous ceux qui le précèdent.

3.3. Les classification à facettes .

Tout en suivant le principe de classification mentionné plus haut, si maintenant on définie une relation R' qui se lit, par exemple, "vu sous l'aspect F_i" ($i = 1 < 2 < 3 < \ldots n$) on aura une classification à facettes qui combine la hiérarchisation des points de vue (facettes) avec l'arrangement des termes par thèmes (focus) dans chaque facette. De cette façon, les classifications de ce type ont une capacité de combiner des termes plus grande que celle des classifications thématiques.

L'inconvénient principal des classifications, même dans le cas des classifications à facettes qui permettent une variété de points de vue, est le manque de capacité combinatoire, c'est pourquoi elles ne sont utilisées généralement que pour décrire le sujet principal d'un documemt (de préférence, une monographie) avec un ou, éventuellement, deux ou trois indices.

Pour combler cette lacune on a recours à des langages combinatoires dont les termes sont prélevés du langage naturel des documents et peuvent être combinés entre eux lors de l'indexation et de l'interrogation.

3.4. Les vedettes-matière.

Le premier de ce type de langages est connu comme des "vedettes-matière" qui sont classées par ordre alphabétique. Chaque terme y est indépendant et constitue une vedette à part munie des renvois d'orientation à d'autres qui lui sont associés d'une certaine manière. Pour un document il faut le mettre sous autant de vedettes que de thèmes (ou matières) sur lesquels il porte. La combinaison des vedettes entre elles s'effectue au moyen des sous-vedettes qui représentent à la fois plusieurs matières concernées. Ainsi, pour garantir la possibilité de retrouver sous une même vedette tous les documents indexés avec une combinaison quelcoque des termes entrés dans ses sous-vedettes, il faut qu'elle comprenne toutes les combinaisons possibles de ces termes. Soit n leur nombre, celui des combinaisons sera :

$$\sum_{m=1}^{m=n} C_n^m = C_n^1 + C_n^2 + C_n^3 + \dots C_n^n = 2^n - 1$$

Pour cette raison, on n'utilise qu'un petit nombre de vedettes pour chaque document (+) et des vedettes assez

---

(+) On aura alors le nombre de combinaisons p à p :

$$C_m^p = \frac{m\,(m-1)\dots(m-p+1)}{p!} = \frac{m!}{p!(m-p)!}$$

générales, ce qui n'en permet pas une description fine du contenu. Ce type de langage est utilisé surtout pour des fichiers manuels des bibliothèques /1/.

x x

x

CHAPITRE 4. LES LANGAGES POST-COORDONNÉS

Aucun des langages du type pré-coordonné n'assure la possibilité de rechercher des documents d'après une équation de recherche qui n'a pas été établie au préalable. Pour cette raison, depuis une trentaine d'années les langages combinatoires à post-coordination ont connu une utilisation de plus en plus répandue, en particulier à l'avènement de l'informatique dans le domaine de documentation quoiqu'ils soient également utilisables pour les systèmes manuels.

4.1. La liste de descripteurs.

La plus simple forme de ces langages est la liste de descripteurs qui sont des termes prélevés du langage naturel des documents, controlés et homologués ensuite comme termes préférentiels à l'usage de l'indexation et de l'interrogation et arrangés par ordre alphabétique.

La différence essentielle entre la liste de descripteurs et les vedettes-matière en tant que langages combinatoires est que ces dernières n'assurent la combinaison des termes que lors de l'indexation, tandis que la liste permet, même au cours de l'interrogation, à l'utilisateur de formuler sa question de façon qu'il lui convient. Le critère de choix des documents lors de la recherche pour les vedettes-matière, tout comme pour les classifications, est l'identité des termes décrivant leur contenu à ceux donnés dans la question, tandis que pour les langages post-coordonnés, en l'occurence la liste de descripteurs, c'est la conformité des modèles de recherche des documents à une équation de recherche "sur mesure". En voici le principe :

Soit L une liste de descripteurs,

$$L = \{t_1, t_2, t_3 \ldots t_m\}$$

$t_1, t_2, t_3 \ldots t_m$: descripteurs,

Soit encore D l'ensemble de documents dans lequel la

la recherche est faite :

$$D = \{d_1, d_2, d_3 \ldots d_n\}$$

où d_1, d_2, $d_3 \ldots d_n$ sont des documents distincts dont les modèles de recherche sont formés par des éléments $t_i \in T$.

Une relation R_{t_i} définit dans D une classe d'équivalence D_{t_i} telle que :

$$\forall d_i \in D_{t_i} : d_i \text{ contient } t_i \text{ dans son modèle de recherche}$$

Soit, par exemple, Q une question exprimée par l'équation :

$$Q = t_1 \cap (t_2 \cup t_3)$$

La recherhhe d'après Q consiste à réaliser les opérations données en substituant D_{t_1}, D_{t_2}, D_{t_3} à t_1, t_2, t_3 respectivement, dans l'équation. Le résultat en sera :

$$d_Q = D_{t_1} \cap (D_{t_2} \cup D_{t_3})$$

qui représente un sous-ensemble de D dont les éléments sont caractérisés par un modèle de recherche en conformité à l'équation donnée.

Ces opérations s'effectuent dans la pratique au moyen d'un mécanisme appelé "fichier inversé" dont le principe est le même pour un système manuel ou un système informatisé.

Dans un système manuel, le fichier inversé est composé de fiches à coincidence optique, à chacune desquelles est associé un descripteur, et à chaque document contenant , entre autres, ce descripteur est donné un numéro (son adresse) qui est représenté sur la fiche par une perforation dont les coordonnées correspondent respectivement aux valeurs numériques du numéro : l'abscisse aux centaines et l'ordonnées aux dizaines et aux unités. De cette façon, on

peut représenter n'importe quel nombre de 1 à 10 000 (si la fiche compte 100 lignes et 100 colonnes) ou de 1 à 2 500 (si elle contient 50 lignes et 50 colonnes) dans une case déterminée. De ce fait, chaque fiche représente une classe d'équivalence des documents indexés à l'aide du descripteur concerné.

Lors de la recherche documentaire, l'équation de recherche établie, on sélectionne les fiches portant chacune un descripteur donné dans l'équation et on les superpositionne pour réaliser l'opération d'intersection. Tous les documents ayant pour thèmes l'ensemble de ces descripteurs seront indiqués par des perforations correspondantes la coincidence desquelles sera signalée par exposition contre une source de lumière, d'où l'appellation "coïncidence optique". Puisque l'intersection et l'union sont distributives l'une par rapport à l'autre on peut les réaliser dans n'importe quel ordre sur les termes de l'équation se trouvant entre les parenthèses.

4.2. Thesaurus.

4.2.1. Définition et fonction.

L'avènement de l'informatique documentaire a permis à l'utilisateur de dialoguer avec le système lors de la recherche pour raffiner ses questions au fur et à mesure des réponses respectives du système. Ce mode de recherche en conversationnel s'effectue grâce à la possibilité d'élargir ou de rétrécir l'équation de recherche soit au niveau des opérateurs logiques - ce qui est aussi possible dans le cas d'une liste de descripteurs, mais qui risque d'entraîner un règlage trop "à coup" - soit au niveau des termes en les remplaçant par d'autres qui sont plus génériques ou plus spécifiques ou qui leur sont associés de manière ou d'autre. Cela exige une structuration plus profonde du langage, c'est-à-dire il faut transformer une liste de descripteurs en une structure dans laquelle les

les relations hiérarchiques et associatives entre les notions désignées par les termes seront représentées explicitement. Cette structure, connue sous l'appellation de "Thesaurus" dans la documentation, a pour but de servir "d'instrument de contrôle de la terminologie utilisé en transposant en un langage plus strict (langage documentaire, langage d'information) le langage naturel employé dans les documents et par les indexeurs ou les utilisateurs"/5/.

Un thesaurus est donc un dictionnaire normatif, ou plutôt, sa symétrie. En fait, en nous référant encore une fois au triangle d'Ogden et Richards nous verrons que la relation signifiant-signifié est réciproque et réversible :

Sa ⟷ Sé

Cela veut dire que si l'on peut partir du signifiant pour trouver le signifié comme avec un dictionnaire ordinaire, on peut, inversement, partir du signifié pour avoir le signifiant en utilisant un thesaurus.

Le travail de l'indexeur ou la démarche de l'utilisateur consiste à choisir la notion qui lui semble la plus appropriée parmi celles qu'implique un terme du langage naturel et de l'exprimer par un terme du langage documentaire. Le thesaurus lui vient en aide en lui offrant même la possibilité d'en choisir un qui est plus générique ou plus spécifique (le deuxième cas est pour l'utilisateur seulement). A cet effet, le thesaurus doit avoir une structure hiérarchique et contenir également des termes du langage naturel qui servent de points d'accès et les renvoyer à des termes homologués que sont les descripteurs, les premiers étant des "non-descripteurs"(+).

(+) A l'heure actuelle, nous "préférons" les thesauri à termes préférentiels en raison de la possibilité de mise à jour manuelle.

4.2.2. Relations dans un thesaurus .

Les relations entre descripteurs (signifiants) dans un thesaurus seront examinées sur la base des composantes signifiant (Sa), signifié (Sé) et référent (Re) (voir diagramme, p. 10). Ce sera une démarche tout à fait nominale car, en fin de compte, toutes ces relations s'articulent autour le sighifié comme point d'appui. A cet égard, DEWEZE /2/ a proposé un outil d'analyse plus fin basé sur la "configuration sémique" qui est l'ensemble de sèmes, ou éléments de signification, représentant des propriétés ou des caractéristiques élémentaires d'une notion ou d'un objet. Cette approche dépasse en profondeur notre sujet, mais nous nous y référerons, éventuellement, au cours de notre analyse.

4.2.2.1. Relations basées sur le signifiant.

4.2.2.1.1. Homographie.

C'est le cas où à un Sa correspondent des Sé tels que :

$$Sé_1 \cap Sé_2 \cap \ldots Sé_i = \emptyset$$

Ex. 1. KINH (mot vietnamien)

Sa/KINH/ Sé → 1 (capitale)
2 (peur)
3 (prières)

Ex. 2. SON

Sa/SON/ Sé → 1 (vibration)
2 (possessif)
3 (enveloppe de grain)/2/

L'ambiguité dans ce cas peut être levée par une analyse morpho-syntaxique ou par indication d'appartenance à des champs sémantiques distincts.

4.2.2.1.2. Polysémie.

Dans ce cas, à un Sa correspondent des Sé tels que :

$$Sé_1 \cap Sé_2 \cap \ldots Sé_i = Sé_D \neq \emptyset$$

Ex.1.BẢNG (mot vietnamien)

Sa/BẢNG/ ⟶ Sé < 1 (tableau), 2 (pancarte), 3 (table)

Ex. 2. SIEGE

Sa/SIÈGE/ ⟶ Sé < 1 (meuble), 2 (fondement), 3 (maison mère) /6/

Entre ces termes il ya quelque chose de commun. C'est l'ensemble des sèmes qui caractérisent un thème dominant qu'est l'idée d'affichage dans le premier exemple et l'idée de fondation, d'assise dans le deuxième.

L'ambiguïté dans cette relation peut être levée aisé aisément, si les polysémies appartiennent à des champs sémantiques bien distincts, par des spécifications correspondantes. Si ce n'est pas le cas, il faut leur affecter d'autant de marqueurs que de champs sémantiques susceptibles de l'accueillir /2/.

4.2.2.1.3. Syntagme.

L'association de plusieurs Sa distincts fait naitre un nouveau Sa auquel correspond un Sé qui est différent (ou plus spécifique) de la somme des Sé auxquels correspondent les Sa constitutifs :

$$Sa_1 + Sa_2 + \ldots Sa_n \longrightarrow Sa_w \longleftrightarrow Sé_w$$

a. $Sé_w \neq \bigcup_i Sé_i$; $Sé_i \longleftrightarrow Sa_i$

Ex. 1. BUỒNG (chambre), TỐI (obscure)

Sa/BUỒNG TỐI/ Sé (caméra obscura)

Ex. 2. OEIL-DE-BOEUF

b. $Sé_w \subset \bigcup_i Sé_i$

Ex. 1. CHUỐI banane), NGỰ (royale)
CHUỐI NGỰ (espèce de banane très savoureuse)

Ex. 2. CANNE A SUCRE

Dans les deux cas (a et b) le syntagme doit être maintenu tel quel et répertorié dans le thesaurus comme descripteur composé s'il n'a pas de synonyme qui a été homologué (voir § suivant).

4.2.2.2. Relations basées sur le signifié.

4.2.2.2.1. Synonymie.

Cette relation est caractérisée par le fait qu'à un Sé correspondent des Sa différents tels que:

$$Sé \longleftrightarrow Sa_1 R Sa_2 R Sa_3 \quad ; \; R : \text{synonymie}$$

Les signifiants Sa_1, Sa_2, Sa_3 sont considérés comme rigoureusement synonymes si, et seulement si, ils représentent exactement le même Sé. Dans cette condition R sera une relation d'équivalence :

$$Sa_1 \; R \; Sa_1$$
$$Sa_1 \; R \; Sa_2 \longrightarrow Sa_2 \; R \; Sa_1$$
$$Sa_1 \; R \; Sa_2, \; Sa_2 \; R \; Sa_3 \longrightarrow Sa_1 \; R \; Sa_3$$

La synonymie définit donc dans le thesaurus une classe d'équivalence représentée par le Sa choisi comme descripteur.

4.2.2.2.2. Quasi-synonymie.

Si nous avons $Sa_1 \neq Sa_2 \neq Sa_3$ tels que $Sé_1$, $Sé_2$, $Sé_3$ sont deux à deux différents même d'un sème :

$$Sé_1 \triangle Sé_2 \neq Sé_2 \triangle Sé_3 \neq Sé_3 \triangle Sé_1 \neq \emptyset$$

la relation ne sera pas transitive et, de ce fait, ne reste plus une relation d'équivalence. Dans ce cas nous aurons une relation de ressemblance, qui est une relation associative (voir § 4.2.2.3.3.1), ou au mieux, une relation de similitude (relation d'équivalence floue)/2/ qui peut etre considérée, dans certain contexte, comme une relation de synonymie, les Sa en question étant des quasi-synonymes.

Les cas de synonymie véritable sont extrêmement rares dans un corpus donné /2/, ainsi dans /5/ est-il spécifié que "les termes considérés comme équivalents (similaires ou de signification presque identique) peuvent être réunis dans des catégories d'équivalence, de telle sorte que les termes équivalents correspondent à une seule et même notion. Dans la recherche documentaire, tous les documents associés à la catégorie d'équivalence doivent être retrouvés, même si un seul terme est utilisé comme descripteur. Il faut distinguer :

- Les synonymes, c'est-à-dire les termes qui ont le même sens ou presque dans une discipline particulière, et
- Les quasi-synonyme, c'est-à-dire les termes dont la signification peut être différente dans le vocabulaire utilisé et le domaine concerné, mais qui peuvent être considérés comme synonymes pour les besoins du système de documentation considéré".

4.2.2.3. Relations basées sur le référent.

Comme il a été montré dans le diagramme (p.10), le signifiant sert à référer à un objet (chose ou phénomène) dans l'univers réel. Cet objet, ou le référent, peut se trouver en certaines relations avec d'autres par des facteurs logico-objectifs. C'est sur la base de ces relations extra-linguistique qu'on réunit des signifiants avec les signifiés correspondants dans divers groupes lexico-sémantiques qu'on appelle "paradigmes".

D'après LAUREILHE /7/ le paradigme (+) est un ensemble de mots ayant un élément commun, qui peut être un rapport associative, parfois même une simple association mentale.

Dans le cas d'un paradigme lexical l'élément commun est le lexème qui admet des constructions dérivées par affixation au moyen des grammèmes différents. La présence d'un lexème commun est évidente pour les relations entre les termes d'un paradigme de ce type dans une langue donnée,

---

(+) J.G. Gardin utilise également la notion de paradigme mais avec une connotation différente /2/.

surtout dans une langue où, en général, le générique précède le spécifique comme le français ou le vietnamien.

En ce qui concerne les relations paradigmatiques sur le plan sémantique, l'association mentale engendrée par le contexte réel des référents donne lieu à des relations suivantes:

- Relation générique-spécifique (association par filation),
- Relation partitive (association par partition),
- Relations associatives,(association par analogie, par antonymie et par co-occurence).

4.2.2.3.1. Relation générique-spécifique.

C'est une relation d'ordre (inclusion). Soit :

$$Ga = \{Sa_1, Sa_2, Sa_3 \ldots\}$$

si $Sé_1 \cap Sé_2 \cap Sé_3 = Gé \neq \emptyset$

on aura:

$$Sa_1, Sa_2, Sa_3 \subset Ga$$

L'ensemble Ga est nécessairement définissable en compréhension, c'est-à-dire, par indication d'une propriété commune (en l'occurence, générique) à tous les éléments qui, à leur tour, peuvent constituer des sous-ensembles à part par affectation des propriétés spécifiques correspondantes, et ainsi de suite. De ce fait, l'intersection des Sé n'est autre que cette propriété commune qui définit dans le thesaurus une classe d'équivalence dont le réprésentant est Ga en tant que terme générique, et les éléments Sa_1, Sa_2, Sa_3... sont des termes spécifiques. Il est à noter que dans cette classe les spécifiques ne sont mutuellement remplacables qu'au niveau du générique.

Ex. LIVRE (+), PERIODIQUE (+), BIBLIOGRAPHIE

Sé/LIVRE/ ∩ Sé/PERIODIQUE/ ∩ Sé/BIBLIOGRAPHIE/= Sé/PUBLICATION/

Donc,

LIVRE, PERIODIQUE, BIBLIOGRAPHIE ⊂ PUBLICATION

(+) Il sera peut etre intéressant de signaler que LIVRE, PERIODIQUE et MANUEL sont polysémiques.

Une articulation plus fine nous donne, par example :

Sé /MANUEL/(+)∩ Sé/GUIDE/∩ Sé/ANNALE/∩ Sé/DICTIONNAIRE/
= Sé/LIVRE/

Cela veut dire que :

MANUEL, GUIDE, ANNALE, DICTIONNAIRE ⊂ LIVRE

Lors de l'indexation ou l'interrogation on peut choisir un terme générique si le terme spécifique désiré manque ou si l'on veut exprimer une notion plus générale. L'interrogateur a, en outre, la possibilité de choisir un terme plus spécifique s'il veut une notion plus restreinte. D'autre part, il faut tenir compte de la polyhiérarchie dans cette relation, parce qu'un terme spécifique peut, sous différents aspects, être inclu dans différents termes génériques. Dans ce cas, il faut préciser bien les champs sémantiques correspondants :

Ex. PETROLE (Pétrochimie)
PETROLE (Energétique)

4.2.2.3.2. Relation partitive (ou tout-partie).

Par opposition à la relation générique-spécifique, la relation tout-partie n'est pas une relation d'inclusion mais plutôt celle d'appartenance. Soit :

$$Ga = \{Sa_1, Sa_2, Sa_3 \ldots\}$$

si $Sé_1 \cap Sé_2 \cap Sé_3 \ldots \neq Gé$

on a

$$Sa_1, Sa_2, Sa_3 \ldots \in Ga$$

L'ensemble Ga est définissable seulement en extension, c'est-à-dire, par simple énumération des éléments qui n'ont autre propriété commune que l'appartenance à ce même ensemble, ce qui s'explique par le fait que l'intersection des Sé correspondants est différente du Gé.

Il ne s'agit donc pas d'une hiérarchisation du point de vue sémantique mais plutôt d'une association basée sur la décomposition physique du référent qui, de ce fait, n'est plus un "tout" mais plutôt l'ensemble des éléments qui en font "partie".

Ex. GUIDON, ROUE, PEDALE = ensemble des pièces d'une bicyclette,

Sé/GUIDON/ ∩ Sé/ROUE/∩ Sé/PEDALE/ ≠ Sé/BICYCLETTE/

4.2.2.3.3. Relations associatives.

Ces relations, par distinction des autres relations associatives qui sont génériques-spécifiques ou partitives, existent entre les signifiants par une association quelconque. Soient:

Sa_1 et Sa_2, tels que :
$Sé_1 \cap Sé_2 = Ré \neq \emptyset$
$Sé_1 \Delta Sé_2 \neq \emptyset$

4 Nous allons voir les relations engendrées par différentes associations.

4.2.2.3.3.1. Association par analogie.

L'intersection $Ré \neq \emptyset$ veut dire qu'entre les deux Sé il ya des sèmes communs qui rendent témoignage de leur similitude ou ressemblance, mais en meme temps leur différence symétrique $\Delta \neq \emptyset$ signifie qu'un certain facteur existe qui les différencie . En fonction d'une pondération donnée à Ré et à Δ respectivement, on peut déterminer s'il s'agit d'une relation de similitude, qui est transitive et de ce fait peut passer pour relation d'équivalence dans un contexte défini, ou d'une relation de ressemblance qui n'est pas transitive et reste toujours une relation associative banale (+).

(+) A cet égard, DEWEZE /2/ a recours à des sous-ensembles flous pour résoudre le problème, mais son approche utilise des instruments sophistiqués qui dépassent le sujet de cette étude.

Ex. BROUILLARD - FUMEE (relation de ressemblance)
DISCRETISATION - QUANTIFICATION (rel.de similitude)

4.2.2.3.3.2. Association par antonymie.

Dans la relation suggérée par cette association la différence symétrique $Sé_1 \Delta Sé_2$ représente un sème qui peut prendre l'une parmi deux valeurs contraires.

Ex. EMETTEUR - RECEPTEUR
CHARGE NEGATIVE - CHARGE POSITIVE

4.2.2.3.3.3. Association par co-occurence.

Dans cette relation l'intersection $Sé_1 \cap Sé_2 \cap \ldots = Ré \neq \emptyset$ peut suggérer une co-occurence temporelle ou spatiale qui implique deux (ou plusieurs) objets ou phénomènes :

Ex. PATIENT - MEDECIN - HOPITAL
ELECTRICITE - MAGNETISME

Quand $Sé_1 \cap Sé_2 = Ré \neq \emptyset$ signifie un phénomène, $Sé_1 \Delta Sé_2 \neq \emptyset$ peut en indiquer la cause et l'effet respectivement, dans ce cas nous avons une relation de causalité :

Ex. FEU - FUMEE
PESANTEUR - CHUTE

Les relations associatives peuvent aussi être établies en recherchant les co-occurences des termes dans un corpus.

Lors de l'élaboration d'un thesaurus, il convient de chercher à formuler toute relation associative possible en vue d'une meilleure recherche documentaire,"une réponse à une question pouvant toujours être trouvée à une notion apparentée /7/.

4.2.3. Présentation d'un thesaurus.

Un thésaurus est généralement composé de trois parties:

4.2.3.1. La liste alphabétique structurée.

C'est la partie principale, parfois unique, du thesaurus, dans laquelle les descripteurs et les non-descripteurs considérés comme leurs synonymes dans le système sont arrangés par ordre alphabétique.Entre les termes

qui y figurent il y a des renvois croisés en fonction des relations réciproques qui les relient.

Dans cette liste, chaque descripteur fait l'objet d'un article distinct présenté, en général, sous la forme suivante :

$$d_i \ [t_{i1}, t_{i2}, t_{i3}, t_{i4}, n]$$

où

d_i désigne le descripteur en question (souvent mis en évidence par l'emploi des caractères de haut de case /5/)

t_{i1} - l'ensemble en ordre alphabétique des non-descripteurs considérés comme synonymes de d_i. Tout élément en est précédé par l'abréviation EP (employé pour), ou plus récemment, selon /8/ par le symbole =,

t_{i2} - l'ensemble en ordre alphabétique des descripteurs liés à d_i par une relation hiérarchique (générique-spécifique ou tout-partie) qui est désignée par l'abréviation TG (terme générique), ou selon /8/ par le symbole $<$, placé avant chaque élément. Dans les thesauri où on fait la distinction entre les relations générique-spécifique et tout-partie, la première est marquée par l'abréviation TGG (terme général générique) et la deuxième par TGP (terme général partitif)/5/.

t_{i3} - l'ensemble en ordre alphabétique des descripteurs liés à d_i par une relation hiérarchique d'infériorité qui se traduit par l'abréviation TS (terme spécifique) ou TSG (terme spécifique générique) par opposition à TSP (terme spécifique partitif quand il s'agit d'une relation partitive, voir § précédent). Ces abréviations sont remplacées par le symbole $>$ proposé dans /8/,

t_{i4} - l'ensemble en ordre alphabétique des descripteurs liés à d_i par une relation associative autre que celles d'équivalence et de hiérarchie. Chaque élément y est précédé par l'abréviation TA (terme associé) ou par le symbole —— selon /8/.

n - l'ensemble de notes d'application servant principalement à limiter l'emploi du descripteur parindication du champs sémantique auquel il appartient ou à éliminer les homographies et polysémies.

N'importe quel ensemble mentionné ici peut n'avoir qu'un élément ou peut etre vide.

Dans cette liste les non-descripteurs sont renvoyés aux descripteurs correspondants par l'abréviation EM (employez), ou par le symbole ——→ /8/.

4.2.3.1. La liste de la hiéarchie (ou "carte sémantique")

C'est un système de champs sémantiques que sont de grandes divisions, facettes ou thèmes dans lesquels sont regroupés les descripteurs avec indication du niveau hiérarchique (par différents moyens soit un décalage vers la droite, soit des plots carrés en nombre croissant selon les niveaux ou des indices numériques et d'autres - voir annexe 1).

Par rapport à la liste alphabétique structurée qui n'indique les relations entre les termes qu'au niveau d'un descripteur, ce système offre une visualisation immédiate de la hiérarchisation au niveau d'un champ sémantique tout entier, ce qui facilite le choix des termes dans l'environnement de leurs relations réciproques.

La mise en évidence des relations hiéarchiques peut s'effectuer au moyen d'une présentation graphique que sont des schémas fléchés dont chacun représente un champ sémantique. Les relations y sont indiquées par la position relative des descripteurs et par des flèches qui les

tique. Les relations y sont indiquées par des flèches unissant les descripteurs dont les positions relatives au descripteur central (qui donne le nom du champ sémantique) représentent leur affinité respective à celui-ci. Les termes associés se trouvent en marge avec des renvois indiquant les numéros des champs sémantiques correspondants. Le schéma est en général divisé en carrés dont les co-ordonnées servent à renvoyer les descripteurs en question à la liste alphabétique. Quelques exemples des schémas fléchés sont donnés dans les annexes 2 et 3. Une autre variante des schémas fléchés se trouve sous forme circulaire (voir annexe 4).

4.2.3.3. La table des descripteurs et des non-descripteurs.

C'est une liste alphabétique des termes qui sert d'outil de transposition du langage naturel en langage documentaire avec des renvois à la liste structurée ou à celle de la hierarchie.

\+ \+

\+

CHAPITRE 5. REFLEXIONS SUR LES PARTICULARITES DE LA LANGUE VIETNAMIENNE A PROPOS DE L'ELABORATION D'UN THESAURUS

5.1. Les aspects morphologiques.

La langue vietnamienne est connue comme une langue monosyllabique, c'est-à-dire chaque mot y est composé d'un seul syllabe.

Ex. 1. /THAN/ (charbon)
/DI/ (aller)

5.1.1. Le mot et le lexème.

Cependant, cela ne semble vrai que si le mot est défini, dans l'écriture latinisée actuelle, comme une suite de caractères séparée par des blancs ou, dans l'écriture idéographique ancienne, comme un symbole à part (par exemple, /南 (NAM)/ = le Sud).

Une analyse morpho-sémantique plus fine révélera que dans la plupart des cas le mot ainsi défini n'est qu'une base lexicale (morphème lexical ou lexème).

Comme dans toute autre langue, un lexème peut lui-même constituer un mot entier (Ex.1), mais dans notre langue il est le plus souvent composé avec d'autres pour former des mots, et ce simplement par juxtaposition mais non par concaténation.

Ex. 2.

/THIÊN/ (ciel)
/VĂN/ (observer) → /THIÊN VĂN/ (observation du ciel) → /ASTRONOMIE/

/VẬT/ (choses)
/LÝ/ (raison) → /VẬT LÝ/ (raison des choses) → /PHYSIQUE/

Ainsi, dans notre lexique à côté des mots "tout faits"

monosyllabiques, se trouvent quantité de lexèmes disponibles à entrer en jeu pour former des mots qui sont en général des unités simples à deux syllabes. Les lexèmes, ces atomes polyvalents du lexique, ne se trouvent que très rarement en état libre sauf en tant que noms propres.

De ce fait, il s'ensuit que la langue vietnamienne n'est monosyllabique qu'au niveau du lexème mais non au niveau du mot comme l'écriture syllabique le donne à penser.

5.1.2. Formation des mots.

La disponibilité d'un bon nombre de lexèmes qui sont généralement polysémiques donne lieu à une grande capacité de formation de mots par différentes voies.

5.1.2.1. Mot simple monosyllabique (Msm) ;

L ⟶ Msm (Ex. 1)
L = lexème

5.1.2.2. Combinaison :

1. : L + L ⟶ Msb (mot simple bisyllabique) (Ex. 2)
2. : Msm + Msb ⟶ Mc (mot composé)

Ex.3 : /NHÀ + VẬT LÝ/ (physicien)

3. : Msb + Msb ⟶ Mc

Ex.4 : /VẬT LÝ + THIÊN VĂN/ (astrophysique)

5.1.2.3. Affixation :

Dans la langue vietnamienne un mot n'a qu'une seule présentation morphologique, les moyens de suffixation tels que déclinaisons, désinences, conjugaisons etc. étant inexistants sauf quelques suffixes de verbalisation et de nominalisation qui, toujours en raison du syllabisme, s'écrivent séparément de la base :

B + S ⟶ M (morphologiquement invariable)
B = base
S = suffixe
M = mot

Ex. 5.

B = /TỰ ĐỘNG/ (automatique/

B + S = /TỰ ĐỘNG HÓA/ (automatiser)
(automatisé(e),-(és),-(eés)
(automatisation(s)

Il n'y a donc pas de problème de réduction des variantes morphologiques à la forme de base (substantif singulier ou pluriel ou phrase nominale) comme dans les thésauri d'autres langues / /.

En ce qui concerne la préfixation, dans la langue française ou anglaise et d'autres, le préfixe s'attache à la tête de la base. Mais, en vietnamien, il la précède directement sans s'y coller or la syntaxe spécifie que l'attribut suivra toujours le mot qu'il qualifie (la langue étant du type "générique-spécifique", par opposition à l'anglais et à d'autres qui sont du type "spécifique-générique"), ce qui peut conduire à des ambiguïtés, sauf le cas où les préfixes ont un sens incomplet, comme /BÁN/ (semi-), /VÔ/ (in-), /TÁI/ (re-), /PHẢN/ (anti-) etc. Reprenons Ex.5 pour analyser l'équivalent vietnamien du mot francais /ASTROPHYSIQUE/:

/ASTROPHYSIQUE/ = /VẬT LÝ + THIÊN VĂN/

P + B P + B ?

P = préfixe ou M + A ? ; A = attribut

L'identification du rôle des composantes d'un mot est nécessaire à l'arrangement par thème des descripteurs dans un thesaurus. Dans la majorité des cas on peut trouver la solution en se basant sur le fait que presque tous les préfixes en langue vietnamienne sont monosyllabiques même s'ils sont dérivés des mots polysyllabiques.

Ex. 6. /SINH/ (la vie, levivant)
/HỌC/ (théorie, science) /SINH HỌC/ (science de la vie) ou (biologie)

/SINH/= P = /BIO-/

la formule P + B⟶M nous donne :

/SINH + TỔNG HỢP/ (biosynthèse)
/SINH + QUYỂN/ (biosphère)

Mais quand des règles euphoniques ou même sémantiques l'empêchent, au lieu du préfixe il faut utiliser la forme originale en tant qu'attribut:

Ex.7. /KỸ THUẬT + SINH HỌC/ ("technologie biologique") au lieu de (biotechnologie)

Parfois les deux formules co-existent:

Ex.8.

P + B /SINH TỔNG HỢP/
ou (biosynthèse)
M + A /TỔNG HỢP SINH HỌC/

Mais des ambiguités peuvent surgir : /LÝ SINH HỌC/ selon la règle aurait dû représenter la formule P + B ⟶ M qui donnerait l'équivalent français /PHYSICO-BIOLOGIE/, mais en réalité il signifie /BIO-PHYSIQUE/. Cela résulte du fait que /SINH LÝ/ a eu déjà un sens à part qui veut dire /PHYSIOLOGIE/. Par analogie à /LÝ SINH HỌC/ on a /HÓA SINH HỌC/ au lieu de /SINH HÓA HỌC/ qui n'a rien pour homographe. Dans ces cas on peut s'aider d'une analyse contextuelle ou sémantique pour situer les termes correctement dans leurs environnements sémantiques.

5.1.3. Règles d'écriture.

La combinaison des lexèmes (y compris les affixes) au sein d'un mot par simple juxtaposition sans autre signe lexical que le blanc posera la question de définition d'un mot pour le moins du point de vue informatique.

La concaténation semble apporter une solution au problème mais en même temps elle peut en créer d'autres. Par exemple, le mot /THỊ THÀNH/ (ville, urbain) deviendra sous forme accolleé /THỊTHÀNH/ qui peut s'interpreter comme /THỊT HÀNH/ (oignon et viande) (!)

Heureusement, de tels cas sont assez rares, mais l'inconvénient principal consiste à accoutumer les gens à des "mots-serpents", comme :

/QUÀNLẠCSINHĐỊAẢNHIỆTĐỚI/ (géobiocénose subtropicale)

Et cela sans compter les problèmes imposés par les signes diacritiques et les accents comme nous le verrons plus loin.

Le trait d'union donc présente une alternative plus acceptable sinon unique, mais dans ce cas, pour éviter des ambiguïtés il ne sera plus utilisé comme signe de concaténation à la fin des lignes d'un texte traité sur des matériels informatiques, ce qui n'implique pas de grandes difficultés car le plus long mot monosyllabique chez nous ne compte que sept caractères - /NGHIENG/ (incliné).

5.1.4. Les signes diacritiques et les accents.

Se basant sur les lexèmes monosyllabiques pour former son lexique la langue vietnamienne en a besoin un très grand nombre. A cet effet, les consonnes restant comparables en quantité à ceux dans d'autres langues, le système de voyelles s'est vu développer considérablement grâce à bien de diphtongues et même de triphtongues et, surtout, à des signes diacritiques qui donnent, par exemple, trois variantes du voyelle "a" /a, ă, â/ et trois du voyelle "o" /o, ơ, ô/. Cela donne lieu à une grande variété de syllabes. Mais étant donné que la plupart de leurs combinaisons sont à raison de deux par deux, sans compter des contraintes sémantiques et euphoniques, il faut un autre moyen pour les multiplier. On a donc recours à des accents qui représentent cinq tons différents, le ton normal non compris, pour la majorité des syllabes. Muni de différents accents un syllabe donne lieu à des lexèmes ou des mots ayant des sens totalement changés:

Ex. 9. /Mai/ (demain)
/MÀI/ (affûter)

/MÁI/ (toit)
/MẠI/ (vendre)
/MẢI/ (s'absorber)
/MÃI/ (toujours, sans cesse)

De ce fait, l'élaboration d'un thesaurus à utiliser sur des matériels informatiques se heurtera à des problèmes engendrés par ces moyens morphologiques qui ne se prêtent pas à une simple élimination sans causer de graves ambiguïtés, même au niveau de la phrase.

Ex. 10. /NHÀ MÁY CƠ KHÍ GIA LÂM/ (usine de construction mécanique de Gia lâm); débarrassée des signes diacritiques et des accents cette phrase devient /NHA MAY CO KHI GIA LAM/ qui peut se lire /NHÀ MÀY CÓ KHỈ GIÀ LẮM/ (chez toi il y a un bon vieux singe).

Bien que le contexte puisse y venir en aide dans certains cas, il faut en trouver des moyens de distinction réguliers. Une première solution, pratiquée parfois dans les PTT, consiste à substituer des lettres inexistantes dans notre alphabet (w, f, j, z...) et des consonnes qui ne se trouvent jamais à la fin d'un mot (s, x, q...) à ces signes et accents. Une deuxième pourra faire appel, au lieu de tels caractères alphabétiques, à des chiffres qui se laissent placer soit à la tête soit à la queue du mot. Une approche intermédiare aura recours aux caractères alphabétiques et numériques à la fois.

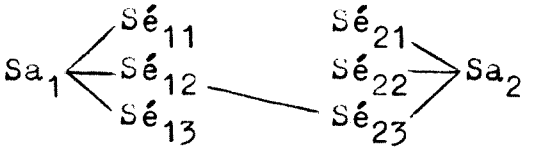
De toute façon, une solution plus ou moins satisfaisante du problème présuppose une analyse minutieuse et aboutira à des compromis inévitables des facteurs en jeu - on ne peut tout gagner sans rien perdre.

5.2. Des problèmes sémantiques.

5.2.1. La homographie et la polysémie.

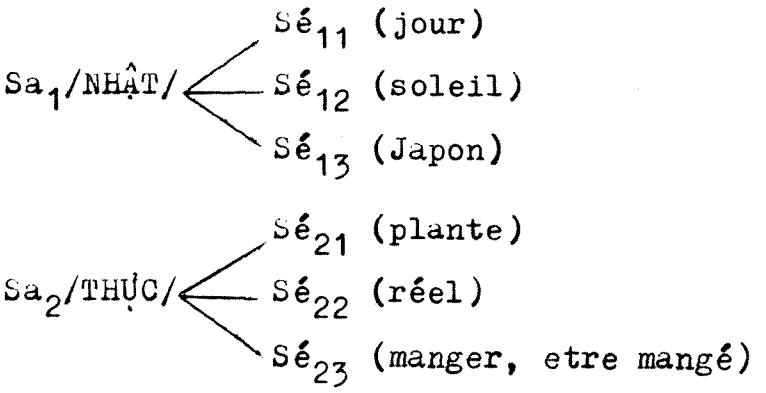
Etant donné des lexèmes monosyllabiques polysémiques et morphologiquement invariables en tant qu'unités struc-

turales du lexique il semble que la langue vietnamienne soit remplie de homographies et de polysémies. Mais, en réalité, cela n'est vrai qu'au niveau des mots monosyllabiques. Une fois entrés dans une combinaison acceptée (le plus souvent bisyllabiques) les lexèmes en question ne peuvent retenir qu'un de leurs sens qui est le plus approprié dans cette fusion. Reprenons les symbole utilisés dans les chapitres précédents pour en illustrer le mécanisme:



$$Sa_1 + Sa_2 = Sa_3 \longrightarrow S_{12} + Sé_{23},$$

Ex.11.



NHẬT + THỰC = NHẬT THỰC (soleil mangé, ou éclipse solaire)

D'autre part, nous pouvons retrouver /THỰC/ dans /THỰC TẾ/ (réalité),/THỰC VẬT HỌC/ (botanique) ou /THỰC PHẨM (denrée alimentaire), et /NHẬT/ dans /NHẬT HOA / (couronne solaire), /NHẬT TRIỀU/ (marée diurnale) ou /NHẬT THUỘC/ (domination japonaise).

Il faut donc, lors de l'établissement d'un thesaurus, ne tenir compte particulièrement que des termes monosyllabiques pour en débarrasser éventuellement des homographies et des polysémies par des moyens mentionnés dans les paragraphes 4.2.2.1.1. et 4.2.2.1.2.

5.2.2. La synonymie.

Tout comme dans les autres langues, les cas de synonymie véritable en vietnamien sont rares, qui s'observent surtout dans la langue courante et représentent des variantes linguistiques plutôt de caractère dialectal. Ils sont plus rares encore dans la terminologie scientifique et technique qui a été développée pendant la dernière vingtaine d'années où les moyens de communications sont assez répandus pour que l'on puisse se mettre en accord à propos des termes nouvellement introduits dans la littérature spécialisée. De toute façon, on peut y encore en trouver des exemples, particulièrement dans les domaines de connaissance qui sont en cours de pleine évolution comme, par exemple, l'informatique et la génétique.

Ex. 12.

/TIN HỌC/
/KHOA HỌC THÔNG TIN/ } /INFORMATIQUE/
/ĐIỆN TOÁN/

/ MẠCH TỔ HỢP/
/MẠCH TÍCH PHÂN/ } /CIRCUIT INTEGRE/
/MẠCH RẮN/

ou

/ĐỘT BIẾN/ } /MUTATION/
/BIẾN DỊ/

/
/ARN-thông tin/ } /ARN-messager/
/ARN-truyền tin/

A cet égard, le concepteur d'un thesaurus aura une double fonction. En faisant le choix des termes du langage

scientifique de sa spécialité pour les homologuer et structurer il contribue en même temps à l'unification et à la normalisation terminologiques au mileu de ses collègues qui seront des utilisateurs de sont oeuvre.

5.2.3. Les paradigmes sémantiques.

Comme il a été dit plus haut, les paradigmes sémantiques reflètent des relations logico-objectives existant entre les choses et les phénomènes dans l'univers réel. Ils sont donc en général indépendants des systèmes linguistiques dans lesquels ils sont exprimés, ce qui veut dire que, par exemple, un prunier est toujours un arbre fruitier bien qu'il pousse en France ou au Vietnam, ou la notion de fumée est associée à celle de feu n'importe qu'elle est exprimée en vietnamien ou en anglais.

De ce fait, lors de l'élaboration d'un thesaurus en langue nationale il convient de se référer à dee travaux existants en autres langues (de préférence, en l'occurence, celles où le générique précèdent le spécifique comme dans notre langue) pour établir des relations hiérarchiques et associatives entre les descripteurs choisis.

Néanmoins, à ce propos la laggue vietnamienne nous offre dans certains cas une convenance dont on peut faire l'usage en établissant des relations génériques-spécifiques. C'est la disponibilité, pour bien de catégories d'objets, des termes génériques qui précèdent tous les spécifiques contenus dans les notions respectives qu'ils représentent.

Ex. 12.

/CÁ/ (poisson)	:	terme générique
/CÁ CHÉP/ (carpe)	:	terme spécifique
/CÁ CHÍCH/ (sardine)	:	-id-
/CÁ THU/ (morue)	:	-id-

Ex. 14.

/MÁYY (machine)	:	terme générique

/MÁY KHOAN/ (perceuse) : terme générique
/MÁY MÀI/ (affuteuse) : -id-
/MÁY PHAY/ (fraiseuse) : -id-

De tels cas se rencontrent beaucoup plus fréquemment en vietnamien qu'en français ou en anglais, où on a, par example MACHINE A ECRIRE, MACHINE A COUDRE, ou PRINTING MACHINE, WASHING MACHINE ...

\+ +

\+

CONCLUSION

Nous avons essayé dans cette étude de présenter très sommairement l'indispensabilité d'un langage documentaire dans un système d'information et de documentation, les différents types de langages et leurs principes généraux de conception et d'utilisation. Il est clair que dans un travail de cette envergure il n'est possible que d'énumérer des problèmes plutôt que de les résoudre d'une manière plus ou moins satisfaisante.

De toute façon, si nous sommes arrivés, dans certaine mesure, à décrire les langages documentaires dans l'ensemble de leur évolution qui tend vers une capacité combinatoire et significative de plus en plus grande que représentent les thésauri, ce serait une contribution, bien que très modeste, à l'effort commun chez nous actuellement de créer de tels instruments à l'intention des documentalistes et des utilisateurs de l'information du pays, dans la perspective d'une informatisation ultérieure pour mieux servir le développement scientifique et technologique.

BIBLIOGRAPHIE

/1/. GUINCHAT, C; MENOU, M. : Introduction générale aux sciences et techniques de l'information et de la documentation; les Presses de l'UNESCO, 1981, 402p.

/2/. DEWEZE, A.: Réseaux sémantiques - essai de modélisation-application à l'indexation et à la recherche de l'information documentaire (Thèse de Doctorat de l'université en Sciemces Mathématiques), Université Claude Bernard, Lyon, oct. 1981, 451p.

/3/. MIKHAILOV, A.I et al.: Les fondements de l'informatique - en russe : Osnovy Informatiki; Maison d'Edition Nayka, 1968, 755p.

/4/ UNESCO : Principes directeurs pour l'établissement et le développement de thésaurus scientifiques et techniques monolingues destinés à la recherche documentaire; 1971.

/5/. ISO 2788-1974 (F) : Documentation - Principes directeurs pour l'établissement et le développement de thésaurus monolingues - 1. Norme Internationale ISO 271; 1974, p.1-14.

/6/. LONG, B. : Linguistique et indexation -"Documentaliste", vol.17, N3, mai-juin 1980, p.99-106.

/7/. LAUREILHE, M.T. : Le thésaurus - son role, sa structure, son élaboration; ENSB, 1981, 88p.

/8/. AFNOR : Norme francaise Z47-103 "Thésaurus monolingues et multilingues, symbolisation des relations", 1980.

ANNEXE 1

(extrait de /7/)

Le "Thesaurus INIS" de l'AGENCE INTERNATIONALE DE L'ENERGIE ATOMIQUE met des numéros indiquant le niveau, outre le décalage des termes vers la droite:

Exemple:

Rayonnement cosmique
- TG1 Rayonnements ionisants
  - TG2 Rayonnements
- TS1 Composante dure
- TS1 Composante molle
- TS1 Rayonnement cosmique solaire
  - TS2 Gerbes cosmiques
    - TS3 Grandes gerbes atmosphériques
- TA Détection des rayons cosmiques

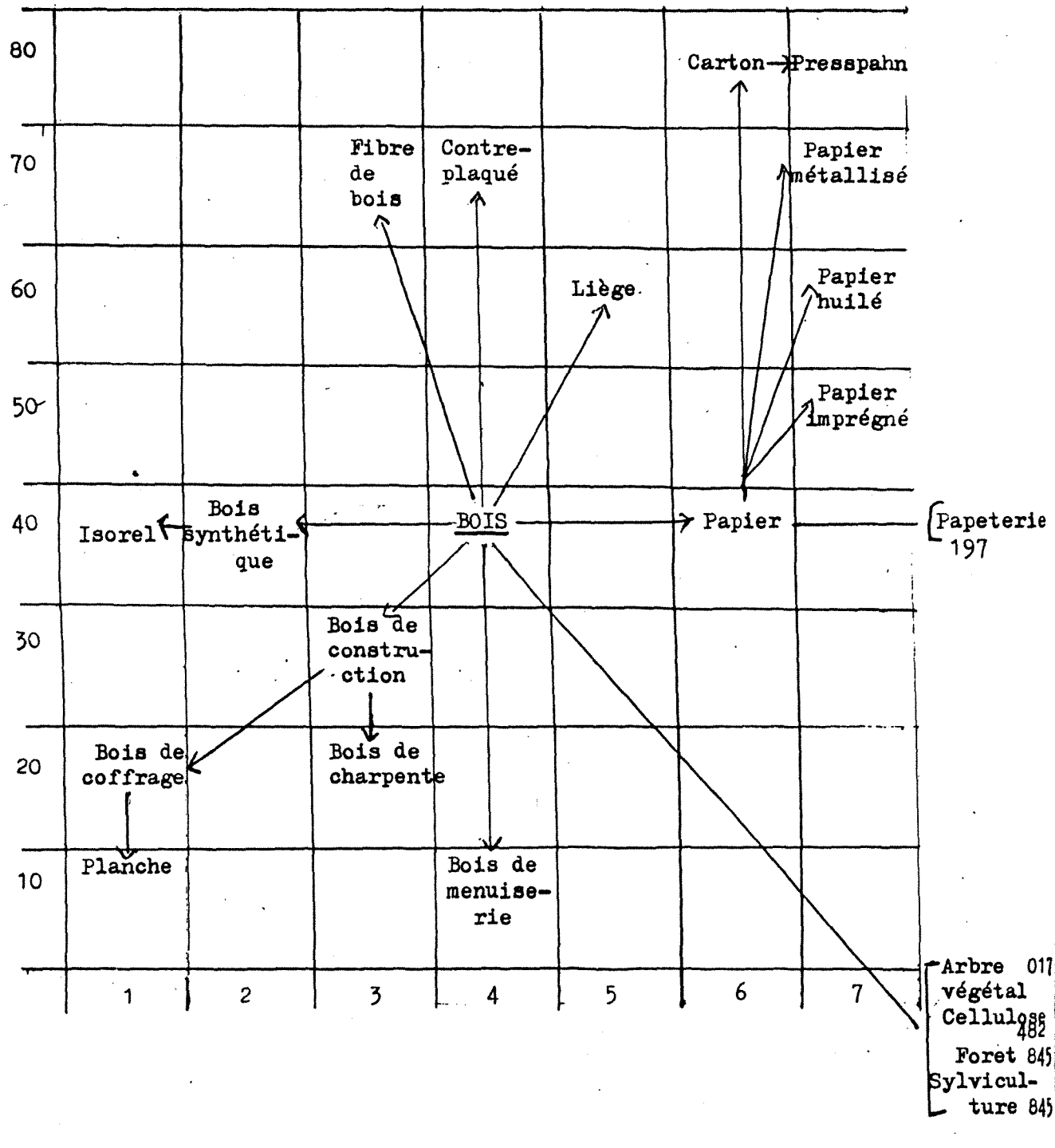
Ce procédé est assez souvent employé. Le "Thesaurus pour l'électricité et l'électronique"THESEE , emploie un autre procédé et indique la hiérarchie par des plots carrés en nombre croissant selon les niveaux.

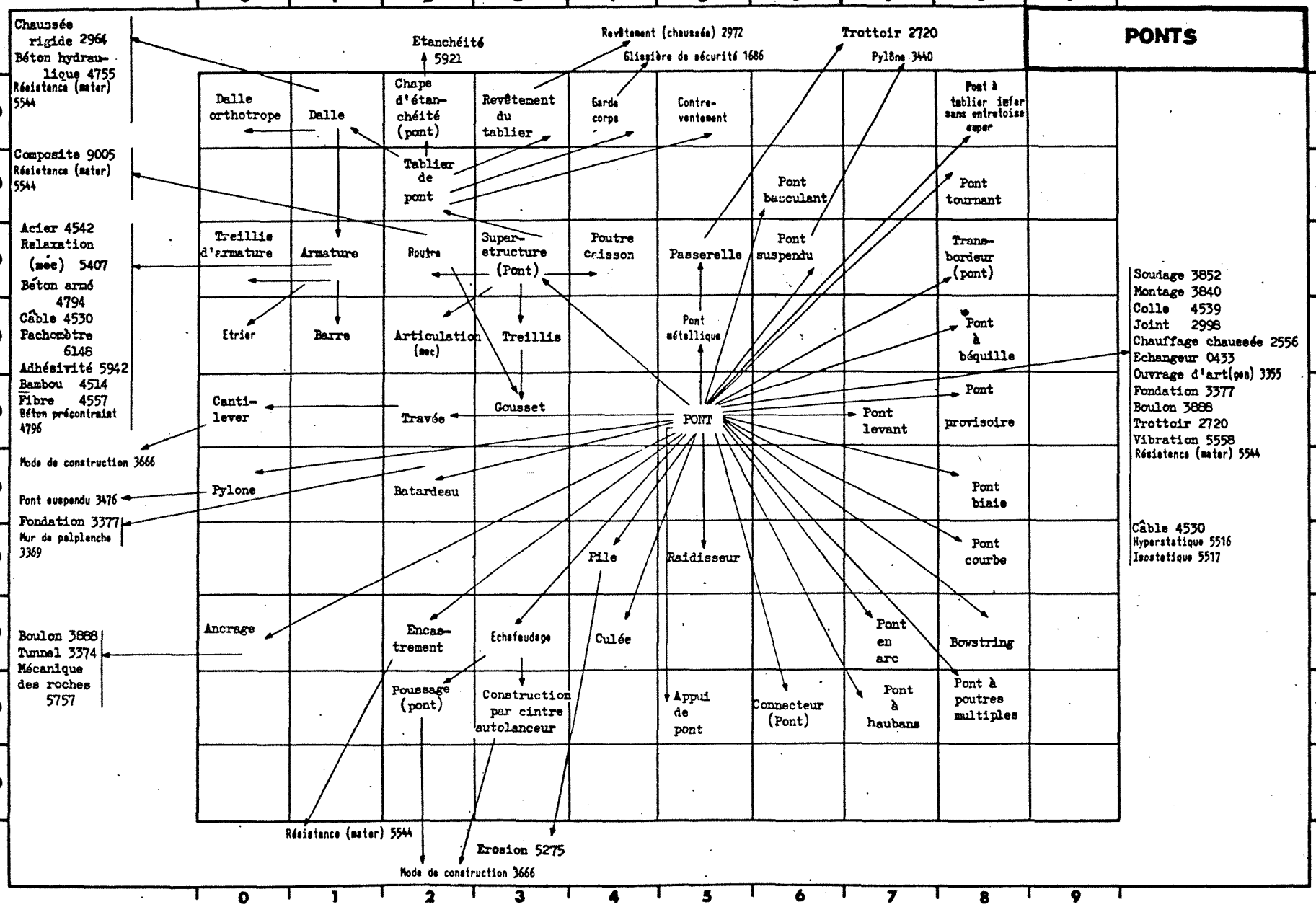
Aimant
- ■ Electroaimant
- ■ Aimant permanent
- ■ Electroaimant supraconducteur

Alimentation électronique
- ■ Convertisseur d'énergie électrique
  - ■ ■ Convertisseur alternatif-continu
  - ■ ■ Convertisseur continu-continu
  - ■ ■ Convertisseur de fréquence
    - ■ ■ ■ Diviseur de fréquence
    - ■ ■ ■ Multiplicateur de fréquence
  - ■ ■ Groupe convertisseur
  - ■ ■ Convertisseur statique

ANNEXE 2

Extrait du "Thesaurus ELECTRICITE DE FRANCE". Schéma 515 : BOIS.




PONTS
PONT
Chaussée rigide 2964
Béton hydraulique 4755
Résistance (mater) 5544
Composite 9005
Résistance (mater) 5544
Acier 4542
Relaxation (mée) 5407
Béton armé 4794
Câble 4530
Pachombtre 6146
Adhésivité 5942
Bambou 4514
Fibre 4557
Béton précontraint 4796
Mode de construction 3666
Pont suspendu 3476
Fondation 3377
Mur de palplanche 3369
Boulon 3888
Tunnel 3374
Mécanique des roches 5757
Etanchéité 5921
Revêtement (chaussée) 2972
Glissière de sécurité 1686
Trottoir 2720
Pylône 3440
Dalle orthotrope
Dalle
Chape d'étanchéité (pont)
Revêtement du tablier
Garde corps
Contre-ventement
Pont à tablier inf sans entretoise super
Tablier de pont
Pont basculant
Pont tournant
Treillis d'armature
Armature
Poutre
Super-structure (Pont)
Poutre caisson
Passerelle
Pont suspendu
Trans-bordeur (pont)
Etrier
Barre
Articulation (mec)
Treillis
Pont métallique
Pont à béquille
Canti-lever
Travée
Gousset
Pont provisoire
Pylone
Batardeau
Pont levant
Pont biais
Pile
Raidisseur
Pont courbe
Ancrage
Encas-trement
Echafaudage
Culée
Pont en arc
Bowstring
Poussage (pont)
Construction par cintre autolanceur
Appui de pont
Connecteur (Pont)
Pont à haubans
Pont à poutres multiples
Résistance (mater) 5544
Erosion 5275
Mode de construction 3666
Soudage 3852
Montage 3840
Colle 4539
Joint 2998
Chauffage chaussée 2556
Echangeur 0433
Ouvrage d'art(pas) 3355
Fondation 3377
Boulon 3888
Trottoir 2720
Vibration 5558
Résistance (mater) 5544
Câble 4530
Hyperstatique 5516
Isostatique 5517
0 1 2 3 4 5 6 7 8 9
90 80 70 60 50 40 30 20 10 00

ANNEXE 4

Extr. de LANCASTER.- Vocabulary control for information retrieval.

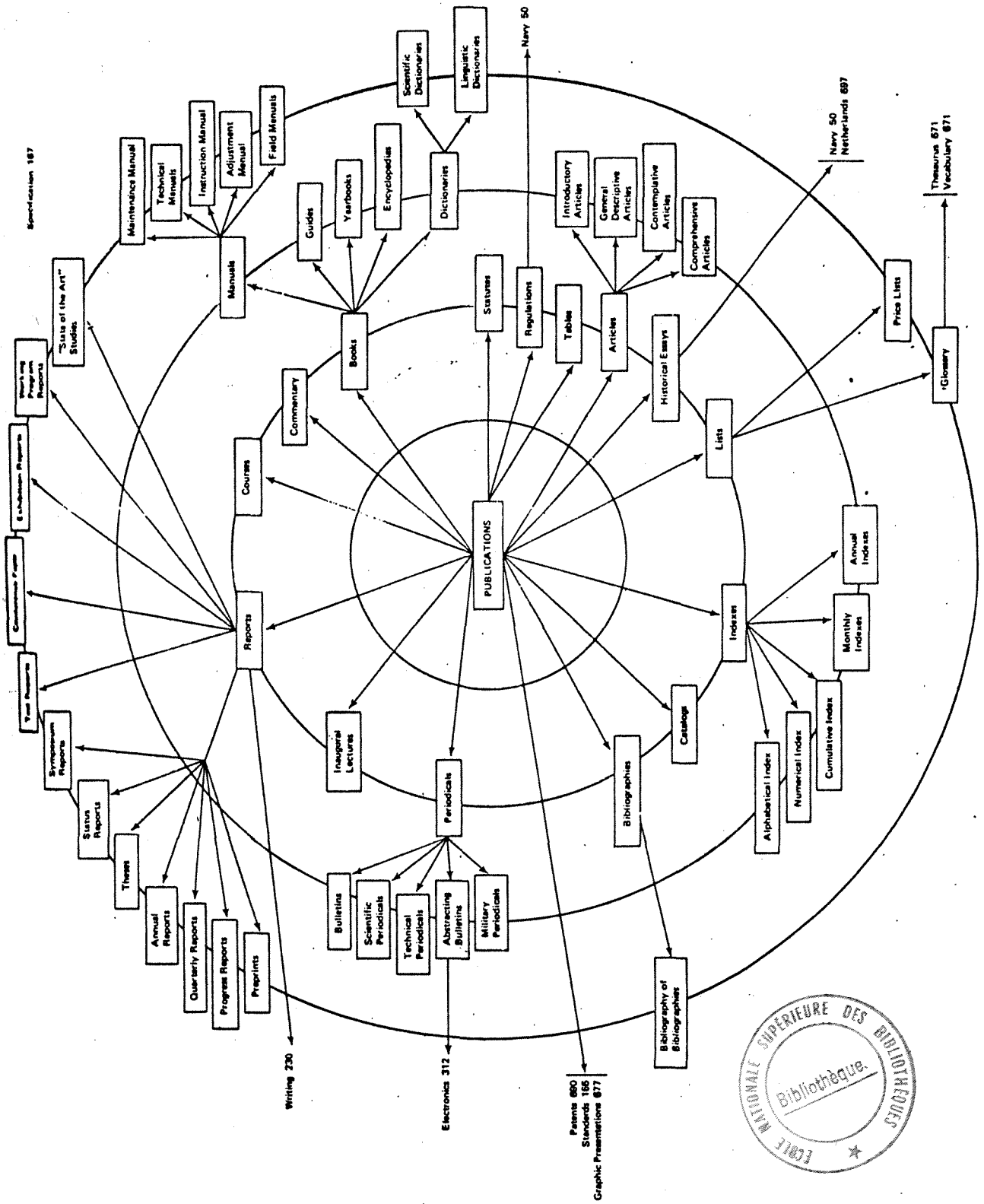
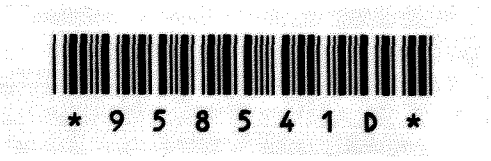


FIGURE 31 Specimen page from *TIDA Circular Thesaurus System.*


9585410