UNIVERSITÉ
LYON 1

D.E.S.S. INFORMATION SPECIALISÉE

BASES DE DONNÉES BIBLIOGRAPHIQUES

Mémoire présenté et soutenu
par QESTERI Serzh
sous la direction de
M DEWEZE André

LYON, Juillet 1984

UNIVERSITÉ
GRENOBLE II

UNIVERSITÉ
LYON I

D.E.S.S. INFORMATION SPECIALISÉE

BASES DE DONNÉES BIBLIOGRAPHIQUES

Mémoire présenté et soutenu
par QESTERI Serzh
sous la direction de
M DEWÈZE André

CPU

LYON, Juillet 1984

Introduction

Les problèmes liés aux bases de données bibliographiques sont nombreux et il est impossible de les aborder. Dans ce travail sont présentés les aspects les plus importants de la création et de la maintenance d'une base de données. L'accent est mis en général sur les problèmes théoriques en donnant les meilleures solutions. Les exemples sont pris du travail pratique fait pendant le cours. Pour la rédaction de ce mémoire ont contribué beaucoup les connaissances documentaires et informatiques données par les professeurs de divers modules. Le sujet qui couvre ce travail est très vaste, c'est la raison que les problèmes ne sont pas traités en profondeur.

Je remercie infiniment mon professeur et dirigeant de mon travail, M A. Deweze, qui a été toujours près de moi pour m'aider pendant la rédaction du mémoire. Ses conseils étaient indispensables pour la réussite de mon travail, et ils m'ont aidé non seulement pour la rédaction du mémoire, mais aussi pour mieux comprendre les problèmes présentés.

Juin 1984

Table des matières

BASES DE DONNÉES BIBLIOGRAPHIQUES

1. Définition et objectifs

La base de données bibliographiques est un ensemble de documents secondaires qui font référence aux documents primaires dans lesquels se trouve le texte integral. Si l'on lui pose une question, le système répond en donnant des références des documents primaires. Ces références comprennent le titre d'article, le revue où il se trouve, l'auteur, la date de publication, le nombre de pages, le résumé, etc. Les questions peuvent être posées à l'aide d'un mot-clé, du nom d'auteur, etc. Ainsi, les bases de données bibliographiques sont appelées aussi référothèques. La construction de ces bases de données est indispensable, parce que chaque année on publie dans les domaines scientifiques et techniques une immense masse de documents primaires et les utilisateurs ne sont pas capables de trouver ce qui les interesse dans la littérature de leur domaine, parue pendant plusieurs semaines et mois. Il est impossible d'accéder directement à un tel grand nombre de documents qui s'accumulent pendant des années, pour retrouver ceux qui traitent un certain problème. Et il faut savoir si ce problème n'est pas traité par des autres, car souvent on perd beaucoup de temps en abordant un problème qui est déjà traité.

Les objectifs d'une base de données bibliographiques créée à partir d'analyse de contenu des documents sont:

- d'informer périodiquement l'utilisateur de ce qui se publie de neuf dans son domaine
- de lui permettre de retrouver les documents qui portent sur un sujet donné.

On peut atteindre le premier objectif par un moyen classique, p.ex. publication de fiches bibliographiques, diffusion de bulletins analytiques. Une méthode récente, la diffusion sélective de l'information, consiste à faire parvenir périodiquement à l'utilisateur des fiches analytiques qui correspondent à son centre d'interêt. Cette diffusion est faite au moyen d'un ensemble de mots-clés ou descripteurs : le profil documentaire. La diffusion sur profil fournit des réfésences bibliographiques et non des informations finales; ces dernières sont des documents qui correspondent aux réfésences. Les products documentaires sont en général établis de façon standardisée pour tous les utilisateurs et reprennent l'ensemble des documents enregistrés pendant une période déterminée. La diffusion sélective consiste en l'extraction des seules nouvelles réfésences qui correspondent au profil d'interêt de chaque utilisateur ou de chaque groupe d'utilisateurs. Le profil est un ensemble de mots-clés, plus au moins structuré, qui décrivent les sujets intéressant l'utilisateur. Ces mots-clés sont comparés à ceux qui figurent sur les descriptions des documents. Lorsqu'il y a coïncidence les documents sont sélectionnés. Les résumés et les réféfences sont envoyés à l'utilisateur. L'utilisateur peut indiquer sur une fiche si le document l'intéresse, s'il désire

une copie, etc. La mise au point des profils nécessite l'intervention des spécialistes de l'information et la participation de l'utilisateur. Le service de la diffusion sélective peut être effectué manuellement pour un petit nombre d'utilisateurs, mais il est répandu avec la généralisation des bases de données bibliographiques lisibles par machine. Ce service fonctionne sur la base d'abonnements qui tiennent compte du nombre de bases bibliographiques utilisées et du nombre de mots-clés qui figurent dans le profil. Les références trouvées sont envoyées à l'utilisateur lorsque la base est augmentée (toutes les semaines ou tous les mois). La diffusion sélective est assurée à un utilisateur particulier (profil individuel) ou à un ensemble de personnes qui ont des intérêts communs (profil de groupe). Ainsi on peut définir quelques dizaines de profils de groupe qui couvrent l'essentiel des besoins d'information de tous les utilisateurs.

Le second objectif est atteint par des moyens traditionels, p.ex. fichiers manuels à accès par matières, bulletins d'index, etc. Maintenant on utilise une méthode récente, la recherche rétrospective dans les bases des données. Elle est une démarche qui consiste à explorer un ensemble de notices bibliographiques en vue d'extraire les références des documents qui répondent à des critères de recherche posés par l'utilisateur : auteur, sujet, matière, date, langue, type de document, etc. La recherche rétrospective peut être effectuée de la création de la base jusqu'à la dernière mise à jour (elle n'est pas différente alors de

l'exécution d'un profil) ou à partir d'une date déterminée. La démarche de la recherche peut être faite par un utilisateur, pour répondre à un préoccupation immédiate, ou par un documentaliste, qui réagit à une demande collective, actuelle ou prévue, d'information sur un sujet particulier. Cette recherche s'effectue en quelques étapes. D'abord, l'utilisateur qui a à résoudre un problème, prend connaissance des informations qui sont disponibles sur le sujet. Ensuite il formule la question, c.-à-d. énonce des concepts sur lesquels est recherchée l'information et traduit ces concepts en descripteurs. L'utilisation d'un thésaurus est indispensable pour réaliser la traduction (grâce aux relations d'équivalence) et pour étendre le champ de la recherche à tous les autres descripteurs qui peuvent caractériser des documents utiles (grâce aux relations hiérarchiques et associatives). Après la traduction, on construit une équation sur la base de la question: on établit des relations syntaxiques, p.ex. logique booléenne, pondération, voisinage. L'étape suivante est l'extraction des références. L'exploration du fichier de recherche se fait par comparaison des descripteurs de la question et de ceux des documents. Le filtrage des références est fait par l'utilisateur qui examine les informations et élimine ceux qui ne sont pas pertinentes.

2. Analyse de contenu

L'analyse du contenu est un ensemble d'opérations par lesquelles on décrit ce dont traite un document ou une question

et les produits qui résultent de ces opérations : indexation et résumé. Un document peut faire l'objet de plusieurs analyses à niveaux différents : l'attribution d'un indice de classification, l'indexation par une dizaine des descripteurs, le résumé en quelques centaines de mots. Cette analyse de contenu se situe lors de la production du document primaire (résumé d'auteur), ou avant le stockage de l'information ou lors de la recherche de l'information et de l'exploitation des réponses. Les objectifs que permet d'atteindre l'analyse sont : rendre compte du contenu des documents, opérer des choix tels qu'éliminer ou conserver un document, ranger matériellement les documents et les stocker pour retrouver (introduire dans le fichiers les références des documents sous les rubriques). L'analyse du contenu d'une question est faite pour préciser son champ, d'expliciter et de classer les sujets, de les exprimer en termes précis et de traduire ces termes en mots les plus appropriés du langage documentaire. L'analyse est utilisée parce que les documents originaux sont volumineux, et l'auteur et l'utilisateur n'emploient pas le même vocabulaire. Aussi faut-il avoir une coïncidence entre la formulation des questions et la représentation du contenu. Les différentes analyses du contenu sont basées en fonction de ces facteurs : nombre de termes, organisation de ces termes entre eux, précision ou spécifité de termes et nombre de documents primaires. Les produits plus fréquemment utilisés sont : la classification (indice pouvant correspondre à un descripteur), l'indexation (un ou plusieurs descripteurs liés ou non au moyen

d'une hiérarchie), le résumé (condensation du contenu dans un langage naturel), extraction de données (éléments naturels directement utilisables).

L'analyse du contenu doit présenter quelques qualités qui sont essentielles. Elle doit être pertinente, c.-à-d. représenter le document aussi bien et aussi complétement que possible. Elle doit être aussi précise, aussi peu ambiguë que possible pour satisfaire à la première règle, la pertinence. La cohérence est une qualité très importante, car on se trouve devant une pluralité d'auteurs et d'utilisateurs, et les analyses sont réalisées par une pluralité des personnes et parce que les informations doivent être utilisées aussi longtemps que possible. Ainsi, les descriptions doivent être stables et homogènes, et il faut que les mêmes notions ou objets soient exprimés de la même façon. Une autre qualité est le jugement: une analyse doit être objective ou neutre, on ne doit pas introduire dans la description des éléments qui ne figurent pas dans le document.

2.1. L'indexation

L'indexation est l'opération qui consiste à déterminer les termes les plus appropriés pour représenter le contenu du document. Ceux-ci sont exprimés avec le vocabulaire du langage documentaire (p.ex. thésaurus) et on les utilise pour construire les fichiers qui servent à la recherche. C'est une opération centrale pour le stockage et la recherche des informations. Les questions doivent être décrites dans les mêmes termes que les documents, car il faut comparer le contenu des deux ensembles pour trouver les documents qui répondent

à la question. Les produits de l'indexation sont des index, listes de termes significatifs, etc. Ils sont exprimés dans le document primaire ou dans les publications secondaires et aussi incorporés dans les fichiers, qui permettent de sélectionner les documents en fonction des sujets. Le niveau de l'indexation dépend des besoins et des possibilités du système et en général porte sur les sujets principaux. Elle peut être une indexation générique et peut identifier plusieurs sujets. Le plus fréquemment elle porte sur l'ensemble des sujets traités dans le document, identifiés par des termes généraux (indexation moyenne qui comporte une douzaine de descripteurs), mais aussi sur tous les sujets, décrits assez finement (plus d'une douzaine de descripteurs) qui est une indexation en profondeur.

Les opérations fondamentales de l'indexation suivent le modèle de l'analyse du contenu. Les plus importantes sont:
- la prise de connaissance du contenu du document qui est réalisée par une lecture rapide du résumé d'auteur, de l'introduction, des conclusions, des titres et de paragraphes
- le choix des concepts à représenter, fondés sur les règles: la sélectivité (il faut retenir les concepts pour lesquels le document apporte une information susceptible d'intéresser les utilisateurs) et l'exhaustivité (tous les concepts utiles doivent être retenus)
- la traduction des concepts choisis en descripteurs du thésaurus; on applique la règle de la spécificité qui peut être verticale et horizontale
- l'incorporation des éléments syntaxiques (liens, rôles, poids).

Pour l'indexation on utilise, en général, de 5 à 20 descripteurs (une moyenne de 8 à 12) et elle demande 5 à 15 minutes du travail selon la longueur du texte, sa complexité, la profondeur d'indexation et la familiarisation de l'indexeur avec la matière et avec le thésaurus. L'exhaustivité est conditionnée par la richesse du thésaurus, par le type d'indexation et par le comportement du documentaliste.

2.1.1. Indexation en langage naturel

Les langages naturels (parlés) ont des caractéristiques qui ne permettent pas les utiliser tels quels pour le traitement d'information. Ils sont adaptés à des modes de communication dans lesquels se fait un dialogue, auquel le temps et l'espace prennent une part importante. La recherche dans un fichier ou la diffusion des informations dans un produit documentaire sont des modes de communication différentes. Pour pouvoir être retrouvées rapidement et sûrement, les informations doivent être exprimées dans le plus petit espace possible et de façon non ambiguë. Les langages documentaires font une condensation et une simplification du langage naturel (langage contrôlé) en retenant une petite partie des mots et quelques formes, mais ils doivent préserver la richesse de l'information originale.

Cependant, le traitement de l'information en langage naturel est possible grâce à l'utilisation de l'informatique qui réduit le temps de recherche. Les systèmes se contentent des descriptions fournies par les auteurs, qui sont enregistrées et puis comparées avec les questions. Le langage naturel doit être assez précis, ce qui est le cas pour le lan-

gage scientifique et technique, mais moins pour les sciences sociales. Les types des mots qui composent un langage naturel n'ont pas la même valeur informative; il existe des substantifs, des adjectifs, des verbes, des adverbes, des conjonctions qui précisent les substantifs ou les mettent en relations entre eux. Les types des mots évoluent de façon propre, certains sont très fréquents et n'ont pas une grande valeur informative, d'autres sont des concepts centraux dans un domaine et peuvent s'appliquer à toute information. La grammaire associe les mots selon des règles pour exprimer les idées et elle se traduit par une transformation de certains mots: suffixes pour le pluriel, la personne et la mode des verbes, etc. Il existe des relations entre les mots, mais implicitement et on fait des efforts de les rendre explicites. On distingue certains types de relations.

Les relations hiérarchiques: un terme désigne un objet ou un phénomène particulier qui appartiennent au même ensemble et cet ensemble est désigné par un terme différent (il existe des relations générique/spécifique et tout/partie).

La polyhiérarchie: il existe des mots qui appartiennent à plusieurs ensembles, et certains mots sont dérivés du même mot ou radical et ont une relation hiérarchique ou de voisinage s'ils désignent des concepts qui appartiennent à des groupes différents.

La synonymie: les mots ont une relation d'équivalence (p.ex. un terme ancien et un terme nouveau, un terme vulgaire et un terme scientifique, etc., terme qui ont pratiquement la même signification).

L'antonymie : des mots qui expriment des concepts ou objets contraires.

La polysémie : des mots qui ont la même forme mais qui couvrent des réalités différentes ou une utilisation dans des domaines différents avec une signification particulière.

Relations de voisinage : des mots se rapportant à des objets ou à des phénomènes qui ont des relations communes.

L'indexation en langage naturel à l'avantage d'être très simple dans l'utilisation informatique, mais il présente des inconvénients, p. ex. le nombre d'entrées est très grande parce que il existe des formes nominales, adjectivales, verbales, etc. qui se réfèrent à la même notion, l'intercalation alphabétique des mots qui expriment d'autres notions, éclatement de mots composés après la suppression du trait d'union ou des syntagmes après l'élimination des mots vides, etc.

- L'antidictionnaire

L'antidictionnaire est un ensemble des mots qui ne sont pas significatifs (mots vides) pour effectuer la recherche rétrospective. Ils ne figurent pas évidemment dans le lexique des descripteurs et servent essentiellement lors de la création des fichiers inverses par traitement du texte (titre et résumé) en langage naturel. L'antidictionnaire est constitué à partir des catégories lexicales déterminées ou de certains éléments qui appartiennent à d'autres catégories lexicales. Les catégories lexicales déterminées sont : articles, pronoms, prépositions, conjonctions, adjectifs (indéfinis, démonstratifs, possessifs, relatifs), verbes auxiliaires. Certains formes lexicales en-

trent en conflict avec des formes de mots significatifs. Les autres catégories lexicales correspondent à des notions non significatives pour la recherche rétrospective, p. ex. des substantifs, verbes, adjectifs qualitatifs qui expriment des conditions opératoires. Souvent l'introduction de tels mots dans l'anticlictionnaire est inopportune : un substantif ou un verbe peut être considéré comme mot vide dans une phrase non significative et comme notion significative dans une autre.

2.1.2. Indexation en langage documentaire

Un langage documentaire est constitué de quelques éléments :

- Des mots qui servent à décrire les informations : les descripteurs qui sont tirés du langage naturel et réduits à une forme grammaticale unique et invariable (substantif singulier). Ils sont simples ou composés.
- Des mots du langage naturel qui ont des relations avec les descripteurs et qui sont répertoriés avec un renvoi au descripteur correspondant. Ils sont contrôlés par le langage documentaire et ne sont pas utilisés pour décrire les informations
- Des relations entre les descripteurs - relations hiérarchiques, d'équivalence ou de voisinage, qui permettent de regrouper les notions sous un terme, d'élargir ou de préciser une recherche. Les relations sont signalées par les codes : TG (terme générique), TS (terme spécifique), TS 1, 2, 3, ... (terme spécifique de niveau 1, 2, 3, ...), EM (employer : terme rattaché à un descripteur), EP (employé pour : les mots du langage naturel rattachés à un descripteur), VA (voir aussi : les descripteurs voisins).
- Des notations, numériques, alphanumériques, alphabétiques, sym

boliques, par des syllabes qui permettent d'identifier les descripteurs et de les faire figurer sous cette forme plus courte sur les notices bibliographiques et dans les fichiers.

- Des notes explicatives générales, qui précisent le sens dans lequel un élément du langage doit être employé, un descripteur ou un groupe de descripteurs.

- Des éléments de syntaxe, par l'ordre de présentation des descripteurs, par l'utilisation de mots ou signes qui permettent de les lier entre eux, par l'emploi d'une grammaire limitée.

- Des graphiques qui montrent les descripteurs et leur relations.

La liste des descripteurs figure toujours dans le langage. La présentation du langage se fait sous forme de document imprimé ou sous forme de fichiers lisibles par machine. Elle comporte :

- Une introduction avec quelques explications sur le contenu, l'organisation, les moyens employés et la façon d'utiliser le langage.

- Une liste des descripteurs : une liste alphabétique ou une liste systématique. Elles font apparaître, ou non, les relations entre les descripteurs. Les langages sont présentés de deux façons : la liste alphabétique vérifie l'existence d'un descripteur, la liste systématique son sens ou sa valeur en fonction de la catégorie à laquelle il appartient.

- Des représentations graphiques, qui montrent les relations entre les termes sur la base de la liste systématique. La perception du langage est facilitée beaucoup et elles pre-

nnent la forme des cercles concentriques, des cartes carrées, etc.

Les mots utilisés peuvent être des mots simples (langage uniterme), des mots simples et composés (plus fréquemment), des mots composés dans l'ordre du langage naturel ou dans l'ordre inversé dans les listes de vedettes-matière. Il existe des langages complètement hiérarchisés et précoordonnés (les classifications) et langages sans hiérarchisation ou à hiérarchisation discontinue et combinatoire (liste des descripteurs et thésaurus).

2.2. Le résumé

Le résume (ou la condensation) est l'opération qui consiste à rédiger une description du document qui permet de diminuer le volume et fait de ressortir les aspects qui intéressent l'utilisateur. Les résumes, comme produits, sont des textes courts qui accompagnent le document ou le remplacent. L'utilisation du résumé a pour objet: la diffusion de l'information, la sélection de l'information par l'utilisateur et la recherche de l'information. Il existe plusieurs types de résumés qui se différencient par leur longueur, le détail plus au moins grand dans lequel est présenté le contenu du document, la présence ou l'absence d'appréciation ou de critique, le fait que le document est considéré dans la totalité ou dans des aspects qui intéressent l'utilisateur, le fait que le résumé est rédigé par l'auteur ou une autre personne, le langage utilisé (p.ex. langage naturel ou langage artificiel). Le types principaux du résumé sont:

- La notation du contenu est un titre amélioré qui com-

porte l'ensemble des thèmes qui ont servi à indexer le document (10 à 50 mots).

-Le résumé signalétique est court et indique de quoi traite le document (50 à 200 mots).

-Le résumé analytique est un peu plus long et indique non seulement de quoi traite le document, mais présente aussi les principales conclusions et les informations significatives (100 à 500 mots).

-Le résumé critique fournit, en plus d'une analyse, un commentaire du document, mais il est très rare en documentation.

-Le digest est un résumé assez long et peu utilisé en documentation (il présente de 10 à 20% du texte).

L'état de la question (the state of art) résulte de la synthèse de plusieurs documents qui portent sur le même sujet et il peut être rédigé par le service de documentation.

La rédaction du résumé peut être effectuée selon ces modalités :

- Vocabulaire et syntaxe libre : dans ce cas on choisit les mots et les structures de phrases sans contrainte ; on peut extraire des phrases du texte à résumer.
- Vocabulaire contrôlé et syntaxe libre. Dans ce cas on doit exprimer les concepts essentiels au moyen de descripteurs du thésaurus et les lier par des mots vides ; on est libre de choisir la structure de phrases.
- Vocabulaire et syntaxe contrôlés : les mots utilisés, même les articles et prépositions, proviennent du thésaurus de descripteurs et d'anticlictionnaire ; les phrases sont construites suivant un nombre de structures canoniques. Cette mé-

thode permet de lier les descripteurs de façon significative et de procéder à une traduction automatique du résumé.

Le contenu essentiel du résumé est constitué par la synthèse du document et il indique quels sont le sujet ou les sujets traités, la nature du document, le but du travail, les méthodes employés, les résultats obtenus, les conclusions d'auteur, le lieu, la date et les circonstances du travail et quelquefois une appréciation de l'importance du document. La lecture du résumé permet à l'utilisateur de connaître avec précision le document et de déterminer s'il lui est nécessaire de se reporter au document original. Souvent le résumé peut remplacer le document pour obtenir des informations élémentaires, p.ex. sur un document produit dans une langue étrangère.

Le résumé doit présenter quelques qualités:

- La concision, indépendemment de la longueur du résumé. On évite les expressions et les périphrases remplaçables par des mots, mais pas aux dépens de la précision. On évite aussi de vider la description de son substance en employant des termes et des phrases très générales, qui condensent le texte, mais peuvent s'appliquer à un autre. On doit utiliser des expressions exactes et spécifiques.
- Le résumé doit se suffire à lui-même: la description doit être complète et intelligible sans avoir besoin de se référer à un autre document.
- L'objectivité: il ne faut pas accepter les interprétations ou appréciations personelles de l'auteur. Le document doit être décrit tel qu'il est, en fonction des besoins de l'uti-

lisateur. Dans le résumés critiques, les éléments objectifs doi
vent être présentés explicitement. Il ne doit pas être vide,
c.-à-d. de ne pas être une paraphrase du titre.

La langue des résumés est en général celle d'auteur
du document primaire. Il convient donc, soit de le conver
tir (traduire) dans la langue acceptée comme langue de
travail pour la base de données que l'on veut créer, ou
les conserver dans la langue d'origine si celle-ci est
considérée comme suffisament connue (p. ex. anglais et
français). On ne doit pas employer la première persomne
et l'expression doit être claire. Les termes doivent être
intelligibles pour les utilisateurs; en particulier, les ab-
réviations et les symboles doivent être ceux qui sont d'usa
ge courant et bien établis. La présentation matérielle
du résumé suit des règles strictes, parce que il est inclu
dans des publications ou introduit dans des fichiers. Le
résumé demande beaucoup de temp (un quart d'heure à
une heure par document) et il est l'opération la plus
coûteuse de la chaîne d'enregistrement (il représente 50%
du coût). Il y a intérêt, partout où c'est possible, de re-
prendre les résumés d'auteurs ou les résumés publiés pa
des sources secondaires.

Exemple de résumé d'un document

Titre : Automatisme et documentation

La documentation automatique fait immédiatement associe
documentation à ordinateur et informatique. On montre ici
qu'il existe des équipements apportant des solutions possi-
bles et offrant de nombreuses utilisations de l'automatis-

me pour aider au traitement de l'information. On présente divers matériels pouvant répondre à trois domaines des techniques documentaires faisant appel à des moyens automatiques : la sélection et la recherche d'informations (appareil Miracode et le filesearch), la recherche d'images sur support micrographique (système CARD, COMPTCARD, MTC) et la transmission à distance des images (système Sanders Diebold) permettant dans certains cas, d'obtenir une copie papier du document reçu sur l'écran.

3. Le thésaurus

Le thésaurus d'un langage documentaire présente les unités sémantique qui composent ce langage et indique les relations sémantique entre ses unités. On distingue trois sortes d'unités sémantiques :

- les descripteurs (ou mots-clés, mots-vedettes, vedettes-matières) sont des mots ou des expressions qui désigent des concepts lesquels peuvent caractériser le contenu du document et de la question, et qui sont utilisés dans le système
- les non-descripteurs (ou termes interdites, termes équivalents, synonymes, etc.) sont aussi des mots ou des expressions qui désignent des concepts dans le langage naturel, mais on a décidé par convention de ne pas les utiliser pour caractériser les documents et les questions ; les concepts correspondants sont représentés par les descripteurs qui ont le sens le plus proche
- les mots-vides (ou termes non significatifs : leur liste est l'antidictionnaire) ne sont pas des concepts et ils n'apportent pas de signification dans les documents ou les questions.

Cette liste des unités sémantiques dans le thésaurus délimite le vocabulaire utilisé pour représenter le contenu du document et de la question dans le système.

3.1. Les relations sémantiques

L'objet des relations sémantiques est de préciser le sens des descripteurs et d'aider l'utilisateur à trouver les descripteurs adéquats qui représentent les documents et les questions. Il existe deux grands types de relations sémantiques:
- les relations qui lèvent les ambiguïtés du langage naturel et traduisent ce-ci en langage documentaire (avec les descripteurs)
- les relations qui enrichissent la représentation des documents et des questions en trouvant les descripteurs les plus adéquats.

Les relations de traduction aident à resoudre divers problèmes, p.ex. de synonymie (un même concept est exprimé dans la langue naturelle de multiple façons: regroupement sur un seul descripteur), de polysémie (un même mot exprime plusieurs concepts: dissociation dans la représentation du document). Ils sont quatre: trois lèvent les polysémies et une les synonymies.

- Relations d'inclusion. L'ensemble des disciplines dans un thésaurus est subdivisé en un nombre de groupes (domaines) qui contiennent un nombre limité des descripteurs. L'appartenance du descripteur à un groupe permet de préciser son sens. On marque cette relation par l'indication "dom" (domaine) entre la désignation du descripteur et celle du groupe. Il existe deux types principaux de groupes: les thèmes (les descripteurs sont groupés par matières) et les facettes

(les descripteurs sont groupés par catégories linguistiques).

- Relations de qualification. Si deux concepts polysémiques figurent dans un thésaurus, on les distingue par un qualificateur qui précise leur sens; le qualificateur est mis entre parenthèses.

- Relation de définition (note d'application). Si un concept est exprimé par un descripteur qui n'a pas un sens précis, on détermine ce descripteur par une définition, en utilisant la mention déf, et elle ne fait pas partie de l'énoncé du descripteur.

- Relation d'équivalence (préférentielle, de substitution). Elle a pour but de lever les ambiguïtés de la synonymie et elle aide à traduire les concepts de la langue naturelle en descripteurs. L'équivalence est une relation réciproque et se marque par l'indication : EM (employer), VOIR, UTILISER placée entre le non-descripteur et le descripteur, ou E.P. (employé pour), SYN, EQUIV, UTILISE POUR placée entre le descripteur et le non-descripteur. Il existe plusieurs cas d'équivalence : synonymie vraie, nom de marque, sigle, termes équivalents (quasi-synonymes), termes antonymes, orthographes différentes, jargon, etc. Le choix du descripteur et des non-descripteurs se fait sur base de la fréquence d'utilisation dans le langage naturel: le descripteur est le terme le plus couramment utilisé. Le non-descripteur renvoie à un seul descripteur, tandis que le descripteur couvre zero, un ou plusieurs non-descripteurs.

Les relations d'enrichissement son deux et elles concernent les descripteurs seulement.

-Relations hiérarchiques montrent la hiérarchie des descripteurs dans les disciplines d'un thésaurus. Cette hiérarchie est souple, adaptée aux points de vue des spécialistes des disciplines et à leurs changements. La polyhiérarchie est aussi admise (un descripteur est rattaché à deux ou plusieurs autres qui sont supérieurs). Les relations hiérarchiques sont réciproques et signalées par : T.S. (terme spécifique) qui se situe entre un descripteur générique et son spécifique ; T.G. (terme générique) qui se situe entre un descripteur spécifique et son générique. Il existe deux types de hiérarchie : relation générique (générique / spécifique) signalée par T.G.G. et relation partitive (tout / partie) signalée par T.G.P.

-Relations associatives (d'affinité). Ces relations lient des descripteurs dont le sens est voisin, mais il n'est pas possible de créer une relation d'équivalence. Cependant, une question sur un descripteur peut impliquer la recherche des documents indexés à l'aide de l'autre. Elles sont réciproques, comme les relations hiérarchiques, mais elles sont symétriques aussi. Elles sont signalées par V.A. (voir aussi) placé entre les deux descripteurs. Elles expriment des relations de causalité, d'instrumentation, d'origine ou de destination, entre une qualité et sa mesure, de similarité, de distinction, etc.

3.2. Présentation du thésaurus

Il existe plusieurs façons de représenter une liste d'unités sémantiques (descripteurs et non-descripteurs) et leurs relations sémantiques (équivalences, hiérarchies et associations).

-Le dictionnaire (liste alphabétique complète) : les descripteurs et les non-descripteurs sont rangés alphabétiquement. Sous cha-

que descripteur figure la liste complète de ses relations sémantiques et sous les non-descripteurs se trouve le renvoi au descripteur correspondant. Les relations sémantiques suivent un ordre constant, p.ex. descripteur, un qualificateur, suivi de définition, domaine, équivalence, hiérarchie, association. Les descripteurs sont classés alphabétiquement à l'intérieur de chaque groupe relationnel. Les différents niveaux hiérarchiques qui apparaissent dans le thésaurus sont indiqués par un décalage vers la droite.

-Le lexique est une liste alphabétique simple des descripteurs, sans leurs relations sémantiques, mais le plus souvent les non-descripteurs sont intégrés avec leur renvoi aux descripteurs correspondants.

-Le lexique permuté : chaque mot constructif d'un descripteur ou d'un non-descripteur (mots simples ou expressions composées) est une entrée distincte dans la liste alphabétique. La permutation se fait sur les termes significatifs seulement. Dans le lexique permuté les relations sémantiques d'un descripteur sont présentées une seule fois dans l'entrée principale.

-Le lexique par groupes : le lexique est éclaté par champs sémantiques (thèmes ou facettes); dans chaque groupe les descripteurs sont rangés dans l'ordre alphabétique.

-Schémas fléchés : les listes par champs sont représentées sous forme des schémas fléchés. Ils permettent de visualiser les descripteurs d'un secteur et leurs relations sémantiques à plusieurs niveaux.

-La liste systématique : les seuls descripteurs figurent dans l'ordre qui correspond à une hiérarchie des concepts suivant

le point de vue des auteurs de classification. Il n'existe pas de relations associatives et d'équivalence.

Le dictionnaire, le lexique par groupe et le schéma fléché sont les modes les plus utilisés pour le thésaurus; la liste systématique et le lexique pour les plans de classification.

Indexation des documents

Les liaisons entre les descripteurs qui sont utilisés pour indexer un document varient suivant le système et on peut distinguer :

- la juxtaposition, la liaison la plus simple, mais la plus utilisée, consiste à enregistrer l'un derrière l'autre les descripteurs choisis
- la distinction entre deux groupes des descripteurs : les descripteurs principaux (caractérisent le contenu principal) et des descripteurs secondaires (représentent les concepts moins importants)
- la pondération est une distinction plus profonde qui consiste à attribuer à chaque descripteur un poids (coefficient de pondération : la variation 1 à 3 par exemple) qui marque l'importance du chaque concept dans le document indexé.

4. Structure de fichiers

Après les opérations de description et de représentation de contenu, on obtient une notice bibliographique, ou bordereau d'analyse, qui contient les données caractéristiques du document. Celles-ci sont classées chacune dans une zone particulière : titre, auteur, source, langue, descripteurs, résumé, etc. Organiser les fichiers c'est saisir l'information de la notice, la contrôler et la mettre en forme pour permettre l'organi-

Exemple de thésaurus : entre alphabétique avec liaisons

Analyse documentaire
  TG Enregistrement documentation
Banque de données
  TG Ordinateur
  TS Base de données
Base de données
  TG Banque de données
Bibliographie
  TG Information
Bibliothèque
  TG Documentation
Bibliothèque de Clamart
  TG Documentation
Bibliothèque infantine
  TG Documentation
Centre de documentation
  TG Documentation
Documentation
  TG Information
  TS Bibliothèque
    Bibliothèque de Clamart
    Bibliothèque infantine
    Centre de documentation
    Serveur documentaire
Enregistrement documentation
  TG Information
  TS Analyse documentaire
Europe occidentale
  TG Régions et pays
  TS Portugal
Formation
  TS Pédagogique (Méthode)
    Scolaire (Formation)
Fortran IV.
  TG Programmation (Langage)
Indexation
  TG Information
Information
  TS Bibliographie
    Documentation
    Enregistrement documentation
    Indexation
    Langage
    Lecture
    Recherche documentaire
Langage
  TG Information

sation des produits documentaires.

La saisie de la description bibliographique et de la représentation du document s'effectue en recopiant ces données sur un support, p.ex. on frappe les identificateurs de zone et le contenu des zones sur le clavier d'un terminal d'ordinateur. L'enregistrement est transmis à l'ordinateur soit immédiatement (traitement en ligne), soit après saisie d'un certain nombre de notices (traitement par lots).

Le contrôle effectue plusieurs vérifications, p.ex. la séquence de données, présence des identificateurs de zones et de notices, présence obligatoire ou facultative d'une donnée dans une zone, type de données dans chaque zone (numérique, alphanumérique, alphabétique), longueur d'une donnée de longueur fixe, appartenance de certaines données à une liste préétablie (thésaurus, type de document, langue, etc.). Certaines informations sont incorporées systématiquement à la notice par l'ordinateur, p.ex. date d'enregistrement, données communes à plusieurs notices, numéro d'ordre, identificateurs de zone. Le contrôle se fait par ordinateur qui émet le message d'erreur pour chaque faute, soit en temps différé (quelques heures ou quelques jours après la saisie), soit en temps réel (immédiatement après la saisie sur terminal en ligne avec l'ordinateur). Le contrôle se termine par la correction des erreurs; dans le cas d'une saisie en temps différé au cours d'une deuxième opération de saisie, dans le cas d'une saisie en temps réel au cours de la saisie initiale (le mode interactif: un dialogue entre l'opérateur et l'ordinateur, qui signale les erreurs

et qui sont corrigés par l'opérateur aussitôt).

Au niveau de l'exploitation documentaire on peut distinguer deux types d'information sur une notice : les données (critères) de recherche (les éléments sur lesquels porte la recherche des notices : auteurs, matières, etc.) et les données d'accompagnement (éléments sur lesquels la recherche ne peut pas porter, mais qui sont fournis après la recherche sur les critères : titre, source, etc.). Les critères de recherche et les données d'accompagnement varient selon le fichier d'exploitation : dans un fichier auteur le critère de recherche est le nom d'auteur et l'accompagnement le reste de la notice ; dans un index matières, les critères sont les descripteurs et l'accompagnement le reste de la notice. La mise en forme constitue les couples critère de recherche - accompagnement et organise ces couples en fichiers. Elle comporte quatre phases :

- extraction des critères de recherche : auteurs, descripteurs, organismes, alates, etc.
- duplication de l'accompagnement qui est propre à chacun des critères
- tri sur les critères de recherche pour obtenir les fichiers d'exploitation : index (auteurs, matières, etc.), bulletin (signalétique, analytique), diffusion sélective sur profil
- fusion avec les fichiers enregistrés pendant les saisies précédentes pour obtenir les fichiers cumulés : recherches rétrospectives, index cumulatifs.

4.1. Fichier de saisie

Le fichier de saisie (fichier direct) permet la mise à jour de

la base de données bibliographiques. Les documents sont représentés dans ce fichier par des enregistrements. Chaque enregistrement est constitué par un nombre de zones. Ces zones permettent d'effectuer quelques opérations: édition d'un bulletin analytique, production d'un index auteurs et mots-clés pour ce bulletin, recherche en ligne dans la base de données à partir de clés d'accès (numéro de fiche, mnémonique de revue, auteur, mots-clés), visualisation de l'enregistrement des documents dans une recherche rétrospective. Dans la grille d'enregistrement sont présentées les différentes zones qui constituent en général un enregistrement du fichier. On peut représenter la structure de l'enregistrement de façon conventionelle. La zone est déterminée par sa position dans l'enregistrement (le premier nombre) et sa longueur (deuxième nombre). Cette grille sert à rédiger le bordereau de saisie et à visualiser les documents. Un exemple de grille d'enregistrement et des commentaires sur les problèmes à longueur fixe et de la limitation du nombre de descripteurs se trouvent dans l'annexe 1.*

4.2. Fichier thésaurus

Du point de vue informatique, un thésaurus est composé par des sous-fichiers, entre les zones desquels existent des relations. Ces sous-fichiers sont p. ex. le lexique alphabétique des descripteurs (lexique thésaurus), le fichier des relations hiérarchiques (générique/spécifique, tout/partie), le fi-

* Pris de l'article "Création et maintenance de bases de données bibliographiques" de A. Dewèze.

chier des synonymes, le fichier des relations associatives, etc. Le fichier du lexique alphabétique des descripteurs comprend le code des descripteurs et le libellé de ceux-ci. Le codage dans le fichier des relations hiérarchiques se fonde p. ex. sur la représentation de la position des descripteurs sur une arborescence. Il comprend le code d'arborescence, la position sur l'arborescence et le code du descripteur. La position est indiquée par un nombre des groupes de deux chiffres, égal au nombre de niveaux sur l'arborescence. Les relations entre notions peuvent être décrites par paire de code termes dans le sens TG → TS et TS → TG. Aussi la relation de synonymie et la relation associative peuvent être enregistrées par paire de code termes.

- Mise à jour du thésaurus

Le thésaurus est utilisé comme aide à l'indexation des documents lors de la saisie et des questions lors de la recherche. Il doit être évolutif pour rendre compte des aspects nouveaux des certaines notions. Les éditions sur papier sont onéreuses et c'est pourquoi les thésaurus peuvent être gérés en temps réel.

- Contrôle des chaînages hiérarchiques entre descripteurs

L'indexation peut être, par exemple, à trois niveaux: DOM - qui correspondent à une table des matières, TG - termes génériques (ou mots-clés thématiques) et TS - termes spécifiques qui correspondent à des concepts inclus dans les concepts décrits par les termes génériques.

La relation d'inclusion TG ⊂ DOM est monohiérarchique, chaque TG est inclus dans un seul domaine. La relation

TS ⊂ TG peut être polyhiérarchique. Dans la table TG chaque libellé est accessible par un index et une zone indique le domaine d'inclusion. Lors de la rédaction du bordereau de saisie, on attribue à chaque TS un index prélevé dans la table des termes génériques et cet index est enregistré dans la zone des mots specifiques. La relation explicite TS ⊂ TG ainsi créée est utilisée pour élargir une question lors de la recherche rétrospective.

- Traitement de la polyhiérarchie

Le thésaurus n'est pas distinct de la base de données, car chaque enregistrement de la base contient un petit partie de thésaurus grâce au chaînage TS–TG (par intermédiaire des index de termes génériques). La mise à jour du thésaurus est dinamique est elle présente des inconvénients: établissement des liens inappropriés entre TS et TG et prolifération de polyhiérarchie entre un TS et plusieurs TG. Le premier résulte de l'erreur humaine et est corrigé par l'examen périodique des listes. Le deuxième est inévitable et il convient de circonscrire dans les limites raisonnables. Le lien d'une notion spécifique vers une notion générique peut évoluer. Si l'on propose des aiguillages, il faut que ceux-ci ne soient pas trop nombreux et correspondent à des documents réels.

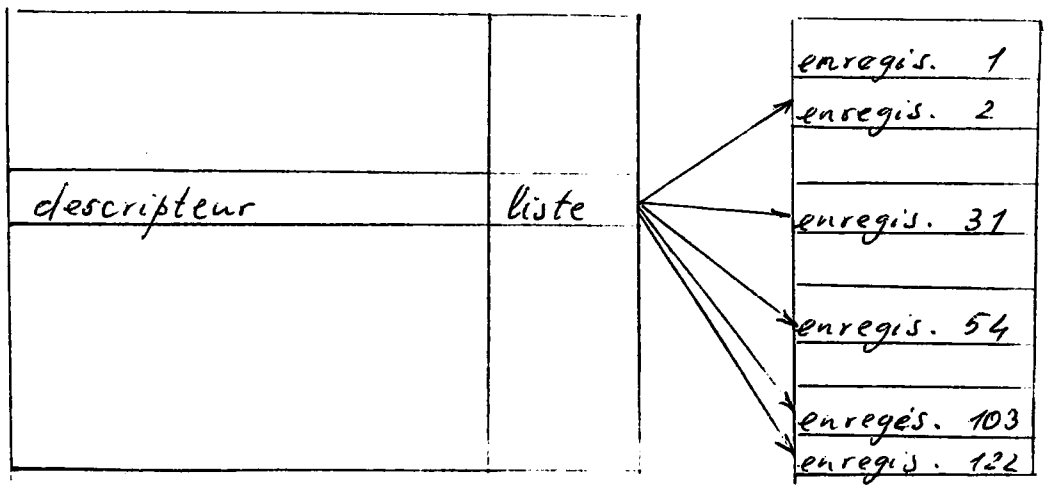
4.3. Fichier inverse

Dans les bases de données bibliographiques l'indexation conduit à la création du fichier inverse qui sert à la recherche rétrospective. Si l'on veut retrouver les documents qui ont un descripteur donné, on peut utiliser une table d'accès à

l'information. Cette table comporte autant de lignes que de descripteurs différents dans toute la base. En général ces descripteurs sont organisés par ordre alphabétique. À chaque descripteur correspond la liste des numéros de tous les documents qui comportent le descripteur. Cette table d'accès est le fichier inverse. Ce fichier sert à fournir tous les renseignements qui permettent de sélectionner directement les enregistrements répondant à la question posée. On dispose donc une instruction de lecture en accès direct, on obtient rapidement tous les enregistrements sélectionnés par la clé fournie. Ainsi, la consultation du fichier direct est remplacée par la consultation du fichier inverse qui est moins important et par l'appel dans le fichier direct des enregistrements sélectionnés. Une telle méthode de consultation est coûteuse au point de vue place et ne procure pas une rapidité supérieure aux autres méthodes. Mais elle permet l'accès direct à un fichier suivant un critère quelconque et suivant plusieurs critères aussi. Les enregistrements de la base de données sont fournis dans un ordre quelconque et le fichier inverse est constitué séquentiellement par modifications successives. Chaque enregistrement du fichier inverse a une longueur variable, parce que la fréquence d'utilisation des valeurs d'une même clé est variable.

Le fichier inverse est constitué par l'entrée alphabétique des descripteurs (termes génériques, termes spécifiques et mots du lexique contrôlé) qui sont suivis des numéros de documents dans le fichier direct. Lorsqu'on crée un document, le fichier inverse est mis à jour automatique-

ment pour les termes génériques et les termes spécifiques et après validation pour les mots du lexique contrôlé. Cette validation se fait à l'aide du résumé de la fiche. Un terme générique ne peut être terme spécifique, ni réciproquement. Les unitermes du lexique contrôlé peuvent être identiques à un uniterme de terme générique ou terme spécifique, ou inclus dans un syntagme terme générique, terme spécifique. Les éléments qui constituent le fichier inverse sont : le lexique thésaurus brut qui est une liste alphabétique de tous les mots-clés (unitermes ou syntagmes), le lexique contrôlé, les mots thématiques (termes génériques: unitermes ou syntagmes). Le fichier inverse peut être enregistré selon diverses méthodes qui dépendent de la dimension du lexique et de celle du fichier de références. Ces méthodes sont présentées dans l'annexe 2*.



* Pris de l'article "Création et maintenance de bases de données bibliographiques" de A. Deweze.

5. Interrogation des bases de données bibliographiques

5.1. La recherche de l'information

La recherche de l'information est un ensemble d'opération qui a pour objet de fournir à l'utilisateur les renseignements qui répondent à ses questions. Ces questions peuvent être occasionelles ou permanentes. Les questions occasionelles portent sur une recherche rétrospective, car on identifie toutes les sources, enregistrées au préalable, qui se rapportent sur un sujet donné. Les questions permanentes impliquent une recherche d'information courante, car on identifie les sources qui sont enregistrées au cours de périodes de temps successives. La recherche de l'information recouvre la recherche des documents ou des sources et aussi celle des données auxquelles les premiers donnent accès. Ces opérations se situent au milieu de la chaîne documentaire et elles préparent la diffusion de l'information.

La recherche de l'information suit une certaine procédure. Lorsque l'utilisateur demande une information, un dialogue s'engage entre lui et le spécialiste, et la question est reformulée plus précisement, ce qui permet de définir une stratégie de recherche. Le spécialiste traduit les termes de la recherche dans les termes du langage documentaire et établit des critères de recherche. S'il utilise un thésaurus, il sélectionne les descripteurs et on obtient un groupement. P. ex. si l'utilisateur demande des informations sur la création des bases de données bibliographiques, les descripteurs seront : Création, Conception, Elaboration, et le groupement : Création ou Conception ou Ela-

boration. Pour chaque critère de recherche il cherche les descripteurs pertinents : Base de données, Banque de données, et le groupement sera : Base de données ou Banque de données. Chaque groupe des descripteurs est lié au groupe suivant par ET, car les documents doivent traiter de tous ces points. Le spécialiste peut indiquer, au moyen de SAUF, que les documents qui portent sur les réseaux ne sont pas demandés. On utilise dans ce cas la logique booléenne qui est très utilisé dans les procédures de recherche. Les opérations de coordination permettent d'établir les équations de recherche. Les documents obtenus sont triés et on retient les références qui sont intéressantes ou qui peuvent être consultées sans limitation ; on élimine les documents qui font double emploi, car l'objectif est de sélectionner les documents véritablement pertinents.

Il existe d'autres procédures qui permettent de raffiner et de compléter l'équation de recherche, p. ex. la pondération, le voisinage, la troncature, l'extension, la comparaison numérique. La procédure de recherche peut prendre des formes très diverses, mais les plus importantes sont: la recherche directe et la recherche déléguée. La recherche directe est faite par l'utilisateur lui-même auprès des sources qui sont à sa disposition, et la recherche déléguée par le spécialiste pour le compte et à la demande de l'utilisateur. Les étapes les plus importantes sont:
La prise de conscience par l'utilisateur d'un besoin d'information et définition de celui-ci (sujet, délai, type de document, langues, etc.).

Identification des sources auxquelles l'utilisateur peut s'adresser.
Communication de la demande et si nécessaire une discussion sur la demande entre l'utilisateur et l'informateur.
Formulation de la question dans le langage documentaire et détermination des stratégies et équations de recherche.
Recherche des citations dans le sous-système de recherche de l'information. Si nécessaire, modification de la stratégie de recherche.
Rassemblement des références bibliographiques.
Filtrage : sélection des références les plus importantes en fonction des spécifications de la demande et des caractéristiques principales secondaires.
Communication des résultats de la recherche à l'utilisateur.
Vérification par l'utilisateur de la validité de la réponse.
Communication à l'utilisateur des documents primaires retenues.

Pour améliorer les questions on utilise diverses méthodes l'ajout ou la suppression d'un descripteur, le remplacement d'un descripteur plus générique ou par descripteurs plus spécifiques, l'ajout, la suppression ou la transformation d'un opérateur logique (modification de l'équation de recherche), l'utilisation des troncatures, le découpage des questions en plusieurs sous-ensembles.

5.2. Utilisation du logiciel d'interrogation

Lorsqu'on fait la recherche et visualise les résultats de l'interrogation, on suit quelques phases successives qui en

général sont :

- le choix de la base qui peut répondre aux besoins
- le choix du serveur qui diffuse la base
- la procédure de connexion
- la formulation des concepts recherchés qui est faite par les mots-clés ou descripteurs
- la combinaison des descripteurs au moyen d'opérateurs logiques
- la combinaison des descripteurs au moyen d'opérateurs de distance
- la recherche d'une chaîne de caractères dans le texte
- la visualisation des résultats en ligne
- l'impression des résultats en différé
- la sauvegarde du questionnaire
- la procédure de déconnexion.

Pour illustrer ces phases on peut utiliser le logiciel QUESTEL. La connexion se fait par le réseau TRANSPAC : après l'affichage du message TRANSPAC 0000000, on tape le numéro d'appel du serveur. Une fois la communication établie, sur écran est affiché le message :

COM

TELESYSTEMES QUESTEL ST00TR00

11.04 * 11.35 * PLEASE LOGIN :

On tape le numéro d'utilisateur et sur écran est affiché :

PASSWORD :

On tape le mot de passe d'utilisateur et on attend l'affichage du message :

.. INFO, .. MENU, .. BASE ?

Pour sélectionner une base de données on tape :

?.. BA PASCAL

Sur écran est affiché le message :

BASE CONNECTEE : PASCAL 03/04/84

COMMANDE, OU ETAPE DE RECHERCHE 1

5.2.1. La recherche dans la base bibliographique

Après le choix de la base le système signale qu'on peut poser la première question, p.ex. :

? BASE DONNEE

On adopte la convention : les commandes ou questions sont mises dans un quadrangle. La touche RET ne figure pas, car elle est utilisée après chaque question. Avec QUESTEL la question est frappe immédiatement après le ? affiché par le système. L'énoncé de la question est appelé ETAPE DE RECHERCHE, la réponse donnée par l'ordinateur est :

1 RESULTAT 2356

COMMANDE, OU ETAPE DE RECHERCHE 2

Ensuite on peut frapper la deuxième question :

? BANQUE DONNEE

Des questions sont posés pour tous les mots cherchés. Les résultats des questions sont combinés pour rechercher les références qui contiennent ces mots. En utilisant l'opérateur ET, on peut obtenir des ensembles, l'intersection desquels donne un ensemble qui constitue le résultat recherché. Pour obtenir le résultat attendu de la recherche, il faut utiliser l'opérateur ou qui réunit les ensembles. Cet opérateur sert à élargir la recherche. C'est la raison qu'on utilise d'abord la réunion des ensembles de documents

et ensuite on effectue l'intersection des ensembles. Par ex
ample :

? BASE DONNEE OU BANQUE DONNEE

1 RESULTAT 5503

COMMANDE, OU ETAPE DE RECHERCHE 2

? 1 ET (CREATION OU CONCEPTION OU ELABORATION)

2 RESULTAT 915

COMMANDE, OU ETAPE DE RECHERCHE 3

Le logiciel permet de combiner deux ou plusieurs termes en une seule question, mais cette combinaison peut conduire à certains inconvénients. Lorsqu'on utilise les opérateurs ET et OU dans une même question il faut utiliser les parenthèses, parce que l'interprétation que fait le logiciel peut être différente de celle de la question posée.

Pour exclure les documents qui traitent une notion particulière on utilise l'opérateur SAUF, par exemple :

? BASE DE DONNEE OU BANQUE DONNEE

1 RESULTAT 5503

COMMANDE, OU ETAPE DE RECHERCHE 2

? 1 ET (CREATION OU CONCEPTION OU ELABORATION)

2 RESULTAT 915

COMMANDE, OU ETAPE DE RECHERCHE 3

? 1 ET MAINTENANCE

3 RESULTAT 65

COMMANDE, OU ETAPE DE RECHERCHE 4

? 3 SAUF RESEAU

4 RESULTAT 59

On n'utilise pas cet opérateur pour combiner des notions, par

ce que son utilisation peut conduire à éliminer les documents qui contiennent une notion cherchée.

-Utilisation des troncatures et masques

Le fichier inverse est constitué par la liste alphabétique de tous les mots que le système rencontre pendant la lecture des enregistrements. Mais, dans le texte les mots peuvent se présenter dans ses variantes morphologiques, variantes orthographiques ou dans formes non fixes. Le système recherche dans l'index l'entrée qui correspond à la chaîne frappée seulement. On trouve dans ce cas un nombre réduit de documents, qui peut correspondre à un silence important. L'utilisation de la troncature simplifie la question et permet de prendre en compte les variantes morphologiques. On frappe la chaîne de caractères qui correspond au radical suivi d'un caractère qui remplace la chaîne à droite du radical : p. ex. avec QUESTEL le +. Le système répond en donnant tous les variantes : on obtient ainsi toutes le références qui comprennent ces variantes. Le bruit dans ce cas est inévitable : la troncature illimitée entraine un bruit d'autant plus important qu'elle s'applique à droite d'une chaîne courte. La troncature limitée élimine cet inconvénient : on définit le nombre des caractères à droite du radical. La troncature limitée avec QUESTEL est exprimée par une série de ? qui indiquent le nombre des caractères remplacés, p. ex. RESEAU ?. Il permet aussi une troncature à gauche.

On peut masquer un ou plusieurs caractères à l'intérieur d'un mot pour obtenir des variantes orthographiques:

à chaque # correspond un caractère, mais pour obtenir tous les variantes on utilise le ?.

Si l'on n'utilise pas la troncature, on obtient des silences, et son utilisation inconsidérée implique des bruits. Il est utile de consulter la liste des termes dans l'index:

```
? BIBLIOGRAPHI+
   T1 BIBLIOGRAPHIA  /UT
   T2 BIBLIOGRAPHIC  /UT
   T3 BIBLIOGRAPHIC ACTIVITY  /DE
   . . . . . .
   T14 BIBLIOGRAPHICO  /UT
   T15 BIBLIOGRAPHICS  /UT
   AUTOMATIQUE (A) /SELECTIONNER (STI)/CONTINUER L'EDITION
                                          (O/N)?
```

- Utilisation des opérateurs de comparaison

Souvent on limite la recherche par comparaison avec certaines valeurs, p. ex. on désire de rechercher les documents publiés après une date donnée ou au cours d'une période donnée, etc. Dans QUESTEL les opérateurs sont: = (égal à), < (inférieur à), <= (inférieur ou égal à), # (différent), > (supérieur à), >= (supérieur ou égal à). Par exemple, on cherche les documents publiés après l'année 1978:

```
? TX /DP > 1978
```

Quelquesfois on souhaite avoir une vue des étapes, des questions et du nombre de références qui correspondent à chaque question. Pour visualiser l'historique on utilise la commande ?..HI. Par exemple:

?1.. HI

ETAPE	FREQ	
1	5503	BASE DONNEE OU BANQUE DONNEE
2	915	1 ET (CREATION OU CONCEPTION OU ELABORATION)
3	65	1ET MAINTENANCE
4	59	3 SAUF RESEAU
5	3	4 ET BIBLIOGRAPHI+

5.2.2. Visualisation et impression des résultats

La visualisation des documents est faite en ligne au moyen de l'écran, de l'imprimante utilisée seule ou en recopiant l'écran. L'impression peut être effectuée en différé : le serveur envoie les résultats dans un certain délai. La visualisation en ligne est utilisée pour un nombre limité de références (une vingtaine au maximum). La visualisation et l'impression des résultats demande la définition des paramètres : numéro d'étape ou de question, format de visualisation (tout ou partie des zones), nombre des références. Les citations sont visualisées dans l'ordre inverse de leur introduction dans la base. Pour l'impression on obtient le format maximal. Si l'on n'indique pas le numéro de question, c'est le résultat de la dernière question qui est visualisé. Lorsqu'on utilise la commande .. VI des résultats sont visualisés dans une liste déterminée des zones:

?..VI

-1- 3123246 C. PASCAL

NO : 83-X-0378869

AU : HATZOPOULOS M. ; KOLLIAS J.G.

AF : MICHIGAN TECHNOL. UNIV., DEP. MATH. COMPUTER SCI./HOUGHTON

ET : MAINTAINING THE INITIAL PERFORMANCE OF NETWORK STRUCTURE

SO : ANGEW. INFORM. // ANGEWANDTE INFORMATIK ; ISSN 0013-5704 ;
DEU ; OA. 1982 ; NO. 5 ; PP. 290-294 ; ABS, GER ; BIBL. 8 REF ; LOC.
CNRS-2890

L'impression en différé est pratiquée lorsque le nombre de références est très grande. Le format est le format maximal. La commande comprend les mêmes paramètres que celle de visualisation. Dans QUESTEL la commande est : ?.. ED, et pour sortir de ce mode on frappe : ? FIN, alors le système donne son adresse avec le message : VALIDATION COMMANDE ET ADRESSE (O/N) pour confirmer ou annuler l'impression.

La commande de se déconnecter est suivie des paramètres de sortie : .. ST SV (sauvegarde des étapes) ou .. ST FI (sans sauvegarde).

Le serveur TELESYSTEMES permet la commande en ligne des documents primaires auxquels font référence les bases de données. La commande avec QUESTEL est :

.. OR ET 3 2 5 8

où sont commandés les documents qui correspondent aux références 2 5 8 de l'étape 3.

Conclusion

Les problèmes traités sont des étapes très importantes de la chaîne documentaire. L'analyse du contenu, qui recouvre un certain nombre d'opérations, suppose une bonne connaissance du sujet traité et une définition précise du niveau d'information à conserver. On connaît bien l'intérêt d'un résumé, faciliter l'enregistrement dans la mémoire et réduire le temps de consultation. Ces opérations présentent le document et l'information qu'il contient par une notice qui est mise en mémoire. À partir de la mémoire s'effectuent les opérations de recherche documentaire. La recherche documentaire et la diffusion de l'information sont la raison d'être du centre de l'information et de la documentation.

Des diverses opérations peuvent être réalisées automatiquement par ordinateur : entrée et sélection des données bibliographiques, contrôle et vérification, le thésaurus, saisie des données dans les fichiers et recherche documentaire, édition de produits documentaires et réponses aux questions. Maintenant on fait des essais de condenser et traduire automatiquement les textes. Aussi avec l'ordinateur on peut gérer l'acquisition, la commande des documents. Tout cela est effectué à une vitesse très élevée et supprime les doubles emplois et les travaux manuels répétitifs.

Bibliographie

CHAUMIER (Jacques). Les techniques documentaires. Paris : PUF, 1979

DEWEZE (André). L'accès en ligne aux bases documentaires. Paris : MASSON, 1983

DEWEZE (André). Création et maintenance de bases de données documentaires, Grenoble : Merlin Gerin, 1982

DEWEZE (André). "Eureka" : système d'interrogation de bases de données bibliographiques en libre service. Grenoble : Merlin Gerin, 1982

GUINCHAT (Claire) ; MENOU (Michel). Sciences et techniques de l'information et de la documentation : introduction générale. Paris : Les Presses de l'Unesco, 1981

HURTUBISE (Rolland). La gestion de l'information, Paris : Les éditions d'organisation, 1977

JOUFFROY (Claude) ; LETANG (Charles). Les fichiers. Paris : DUNOD, 1977

LANCASTER (Frederick Wilfrid). Information retrival systems. New York : "A Wiley-Interscience publication", 1978

MARTIN (Daniel). Bases de données. Paris : DUNOD, 1981

VAN SLYPE (Georges). Conception et gestion des systèmes documentaires, Paris : Les éditions d'organisation, 1979.

CHAMP		LIBELLÉ	(POSITION, LONGUEUR)		
Z.0		LABEL DISQUETTE	(1,11)		
Z.1		MATRICULE	(12,22)		
Z.2		ORGANISME DÉTENTEUR	(34,8)		
Z.3		MATRICULE INTERNATIONAL	(42,16)		
Z.4		NUMÉRO DE FICHE	(58,13)		
	4.1	numéro bulletin			(58,3)
	4.2	rubrique			(62,4)
	4.3	numéro chronologique			(67,4)
Z.5		AUTEURS	(71,72)		
	5.1				(71,24)
	5.2				(95,24)
	5.3				(119,24)
Z.6		APPARTENANCE	(143,76)		
Z.7		TITRE EN FRANÇAIS	(219,160)		
Z.8		LANGUE	(379,11)		
Z.9		TITRE EN ÉTRANGER	(390,63)		
Z.10		RÉFÉRENCES PÉRIODIQUE	(453,77)		
Z.11		NATURE DOCUMENT	(530,7)		
Z.12		INDICATIONS COMPLÉMENT.	(537,70)		
Z.13		MOTS THÉMATIQUES	(607,188)		
				index	libellé
	13.1			(607,6)	(613,41)
	13.2			(654,6)	(660,41)
	13.3			(701,6)	(707,41)
	13.4			(748,6)	(754,41)
Z.14		MOTS-CLÉS	(795,188)		
				index	libellé
	14.1			(795,6)	(801,41)
	14.2			(842,6)	(848,41)
	14.3			(889,6)	(895,41)
	14.4			(936,6)	(942,41)
Z.15		RÉSUMÉ	(983,900)		

FIG. 3 – TRACÉ D'ENREGISTREMENT DU FICHIER «RÉFÉROTHEQUE»

5.4. SPÉCIFICATIONS DES CONTROLES DE SAISIE

Champ 0 – Label «bulletin» (1,11)*

En tête de chaque enregistrement, champ de longueur fixe comprenant onze caractères, exemple: 504-08-78-G

504- numéro de bulletin suivi d'un tiret généré par programme
08- mois d'édition du bulletin suivi d'un tiret généré par programme
78- année d'édition du bulletin suivi d'un tiret » »
G domaine traité dans la base,
soit G gestion et économie
ou T documentation technique
ou B brevets
ou R recherche

Champ 1 – Matricule (12,22)

Champ obligatoire. Cadrage à gauche, troisième caractère = espace. Il doit commencer par:

- le mnémonique, prélevé dans le fichier «revues», suivi de la date et de la première page de l'article.
- sinon, B. C. L. N. R. T. BR CT TH TP : selon le type de document (bibbliographie, catalogue, livre, norme, etc.).
- sinon, matricule libre, cadré à gauche.

Ce champ constitue une clé secondaire non duplicable (contrôle correspondant).

Champ 2 – Sigle organisme détenteur (34,8)

Ce champ peut être vide. Sinon on cadre à gauche.

Champ 3 – Matricule international ou organisme émetteur (42,16)

Ce champ peut être vide. Sinon, on cadre à gauche. On peut y faire figurer l'ISSN, l'ISBN, le numéro d'un brevet, etc.

Champ 4 – Numéro de fiche (58,13)

Ce champ est obligatoire. Il comprend trois zones:

- numéro de bulletin (généré par programme à partir du champ 0)
- rubrique (prélevée dans la table des rubriques)
- numéro de fiche (incrémenté par programme)

Exemple: 098-0526-0051 (les tirets sont générés par programme)
n° bulletin *rubrique* *n° fiche*

Champ 5 – Auteurs (71,72)

Ce champ peut être vide. Il est subdivisé en trois zones, de vingt-quatre caractères chacune. Cadré à gauche, doit commencer par A ... Z en majuscules ou une parenthèse (cette parenthèse est utilisée pour la présentation des entreprises en tête de l'index «auteurs»).

* (1,11): le premier nombre avant la virgule indique le début du champ sur l'enregistrement de 1882 caractères - le deuxième nombre après la virgule indique la longueur maximale du champ.

Champ 6 – Appartenance (143,76)

Ce champ peut être vide.

Champ 7 – Titre en français (219,160)

Champ obligatoire. Cadrage première ligne à gauche. Il peut se poursuivre sur deux lignes de 80 caractères.

Champ 8 – Langue (379,11)

Ce champ peut être vide. S'il comprend des caractères, le champ 9 doit également en comprendre.

Champ 9 – Titre en langue étrangère (390,63)

Ce champ peut être vide, sauf si le champ 8 comprend des caractères.

Champ 10 – Références du périodique (ou éditeur) (453,77)

Le début de ce champ doit être extrait du fichier «revues» si on a enregistré un mnémonique ou L. (livre) dans le champ 1: dans ce cas, on cadre le reste du texte immédiatement après le point-virgule. Pour les deux autres cas d'enregistrement du matricule (champ 1), le champ 10 est introduit entièrement à partir du clavier.

Champ 11 – Nature du document (530,7)

Ce champ peut être vide. Il est utilisé pour indiquer la nature du document (rapport, actes, colloque, etc.); ou, lorsqu'il s'agit de sources secondaires (bulletins bibliographiques du CNRS, le BOPI, US Government Research Reports), le sigle de cet organisme. Dans ce cas, le champ 12 est utilisé pour indiquer la localisation de la notice signalétique dans la source secondaire citée.

Champ 12 – Indications complémentaires (537,70)

Ce champ peut être vide. Si le champ 11 n'existe pas, il n'y a pas de champ 12.

Champ 13 – Mots thématiques (607,188)

Il se subdivise en 4 zones, chacune comprenant 47 caractères au maximum. La première zone est obligatoire. Les zones 2 à 4 sont facultatives. Pour obtenir une zone, l'opératrice forme l'INDEX (5 caractères alphabétiques pour le bulletin G; 3 caractères numériques plus un point ou un caractère alphabétique pour le bulletin T), dont la présence est vérifiée dans la table des mots thématiques G ou T* (selon la lettre qui aura été frappée dans le champ 0), le libellé est ensuite imprimé automatiquement.

Champ 14 – Mots-clés (795,188)

Ce champ peut être vide. Son organisation est identique à celle du

* pour les bulletins B et R, on utiliserait la même table T.

Champ 6 – Appartenance (143,76)

Ce champ peut être vide.

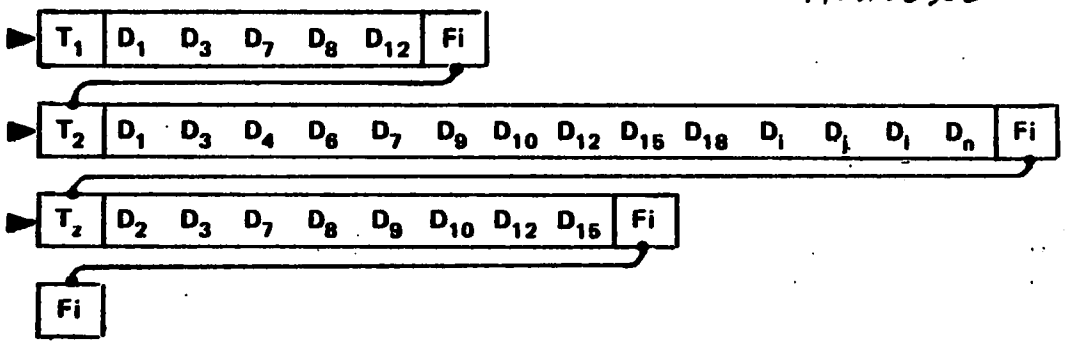
Champ 7 – Titre en français (219,160)

Champ obligatoire. Cadrage première ligne à gauche. Il peut se poursuivre sur deux lignes de 80 caractères.

Champ 8 – Langue (379,11)

Ce champ peut être vide. S'il comprend des caractères, le champ 9 doit également en comprendre.

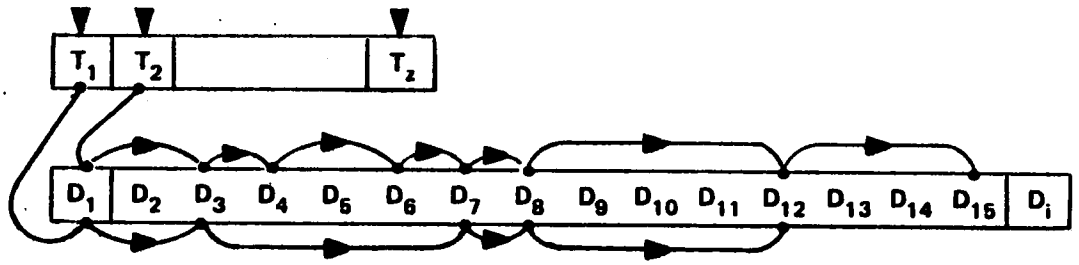
Enregistrements de longueur variable



Avantages: peu de place sur disque, comparaison logique (ET, OU, SAUF) aisée entre D contenant les termes T.

Inconvénients: réorganisation nécessaire à chaque mise à jour.

Pointage de T_1, T_2, T_z vers D_j, D_k, D_n



Avantages: pas de réorganisation des D_1, D_n à chaque mise à jour.

Inconvénients:

- encombrement disques plus important: il faut prévoir avec chaque D un nombre de T égal au nombre maximal de termes d'indexation (entre 20 et 30);
- risque de temps accru pour la comparaison logique (temps de pointage des D successifs pour un T donné avant comparaison avec les D correspondant à un autre T).

Solution mixte