Systèmes d'accès à des ressources documentaires : vers des anté-serveurs intelligents

Hervé Le Crosnier

UNIVERSITE DE DROIT, D'ECONOMIE ET DES SCIENCES D'AIX-MARSEILLE
FACULTE DES SCIENCES ET TECHNIQUES DE SAINT-JEROME

THESE

Présentée par

Hervé LE CROSNIER

Pour obtenir le grade de Docteur en Sciences
de l'Université de Droit, d'Economie et des Sciences

Spécialité : Sciences de l'Information et de la Communication

SYSTEMES D'ACCES A DES RESSOURCES DOCUMENTAIRES -VERS DES ANTE-SERVEURS INTELLIGENTS-

Soutenue le 21 Décembre 1990 devant la commission d'examen à 10H à la Bibliothèque Universitaire - salle CRRM.

B. MIEGE
-Y. LE COADIC
H. DOU

P. HASSANALY
B. VITTORI
J.G. CAILLOUX

Je tiens à remercier Farina **Hassanaly** d'avoir immédiatement accepté d'examiner mon dossier de recherche pour me permettre de présenter cette thèse. Je suis très touché par la gentillesse et le tact dont elle a fait preuve au cours de notre travail en commun, et pour la qualité de son travail de direction de recherche, ses incitations et ses encouragements permanents.

Je remercie **Henri Dou** pour avoir défendu mon dossier et avoir accepté de co-diriger cette thèse. Sa confiance dans le travail en cours m'a été d'un grand soutien.

Bernard Victorri est pour une large part, moralement et scientifiquement, dans le travail qui est présenté ici. Il a su depuis de longues années me donner confiance et m'encourager à m'investir dans des activités de recherche. Il sait toute l'amitié que je lui porte.

Je remercie **Yves Le Coadic** pour ses encouragements depuis plusieurs années. La qualité de ses critiques permet d'aller de l'avant autant que la générosité de son soutien.

Je remercie **Bernard Miège** d'avoir bien voulu présider le jury de thèse malgré ses nombreuses occupations. Ses remarques sur la question du traitement documentaire des images sont une incitation à poursuivre le travail avec plus de précision.

Il n'est pas possible de citer ici tous ceux qui m'ont aidé, soutenu et encouragé. J'ai cependant une pensée particulière pour le personnel de la Bibliothèque Scientifique de l'Université de Caen sans qui ce travail n'aurait pas pu voir le jour et pour le personnel du Centre de Formation aux Carrières des Bibliothèques, du Livre et de la Documentation de l'Université de Caen qui m'a accueilli avec générosité et m'a donné tous les moyens dont j'avais besoin pour ce travail.

Une pensée pour Anne, Matthieu et Camille.

Avant-propos

Cette thèse traite des modèles de systèmes documentaires. L'informatique, en pénétrant dans toutes les sphères de la production et de la circulation du savoir offre de nouvelles perspectives de diffusion et d'organisation des fonds documentaires. L'informatique permet à l'information de se répandre en dehors des contraintes de l'espace et du temps. L'accès à un ordinateur passe par un réseau mondial, et se déroule en temps réel. Pourtant, nos capacités à lire, à organiser les informations et à juger de la pertinence des documents qui nous sont proposés restent lentes et soumises à l'espace de l'écran. Dès lors, on doit attendre des systèmes documentaires informatisés qu'ils nous aident à chercher et découvrir les documents ou les informations dont nous avons besoin.

Si dans un premier temps l'aspect massif de l'information disponible en ligne l'emportait (avoir toute la connaissance du monde à portée de clavier), c'est aujourd'hui l'aspect d'aide efficace à la sélection qui prédomine. Dans les deux cas, les techniques de présentation et de sélection sont des enjeux déterminants à l'échelle politique, sociale, économique et culturelle. Sélectionner, c'est mettre en avant des critères qui peuvent conduire des idées ou des langues à l'isolement. Tout proposer, c'est aussi provoquer un effet de surinformation qui tend à rendre toutes les informations équivalentes et passagères.

Les systèmes d'information électronique interactifs commencent à connaître un réel succès. L'exemple du vidéotex français est là pour le montrer. Pourtant, les méthodes employées par les grands serveurs d'information restent datées. Les conceptions de base ont plus de trente ans, et chacun sait depuis que l'utilisateur connaît toujours de grandes difficultés à exprimer son problème dans les termes que reconnaît le système. La diffusion d'information se fait par le biais de groupes économiques importants, sur des réseaux monopolistiques. Les principales sources d'information sont regroupées dans les mains de quelques producteurs *(Dow Jones, Chemical Abstracts, Duns and Bradstreet,....).* Il y a un enjeu important à l'amélioration des instruments permettant d'aborder les informations circulant sur le réseau informatique. Pourtant, chaque amélioration vient renforcer le pouvoir des producteurs déjà dotés d'une force immense.

L'information est devenue une denrée vitale pour les sociétés développées. La recherche scientifique, le développement technique, le positionnement commercial, dépendent largement de la rapidité d'accès à l'information. Soumettre les sources d'information à la mainmise des producteurs de certains pays, aujourd'hui des Etats-Unis, est un facteur de déstabilisation politique et stratégique. L'exemple de l'affaire des dépôts de brevets concernant les tests de dépistage du SIDA qui a opposé le professeur Gallo à l'Institut Pasteur est à ce titre des plus significatifs.

Peut-on sur ce point faire une confiance absolue à la liberté de circulation de l'information, nouveau cadre de la société démocratique ? Il serait plaisant d'y croire, comme vérification de la thèse d'une fin de l'histoire. Mais cela serait plutôt naïf dans un monde où la guerre commerciale et la guerre des images font rage pour la plus haute place du podium. Pourtant, le continent européen tend à abandonner de larges part de son indépendance informationnelle : adhésion de nombreuses bibliothèques françaises au réseau privé d'origine américaine OCLC ; distribution en Europe du réseau STN, proposant avec une forme de dumping ouverte, les principales banques de données scientifiques à partir d'un ordinateur basé dans l'Ohio ; problèmes des agences de presse nationales, notamment de l'AFP en France ; abandon de toute production de banques de données dans le domaine de la chimie ; prédominance de l'anglais dans les travaux scientifiques européens...

Ces remarques permettent de dessiner un paysage de l'information qui n'est nullement déconnecté des préoccupations économiques, sociales, politiques et stratégiques des sociétés modernes. Il semble souhaitable de rappeler cette situation en avant-propos à un exposé sur les modèles de recherche documentaire informatisée. Il n'y a pas de modèle en-soi, ni de conception d'une activité documentaire indépendante des buts poursuivis par la société.

La recherche documentaire manipule des éléments fondamentaux d'une activité sociale :

- les bases culturelles : écrits, langue des documents et langue de manipulation des documents
- les bases économiques : connaissance des marchés, connaissance de la concurrence mondiale, nouveaux produits

- les bases politiques : actualité dans le monde, mémoire des actions récentes comme des plus anciennes

- les bases sociales : constitution de réseaux d'utilisateurs, circulation de l'information dans les organisations (journaux électroniques d'entreprises, syndicats, partis politiques...)

- les bases stratégiques : connaissances techniques (brevets, transferts de technologies...), connaissances fondamentales, veille technologique.

Mais il ne faut pas, en sens inverse, déduire mécaniquement des solutions globales ou globalisantes de la stricte remarque que l'homme est un animal communicant et que l'ensemble de ses activités peuvent être considérées sous l'angle d'un échange d'information. Les théories de la *"société de l'information"* et ses diverses applications dans la mercatique, le management ou la politique pèchent par excès, par arrogance. Qui peut prétendre que l'époque des grandes découvertes ne fut pas une époque de communication et d'information ? Marins et explorateurs ont produit, parallèlement à la découverte de nouveaux marchés, des textes, des mémoires, des connaissances sur le monde, sur les peuples et les civilisations qui l'habitent, sur les instruments de transport ; ils ont de même encouragé, par l'exigence de leur art, des découvertes scientifiques fondamentales (astronomie, géographie...). Que ces connaissances aient servi à imposer, à l'appui de l'épée, langues, coutumes et modes de vie européens dans de nombreuses parties du globe est aussi un fait historique !

Mais que dire de nos connaissances actuelles, des moyens modernes de circulation (rapidité, sélection de l'information pertinente...) en regard des objectifs d'expansion de la société marchande ? L'information sert principalement des objectifs économiques. Les systèmes d'information visent en conséquence à imposer des modèles, des langues (la prédominance de l'anglais) et dans la foulée des entreprises ou des groupes sociaux. La modernité du processus ne doit pas occulter les racines profondes des activités qu'il sous-tend.

Quel est l'aspect principal de la modernité dans le domaine de l'information ? On pourrait la situer entre deux pôles :

- la rapidité. En remplaçant par une équipe de pigeons voyageurs les coursiers voyageant en chemin de fer entre les diverses bourses du continent européen, Charles Havas a introduit une nouvelle donne de l'information : la

vitesse. Ce que la stratégie militaire avait mis en œuvre depuis un demi-siècle (des télégraphes de Claude Chappe qui aidèrent l'armée républicaine à Valmy, à la stratégie de mouvement des armées napoléoniennes) venait de pénétrer ce que l'on n'appelait pas encore la société civile. Il faudrait dès lors accélérer sans cesse la circulation de l'information : mise en place du réseau télégraphique, du téléphone, des réseaux hertziens, des satellites, des réseaux informatiques. Il faudrait, de même, que l'ensemble des informations puissent circuler avec la même vitesse : texte codés (Morse), parole (téléphone), son (radio), image (bélinographe), image animée (télévision). Avec l'avènement du *"tout numérique",* on s'oriente vers une circulation immédiate de flux de 0 et de 1 orchestrant l'ensemble des besoins informationnels.

- la sélection. Si l'information peut arriver rapidement de l'ensemble des points du monde, il importe de savoir choisir parmi la masse informationnelle la part, en regard toujours très faible, qui intéressera tel ou tel utilisateur. Plus qu'une hypothétique *"explosion documentaire",* qui n'est en réalité que l'écume des choses, il semble important de considérer l'activité informationnelle du point de vue de l'utilisateur. Car pour chaque utilisateur particulier, la masse de documents pertinents est nécessairement limitée, et certainement de volume constant ou du moins en faible expansion. Manfred Kochen, un des fondateurs des sciences de l'information, écrivait ainsi : *"Au fur et à mesure que la production documentaire croît, elle se fragmente en collections spécialisées de taille à peu près constante. Le nombre d'auteurs - lecteurs croit à peu près dans les mêmes proportions. Cette communauté elle-même se fragmente en groupes eux aussi de taille constante."* ([KOC63], cité par [SWA90]). Cette manière d'aborder avec plus de recul le problème documentaire, en situant le point nodal vers l'utilisateur, qui ne recherche toujours qu'une quantité limitée d'information, est en partie contradictoire avec le point de vue des agences de gestion de l'information (bibliothèques, centres documentaires, producteurs d'information -agences de presse, banques de données...-) qui doivent pour leur part satisfaire pleinement des groupes d'utilisateurs plus nombreux et plus spécialisés.

En particulier, cette conception impose l'idée qu'il existe une fonction sociale déterminante de filtrage de l'information à destination de chaque utilisateur particulier. L'objectif de ce secteur professionnel est de mieux organiser le savoir afin d'en permettre la sélection efficace (à la fois complète et précise) par chaque utilisateur. Cette fonction est productrice et garante d'un lien

entre la production documentaire et le monde réel, en assurant une réorganisation des connaissances par recomposition des savoirs et non par accumulation des documents. C'est par exemple le rôle déterminant, dans le domaine scientifique des auteurs de *"revues de mise au point"* (reviews). C'est aussi le travail des documentalistes et des producteurs d'information qui cherchent à organiser une meilleure sélection des informations dans les banques de données, c'est-à-dire une sélection basée sur la pertinence des réponses pour chaque utilisateur particulier.

Bien entendu, cette fonction est aussi lourde de risques, en particulier celui d'éliminer pour des motifs divers (politiques, culturels...) des fragments entiers du corpus de la connaissance. Mais on doit remarquer que ce risque est inhérent à la nature de tout système d'information. Les bibliothèques avaient leurs *"enfers*. On connaît le *"silence radio"* porté sur certains faits d'actualité, ou la tyrannie des heures de grande écoute qui tend à imposer un modèle affadi de production cinématographique. On constate la disparition de certains genres littéraires, comme la poésie qui quittent les étals des libraires... Proposer 543 réponses à une requête d'utilisateur est aussi un moyen de noyer et d'enfermer le savoir. Encombrer les rayons d'une bibliothèque avec des ouvrages anciens et démodés est plus rébarbatif qu'incitateur. Il convient de distinguer la relativité du savoir, qui se cultive en développant l'histoire des sciences, une discipline qui a besoin de disposer de tous les documents, et la progression du savoir, qui requiert avant tout des capacités de synthèse (le savoir se développe par les questions, et non par les réponses, enseignait Bachelard).

Vitesse et sélection sont deux caractéristiques des médias interactifs (vidéotex, banques de données, banques d'images, hypertextes...). Dans ce cadre, il convient de réduire notre approche de l'ensemble des systèmes d'information à l'étude des systèmes documentaires interactifs. Une telle réduction n'est pas limitative, car de nombreuses remarques peuvent s'élargir aux systèmes informationnels diffusés (médias de la filière presse édition ou de la filière audiovisuelle) ou aux systèmes informationnels interpersonnels (réseaux téléphoniques, messageries, communication d'entreprise, rencontres organisées, salons, congrès...). L'avantage des systèmes documentaires comme modèles des relations de conservation/sélection/transmission de l'information tient à la qualité des fonds considérés (masse d'information de plusieurs millions de documents dans les banques de données, pratique multiséculaire des bibliothèques), à la

diversité des types d'informations (texte, références, données, images,...) et à la place stratégique de la recherche d'information dans les processus de prise de décision.

Cette thèse est divisée en trois parties :

- la première partie est consacrée à l'étude de la modélisation de l'activité documentaire.

- la seconde partie décrira une application pratique dans le cadre de la bibliothèque scientifique de l'université de Caen, la réalisation d'un catalogue informatisé.

- la troisième partie cherchera à envisager les axes de travail qui découlent de l'analyse générale des systèmes documentaires et de la constatation établie par la seconde partie que l'on peut améliorer les conditions d'accès aux informations dans des applications concrètes. Dans ce cadre seront présentés les travaux que je réalise au sein de l'équipe de recherche en sciences de l'information de l'université de Caen.

La première partie nous permettra d'étudier tour à tour :

- un modèle général des systèmes documentaires qui prenne en compte dans la définition du système lui-même l'ensemble des activités documentaires, depuis la description des documents, l'aide apportée à l'utilisateur pour formuler ses questions et la production de résultats répondant aux besoins documentaires.

- l'environnement des systèmes documentaires. Pour comprendre l'enjeu des modèles de recherche documentaire, il convient de placer ceux-ci dans le cadre plus général de la production, de la diffusion et de l'utilisation des résultats d'une recherche documentaire. Il existe plusieurs approches du processus de la recherche documentaire (recherche en flux, pour alerter, ou recherche rétrospective) et des réseaux d'utilisation de l'information (systèmes interactifs d'aide à la décision, bureautique, communication interne...). Il convient aussi d'examiner les moyens que nous offre la production documentaire et surtout son organisation en banque de données pour analyser les évolutions des sciences et des techniques (bibliométrie, veille technologique).

- les modèles d'interaction entre l'utilisateur et le fonds documentaire. Comment le fonds documentaire électronique est-il perçu par l'utilisateur, comment peut-on lui proposer les résultats de sa requête pour l'aider à concevoir le contenu du système, et répondre de son point de vue particulier aux exigences de vitesse, de sélection et de pertinence formulées plus haut ? Comment les nouveaux instruments et modèles informatiques permettent des nouvelles approches de cette relation (méthodes à jugement de pertinence, hypertexte...) ?

- les modèles de description des documents dans les systèmes. Il faudra aborder ainsi les problèmes de vocabulaire, d'indexation, la constitution de langages documentaires. On pourra distinguer le document de sa représentation, et définir ainsi les modèles d'espace documentaire. Cette partie sera aussi un moyen d'aborder la dynamique des systèmes informationnels, et la nécessaire évolution des vocabulaires comme la réorganisation des fonds documentaires, (élimination, recentrage autour des intérêts manifestés par les utilisateurs). L'étude des incohérences de l'indexation et des formules d'indexation automatique permettra d'envisager les méthodes de la description documentaire.

- les méthodes de traitement informatique permettant de mettre en regard les requêtes d'un utilisateur et les résultats présentés, et d'approcher au mieux les modèles d'interaction décrits plus haut. On étudiera ainsi les modèles booléens à base de fichiers inverses, les modèles vectoriels et probabilistes, les modèles hypertextes et les modèles connexionnistes. Les traitements à base de machines parallèles et de signatures de documents seront abordés car ils représentent certainement des voies de développement capables de porter sur le marché documentaire les modèles théoriques, notamment en assurant un temps de traitement compatible avec la recherche en temps réel.

La seconde partie permettra d'aborder :

- les problèmes généraux de la mise en place des catalogues informatiques de bibliothèques, en particulier les distinctions entre les notions de localisation des informations et les notions de marché des notices bibliographiques. Nous nous appuierons sur l'étude menée par le Ministère de la Culture concernant un nouveau *Schéma directeur de l'information bibliographique*.

-la réalisation d'un catalogue informatisé orienté vers les besoins de recherche documentaire des utilisateurs d'une bibliothèque scientifique. Ce catalogue est réalisé avec les moyens traditionnels des documentalistes, le logiciel documentaire *Texto* et son langage de manipulation des données associé *Logotel.*

La troisième partie sera plus orientée vers la recherche de perspectives destinées à donner à l'utilisateur des moyens de s'approcher dans des applications concrètes des méthodes décrites dans la première partie. On étudiera de ce point de vue :

- les recherches les plus significatives pour utiliser la formalisation en systèmes experts des pratiques documentaires. On traitera de la notion d'apprentissage dans les systèmes documentaires, qui permet d'envisager la construction de modèles évolutifs, s'appuyant sur les apports cognitifs des utilisateurs. Un système documentaire ne peut pas être considéré comme un produit diffusé, établi une fois pour toutes, mais au contraire comme un produit interactif, qui change avec l'évolution de la documentation, avec l'invention linguistique, et avec l'évolution des besoins des utilisateurs.

- le développement d'un système à jugement de pertinence, utilisant un algorithme qui pourrait être facilement porté sur des machines parallèles. Je présenterai le travail de description des informations dans un espace sémantique référentiel qui est étudié à Caen. Ce logiciel QUID permet de proposer des documents classés, mais aussi des pistes de recherche. L'application décrite concerne une banque de données des banques de données, qui montre que ce système peut permettre une recherche sur des documents complexes (on ne peut réduire la description d'une banque de données à quelques mots-clés).

- la réalisation d'anté-serveurs intelligents capables de s'appuyer sur les banques de données telles qu'elles sont aujourd'hui produites et diffusées (indexation aléatoire, interrogation booléenne) mais offrant à l'utilisateur un nouveau visage de l'accès à l'information. L'objectif de ce chapitre sera de décrire les fonctionnalités d'un tel anté-serveur, en s'appuyant sur mon expérience professionnelle d'intermédiaire en information, et en rapportant à une situation envisageable à très court terme les notions générales ouvertes par l'étude des modèles de systèmes documentaires.

Table des matières

Première partie

Les systèmes documentaires

II - f_q : QU -> EQ : L'utilisateur face au système documentaire

1 - L'évaluation des systèmes documentaires

l.a - Du besoin documentaire à la formulation de la requête

. une méthode par essais et par erreurs

. la reformulation

. butinage et navigation

l.b - La notion de pertinence

Le - Les critères d'évaluation

. évaluation ergonomique

. évaluation documentaire

2 - La formulation de la question par l'utilisateur

2.a - Convivialité

2.b - Les limites des systèmes commerciaux

2.c - Utiliser le jugement de l'utilisateur

3 - L'interface de manipulation des informations

3.a - Les styles de dialogue homme/système

3.b - La lisibilité des fonctionnalités

3.c - Les signes de l'échange

3.d - La tolérance aux erreurs

III - f_i : DOC -> ED : La fonction d'indexation

1 - Une approche générale de l'indexation

2 - L'indexation manuelle

2.a - Règles et méthodes

2.b - Une activité inconsistante

3 - L'indexation automatique : remarques générales

4 - L'indexation en texte intégral

Table des matières

Table des matières

Deuxième partie

Réalisation d'un catalogue informatisé

I - Le choix d'une hypothèse de travail

Table des matières

Table des matières

Troisième partie

Quelques travaux de recherche

I - Trois axes de travail significatifs

1 - L'expertise
- l.a - Présentation de I^3R
- l.b - Présentation de EURISKO
- l.c - Les leçons de l'expertise

2 - L'apprentissage
- 2.a - L'apprentissage à court ternie : la reformulation
- 2.b - L'apprentissage à long terme
  - . construction interactive d'outils documentaires
  - . modification interactive de l'indexation

3 - La représentation
- 3.a - Modifier la structure des documents
- 3.b - Une approche topologique
  - . représenter la polysémie
  - . QUID : la notion de référentiel sémantique

II - La conception des anté-serveurs

1 - Typologie des besoins de l'utilisateur

2 - Les fonctionnalités d'un anté-serveur
- 2.a - Echanges avec le terminal de l'utilisateur
- 2.b - Echanges avec les serveurs

Table des matières

Quatrième partie

Conclusion

Bibliographie

Table des matières

première partie

Les systèmes documentaires

I - Introduction

Un système documentaire est une entité complexe qui remplit les diverses fonctions d'acquisition, d'analyse, de stockage et recherche et de diffusion de l'information. De nombreuses références décrivent les fonctions et les méthodes propres aux systèmes documentaires informatisés. On consultera par exemple [VRI75], [SAL83], ou [KRA85]. Nous n'entrerons pas dans le détail des opérations d'un système documentaire quand elles concernent des objets matériels (par exemple les acquisitions des bibliothèques, ou les systèmes de prêt et de classement des documents).

Nous nous limiterons aux applications informatisées. On peut considérer qu'il existe une relation profonde entre les deux systèmes : toute opération du niveau matériel doit trouver un correspondant dans la représentation informatisée. L'acquisition d'un document correspond par exemple à l'entrée d'un document électronique dans le système informatisé (soit une référence bibliographique -cas général-, soit une copie en texte intégral, soit une copie en mode image -système d'archivage, fonds iconographiques-). Le classement des ouvrages correspond à l'étape de classification informatique, par exemple par des systèmes d'indexation hiérarchisés (plan de classification ou classifications universelles -Dewey, CDU) ou en classification automatique (notion d'agrégats de documents). La diffusion de l'information recouvre les méthodes d'accès aux livres (libre accès, signalisation, catalogage) comme les méthodes de recherche documentaire informatisée. La limitation au système informatisé permet d'une part une modification plus facile des représentations de documents, et d'autre part permet de prendre en compte des opérations uniquement intellectuelles sur les documents (indexation, classification, dissémination d'une image du document et non du document lui-même). En sens inverse, toute manipulation de la description informatique d'un document, ou de la description d'un document informatique (il convient d'envisager le cas où le document n'existe que sous forme électronique) pourrait retrouver son équivalent dans le monde des objets matériels. Ainsi, la modification des agrégats documentaires correspond à la

notion de "classement par centre d'intérêt" des bibliothèques [ROY87] ; les opérations de balayage ou de butinage *(browsing)* de références correspondent à l'attitude des utilisateurs de bibliothèques devant les rayonnages.

1 - Un modèle général des systèmes documentaires

A la suite de Cater [CAT86], on peut définir un système d'information comme un octuplet ordonné :

$$\mathbf{SD := < DOC, QU, ED, EQ, EP, f_i, f_q, f_p >}$$

avec :

- DOC : un ensemble fini de *documents,* ici de documents électroniques, utilisés comme objets de la recherche documentaire.

- QU : *l'ensemble des requêtes* possibles, exprimées en formulation libre, qui peuvent être posées au système.

- ED : *l'espace documentaire,* c'est-à-dire la représentation abstraite des documents qui sert lors de la comparaison entre la question et le document.

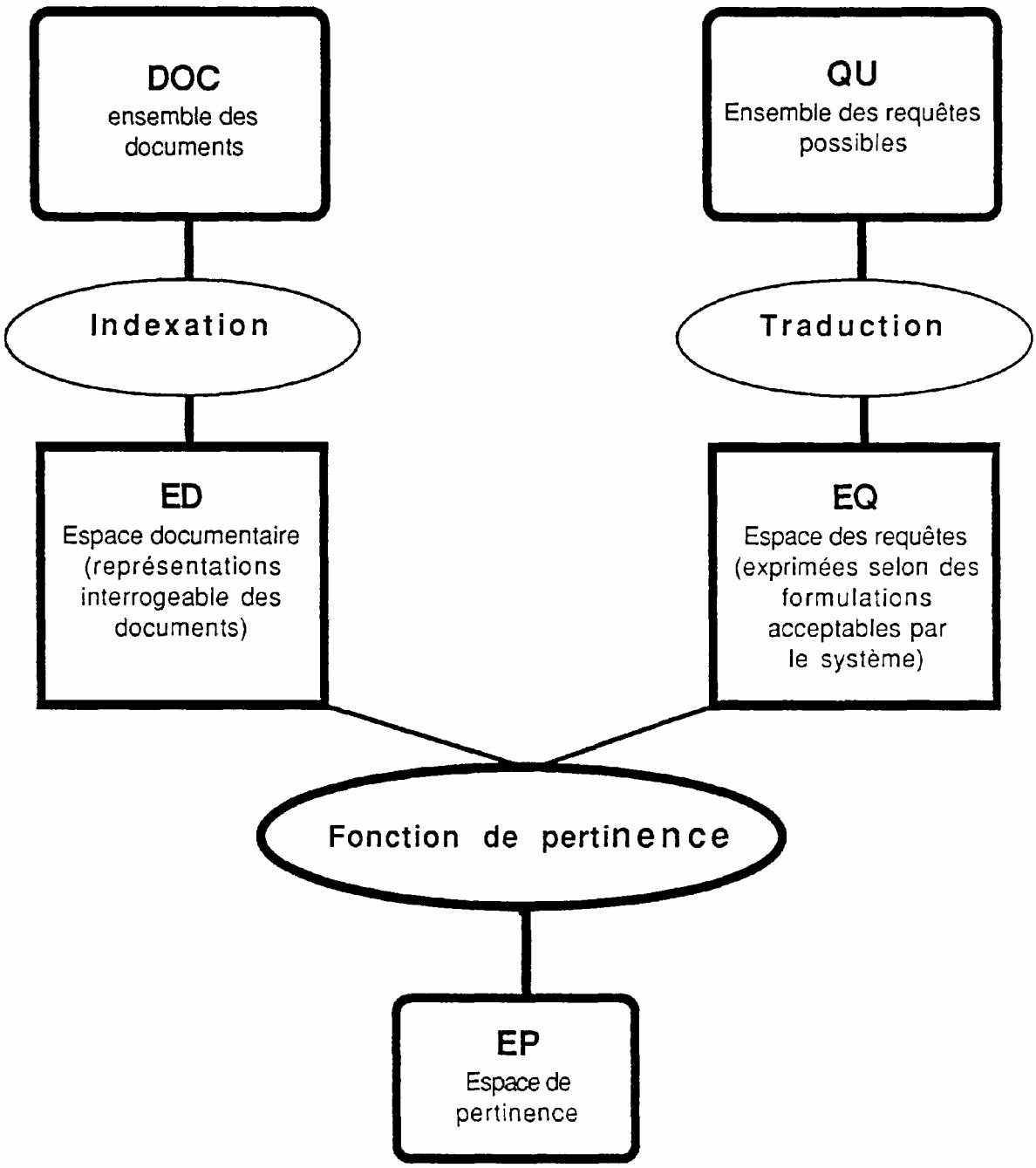
- EQ : *l'espace des questions,* représente les questions acceptables par le moteur de recherche, qui effectue la comparaison entre les questions et les documents.

- EP : *l'espace de pertinence,* constitué par un ensemble de valeurs, en général rapportées à l'intervalle fermé [0 1], permettant d'indiquer le degré de pertinence formelle d'un document par rapport à une question

- f_i : DOC -> ED est la *fonction d'indexation,* qui permet de décrire les documents dans l'espace documentaire

- f_q : QU -> EQ est la *fonction de traduction* qui permet de rapporter les questions exprimées en formulation libre sous une forme qui puisse être utilisée par le moteur de recherche

-f_p . ED X EQ -> EP est la *fonction de pertinence formelle,* qui permet de déterminer le coefficient de satisfaction des documents en regard de la question posée.



Chacun des constituants d'un système documentaire mérite une analyse particulière. Toutefois, les points clés sont la fonction d'indexation qui renvoie aux méthodes et au langage documentaire choisis, et la fonction de pertinence, dont la définition influe sur les résultats de la recherche documentaire.

Cette approche permet de modéliser toutes les opérations d'un système documentaire. Elle servira de trame à l'ensemble de ce document. Il convient donc de préciser le sens exact donné aux divers éléments.

1.a - DOC, l'ensemble des documents

Pour qu'un système documentaire soit cohérent, on suppose d'abord que les documents sont regroupés pour des raisons déterminées et connues. La collection documentaire est sensée avoir une raison d'être. Nous nous limiterons à des systèmes informatisés, dans lesquels les documents existent sous une forme électronique.

On distingue plusieurs types de documents électroniques autour desquels sont constitués des systèmes documentaires :

- références bibliographiques : références de livres (catalogues de bibliothèques, catalogues de livres disponibles) ; références d'articles (bibliographies scientifiques, techniques et économiques obtenues par le dépouillement de la presse spécialisée), références de brevets ou de documents non publiés (littérature grise, thèses, documents administratifs).

- texte intégral de documents : il s'agit d'une forme codée d'un texte complet. Le "texte intégral" se distingue de l'image de texte, qui pourrait être une copie numérique de la page de texte imprimée (télécopie, archivage...). Dans un cas, le système informatique peut accéder à chacun des mots du texte pour lui faire subir divers traitements (réalisation de fichiers inverses, traitements linguistiques, résumé automatique...), dans l'autre, le texte est équivalent à une image uniquement destinée à la reproduction du texte original sur écran ou sur papier (imprimante laser, télécopie).

On trouve dans le domaine commercial de très nombreuses banques de données en texte intégral produites à partir des bandes de photocomposition

(articles de presse, journaux spécialisés, encyclopédies...), à partir du document électronique original (agences de presse, bureautique) et éventuellement, mais la technique doit encore faire ses preuves, à partir de systèmes de reconnaissance de caractères qui transforment une image de texte en texte intégral.

Un problème majeur posé par les documents en texte intégral est le traitement apporté aux divers signes diacritiques des langues non anglo-saxonnes (accents, cédille, tilde...), aux signes des formules mathématiques ou chimiques et aux éléments typographiques. On trouve en général des banques textuelles appauvries, c'est-à-dire avec un jeu de caractères restreint et l'absence de distinctions typographiques. Cette limitation devrait cependant être dépassée, notamment avec l'adoption de règles élargies pour le codage des textes (à l'instar des spécifications du CCITT pour le télétex) et avec la généralisation de postes de travail adaptés à la visualisation du texte enrichi (imprimantes laser, écrans graphiques avec des langages de description de page).

- données sources : l'élément de base du système d'information est constitué par des données que l'utilisateur recherche directement, en général sans faire appel à des descripteurs. On peut distinguer dans ce cadre les annuaires, les répertoires, les catalogues d'objets, les données scientifiques, économiques, statistiques ou financières, les horaires de transport...On se situe alors à la limite des systèmes documentaires et plutôt en présence de systèmes d'information d'entreprise, basés sur des systèmes de gestion de bases de données.

- graphiques et images : le document de base du système n'appartient plus au monde textuel, même si la fonction d'indexation utilise des données textuelles pour assurer la comparaison document/question. On doit distinguer au niveau technologique les images numériques (décrites point par point) et les images analogiques (décrites en mode vidéo, par exemple sur des vidéodisques) et au niveau conceptuel les images fixes et les images animées (le document est alors constitué par la "séquence").

Le post-traitement des images pour assurer la lecture par l'utilisateur est indépendant de la logique du système documentaire, mais relève d'autres apports technologiques : lecteurs individuels de média image (Vidéodisques, CD-I, DVI)

ou réseaux de télétransmission des images (RNIS, réseaux câblés interactifs). L'aspect documentaire des banques d'images doit être distingué de l'aspect technologique, même si des progrès dans ce domaine (stockage, rapidité d'accès...) influent sur les capacités à inventer des systèmes documentaires plus adaptés à la lecture des images.

- documents structurés : un nouveau type de document électronique est en train d'apparaître, qui est constitué par la collection structurée de divers documents élémentaires.

On trouve ainsi des documents hypertextes, qui comportent des éléments d'information multimédia (texte, image, son, logiciels...) reliés entre eux par des liens informatiques qui permettent une circulation rapide d'un point à l'autre du document, chaque usager définissant sa propre circulation entre les éléments d'information. Un document hypertexte peut être considéré sous deux aspects, d'une part il est lui-même un système documentaire particulier, avec des modes de circulation entre les documents élémentaires (les nœuds de l'hypertexte) et d'autre part il constitue un "document", lui-même décrit dans un système documentaire plus vaste, conjointement à d'autre documents (d'autres hypertextes, ou d'autres documents textuels ou imagés).

Un autre type de document structuré est constitué par les documents balisés dans lesquels des balises permettent de distinguer les diverses parties logiques, et les indications typographiques afférentes. Deux normes semblent s'imposer dans ce domaine :

. SGML *(Standard Generalized Markup Language)* adaptée à l'édition

. ODA *(Office Document Architecture),* qui est une norme plus orientée vers la bureautique et les télécommunications.

La capacité de systèmes documentaires à inclure des documents structurés comme éléments objets de la recherche documentaire est un pas important vers une définition généralisée adaptée aux produits informationnels qui paraîtront dans les années qui viennent. Les applications dans le domaine de la bureautique et de l'archivage sont certainement les premières applications de cette généralisation. On peut aussi concevoir des systèmes d'information servant à la

sélection et à l'accès à des logiciels sur des serveurs de logiciels, ou des banques de données de produits éducatifs (logiciels d'EAO, multimédias, hypertextes éducatifs).

En regard de la composition de son fonds documentaire, un système documentaire doit être capable de gérer les opérations classiques :

- acquisition de nouveaux documents
- modification des documents déjà présents dans le système
- élimination de documents.

Ces opérations doivent être réalisées en conservant la cohérence de l'ensemble documentaire, en assurant la capacité à retrouver les nouveaux documents ou les documents modifiés. L'insertion d'un nouveau document doit conserver l'intégrité d'ensemble du système documentaire. De ce point de vue, l'opération d'insertion peut être longue et coûteuse en temps machine.

Dans la suite de notre travail, la notion de *"document"* restera très générale, de façon à couvrir tous les types de documents présentés ci-dessus. Dans la pratique, un document électronique est composé de deux parties :

. ce que voit l'utilisateur et qui correspond à l'objet documentaire susceptible de répondre à un besoin documentaire : référence bibliographique, texte, image, son...

. ce qui permet la recherche du document à l'intérieur du système documentaire, que nous appellerons la représentation du document.

Quel que soit l'objet documentaire, la partie déterminante du point de vue de la modélisation du système documentaire est constituée par sa représentation. Dans le système, l'objet documentaire est réduit à une adresse, dont la connaissance permet de déclencher la lecture : numéro d'un document textuel et position dans la mémoire magnétique, numéro d'image d'un vidéodisque...Qu'importé alors la nature du document si on peut réduire sa représentation à un ensemble de mots. C'est l'hypothèse que nous suivrons, même si en seconde approximation, on pourrait affiner cette conception en pointant les différences qui existent entre le contenu informatif des images et des textes.

I.b - QU, l'ensemble des requêtes

Un système documentaire est opérationnel dès lors que l'on peut inférer de son existence même un ensemble composé des requêtes qui peuvent lui être adressées. Ainsi, on ne demande pas au service de renseignements de la SNCF le prix d'un billet d'avion. Malheureusement, cette situation est loin de se retrouver dès qu'on aborde les systèmes documentaires généralistes. Par exemple [LEL89], les requêtes posées à un catalogue de bibliothèque dépassent de loin les capacités du fonds documentaire (au sens ci-dessus, c'est-à-dire des documents électroniques du système, ici les références d'ouvrages, et non le contenu des ouvrages eux-mêmes). L'utilisateur pense en fait s'adresser à un bibliothécaire électronique, dont il attend d'ailleurs patience et compréhension, plutôt qu'à une nouvelle mouture du catalogue sur fiche.

Si l'on admet cette probabilité de devoir répondre à des questions inattendues, un système documentaire doit :

. accepter toutes les formulations des questions,

. reconnaître celles auxquelles il peut répondre, et les traiter pour assurer la recherche documentaire,

. reconnaître celles auxquelles il ne peut pas répondre et envoyer une réponse coopérative, c'est-à-dire éclairant l'utilisateur sur le contenu de la banque de données et le type de questions correspondant aux objectifs du système. Il s'agit d'un enjeu de recherche passionnant. Des recherches sont menées actuellement pour réaliser des interfaces coopératives de ce type dans le domaine des systèmes de gestion de bases de données ([PUJ89]). Par exemple, pour qu'un système d'information géographique puisse répondre à une question telle que *"Quelle est la capitale de la Basse-Normandie ?"* par l'indication : *"Les régions n'ont pas de capitale"* ou *"La Basse-Normandie est constituée de trois départements dont les préfectures sont..."*. De telles situations sont largement plus complexes dans les systèmes documentaires.

Dans la suite de ce travail, nous supposerons que les questions posées correspondent à l'objectif du système documentaire. En revanche, l'ensemble QU

des requêtes représentera la possibilité offerte à l'utilisateur de formuler sa question comme il l'entend. En ce sens, si le système informatisé fonctionne sur la base d'un "langage de commande" que l'utilisateur ne connaît pas, l'ensemble QU représentera la question telle que l'utilisateur la présente à l'intermédiaire en information qui réalise la fonction f_q de traduction.

I.c - ED, l'espace documentaire

Imaginons que nous ayons un besoin d'information, sachant que la réponse se trouve dans un livre en notre possession. Il devient logique de le feuilleter, et de juger ainsi directement quel chapitre et quel paragraphe répondent à notre besoin. Cette même opération devient impossible dès lors que le nombre et la taille des documents sont importants. Il n'est pas possible de balayer tous les livres d'une bibliothèque pour trouver une réponse. Il nous faut donc posséder de chaque document une version réduite, plus facile à manœuvrer pour effectuer la recherche documentaire.

L'ensemble de ces représentants de documents constitue l'espace documentaire. La structure de l'espace documentaire doit permettre une représentation efficace :

- évitant les redondances : utilisation d'un vocabulaire contrôlé et des instruments permettant de passer des termes rejetés aux formes choisies (thésaurus, listes d'autorité)

- regroupant des informations similaires dans un même terme d'indexation (regroupant sous "nouvelle pauvreté" des expressions comme "nouveaux pauvres", "quart-monde", "personnes les plus défavorisées"...)

- permettant d'exprimer selon quel objectif tel sujet est traité dans le document, en quel endroit, à quelle époque... On a en général recours à des descriptions articulées, soit parce que les descripteurs ont une syntaxe définie (ordre de succession des informations), soit parce que les représentants des documents sont ventilés dans une description structurée, un type d'information étant affecté à un champ précis de la description.

Il existe de nombreuses approches de l'espace documentaire :

- on peut estimer que les termes décrivant les documents doivent appartenir à un vocabulaire défini, contrôlé, et on construit des instruments pour repérer ce vocabulaire et assurer des liaisons entre les termes de ce vocabulaire. Cette logique est dominante dans les milieux de la documentation. Elle tire sa légitimité historique des catalogues sur fiches, disposant d'un seul point d'accès
aux documents, la "vedette matière" qui devait donc assurer seule la description du document. Souvent composés d'expressions, et éventuellement articulés (vedette, sous-vedette, indication de lieu, de date, de forme), les descripteurs sont regroupés dans des listes d'autorité, et dans un cadre documentaire plus restreint permettant l'élaboration et l'entretien d'un vocabulaire spécialisé, dans des thésaurus.

- on peut estimer au contraire que les mots d'un texte ou éventuellement de ses parties les plus lourdes de signification (titre, résumé) suffisent à décrire le document et à donner un aperçu de son contenu suffisant pour la recherche documentaire. Le premier article ouvrant cette voie fut écrit par Swanson en 1960 [SWA60].

- on peut aussi concevoir que les mots ne suffisent pas à décrire des concepts, et on adopte des descriptions codées. Codes de produits, codes de secteurs d'activité professionnelle, codes de concepts biologiques sont couramment employés dans les banques de données.

- les méthodes de classification permettent aussi l'élaboration d'un espace documentaire particulier, obtenu par regroupement de documents autour d'un même thème. Le regroupement peut être défini a priori par exemple dans les grandes classifications universelles (Dewey, CDU...qui classifient la connaissance) ou obtenu automatiquement par regroupement des documents présents dans le système ayant de nombreux attributs en commun (constitution d'agrégats).

En marge du vocabulaire permettant la description documentaire, l'espace documentaire peut accepter une pondération des informations. Il permet alors de dire d'un document qu'il traite principalement de tel sujet, mais évoque tel autre et survole un troisième.

l.d - EQ, l'espace des questions

Pour organiser la recherche documentaire, il convient de projeter les questions des utilisateurs dans un espace de représentation qui est comparable à l'espace documentaire ci-dessus. La comparaison entre les questions et les documents s'établit entre ces deux projections.

Dans un système donné, les opérations et les choix créant l'espace documentaire doivent être utilisés pour créer l'espace des questions. Ainsi, un espace documentaire défini uniquement par un thésaurus ne pourra être utilisé avec efficacité qu'avec une question constituée de termes extraits de ce thésaurus. C'est par exemple le cas de l'interrogation de la banque de données *Medline,* qui pour être efficace doit absolument utiliser les termes du thésaurus *MeSH.* De même, les bibliothèques anglaises utilisent principalement une indexation matière reposant sur la *Classification Décimale Dewey.* Il faut donc rapporter toutes les questions des utilisateurs aux indices *Dewey* correspondants pour pouvoir espérer obtenir une réponse en consultant un catalogue.

Mais pour que la comparaison entre les questions et les documents soit possible, il faut de plus que les liens sémantiques entre les termes choisis soient plus structurés que les divers connecteurs linguistiques. Si l'ensemble QU des requêtes contient des expressions en formulation libre, au niveau de l'espace des questions, il convient de rapporter le sens des expressions à l'articulation de termes par des opérateurs connus du système. En général, les questions acceptables sont formulées par des équations booléennes, éventuellement élargies (opérations de proximité des termes). Certains systèmes imaginent cependant une répartition des questions dans une grammaire plus complète, utilisant d'autres opérateurs (liens de causalité, de mouvement...) [ZAR88].

1.e - EP, l'espace de pertinence

L'espace de pertinence définit la manière dont un document peut répondre à une requête. On peut considérer que le choix est décrit par une alternative simple : un document répond ou ne répond pas à une question. Dans ce cas, l'espace de pertinence est réduit à l'ensemble {0, 1} composé de deux éléments.

Cette logique binaire n'est certainement pas la mieux adaptée à l'univers documentaire, pas plus d'ailleurs qu'à l'univers en général. Très peu de notions sont définies par des frontières fixes. Dans une interview, Lotfi A Zadeh, le père de la théorie de la logique floue, décrivait le problème ainsi : *"Dans les systèmes logiques à deux valeurs, toutes les classes sont considérées comme ayant des frontières précisément définies. Ainsi, un objet est membre d'une classe ou non. Ceci est vrai si l'on parle de choses comme mortel ou immortel, mort ou vivant, mâle ou femelle etc... Voici des exemples de classes qui ont des frontières définies. Mais ce n'est pas le cas de la majeure partie des objets du monde réel. Par exemple, si vous considérez des caractéristiques ou des propriétés comme "grand", "intelligent", "fatigué", "malade" et ainsi de suite, vous ne trouvez pas de frontières définies. La logique classique à deux valeurs n'est pas faite pour traiter ce type de propriétés où il est question de degré"* ([ZAD84]).

Pour dépasser les limites de la logique booléenne à deux valeurs, on peut définir un espace de pertinence permettant de classer les documents suivant leur proximité avec les besoins documentaires exprimés dans la question. On utilise en général l'intervalle réel [0, 1].

1.f - f_i, la fonction d'indexation

La fonction d'indexation est un élément clé de la recherche documentaire. Nous y reviendrons largement. Contrairement à l'habitude des bibliothécaires et des documentalistes français, qui estiment que cette fonction doit et peut être normalisée et ne dépendre que du document concerné (illusion idéaliste), la fonction d'indexation est le produit de nombreux facteurs tant internes au système documentaire que liés à son environnement. L'indexation est une opération d'analyse, bien qu'il soit délicat d'inférer une description documentaire du document lui-même. La fonction d'indexation est aussi une opération de communication : les descripteurs sont affectés au document parce que ce document est susceptible d'intéresser l'utilisateur qui aura utilisé tel descripteur dans sa question.

La fonction d'indexation dépend aussi de la structure de l'espace documentaire choisi. En particulier, la taille du vocabulaire, les liens lexicaux, syntaxiques ou sémantiques établis entre les atomes de ce vocabulaire, la capacité

de l'espace documentaire à accueillir des descripteurs pondérés, la capacité à utiliser des qualificatifs correspondant à diverses facettes d'un problèmes et de spécifier l'angle de vue propre au document, la capacité d'agréger des documents dans cet espace documentaire sont déterminants.

Une fonction d'indexation est une opération très coûteuse :

- effectuée manuellement, elle implique une industrie de documentalistes indexeurs formés aux techniques documentaires et capables de comprendre les documents qu'ils doivent traiter (des spécialistes du domaine couvert).

- effectuée par un ordinateur, et en fonction du degré de complexité de l'espace documentaire choisi, elle peut être moyennement utilisatrice de temps et d'espace (indexation en texte intégral), et facile à mettre en œuvre, mais peu efficace, ou au contraire lourde de traitement et d'utilisation de mémoire si on ajoute des fonctions de compréhension, même limitée, du document (indexation automatique) et des fonctions d'agrégation. De plus, la recherche et le développement de méthodes efficaces pour l'indexation automatique n'a pas encore donné des résultats entièrement satisfaisants, et reste donc très coûteuse en "matière grise", ce qui explique que ces systèmes évolués soient encore aujourd'hui des produits de laboratoire.

Enfin, nous y reviendrons, l'indexation, qu'elle soit manuelle ou automatique est une activité inconsistante : la description d'un même document est variable d'un indexeur à l'autre, et jugée différemment satisfaisante d'un utilisateur à l'autre.

1.g - f_q, la fonction de traduction

La question d'un utilisateur, exprimée en formulation libre, doit être ramenée à l'espace des questions. A l'heure actuelle, dans les systèmes commerciaux, cette fonction de traduction est réalisée en dehors du système par des intermédiaires en information.

Définir une fonction de traduction détermine la convivialité du système. La recherche documentaire étant une opération d'approches successives par essais

et par erreurs pour extraire des documents pertinents du système, la fonction de traduction doit intégrer une possibilité de reformulation des questions. La reformulation intervient de deux manières :

- *utiliser les connaissances du système* pour adapter le vocabulaire de l'utilisateur au vocabulaire de l'espace documentaire (utiliser les informations d'un thésaurus, pondérer les termes de la question...)

- *utiliser le jugement porté par l'utilisateur* sur les premiers documents que le système lui présente pour modifier la question afin de l'adapter aux besoins tels qu'ils sont exprimés par ce jugement, qui peuvent être différents des besoins exprimés par la formulation spontanée de la question.

1.h - f_h, la fonction de pertinence formelle

On peut distinguer deux types de pertinence :

- la pertinence formelle d'une *description* de document envers une *question* telle qu'elle est posée. Une marge de manœuvre entre 0, pour un document éloigné du sujet et 1, pour un document qui a toutes les caractéristiques exigées par la question, est nécessaire à ce niveau pour permettre la précision de la recherche, mais aussi la capacité à découvrir des documents légèrement différents de la requête.

- la pertinence pour le Jugement de l'utilisateur. Cette notion, qui associe un *document* à un *besoin documentaire,* est plus complexe à quantifier. Entrent alors en jeu des éléments divers, que l'on peut difficilement inférer de la question posée : la question était mal formulée, l'utilisateur connaît déjà tel document, seuls certains documents recouvrent exactement son propos, ou bien au contraire, des documents éloignés de ses premières préoccupations peuvent lui ouvrir des horizons nouveaux.

La recherche documentaire se précisant par un jeu de va et vient entre l'utilisateur et le système, la pertinence pour l'utilisateur est jugée à l'échelle du fonctionnement global du système. En revanche, une question est traitée par le

système pour évaluer, à l'instant même de la comparaison, la pertinence formelle de chaque document en regard de la requête telle qu'elle est formulée. C'est dans cette opération d'évaluation formelle que l'on peut définir plusieurs modèles de systèmes documentaires, sachant que la comparaison entre documents et questions influent sur la structure même de l'espace documentaire et de l'espace des questions : on ne peut comparer que des éléments comparables.

Etablir une fonction du produit cartésien des espaces documentaires et des questions sur l'intervalle [0 1] c'est définir un instrument de repérage au sein de l'espace documentaire. De plus, cette fonction de repérage permet d'intervenir sur la structure de l'espace documentaire lui-même, en prenant successivement chaque document comme une question dont les termes sont les descripteurs du document. La valeur de la fonction de pertinence formelle devient la valeur de la "distance" entre documents. Cette propriété est importante, nous le verrons par la suite car elle permet :

- d'agréger les documents sans modifier la qualité de réponse

- de considérer qu'une question est capable de modifier la représentation d'un document dans l'espace documentaire par une application inverse tendant à rapprocher les documents jugés pertinents des termes utilisés dans la question.

2 • Systèmes documentaires et systèmes de gestion des données

Les systèmes documentaires ne constituent qu'une faible part des recherches informatiques. De ce fait, on tend souvent à les considérer comme un sous-ensemble des systèmes de gestion des données. Il existe cependant des différences profondes entre ces deux types de systèmes, qui devraient se traduire dans la conception même des systèmes et qui pèsent sur leur évaluation.

David Blair ([BLA84], [BLA90a]) repère quatre différences fondamentales :

- la méthode de réponse à une requête. Dans un système basé sur les données, la réponse est directe, et correspond exactement à la question. Ainsi, le *"salaire d'Untel"* est une donnée précise, facilement retrouvée. De même la

"moyenne des ventes de la région 4 pendant les deux dernières années". Nous sommes en face d'un système déterministe. Seule l'absence d'un attribut peut être la cause d'une réponse négative.

Dans un système documentaire, la réponse à la requête est un ensemble de *"documents"* (au sens évoqué plus haut) qui doivent permettre à l'utilisateur de trouver lui-même une réponse à sa question en traitant ces documents (lecture, analyse, comparaison...). En ce sens, la question posée ne correspond pas nécessairement à une réponse unique, mais au besoin d'établir un faisceau d'indices pour la résolution d'un problème. Des documents contradictoires peuvent, par exemple, être utiles.

- la relation entre la question formelle et la satisfaction de l'utilisateur. Dans un système de gestion de données, une réponse, si elle existe, est nécessairement correcte, c'est-à-dire correspond aux critères posés par l'utilisateur. Au contraire, dans un système documentaire, la réponse est constituée de documents ayant été décrits par les termes de la question. Il y a alors deux éléments d'indétermination qui se cumulent :

. l'utilisateur, en formulant sa question cherche à deviner (établir une probabilité) les termes qui auraient pu servir à décrire son besoin documentaire.

. l'indexeur, en déterminant les descripteurs associés au document, établit un choix de termes qui sont censés être utiles pour décrire tel document (probabilité que ces termes soient employés par un utilisateur désirant ce document).

Ces deux probabilités transforment la recherche documentaire en une opération de prise de décision dans un univers incertain.

- la définition des critères d'évaluation. Dans un système de données, une recherche positive est une recherche *correcte.* Il n'y a pas d'ambiguïté dans le jugement à porter sur les résultats. Au contraire, dans un système documentaire, l'évaluation ne peut porter que sur *l'utilité* pour l'utilisateur des documents extraits [COOP68]. Ce critère devient beaucoup moins objectif, et l'évaluation des systèmes documentaires est dès lors plus difficile. En particulier, on doit

s'interroger sur la capacité du système à aider l'utilisateur à formuler, puis reformuler sa question pour définir ses besoins documentaires précisément, et avec les meilleurs chances de succès, compte tenu de la structure du système documentaire.

- la rapidité de la recherche. Dans un système de données, la rapidité de la réponse est strictement dépendante de la vitesse physique du système : accès aux données, rapidité de calcul, organisation des données... Dans un système documentaire, la rapidité dépend principalement du nombre de décisions logiques que doit prendre l'utilisateur pour définir et préciser son besoin documentaire (choix des termes de la question, jugement des documents extraits, reformulation...), et dans ce cadre des facilités de lecture et de reformulation qui lui sont offertes.

Ces distinctions entre les systèmes documentaires et les systèmes de gestion des données influent sur la conception des systèmes documentaires. On ne peut appliquer directement les recherches sur l'organisation des données aux recherches conduisant à l'organisation d'un système documentaire. Au contraire, la réalisation d'un système documentaire efficace (i.e. apportant une satisfaction de l'usager en regard de ses besoins documentaires, et non seulement en regard de la question formelle qu'il a posée) doit s'appuyer sur la nature *"par essai et par erreur"* de la recherche documentaire. Etant données les deux probabilités à l'œuvre, il est rare qu'une question ne retrouve que des documents jugés pertinents par l'utilisateur. De même, le système peut avoir échoué à présenter à l'utilisateur tous les documents pertinents, éventuellement même les documents les plus importants. La structure générale d'un système documentaire doit incorporer ces limites et s'évertuer à aider à la reformulation des questions en fonction du besoin documentaire. Il faut dès maintenant remarquer que ce n'est pas le cas des systèmes commerciaux actuels, malgré quelques avancées (la fonction ZOOM du serveur de *l'Agence Spatiale Européenne,* ou GET de *B.R.S.-Orbit)*. Les systèmes sont conçus comme des systèmes de gestion de données, sans tenir compte de l'incertitude placée tant dans les termes d'indexation que dans les termes de la question. Ce problème existe indépendamment même de l'analyse de l'interface utilisateur (langage de commande) de ces systèmes.

3 - L'environnement des systèmes documentaires

Le système documentaire théorique tel qu'il a été décrit plus haut reste assez général pour couvrir l'ensemble des systèmes documentaires. Dans le monde commercial, on offre cependant principalement des systèmes basés sur la logique booléenne et sur la formulation des questions par des intermédiaires. On doit constater que ceux-ci offrent déjà des résultats intéressants. Les banques de données et les systèmes documentaires intégrés dans les organisations font d'ores et déjà partie du monde de l'information et du monde de la prise de décision dans les entreprises.

Ce succès des systèmes commerciaux ne veut pas dire rentabilité, car les investissements sont très lourds, au niveau matériel et logiciel chez les serveurs d'information. Du côté des producteurs d'information les coûts en personnel spécialisé sont de même très importants.

Ce succès ne veut pas dire non plus satisfaction totale des utilisateurs. Le réflexe "banque de données" ne vient pas spontanément à celui qui cherche une information, et l'utilisateur s'interroge toujours sur la confiance à porter aux systèmes d'information, notamment sur la complétude des recherches.

Ce succès doit cependant s'interpréter comme la capacité des systèmes commerciaux à franchir l'épreuve de la réalité et à subsister, s'insérant toujours plus dans les réseaux informationnels préexistants (bibliothèques, agences de presse...) et commençant à faire émerger des nouveaux modes de diffusion (vidéotex professionnel, anté-serveurs...). La notion de *"pétrole gris",* lancée dans les années 70 ([ANDE73]) commence à prendre corps. Cet environnement des systèmes d'information interactifs rend d'autant plus significative la recherche en informatique documentaire. L'enjeu devient de concevoir les systèmes de demain, qui sauront s'appuyer sur l'existence d'un ensemble d'expériences sociales (réseau de diffusion des banques de données, pratiques des utilisateurs...) pour organiser une meilleure satisfaction des demandeurs d'information (lecture électronique, convivialité, pertinence des réponses, intégration dans les processus de prise de décision...).

3.a - Les besoins documentaires

Pour saisir les enjeux de la recherche concernant l'amélioration des systèmes documentaires informatisés, il faut repartir de l'utilisateur, et s'attacher à définir les raisons pour lesquelles il a besoin d'informations et ses voies habituelles pour les obtenir.

On peut distinguer plusieurs types d'utilisation de l'information :

- l'information permettant de construire une réflexion sur un domaine nouveau et précis. Par exemple le dossier bibliographique constitué par un scientifique en début de recherche, le dossier préalable d'un service de mercatique pour envisager la chance de succès d'un nouveau produit, le dossier de presse du décideur politique. Ce type de recherche correspond à la *"recherche rétrospective"* des manuels de documentation, même s'il est plus élargi. Il s'agit d'une situation où la question est formulée à l'instant *t,* avec pour objectif de recueillir les informations publiées antérieurement.

- l'information d'alerte. Dans ce cas, l'utilisateur veut être prévenu dès qu'une information paraît dans le domaine qui l'intéresse. C'est la notion de *"profil documentaire"*. Le scientifique tient à suivre les nouvelles publications dans le cours de sa recherche, le responsable mercatique suit le lancement de nouveaux produits par ses concurrents ou dans un secteur défini, le décideur politique suit l'actualité par les dépêches d'agence de presse (recherche d'information précise) ou par les médias (presse, radio, télévision...).

- l'information factuelle et ponctuelle. L'utilisateur a besoin d'une donnée supplémentaire pour conduire sa prise de décision ou sa recherche. Le scientifique cherche des données numériques ou les propriétés de composés chimiques ; le spécialiste des approvisionnements cherche une adresse ; l'homme politique a besoin de l'horaire d'un avion, de choisir un restaurant ou de connaître son indice de popularité de la semaine ; le français moyen veut connaître la valeur de son portefeuille boursier. Ce type d'information relève plus spécifiquement des "Systèmes de Gestion de Données". L'aspect proprement documentaire est cependant utile dès lors que la recherche de données est attachée à un concept et non à un objet. Par exemple, la recherche d'une liste des fournisseurs de tel type

de matériel est une opération documentaire (comment est décrit ce type de matériel ? quelles sont ses diverses dénominations ?...), alors que la recherche de l'adresse de l'entreprise X est une opération de gestion de données.

- l'information de la veille technologique. A partir d'éléments épars sur les recherches d'un pays ou d'une entreprise, sur les actions de propriété industrielle (brevets, marques) de concurrents, sur l'analyse des tendances ou l'établissement d'une cartographie des savoirs, on peut obtenir des informations synthétiques sur l'évolution de domaines entiers. Pour apparaître clairement, ces informations doivent être traitées, notamment au niveau statistique, pour révéler des tendances dont la connaissance est déterminante dans de nombreuses activités [JAK88]. La veille technologique est la capacité à extraire une information nouvelle, élaborée à partir de la constitution d'un ensemble d'informations, sans avoir à lire (i.e. connaître) chacun des éléments d'information.

- l'information culturelle. La capacité de prise de décision est dépendante des savoirs et des savoir-faire accumulés pendant une longue période. Le scientifique connaît et perfectionne les fondements de son domaine, s'ouvre sur les travaux connexes ou les avancées relatives à d'autres domaines, s'interroge sur la place de la science et ses conséquences dans le monde ; il peaufine sa vision philosophique du monde, ce qui lui permet d'être plus efficace dans sa recherche. Le responsable mercatique connaît les évolutions dans les styles de vie, suit les modes et les habitudes, perfectionne sa capacité à deviner les tendances. L'homme politique connaît ses classiques idéologiques et suit les idées nouvelles avant qu'elles se répandent dans le public (et les sondages). Les étudiants lisent romans et essais pour forger leur personnalité propre...La société actuelle, par réaction contre la culture humaniste tend à oublier ou à dévaloriser l'information culturelle dans la pratique professionnelle. Il importe d'en tenir compte malgré tout dans une analyse des systèmes d'information, qui sont a priori basés sur d'autres préoccupations. La notion de savoirs transférables se développe, par opposition à l'apprentissage de savoir-faire étriqués. L'importance des bases culturelles qui permettent plus d'autonomie individuelle dans les entreprises est aujourd'hui largement soulignée, tant pour le cadre [LUS88], que pour les agents de production [PENI90]. La prise en compte de l'information culturelle, dans le cadre professionnel, et plus précisément encore dans le cadre général de l'activité des individus retrouvera une place déterminante.

Toutefois, la diffusion dans les systèmes interactifs de l'information culturelle est difficile. Par exemple, les ouvrages de réflexion sur la science sont totalement partie prenante du fonds d'une bibliothèque scientifique, mais attirer l'attention d'un utilisateur sur leur présence est difficile, surtout dans un système informatique, car l'utilisateur recherche souvent des documents plus spécialisés. L'information culturelle est certainement un modèle opératoire de la capacité à proposer des services d'information diffusée à l'intérieur même des systèmes interactifs : le concepteur du système peut envisager des méthodes d'alerte (au sens où les journalistes alertent l'opinion sur tel ou tel point) qui viennent compléter les méthodes interactives dans lesquelles l'utilisateur focalise sa requête.

L'information n'est que rarement l'objet même du travail documentaire. Elle ne constitue qu'un élément d'un processus général de prise de décision, et à ce titre répond à des critères précis, à chaque moment donné du développement d'un projet. Par exemple, la recherche d'information d'un étudiant préparant une thèse varie selon les étapes du travail. Dans une première étape le besoin principal est celui de documents de synthèse pour s'approprier le domaine de recherche (livre, revues de mise au point, autres thèses...). Puis vient la recherche d'articles portant sur le sujet principal de la thèse : articles de recherche, articles de méthodologie, articles sur les instruments employés. Ensuite, les orientations ayant changé avec la maîtrise du sujet, il convient de déplacer le besoin documentaire à partir de la connaissance des premiers documents jugés pertinents. La précision de la recherche est combinée avec un souhait d'exhaustivité. Dans la phase terminale, le besoin de références précises est plus marqué (références bibliographiques oubliées ou égarées). Enfin, la recherche des articles de personnes précises et la recherche d'organismes susceptibles de proposer des bourses d'étude accompagnent la préparation d'un stage post-doctoral à l'étranger.

Dans une opération de lancement d'un produit nouveau ou dans des opérations de regroupement d'entreprises pour des choix stratégiques, on pourrait définir de même des besoins documentaires qui varient dans le temps, en fonction des étapes du processus de décision : évaluation des produits existants, études de marché sur le secteur visé, articles de recherche et développement,

études sur la concurrence, évaluation des coûts de réalisation, recherche de technologie et de partenaires, évaluation économique et financière des partenaires repérés, recherche d'un réseau de distribution... A chaque étape des informations différentes sont utiles au décideur. Une information d'alerte permet de plus de conserver la fraîcheur des informations recueillies.

Juger l'*utilité* des informations, à chaque étape d'un projet, permet d'éviter le surplus d'information à un instant donné, et permet d'accompagner en permanence l'évolution des projets. Les instruments de ce processus global ne sont pas encore totalement adaptés, et cela ouvre des perspectives nouvelles pour les sciences de l'information. Car l'information dans ce cadre n'est pas seulement une denrée que l'on va puiser dans un fonds documentaire, mais au contraire un bien que l'on recherche, choisit, annote, fait circuler. De ce point de vue, la liaison serveurs d'information - messageries est précieuse. De même, les systèmes d'alerte qui proposent d'actualiser des questions d'utilisateur à chaque mise à jour des banques de données, et même en continu sur les réseaux d'information financière et d'actualité [IWR90b], sont aussi des instruments privilégiés. Enfin, les modèles documentaires qui permettent une navigation dans un stock d'informations, la constitution de chemins personnalisés et la possibilité de transmettre la clé de ces chemins à d'autres utilisateurs dans le cadre d'un travail collectif, à l'image des systèmes hypertextes, sont d'excellents supports pour la prise de décision. A ce titre, ils influent sur la conception d'ensemble des systèmes documentaires.

Dans la suite de cette thèse, nous nous placerons dans l'hypothèse d'un dialogue immédiat entre l'utilisateur et le fonds documentaire. De ce point de vue, la connaissance de l'environnement et des diverses étapes dans l'expression d'un besoin documentaire seront rapportées à un phénomène immédiat. Nous ne prendrons en compte que l'évolution au cours de la recherche de la *formulation* du besoin documentaire. Il semblait cependant important de replacer cette opération dans le cadre plus général de l'intégration d'une recherche documentaire dans un processus d'ensemble.

3.b - La distribution des banques de données

Si l'on définit les banques de données comme *"un ensemble de données relatif à un domaine défini des connaissances et organisé pour être offert aux consultations d'utilisateurs"* (Journal Officiel, 17 janvier 1982), on doit s'intéresser aux divers modes de diffusion, c'est-à-dire aux diverses manières de permettre à l'utilisateur d'appréhender cet ensemble de données. La distribution des banques de données emprunte de nombreux canaux :

- vente de bibliographies imprimées
- distribution en ligne par les serveurs professionnels
- distribution par le réseau vidéotex
- distribution par Disque Optique Compact (CD-Rom)

La vente de bibliographies imprimées reste le principal mode de diffusion des banques de données. L'origine de l'informatisation des services d'information repose sur la volonté de modifier, pour le rentabiliser, le processus de production des bibliographies imprimées. On retrouve cet aspect aussi bien dans l'activité des grands centres documentaires (Chemical Abstracts, PASCAL ou MEDLINE sont des banques de données créées dans ce cadre), que dans celui des bibliothèques (l'accès en ligne aux catalogues informatisés n'est qu'un enjeu récent, passant après la production de microfiches ou de catalogues imprimés).

Le réseau vidéotex, surtout développé en France, permet d'adapter la recherche à certains utilisateurs novices, notamment en favorisant deux points :

- la connexion sans mot de passe avec prélèvement du coût de la recherche sur la facture téléphonique (système dit *"kiosque multipalier")*

- l'établissement de l'équation de recherche est guidé par menu et la lecture des informations suit le rythme des pages écran.

L'ouverture des banques de données au réseau vidéotex se fait en général par l'ouverture d'un frontal spécialisé, qui intervient comme intermédiaire entre l'utilisateur et la banque de données, gérant les écrans et le transcodage, et transposant la question de l'utilisateur dans les formes adaptées à la banque de données (langage de commande). Toutefois, la formulation de la question reste à

la charge de l'utilisateur, ce qui distingue l'accès vidéotex des systèmes intelligents d'anté-serveurs.

Les Disques Optiques Compacts permettent à l'utilisateur de disposer d'une banque de données sur place, sur un micro-ordinateur. De la même manière que l'on peut lire et relire une bibliographie imprimée une fois qu'elle a été achetée, on peut consulter un temps indéterminé les D.O.C., les coûts ayant été couverts une fois pour toutes avec l'achat de la licence d'utilisation.

Sur D.O.C., la banque de données devrait pouvoir être consultée directement par l'utilisateur, en s'appuyant sur l'ensemble des innovations propres au monde de la micro-informatique (souris, menus déroulants, multi fenêtrage...) et en intégrant les acquis de la recherche en informatique documentaire que nous développerons dans cette thèse (jugement de pertinence, classement des documents en fonction de leur pertinence formelle, recherche sur les agrégats...). Malheureusement, la communauté des éditeurs et producteurs d'information ont réduit la diffusion en D.O.C. à une *"édition en livre de poche des banques de données",* pour reprendre l'expression de Julie Schwering [SCH86], [LEC86]. Cela est significatif des lenteurs avec lesquelles les innovations et les recherches dans le domaine documentaire viennent influencer les stratégies des acteurs financiers et commerciaux.

La majeure partie des D.O.C. en vente aujourd'hui sont conçus sur le même modèle que les produits destinés aux intermédiaires en information, avec un simple dépoussiérage de la présentation des écrans. Dès lors, on peut s'interroger sur l'avenir de ce type de produit pour la distribution des banques de données. Le D.O.C., même s'il peut contenir 550 méga-octets, reste d'une capacité faible pour les systèmes d'information, obligeant à découper le fonds documentaire ; de plus, il conduit à des temps de réponse relativement longs. Enfin, les D.O.C. sont relativement chers, et certainement, pour de nombreux utilisateurs, plus onéreux que la connexion en ligne (compte tenu du taux d'utilisation bien entendu) [ZIN90]. La constitution d'une collection de D.O.C. est difficile, car ils sont mis à disposition des bibliothèques sur la base d'une licence d'utilisation, et non d'un achat ferme et définitif. Pourtant, le D.O.C. aurait pu être un merveilleux tremplin pour des idées nouvelles en sciences de l'information (hypertexte, modèles à jugement de pertinence) ce qui aurait

accentué les pratiques sociales de recherche d'information (pas de coût porté sur l'utilisateur). Autant de bénéfices qui se seraient certainement reportés sur l'utilisation en ligne des sources d'information.

Les serveurs professionnels sont issus des recherches pour permettre l'accès en ligne en mode interactif aux informations scientifiques. Le premier programme d'ampleur a été développé pour que tous les participants au projet *Apollo* puissent accéder aux banques de données de la NASA. Des recherches sur ce mode de diffusion de l'information naîtra le serveur *Dialog*. Parallèlement, la *National Library of Medicine* mettra en ligne la banque de données médicale *Medline* en 1971. Le premier serveur français fut confié à Télésystèmes-Questel à la fin des années 70 [MEA88].

Les serveurs professionnels n'ont pas été conçus pour être manipulés directement par les utilisateurs. Ils offrent cependant des moyens d'intervention et de contrôle importants, dans le cadre de la recherche booléenne, pour les intermédiaires spécialisés très formés. Traditionnellement, on distingue quatre acteurs dans la diffusion des informations en ligne :

. *le producteur d'information* collecte, indexe et organise le fonds documentaire

. *le serveur* se charge de permettre l'accès aux informations via le réseau téléphonique. Il a de plus un rôle de formation des utilisateurs et de gestion financière.

. *le réseau de télécommunications* assure la fiabilité des liaisons entre l'utilisateur, représenté en général par l'intermédiaire en information, et le serveur.

. *l'intermédiaire en information,* bibliothécaire ou documentaliste, qui est formé aux langages d'interrogation des serveurs, et qui transcrit les questions des utilisateurs dans une forme acceptable dans la banque de données considérée.

Cette distinction est de moins en moins opératoire, alors que producteurs et serveurs appartiennent souvent aux mêmes sociétés financières. Cette concentration d'ordre monopolistique, de plus en plus sensible, est rendue encore plus urgente par la structure des investissements nécessaires pour suivre le

rythme technologique [BEN89]. Des opérateurs comme *Maxwell* possèdent ainsi des serveurs *(PFDS, Orbit, BRS),* des producteurs d'information *(Pergamon)* et des forces éditoriales importantes, notamment dans le domaine scientifique et dans la presse. Le groupe de presse *Knight Ridder* a racheté le serveur *Dialog,* numéro un mondial, et propose des informations économiques mises à jour en continu, en synergie avec les titres du groupe. A partir du levier financier du *Wall Street Journal, Dow Jones* dispose d'un serveur professionnel et d'actions dans *Télébase,* qui propose des systèmes anté-serveurs (réseau *Easynet).* Certains groupes importants essaient de créer des serveurs à partir de leur monopole sur certains types d'information. Le réseau *STN,* impulsé par *Chemical Abstracts,* tend ainsi à devenir un acteur primordial dans le domaine scientifique et technique, ce qui n'est pas sans poser des problèmes pour l'accès à l'information [IWR90a].

Cette situation monopolistique est confrontée à un autre monopole, en général d'origine publique, celui des opérateurs de télécommunication. L'intérêt de ces derniers est un accroissement de la circulation d'informations sur le réseau, et un renouvellement des postes terminaux d'abonnés. C'est par exemple la stratégie de *France Télécom* en ce qui concerne le minitel. La large diffusion du terminal de base a permis le décollage du réseau vidéotex, ce qui conduit aujourd'hui *France Télécom* à s'orienter vers la location-vente de terminaux plus perfectionnés, mêlant téléphonie, vidéotex, et service de messagerie *(Minitel 12).*

La question de la déréglementation révèle une opposition entre deux types d'opérateurs : ceux des télécommunications et les serveurs informatiques (dans ce dernier groupe, les serveurs d'information restent une quantité négligeable, à côté des serveurs de transaction - banques...- et des serveurs de réseaux professionnels -transports, voyage, tourisme...-). Globalement, les opérateurs de services voudraient bénéficier de tarifs préférentiels sur le transport des informations pour permettre un élargissement de leur marché, alors que les opérateurs de télécommunication voudraient conserver la possibilité de percevoir une quote-part des richesses produites par les nouveaux services pour rentabiliser l'ensemble de l'infrastructure du réseau téléphonique. En langage codé, les opérateurs de télécommunication disent compenser sur les réseaux professionnels le coût du "poste de l'abonné terminal en campagne". En réalité l'enjeu reste le passage aux nouveaux types de commutation temporelle

asynchrone qui en multipliant les possibilités de faire circuler plusieurs types d'informations sur le réseau, y compris l'image animée, renforcera la place sociale du lobby télé communicant. Il existe une véritable guerre de pouvoir entre les télécommunications, l'audio-visuel et l'informatique. A l'heure où la télévision elle-même devient un instrument complexe, intégrant de nombreuses compétences informatiques (mémoire de trame, instruments de sélection...) et pouvant être reliée à des sources d'informations nouvelles par le câble, le marché de la diffusion de l'information devient potentiellement énorme. Cette situation justifie les prétentions de chaque filière. Les expériences de l'information professionnelle en ligne et du vidéotex ne sont que des premiers pas, permettant de marquer des points dans cette gigantesque partie de Monopoly informationnel. Les leçons qui peuvent en être tirées sont aussi à jauger en fonction de cet enjeu. Développer la convivialité des recherches documentaires est le moyen d'apprendre à développer d'autres types de liaisons entre un usager et un système informatique (banques d'images, vidéothèques communicantes, bureautique répartie...).

Les systèmes documentaires en ligne ont l'avantage de posséder une très grande capacité de stockage, permettant de retrouver presque instantanément une information précise parmi des millions de références *(Chemical Abstracts :* 8,5 Millions ; PASCAL : 7 Millions ; *Nexis :* plusieurs centaines de titres de périodiques en texte intégral...). Ces systèmes sont disponibles simultanément pour de nombreux utilisateurs. La remise à jour des informations est immédiate, et commence même à se réaliser en continu pour les banques de données des agences de presse et les banques de données financières *(Reuters, Knight Ridder,...)* [IWR90b]. Les réseaux de transmission deviennent de plus en plus rapides ce qui permet d'envisager des transmissions de graphiques et d'images fixes. Le réseau *Numéris* permet ainsi des transferts à 16 Kbits par seconde pour les données (aujourd'hui en mode message sur le canal sémaphore, mais certainement demain en continu) et de 64 Kbits par seconde pour le son et les images.

Dans le strict domaine des services d'information, on assiste à un couplage entre des services de messagerie et des services de banques de données. Les serveurs professionnels *(Dialog, Questel...)* et les réseaux d'intermédiaires électroniques *(Easynet, Geomail)* proposent de coupler les recherches

documentaires et les services de messageries professionnelles (envoi par un responsable sur plusieurs boîtes aux lettres des informations recueillies). D'autre part, on assiste au développement de terminaux "intelligents", c'est-à-dire basés sur des micro-ordinateurs. Cette architecture permet de récupérer les informations pour un post-traitement en mode local (intégration dans le système d'information interne de l'entreprise, impression de qualité, utilisation avec des Systèmes Interactifs d' Aide à la Décision...). Elle permet aussi le développement de systèmes d'aide à la formulation des questions aux serveurs basés sur le micro-ordinateur de l'utilisateur. Dans cette catégorie, on trouve les logiciels de communication généralistes, des logiciels adaptés proposés par les serveurs *(Dialoglink, STN Express)* qui permettent de préparer les questions en mode local et de disposer de choix par menus, notamment pour l'élaboration des questions structurales en chimie.

3.c - Quelques scénarios pour l'avenir

L'information est un bien dont la valeur, mesurée à l'aune des enjeux économiques, stratégiques et géopolitiques, devient de plus en plus sensible, à la fois pour les gouvernants chargés de réglementer et d'impulser son développement, et pour les utilisateurs, de plus en plus sensibilisés à la place qu'occupé la recherche d'information dans la prise de décision. Dans ce cadre, la supériorité des systèmes en ligne sur les banques de données sur support optique est flagrante. Le support optique est adapté à des documents qui varient peu dans le temps (absence de l'information d'alerte), et à des documents utiles en permanence autour du poste de travail (dictionnaires spécialisés, textes de base d'une profession, normes...). Le D.O.C. aurait cependant pu jouer la carte d'un remplacement de la diffusion des banques de données par des bibliographies imprimées. Plus économe que l'imprimerie, industriellement comme écologiquement, le D.O.C. joue dans ce même registre d'une information achetée comme un bien (possession) et stockée pour longtemps. Pour que ce scénario ait des chances de voir le jour, il faudra cependant que les acheteurs traditionnels d'ouvrages de référence imprimés (bibliothèques, centres documentaires importants...) adoptent le D.O.C., et installent des lecteurs à destination du public. Mais cela ne sera vraiment envisageable que si deux conditions sont réunies :

- une garantie de pérennité de l'information achetée sur un média optique.

Pérennité physique (stabilisation des normes, fonctionnement des appareils de lecture...) mais aussi pérennité du droit de lire les informations et de les diffuser en
mode local dans la bibliothèque. Or aujourd'hui, ce qu'acheté la bibliothèque est une "licence d'utilisation", ce qui limite la confiance que cette profession peut porter à ce média (qui décidera des règles du jeu demain ?)

- possibilité pour l'utilisateur d'obtenir un degré de satisfaction de ses besoins documentaires comparable avec la recherche sur des instruments imprimés. C'est l'enjeu de la recherche en informatique documentaire, telle qu'elle sera étudiée dans cette thèse. C'est la chance du D.O.C. de pouvoir mettre en place certains de ces modèles dès maintenant (hypertexte, modèle probabiliste, jugement de pertinence...). On ne peut de ce point de vue qu'être étonné du peu d'empressement des éditeurs à proposer des D.O.C. qui soient réellement concurrentiels face aux instruments imprimés, se contentant de singer (avec moins de moyens : faible capacité, lenteur) les méthodes de l'information en ligne.

Devant cette situation, on doit penser que la véritable innovation viendra du monde de l'information en ligne. A partir de quelques expériences actuelles, on peut tenter de désigner quelques points forts des évolutions à venir.

- L'évolution du poste de travail. La généralisation des stations de travail *(Next, Mac II*) et des écrans de large dimension est parfaitement adaptée au travail documentaire pour deux raisons :

. la capacité des systèmes d'exploitation à être réellement multitâche permet d'envisager un processus d'alerte et un processus de communication en direct pendant que l'utilisateur est occupé à d'autres tâches sur son ordinateur.

. le multi fenêtrage et l'espace pour étaler sur son écran les divers documents utiles dans le cadre d'un travail, permettant de passer rapidement de l'un à l'autre est un besoin fondamental.

La capacité à travailler sur plusieurs livres en même temps est récente. Elle émerge à la fin du Moyen-âge avec l'apparition de la forme *codex* [MAR88]. C'est une révolution technique qui a permis une modification profonde des

habitudes culturelles. On peut attendre de la révolution similaire des postes de travail des effets aussi profonds. On assiste ainsi au développement de postes de *"lecture électronique",* à la demande de la *Bibliothèque de France,* qui intègrent les besoins ergonomiques de la lecture dans les systèmes informatiques.

Dans le domaine documentaire, la capacité à lire aisément les écrans est déterminante dans la capacité à comprendre et agir dans le cadre du processus itératif de la recherche documentaire. Les modèles à jugement de pertinence, par exemple, ne se conçoivent efficacement que si l'utilisateur peut tranquillement et confortablement lire les premières informations qui lui sont transmises par le système.

- l'évolution des messageries. Les messageries actuelles ne permettent pas aisément le transfert de fichiers, de textes présentés, de lettres correctement accentuées...Le développement des normes de messageries (X 400) d'une part et d'autre part la généralisation des transferts de protocoles entre matériels hétérogènes permettent d'envisager à court terme l'existence de messageries de qualité. Des applications comme les langages de description de page adaptés aux écrans *(Display PostScript)* ou les normes O.D.A. permettront d'enrichir le texte des messages avec la typographie adaptée, garantissant une meilleure communication. Dès lors, les messageries prendront une part importante dans le travail des professions intellectuelles. Il conviendra toutefois de régler le problème des règles sociales de ce média : comportement des individus dans les forums, éviter le courrier superfétatoire, utilisation régulière...

Du point de vue de l'accès à la documentation, la messagerie représente un bon moyen de diffuser l'information d'alerte. Soit directement à partir d'une prescription de l'utilisateur (profil documentaire), soit par l'intermédiaire d'un répartiteur d'information (documentaliste, responsable organisationnel ou intermédiaire électronique). Le couplage messagerie/système documentaire est un axe de développement déterminant, en ce qu'il permet au sein d'une organisation d'utiliser en symbiose un média diffusé (la messagerie) et un média interactif (le système documentaire).

-La mise en place d'anté-serveurs. Le mode d'accès direct de l'utilisateur aux banques de données n'est pas une solution d'avenir. D'une part il

suppose une pratique régulière de la recherche documentaire informatisée (multiples langages, règles d'indexation, logique booléenne...) d'autre part il suppose une lourde gestion des frais et des abonnements aux serveurs. L'avenir est à des systèmes uniques permettant l'accès à toutes les banques de données au travers d'une même interface utilisateur. Ces anté-serveurs *(gateways)* permettront l'immersion dans le réseau des serveurs de banques de données, qui deviendront un nœud supplémentaire du réseau, possédant des qualités particulières (rapidité de traitement et puissance de stockage), quand un autre système se chargera de la relation avec l'utilisateur.

Cette idée peut s'exprimer par la métaphore d'une ligne de flottaison. Régulièrement des pans entiers du système de télécommunication sont engloutis sous la ligne de flottaison, disparaissant aux yeux de l'utilisateur. Ainsi en fut-il des "opératrices" du téléphone, remplacées par les centraux automatiques, puis par des terminaux plus sophistiqués, où il suffit d'appuyer sur une seule touche pour que ça sonne chez le correspondant préenregistré. Ainsi en est-il des serveurs vidéotex, inconnus du public, qui ne s'intéresse qu'aux services proposés. Ainsi en sera-t-il des serveurs professionnels de banques de données. Les anglais utilisent le terme de *"one stop shopping"* pour dénommer cette idée d'un "centre commercial de l'information", regroupant dans une même économie de lieu (un seul code d'accès), de pratique (les "étagères" à information ont toutes la même disposition - i.e. interface) et de facturation, les diverses activités informationnelles d'une organisation ou d'un utilisateur.

On peut distinguer plusieurs stratégies de constitution des anté-serveurs :

. un système électronique traduit les questions d'un utilisateur dans le langage de commande du serveur. C'est la stratégie du réseau *EasyNet* . Les limites de cette méthode sont liées au savoir-faire actuel : la formulation, le choix des termes de la question et des connecteurs booléens est entièrement aux soins de l'utilisateur; il n'existe aucun moyen d'utiliser le jugement de pertinence de l'utilisateur, et la réponse est limitée à 20 documents, en général les plus récents, ce qui est loin de couvrir la variété des attentes des utilisateurs. L'avantage de ce type de système est de permettre à la personne formée dans un langage de commande de voir ses questions traduites en d'autres langages si les banques de données qui l'intéressent sont sur un autre serveur. C'est un avantage

appréciable par rapport aux simples passerelles de reroutage, qui permettent certes d'accéder à un serveur dont on n'est pas l'abonné à partir d'un serveur connu, mais qui ne remplissent pas cette fonction de traduction des commandes.

. le système électronique d'interrogation des banques de données est une activité annexe d'un réseau de messagerie. C'est la stratégie de *Geomail.* Les résultats des recherches sont portés dans la boîte aux lettres d'un ou plusieurs utilisateurs (liste de diffusion).

. le système traite une série de profils correspondant à un type d'utilisateur déterminé. Il s'agit de diffuser une information d'alerte, régulièrement mise à jour pour des groupes d'utilisateurs compacts. C'est la stratégie de *Global Report* pour l'industrie bancaire. Les limites tiennent à l'uniformité des profils, qui ne sont pas adaptés à chaque utilisateur particulier.

. le système fait le lien entre le système interne d'une entreprise et les banques de données. Chaque utilisateur peut accéder, au travers du système dont il a l'habitude, à une information spécifique correspondant à sa fonction dans l'organisme. Plus proche des besoins exprimés, cette stratégie, utilisé par *NEIS (Networked Executive Information System),* rappelle, en version électronique, les lettres d'information spécialisées. On s'abonne en quelque sorte à un profil pré-déterminé.

. sur le poste de travail de l'utilisateur existe un programme fonctionnant en continu qui filtre les informations versées dans les serveurs, notamment dans ce cadre les serveurs d'agence de presse, en fonction des besoins spécifiés par l'utilisateur. L*'Agence France Presse* diffuse ainsi un produit qui permet de recevoir sur un ordinateur le fil de l'agence, et de plus de pointer toutes les nouvelles correspondant à un événement particulier. Le logiciel documentaire pour micro-ordinateurs *Topic* fonctionne ainsi sur le réseau de l'agence *Reuters* [IWR90b]. La question est préparée par un documentaliste (le système *Topic* est relativement complexe car il permet de pondérer les termes de recherche), et une fenêtre de l'écran est affectée au suivi régulier. Avec le développement du chargement en continu des banques de données, ces systèmes ont certainement un avenir important.

. l'utilisation d'un système expert, en mode local ou sur le serveur d'information, pour préparer les questions des utilisateurs en s'appuyant sur les thésaurus de certaines banques de données. *MenUse* de S. Pollit [POL88] permet la recherche dans *Medline* au travers d'une manipulation intelligente du *MeSH* (thésaurus de la *National Library of Medicine). Tome Searcher* de la société britannique *Tome Associates* permet les recherches sur l'électronique dans *INSPEC*

. enfin, une stratégie visant à permettre à l'utilisateur d'aborder les banques de données comme si celles-ci fonctionnaient en dehors du mode booléen, au moins en proposant un classement des réponses par ordre de pertinence et en permettant une reformulation automatique de la question. Même si les difficultés sont grandes en raison de la structure même des fichiers inverses sur les serveurs professionnels qui ne sont pas optimisés pour une recherche de ce type, on peut simuler certains calculs statistiques et certaines évaluations de pondération permettant de classer les documents selon une fonction de pertinence formelle. Cette stratégie, expérimentée depuis plusieurs années [MOR82], se fixe comme objectif de remplacer l'intermédiaire en information, et de plus de proposer un mode d'accès complètement différent aux banques de données. Ces *"intermédiaires informatiques intelligents"* s'appuient sur les recherches en science de l'information (jugement de pertinence, modèle probabiliste, analyse en ligne de mots clés, reformulation...) et en intelligence artificielle (linguistique informatique). Il s'agit aujourd'hui de systèmes expérimentaux, dont les recherches sont encouragées par l'Etat *(Ministère de la Recherche et de la Technologie)* ou par la C.E.E. *(DG XIII).* Participant à un contrat de recherche de ce type *(M.R.T., société Triel, Université de Caen),* nous décrirons plus en détail dans la troisième partie les fonctionnalités de ce type d'anté-serveurs.

On peut s'interroger sur la meilleure stratégie d'implantation des anté-serveurs, notamment en se demandant s'il est préférable que les nouvelles fonctions de recherche "intelligente" soient menées par un ordinateur indépendant des serveurs ou utilisées en frontal des serveurs. La structure informatique basée sur des anté-serveurs indépendants des serveurs de banques de données est certainement la plus souple, car elle permet de disposer de compétences linguistiques, de méthodes de manipulation des documents similaires pour plusieurs sources d'information. Cependant, de nombreuses

recherches menées pour la réalisation des anté-serveurs seront reprises par les serveurs pour proposer des frontaux d'interrogation plus adaptés à l'utilisateur. Dans le cadre de cette compétition, on assiste déjà à des regroupements entre exploitants de serveurs et producteurs d'anté-serveurs *(Cartexpert* en France qui associe *Télésystèmes* et le *Crédit Lyonnais* pour créer un anté-serveur). Il semble que l'impulsion sera cependant donnée par des acteurs qui ne seront pas directement dépendants des serveurs. La logique des anté-serveurs est d'évoluer pour couvrir tous les types de questions, c'est à dire accéder à toutes les sources d'information, et de proposer des services complémentaires : impressions personnalisées, versement suivant des listes d'adresses dans des réseaux de messageries, règlement simplifié (monétique), analyses statistiques pour la veille technologique... A terme, les serveurs deviendront des fournisseurs de services pour ces réseaux à valeur ajoutée, tout en restant accessibles par des spécialistes de la documentation. A l'image du vidéotex professionnel, le marché des recherches d'information via des anté-serveurs est un nouveau marché, qui s'adresse à une autre clientèle que le marché traditionnel des serveurs.

4 - Systèmes documentaires et bibliométrie

La bibliométrie est l'étude quantitative de la littérature scientifique, technique et économique, telle qu'elle est représentée dans les systèmes bibliographiques. Son objectif est d'extraire des données utiles pour comprendre et évaluer l'évolution des sciences, des techniques et de la recherche universitaire à partir de l'organisation en système bibliographique de l'ensemble de la littérature spécialisée [WHI89]. A l'origine de cette activité on trouve la conception que la science et le savoir ne se réalisent et progressent qu'en fonction de la production de documents (livres, articles, brevets...) par les chercheurs. Alors, l'évolution du savoir, les nouveaux domaines couverts, l'étude sociologique de la science et de ses réseaux... s'inscrivent dans la matérialité des documents. Une analyse des documents devrait en conséquence permettre de rendre intelligibles des phénomènes qui échappent aux acteurs eux-mêmes et aux décideurs politiques et économiques.

4.a - La production documentaire scientifique et technique

Limiter la science à ses productions permet un mode d'évaluation de la recherche qui intéresse au plus haut point les organismes de financement. Même si ce n'est pas l'objet de cette thèse, on peut émettre quelques réserves à cette conception :

- y a t il science sans enseignement ? En d'autres termes, le soin porté à la diffusion de la science, par les voies orales en général et par le contact avec les élèves, n'est-il pas un moyen de créer et faire évoluer la recherche fondamentale ? Les élèves, en résolvant les problèmes typiques de chaque secteur scientifique, contribuent à la stabilisation d'un savoir dans une discipline, et apprennent ainsi ce qui est le paradigme dominant à un moment donné dans cette discipline. Ce faisant, cette pratique éducative tend en retour à permettre de mieux maîtriser les éléments cœurs de cette discipline, et à permettre le changement de paradigme dans les disciplines évolutives.

- y a t il égalité devant la production de documents scientifiques, et plus encore devant l'acceptation de cette production dans les systèmes bibliographiques ? Si l'on doit analyser la science à l'aune de la reconnaissance par un certain type de système, il convient de s'interroger pour savoir si les bases de calcul elles-mêmes ne sont pas fragiles, et déjà biaisées. *Science Citation Index (SCISEARCH)* est la banque de données la plus utilisée dans les recherches bibliométriques. Or sa couverture reste faible, principalement concentrée sur les périodiques anglo-saxons, ce qui tend à favoriser le type de recherche qui intéresse les pays développés, et en leur sein les réseaux de grands laboratoires. On pourrait penser qu'il en est ainsi parce que les périodiques non inclus dans *SCI* sont d'une importance faible. Cette affirmation reste largement discutable.

Des chercheurs en bibliométrie indiens [SEN89] ont ainsi calculé le facteur d'impact de *Indian Journal of Malariology,* un journal indien non inclus dans *Science Citation Index.* Le facteur d'impact d'un périodique est une mesure de la fréquence de citation d'un article moyen publié dans ce journal au cours d'une année particulière. Ce facteur est sensé indiquer la croissance ou la décadence d'un journal dans la communauté scientifique. Pour l'année x, il est calculé par le rapport entre le nombre d'articles citables publiés les années x-1 (c_1) et x-2 (c_2)

et le nombre de citations effectuées l'année x (ct_x), soit :

$$I_f = \frac{ct_x}{c_1 + c_2}$$

On estime à 53 000 le nombre de périodiques scientifiques publiés dans le monde, alors que seuls 3 800 d'entre eux sont pris en considération pour le calcul du facteur d'impact. On pourrait croire que les 49 000 périodiques restant seraient ceux qui ont un facteur d'impact très faible. On trouve dans *S.C.I.* des périodiques ayant un facteur d'impact de 0,001 ou moins. Or le *Indian Journal of Malariology* a un facteur d'impact calculé de 0,528, soit bien supérieur à de nombreux périodiques incorporés dans *SCI.* Un exemple qui se répète pour de nombreux périodiques scientifiques et techniques des pays en voie de développement ou des pays ex-communistes.

- y a t il continuité dans la production de savoir par un individu ou un laboratoire, mesurable par le nombre et l'impact immédiat des publications ? N'existe-t-il pas des périodes de latence, pendant lesquelles des nouvelles découvertes apparaissent sans être suffisamment structurées pour faire l'objet de publications ? La course à la publication a dans le domaine scientifique des effets négatifs comparables à l'accélération de la vie d'un livre dans le domaine littéraire. Aux *"fast-books"* correspond la *"fast-science",* annoncée à grand renforts de médias, ignorée sitôt après, mais qui aura produit son effet dans la course à la subvention (cf. l'annonce permanente et répétée de remèdes miracles contre le SIDA). Dans le même ordre d'idées, si les décisions politiques devaient se faire strictement en fonction de l'analyse des publications, on verrait les laboratoires se spécialiser dans les domaines où ils ont des chances d'être publiés. Aucune nouvelle discipline ne verrait le jour, car pour qu'une discipline émerge, il faut former les chercheurs, les laisser tâtonner et hésiter, organiser des enseignements pour assurer le renouvellement des forces de recherche et la diffusion de modèles et points de vue... Ce travail est certainement long, et doit donc faire l'objet de choix, de décisions qui ne peuvent être réduites à la lecture d'analyses sur le passé ou sur le simple présent. L'absence de reconnaissance d'une discipline comme les sciences de l'information et de la communication en France n'est pas étrangère à ce phénomène : quel chercheur accepterait facilement de s'investir dans un domaine où il ne serait pas productif immédiatement, surtout si ce domaine est déjà balisé par une histoire et un réseau de chercheurs ?

Ces remarques permettent de regarder avec un œil différent les recherches bibliométriques. Sans remettre en cause les découvertes que l'on peut faire sur la science publiée, on doit se garder de les extrapoler sur la science réelle, et sur les politiques de développement qui doivent lui être associées. La bibliométrie offre les moyens de tisser une image de la science dominante. Elle s'efforce de distinguer les évolutions dans ce cadre et de pointer au mieux les secteurs en expansion. Ses conclusions doivent permettre aux décideurs de disposer de plus d'informations, mais ne peuvent en aucun cas être utilisées comme seuls instrument de jugement. Or la frontière est difficile à définir. Les lobbies puissants correspondent à des laboratoires puissants et à des réseaux de publications structurés et reconnus. Ils tendent alors à renforcer leur propre puissance en s'appuyant sur la soi-disant objectivité des analyses bibliométriques. Il semble cependant nécessaire de plaider sans cesse pour l'innovation dans le domaine scientifique et technique, innovation qui peut souvent venir en dehors des lobbies dominants. L'exemple récent de la percée de la micro-informatique, née dans les garages de la *Silicon Valley* et rejetée à l'origine par la communauté "sérieuse" de l'informatique [LEC86a], vient rappeler que l'innovation n'est pas un strict produit de tendances décelables, mais correspond parfois à l'irruption de phénomènes ou de découvertes radicalement extérieurs aux réseaux installés de la science et de la technique.

4.b • Les indicateurs bibliométriques

Les indicateurs utilisés par la bibliométrie sont de plusieurs ordres :

. les termes employés pour décrire une recherche : mots-clés, résumés, mots du titre.
. le choix du journal dans lequel l'article est publié.
. le réseau des auteurs et co-auteurs des documents publiés sur un certain sujet
. le réseau des citations et co-citations faites par un document envers d'autres documents. Cet indicateur est certainement le plus utilisé, notamment pour évaluer l'audience d'un travail scientifique. Cela n'est pas sans poser des problèmes de fiabilité de l'indicateur.

Pour déterminer la place respective de ces indicateurs, il convient de revenir au fonctionnement général des publications scientifiques.

Les fonctions des périodiques scientifiques fondamentaux ont été définies dès l'origine en 1665 *(Journal des Scavans, Philosophical transactions of the Royal Society) :*

. rendre public le résultat des recherches pour l'audience la plus large possible
. conserver une trace permanente de toutes les recherches menées
. prendre date (revendication d'antériorité sur un travail de recherche)
. assurer un niveau garanti de qualité scientifique par le système de jugement par les pairs *("referees").*

Ces fonctions encadrent et justifient la pratique des citations. Notamment, un des rôles du *"référée"* consiste à déterminer si un article proposé pour publication traite correctement les publications antérieures et ne s'arroge pas des découvertes déjà connues et publiées par d'autres. Jill Lambert [LAM85] distingue ainsi plusieurs raisons à la citation de documents :

. rendre hommage aux pionniers
. donner crédit aux autres travaux
. identifier les méthodes et les équipements utilisés dans les expériences décrites dans l'article
. conseiller des lectures de base
. corriger son propre travail ou celui des autres
. critiquer les travaux antérieurs
. donner une plus grande valeur à ses recherches en faisant valoir la communauté de pensée avec d'autres chercheurs
. alerter les chercheurs sur des travaux à paraître
. donner accès à des travaux peu cités ou peu décrits dans les répertoires bibliographiques
. authentifier des données ou des constantes
. identifier les publications dans lesquelles des idées ou des concepts ont émergé
. identifier la publication originale de ce qui devient une "loi"

reconnue par la communauté scientifique concernée
. réduire les revendications des autres chercheurs et discuter les antériorités proclamées par d'autres.

Ce fonctionnement du système des citations reste largement idéal. La pratique réelle est entachée de larges biais. Un auteur ne peut citer que les documents qu'il connaît, ce qui restreint le champ couvert, par exemple en fonction des compétences linguistiques de l'auteur ou de sa situation géographique, mais aussi en fonction de la date de publication (il existe souvent un délai important entre la date de rédaction d'un manuscrit et la date où il est réellement publié) [SAN89].

Dans une revue des problèmes posés par l'analyse des citations, Michael et Barbara MacRoberts [MACR89] repèrent ainsi 7 types de problèmes :

. des influences non citées. Les tenants de l'analyse des citations considèrent que ce problème est réparti de manière aléatoire, mais reconnaissent que nul n'a vraiment étudié cette assertion. Or ces auteurs estiment, à la suite d'une étude sur 15 articles concernant l'histoire de la génétique, que la couverture totale des articles cités sur les articles qu'il aurait fallu citer est seulement de 30%.

. des citations erronées, souvent renvoyant à des sources secondaires plutôt qu'à l'auteur principal d'une découverte. Il existerait ainsi une différence entre la fréquence d'utilisation d'une idée ou découverte et sa citation effective.

. des influences informelles non citées. ou qui ne sont pas prises en compte dans les analyses de co-citation (par exemple les remerciements dans les articles, ou les citations portant la mention "communication personnelle").

. l'autocitation qui concerne entre 10 et 30 % des citations. Ce problème est d'autant plus important que certains articles sont écrits par de très nombreux auteurs, et en conséquence fortement cités.

. la différence de nature entre toutes les citations, qui n'est pas prise en compte dans les analyses statistiques. Les intentions d'un auteur dans son

choix de citations sont souvent obscures. Par exemple, les citations négatives sont en général omises, ce qui est étonnant, la science étant sensée progresser par la critique.

. la variation du taux de citation en fonction du type de publication, de la nationalité, du type de spécialité ou de la taille. On constate ainsi que les mathématiques ou les sciences de l'ingénieur, avec une moyenne de 5 à 6 citations sont loin derrière la physique ou la chimie (12 à 15 références) et la recherche biomédicale (18 à 20 citations). Cette différence est encore plus sensible avec la tendance générale à l'accroissement du nombre de citations, vraisemblablement liée à la facilité de recherche (banques de données) et de gestion (systèmes documentaires personnels) de l'information bibliographique.

Le problème est accentué par les questions linguistiques, les pratiques nationales, l'accessibilité des documents dans les réseaux de bibliothèques... Par exemple, le nombre moyen de citations des publications scientifiques soviétiques serait seulement de 50 à 75 % du nombre moyen de citations des articles américains.

. les limitations pratiques à l'analyse des co-citations. par exemple les erreurs dans les citations, l'existence de diverses orthographes du nom d'une même personne (notamment au niveau de l'écriture des prénoms) et la présence d'homonymes. De même, l'analyse des citations ne peut se réaliser qu'à partir des banques de données qui donnent cette information, notamment *Science Citation Index* pour les articles et *Derwent (WPI)* ou les banques de données des organismes nationaux pour les brevets. Or ces banques de données ne sont pas exemptes de limites. On peut considérer ainsi le taux de couverture du *Science Citation Index*. Cette banque de données ne couvre que 10 % de la littérature scientifique, avec un taux très faible des langues non romanes, un choix différent selon les disciplines scientifiques (6 % des périodiques de biologie, mais 14 % de ceux de médecine clinique) et un choix idéologique et politique en économie (Wiener, cité par [MACR89] indique qu'une note apparaissant dans *Commentary* ou *Public Interest* sera prise en compte, alors qu'aucune mention ne sera faite si elle apparaît dans *New Politics* ou *Social Policy)*. Cette banque de données ne repère les citations que par le premier auteur des articles. Cette limite empêche une analyse approfondie des réseaux de co-auteurs et de leur influence sur la pratique des citations.

Ces limites sont souvent reconnues par les partisans de l'analyse des citations, qui cependant estiment qu'elles sont de faible conséquence sur l'image générale de la science. Ceci est certainement vrai pour la science qui concerne les principaux centres de décisions de la recherche, notamment les Etats-Unis et les grands laboratoires des pays développés occidentaux en général. On peut cependant douter de cet optimisme dès lors que l'on considère la science comme un phénomène mondial, et dès lors que l'on cherche à repérer les biais introduits dans la pratique des citations et que l'on veut pondérer et corriger en fonction de ce travail les analyses bibliométriques.

4.c - Quelques résultats fondamentaux

On peut distinguer quelques résultats fondamentaux de la bibliométrie [WHI89]. Globalement, la bibliométrie s'attache à la variation d'un indicateur bibliométrique en fixant les autres. On peut alors définir les études de la lignée de Bradford, de Zipf et de Lotka.

Les études de Bradford considèrent le journal de publication comme l'indicateur principal. La loi de Bradford détermine le nombre de titres de périodiques nécessaire pour couvrir un domaine (les indicateurs fixes sont alors les termes d'indexation définissant le domaine choisi). Les études de Lokta sont basées sur les distributions des auteurs quand d'autres critères sont fixés. Celles de Zipf sont plus proches de la linguistique statistique, et considèrent la distribution des fréquences des mots dans un texte, mais aussi la répartition des termes d'indexation dans une banque de données. Les distributions observées sont en général comparables si on les rapporte à des distributions rang fréquence. Elles permettent alors de déterminer parmi les indicateurs un *"cœur"* et une *"traîne"* (traduction libre de *Core and Scatter).*

. le "coeur" tend à regrouper les quelques éléments qui concentrent les indicateurs (éléments de fréquences supérieures) et définissent un ensemble cohérent entre les divers éléments étudiés. Par exemple, les mots des titres des articles parus dans un journal précis (indicateur fixe : titre du périodique, indicateur étudié : mots du titre) tendent à concentrer les objectifs de ce journal. De même, Derek Price estime que le nombre d'auteurs publiant la moitié des

articles sur un sujet donné est égal à la racine carré du nombre total d'auteurs ayant contribué à l'élaboration de ce sujet.

. la "traîne" au contraire représente les éléments de diversité travaillant le domaine traité. La loi de Bradford représente bien ce phénomène. Si *x* articles sont publiés sur un sujet donné, et que l'on considère les titres des périodiques publiant ces articles comme indicateurs variables, on s'aperçoit qu'un tiers des articles est publié par quelques périodiques, dits les périodiques-"cœurs" de la spécialité. Le tiers suivant est publié dans *n* autres journaux *(n* est largement supérieur au nombre de périodiques cœurs) , mais le dernier tiers est publié dans n^2 périodiques. Cette loi confirme l'expérience des bibliothécaires (Bradford était bibliothécaire lui-même), qui savent qu'aucune bibliothèque ne peut concentrer tous les documents sur un sujet donné, mais seulement s'approcher du "cœur", et faire reposer la fourniture des autres documents sur le Prêt Inter Bibliothèques.

Les objectifs de la bibliométrie sont de plusieurs ordres :

- établir une cartographie des littératures scientifiques. En général, les critères de co-occurrence sont utilisés. La co-citation est principalement utilisée dans les écoles anglo-saxonnes, et l'école française s'attache à la co-occurrence des termes d'indexation [MIC88], [COUR89]. Dans cette opération, chaque article est considéré comme un nœud dans un réseau, indexé par le contexte des co-occurrences (articles cités et citants, ou termes associés). Dans ces études, le domaine traité est l'objet de recherche, alors que les publications sont les indicateurs.

- vérifier l'adéquation entre l'image de lui-même que se fait un groupe de chercheurs et sa pratique de publication. Cet axe de travail sert par exemple à définir les achats des bibliothèques en fonction de critères plus objectifs. Dans ces études, les titres de périodiques sont l'objet des recherches alors que les articles sont les indicateurs [DOU88]. On peut aussi étudier les chercheurs et définir la composition d'un réseau social de recherche par l'étude des co-auteurs et des co-citations.

- étudier l'évolution dans le temps des recherches et des périodiques

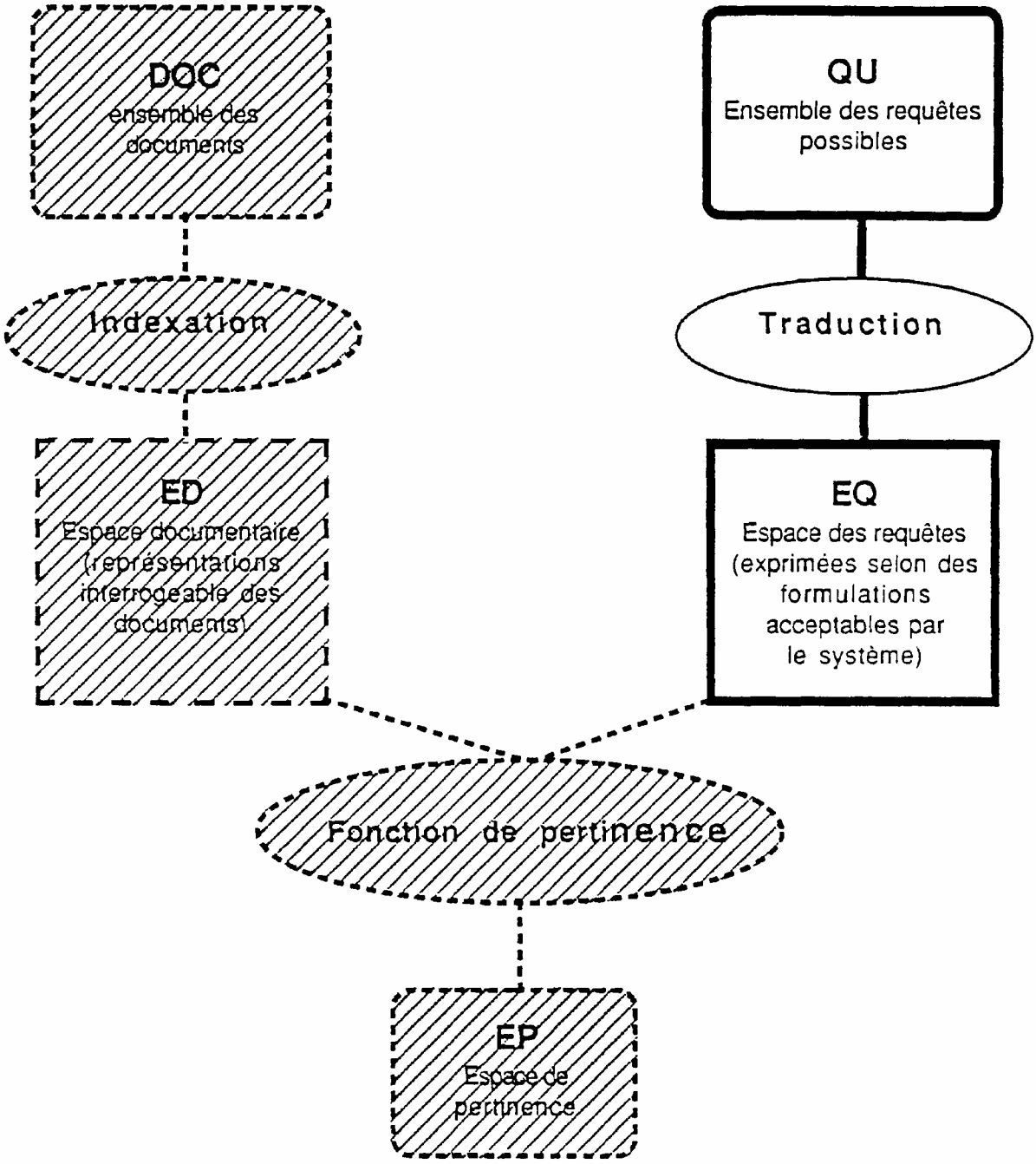
qui leur servent de support. Ces études sont menées selon un axe diachronique (suivi dans le temps d'une production donnée) ou synchronique (évaluation de la production et de son utilisation à un moment donné). On distingue ainsi deux types d'articles : le cas général est représenté par des articles ayant un maximum d'influence (i.e. de citations) dans les deux à trois ans suivant la parution, puis une décroissance rapide. Cependant quelques articles, articles-"cœurs" d'une discipline, rencontrent ce maximum plus tardivement, six ans environ après leur parution, puis connaissent une décroissance lente. Il en est souvent ainsi des articles marquant une nouvelle avancée théorique, ou une nouvelle méthode expérimentale.

- une utilisation particulière des recherches bibliométriques consiste à prendre en compte les découvertes de cette méthode pour indexer les documents et intervenir dans le processus de recherche documentaire. On considère alors qu'un auteur, en citant un article, effectue une forme particulière d'indexation de son texte. Les articles cités représentent alors les concepts agissant dans son propre texte. Nous verrons plus loin les utilisations de cet axe de travail bibliométrique pour l'indexation automatique.

- une nouvelle voie semble aussi se dessiner, qui consiste à utiliser l'indexation et la structuration des banques de données pour proposer des voies de recherche encore inexplorées à partir de recherches documentaires systématiques. Don Swanson a ainsi décrit un processus systématique, par essai et par erreur, permettant de relier logiquement un domaine (représenté par un ensemble de documents) et un autre, sans que des connexions directes aient pu être établies auparavant [SWA89a], [SWA89b]. Dans une expérience de ce type, il a pu mettre en relation un ensemble d'articles traitant d'une maladie du sang (la maladie de Raynaud) et un autre ensemble d'articles traitant des effets thérapeutiques des huiles de poissons. Aucun des articles présents dans ces deux groupes n'étaient reliés par des citations. Ce lien logique, repéré grâce à l'indexation des documents suggère alors une piste de recherche médicale, qui est en voie d'exploration à la suite de la publication de son travail.

II - f_q : QU -> EQ

L'utilisateur face au système documentaire



1 - L'évaluation des systèmes documentaires

Pour pouvoir juger les systèmes documentaires, il convient de définir leur utilité, et par delà, les critères qui permettent d'établir une adéquation entre le fonctionnement du système et les buts qu'il s'est fixé. Or ces éléments ne reçoivent pas de réponse évidente et généralement adoptée par la communauté des chercheurs en sciences de l'information. David Blair [BLA90a] caricature cette situation en la comparant à celle d'un ingénieur de l'automobile qui n'aurait aucun moyen de juger quantitativement des améliorations apportées aux automobiles en termes de consommation, puissance, confort sonore...

1 a - Du besoin documentaire à la formulation de la requête

Avant d'évaluer les systèmes documentaires, il convient de s'interroger sur les besoins documentaires eux-mêmes. Une fois que le besoin documentaire entre dans un cadre de prise de décision tel qu'il a été défini précédemment, il faut alors le formuler devant un système documentaire particulier. Dans un même cadre, la formulation peut varier en fonction de l'objectif immédiat de la recherche. Pour quelles raisons un utilisateur cherche-t-il en ce moment précis de l'information dans un système documentaire ?

On peut par exemple distinguer :

. les recherches "vitales" dans lesquelles l'utilisateur veut obtenir tous les documents relatifs à sa demande. C'est par exemple le cas des recherches d'antériorité de brevets, des recherches juridiques, des recherches militaires...

. en regard, on trouve les recherches d'approche : l'utilisateur veut obtenir quelques éléments d'information pour juger de l'état général d'un secteur et en déduire s'il est valable d'y porter ses efforts. Par exemple le chercheur scientifique qui veut savoir s'il sera efficace de tester une nouvelle hypothèse, ou le responsable de mercatique qui dans un premier temps veut jauger un domaine qu'il ne maîtrise pas encore.

Entre ces deux formes, on trouve toutes les nuances. Il semble donc difficile de définir une règle générale d'utilisation des systèmes documentaires, et par conséquence une règle de fonctionnement obligatoire. Toutefois, l'hypothèse d'une satisfaction de l'utilisateur en regard de sa propre demande à un moment donné reste un instrument d'évaluation précieux.

On peut poser le même problème sous un autre angle en distinguant trois types de besoins [Pedersen, cité par DAC90a] :

. besoin de vérification : *"l'utilisateur veut vérifier ou retrouver de l'information sur des éléments d'information aux caractéristiques connues"*. Par exemple retrouver des données numériques sur des produits chimiques, des valeurs d'actions boursières, des références de documents égarés, des photographies déjà connues... La précision des recherches est alors déterminante, parfois vitale (cas d'un système de documentation technique).

. besoins conscients concernant un sujet : *"l'utilisateur veut clarifier, passer en revue ou approfondir certains aspects d'un sujet bien connu."* C'est le cas du scientifique qui établit une bibliographie, du juriste qui recherche des antécédents pour un litige, de l'ingénieur qui veut exploiter une technique...

. besoins flous concernant un sujet : *"l'utilisateur veut explorer de nouveaux concepts sur des sujets non connus"*. Le système documentaire doit alors l'aider à formaliser sa demande, à la fois pour lui-même et pour obtenir des informations qui vont lui permettre une première prise de décision. A l'heure actuelle, c'est souvent le rôle de "l'interview" avec l'intermédiaire en information qui permet cette première clarification. La capacité à sauter d'une information à une autre dans le cadre du système devient déterminante. La pêche aux idées est une activité essentielle dans la prise de décision. C'est une des missions principales accordées aux hypertextes électroniques. En dehors des systèmes informatisés, cette fonction est remplie par les périodiques professionnels dont la lecture régulière offre en permanence de nouveaux éclairages.

Une fois les besoins documentaires définis par le chercheur, il convient de formuler une requête. Cette requête sera la question telle que le système devra la traiter. On doit cependant se demander si cette requête est l'expression exacte du

besoin documentaire, ou si elle est contrainte par la conception même du système. L'approche du besoin documentaire se fait par étapes, alternant "essais et erreurs".

Pour élaborer sa requête, l'utilisateur doit respecter deux principes [BLA90a] :

. le principe de prédiction *(prédiction criterion)* qui revient à prédire les termes qui sont les plus adéquats pour représenter le besoin documentaire. L'adéquation est représentée par l'existence du terme ou de la combinaison de termes dans la banque de données et la concordance des documents correspondants avec le besoin documentaire.

. le principe de fatigue *(futility point criterion)* qui définit le plus grand nombre de documents électroniques que l'utilisateur peut lire dans le cours de sa recherche documentaire.

Ces deux principes encadrent la recherche documentaire. Si l'utilisateur a correctement prédit les termes représentant son besoin dans le système, il faut de plus que le nombre de réponses soit inférieur à son point de fatigue. Cela même avant d'avoir pu juger si les documents extraits correspondent bien au besoin documentaire. Si le nombre de documents est trop important, il convient de rajouter des termes pour préciser la recherche. On se trouve dans une nouvelle situation où il s'agit de deviner le terme le plus adéquat, alors que celui-ci ne correspond pas forcément au besoin documentaire, et que de nouveau, nous ne savons pas s'il est significatif dans la banque de données, et associé aux documents qui seraient utiles (principe de prédiction). Or ajouter des termes augmente rapidement le nombre de combinaisons entre termes qui sont possibles. Pourtant les systèmes actuels ne permettent pas à l'utilisateur de prévoir correctement ces termes supplémentaires. Les fonctions d'analyse statistique des descripteurs contenus dans un premier lot de documents (ZOOM, GET, MEMTRI suivant les serveurs) sont cependant un pas important dans ce sens : il est plus facile de reconnaître un terme comme adéquat en regard de son besoin documentaire que de deviner un tel terme.

La structure même de la recherche documentaire, telle qu'elle est encadrée par le principe de prédiction et le principe de fatigue, en fait une opération *par*

essais et par erreurs. Dans ce cadre, on appelle reformulation l'opération qui permet de modifier la requête pour s'adapter aux besoins documentaires en fonction du contenu du système considéré.

Il y a deux types de reformulation :

. la reformulation automatique consiste à utiliser les compétences linguistiques du système pour transformer automatiquement la requête de l'utilisateur en une requête plus complète, ajoutant les synonymes ou les formes dérivées des termes de la requête [DEB89], [RAD88], éventuellement corrigeant les fautes orthographiques [PUJ89] ou gérant en mode expert les restrictions ou les élargissements nécessaires pour respecter le principe de fatigue [BAR89].

. la reformulation supervisée correspond à la capacité du système à établir une nouvelle équation de recherche qui soit plus proche des besoins documentaires de l'utilisateur tels qu'ils ont été précisés par une première phase de "jugement de pertinence" sur un lot de documents extraits par une première requête. Le choix de l'utilisateur s'établit au vu des documents. C'est un critère de prédiction plus fiable que de deviner les termes rendant compte du besoin documentaire. Les systèmes à jugement de pertinence *(relevance feedback)* sont ceux qui sont capables d'effectuer automatiquement l'exploitation des choix de l'utilisateur pour opérer une reformulation.

Ces deux types de reformulation peuvent bien entendu coexister dans un même système. Pour autant, la démarche par essais et par erreurs, même associée à des capacités de reformulation automatique, ne peut remplir l'ensemble des besoins documentaires. C'est par exemple le cas si l'utilisateur ne sait pas vraiment comment définir l'objet de sa recherche, ou au contraire si dans le cours même de la recherche il se découvre de nouveaux centres d'intérêt. La méthode correspondant à une démarche de ce type est désignée par les termes de butinage ou de navigation [CANT86]. Dans cette hypothèse, l'utilisateur part d'un document (ou d'un lot de documents) et circule dans le fonds documentaire en fonction des relations d'association qui sont tissées entre les documents. Les relations sont soit des liens hypertextes, soit obtenues par la constitution d'agrégats de documents à l'intérieur du système documentaire. L'agrégation peut s'appuyer sur plusieurs principes (méthode du plus proche voisin, méthode

de Ward...) et sur plusieurs critères (association en fonction des termes d'indexation, en fonction des citations...). Butinage et navigation ne sont pas exempts de problèmes, notamment parce qu'ils induisent souvent une perte de l'objectif principal de la recherche. Cette approche de la recherche documentaire sera traitée dans le cadre du modèle hypertexte, qui est le plus adapté à ce type de démarche.

1.b - La notion de pertinence

Le premier concept élaboré par les chercheurs en sciences de l'information pour évaluer un système documentaire a été celui de la "pertinence". Le terme anglais *relevance* est cependant plus général, représentant *l'à propos* d'un document en regard d'une requête. William Cooper [cité par BLA90, p. 72] dit ainsi : *Relevance...has to do with 'aboutness' (or 'pertinence' or 'topic-relatedness')...*indiquant par cette utilisation de synonymes la complexité et la polysémie de la notion, dont la "pertinence" au sens strict ne représente qu'une partie. Nous conserverons cependant la traduction admise de pertinence, en n'oubliant pas qu'un document peut être *à-propos* sans être indubitablement *pertinent.* Par exemple, un document sur la culture des rhododendrons n'est pas strictement *pertinent* pour une question sur l'entretien des azalées. Il reste cependant *à-propos,* et peut éventuellement satisfaire un utilisateur.

La pertinence représente la qualité d'un document à répondre à la question d'un utilisateur. Dire cela ne nous apprend pas comment juger cette pertinence, Il existe au moins deux éléments d'incertitude qui viennent contrarier cette mesure :

. la question formulée par un utilisateur est-elle la question qui anime sa recherche ? L'expérience des bibliothécaires est claire à ce sujet : il est très rare que la question formelle corresponde à la question réelle. C'est d'ailleurs une donnée générale des interrelations que de ne pas s'afficher à nu. Dès lors, il semble difficile de faire juger la pertinence d'un document par des "experts", disposant à main droite de la question et à main gauche des réponses du système. Seul l'utilisateur peut donner son avis sur la pertinence des documents. Cette remarque implique aussi qu'un système convivial doit, en plus de fournir des

réponses, permettre à l'utilisateur de formuler les bonnes questions. C'est le rôle de la reformulation.

. pourquoi et comment un utilisateur peut-il affirmer qu'un document est "pertinent" pour son besoin documentaire ? Juger de la pertinence est une activité profondément humaine, au sens où elle fait appel à des savoirs et des sentiments non formalisables. Un utilisateur sait dire si tel document l'intéresse, il lui est plus difficile de dire pourquoi. De la même manière que nous savons reconnaître des ressemblances au sein d'une même famille, juger la tonalité d'une voix, goûter un bon vin... sans que ce savoir soit formalisé, ni formalisable [BLA90a, p. 71]. Un utilisateur juge la pertinence non seulement parce qu'il sait reconnaître des éléments d'information dans les résultats proposés par le système, mais aussi parce qu'il pratique en permanence à la fois le domaine de sa recherche, mais aussi le jugement sur les objets de sa recherche. Il sait que telle information est déjà connue, que tel auteur répète toujours la même chose, que tel document est trop ancien... H agit comme expert global.

Offrir à l'utilisateur la responsabilité de juger la pertinence introduit cependant un nouveau biais : la pertinence est alors intimement mêlée à l'*utilité.* Une recherche documentaire n'est jamais une activité isolée. Elle prend place au cœur d'une opération de prise de décision plus large. Le jugement d'utilité est profondément dépendant des connaissances déjà acquises et des objectifs à l'instant de la recherche. L'utilité implique non seulement de savoir si un document est relatif à une question, mais aussi de juger sa qualité, sa crédibilité, sa nouveauté, son importance... L'utilité est comme la pertinence un concept primaire, qui induit des capacités humaines non formalisables et directement dépendantes du contexte.

Ces éléments soulignent une fois de plus la distinction entre les systèmes de gestion de données et les systèmes documentaires. Dans les premiers, les questions d'accès physique aux données sont déterminantes : temps d'accès et répartition des informations dans les mémoires, capacité à poser des questions complexes mais précises (langages de type SQL), capacité à obtenir des présentations de résultats efficaces (tri, calculs sur les données, choix des attributs...). De nombreuses recherches sont tournées vers l'amélioration de cet accès physique aux données. Ces recherches bénéficient aux systèmes

documentaires, mais ne couvrent pas, et de loin, l'ensemble du problème. Dans les systèmes documentaires, c'est l'accès intellectuel aux informations qui est en jeu. De ce point de vue, les opérations physiques de l'ordinateur sont en proportion moins longues que les choix intellectuels. Juger un système documentaire, c'est aussi évalué :

. l'aide du système à la formulation de questions : thésaurus, compétences linguistiques...

. la capacité à présenter les résultats pour une utilisation efficace : passage aisé d'une information courte à l'information générale telle qu'elle est contenue dans le système, capacité à marquer les documents intéressants, lecture facile des informations (en particulier utilisation des signes diacritiques pour les langues européennes)...

. la capacité à reformuler une question pour élargir ou au contraire préciser une première formulation : modèles à jugement de pertinence, compétences linguistiques, proposition de pistes de recherche...

1.c - Les critères d'évaluation

On peut distinguer deux types de critères d'évaluation des systèmes documentaires :

. des critères ergonomiques, qui tiennent à la qualité de l'interface utilisateur. Définir les critères d'une telle évaluation et proposer des moyens d'améliorer l'interface utilisateur sera l'objet de ce deuxième chapitre. Bernard Senach ([SENA90]) propose une revue générale et très complète des méthodes d'évaluation ergonomique des interfaces homme machine.

. des critères documentaires, qui permettent de juger de la capacité du système à fournir les bons documents (tous les bons documents et eux seuls) à l'utilisateur. Il faut s'efforcer, autant que faire se peut, de traduire ces critères documentaires en données chiffrées, afin de comparer les divers systèmes documentaires.

Dans ce cadre, la pertinence, malgré son caractère non déterministe reste un des seuls moyens calculables de juger un système documentaire. Comme souvent dans ce domaine, il faut obtenir des résultats (prendre des décisions) dans un environnement flou et instable (indexation imparfaite, questions insaisissables, jugement imprévisible...). La communauté des chercheurs en sciences de l 'information se base donc sur la pertinence pour établir deux valeurs d'évaluation des systèmes documentaires :

- le taux de couverture, (ou plus simplement la couverture) *(Recall)*
- le taux de précision (ou simplement précision : *Précision)*

Par une recherche documentaire, une banque de données est partagée en quatre ensembles disjoints :

- les documents extraits pertinents
- les documents extraits non pertinents
- les documents non extraits pertinents
- les documents non extraits non pertinents

Soit respectivement E_{pert}, E_{np}, R_{pert}, R_{np} , les cardinaux de ces ensembles. On peut obtenir le tableau suivant représentant la partition de la banque de données.

	Documents	
	Extraits	Non extraits
Pertinents	E pert	R pert
Non Pertinents	E np	Rnp

Le **taux de couverture** est la proportion de documents pertinents extraits par la recherche par rapport à l'ensemble des documents pertinents contenus dans la banque de données :

$$\text{Couverture} = \frac{E_{pert}}{E_{pert} + R_{pert}}$$

Soit en termes probabilistes la condition pour qu'un document pertinent soit extrait par la requête :

$$P\ (D_{ext} \mid D_{pert})$$

En théorie, il est toujours possible d'obtenir un taux de rappel maximum, ne serait-ce qu'en sélectionnant l'ensemble de la banque de données. Dans la pratique, plus on extrait de documents pertinents, plus on est encombré par ceux qui ne le sont pas. Il convient donc de pondérer le taux de couverture par la précision de la recherche.

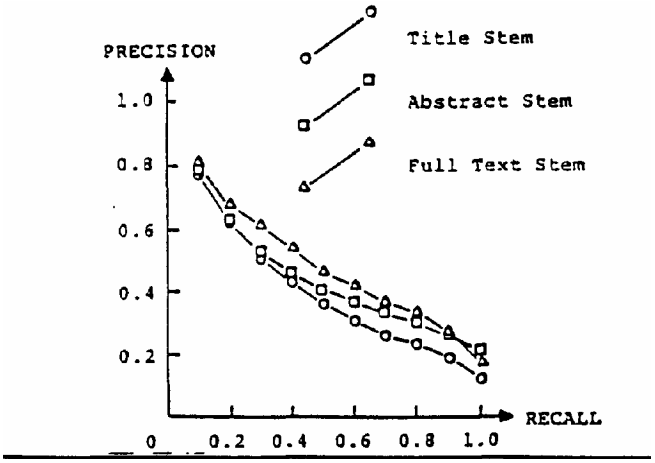
La **précision** est la proportion de documents pertinents parmi l'ensemble des documents extraits :

$$\text{Précision} = \frac{E_{pert}}{E_{pert} + E_{np}}$$

En termes probabilistes, la précision est la condition pour qu'un document pertinent soit présent parmi les documents extraits :

$$P\ (D_{pert} \mid D_{ext})\text{-}$$

Le taux de couverture et la précision varient en général de façon opposée. Le graphique suivant (extrait de [SAL76]) montre le type de liaison qui existe sur une base test entre ces deux valeurs, et les comparaisons qui peuvent être effectuées entre les deux taux suivant les méthodes documentaires employées.



Une autre mesure à été proposée par William Cooper [COOP68] qui se base sur la notion d'utilité des documents pour l'utilisateur : la "Longueur Souhaitée de la Recherche" *(Expected Search Length - ESL*). ESL est le nombre de documents non pertinents que l'utilisateur est susceptible de parcourir avant de trouver le (les) document(s) désiré(s). Cela implique que l'utilisateur n'a besoin que des documents qu'il choisit (des documents importants peuvent lui avoir échappé dans le reste de la banque de données). Cette mesure a l'intérêt de porter l'attention sur la façon dont le système répond aux besoins d'un utilisateur, en laissant à ce dernier le jugement définitif sur le système. Elle évite aussi les difficultés à évaluer le taux de couverture, notamment à connaître le nombre de documents pertinents qui ne sont pas retrouvés. Mais elle a l'inconvénient d'être aussi peu prévisible et fiable que l'ensemble des comportements humains. Un utilisateur exigeant sera moins facilement satisfait qu'un autre. ESL nous donne des moyens de juger globalement un système, mais ne peut convenir comme mesure de comparaison entre méthodes précises (méthodes d'indexation, d'agrégation de documents, de pondération...).

Le taux de couverture est difficile à évaluer en raison de la difficulté à connaître les documents pertinents qui ne sont pas retrouvés (R_{pert}). Cette difficulté peut être levée de deux manières :

. utiliser des banques de données regroupant un faible nombre de documents, ce qui permet de balayer le reste des documents. D'une certaine manière, c'est souvent le choix fait par la plupart des études donnant des valeurs du taux de couverture. Malheureusement, dans le domaine documentaire, le facteur d'échelle est déterminant. Connaître des éléments sur la réaction d'une petite banque de données ne nous apprend pas comment le système réagit devant une banque plus large. C'est un problème qui est souvent rencontré lors de formations d'utilisateurs sur des extraits de banques de données tels qu'ils sont proposés par les principaux serveurs *(Dialog ONTAP, ZPASCAL* sur *Questel...).* Savoir manipuler une petite banque de données (i.e. poser les bonnes questions et connaître le langage de requête) ne nous apprend pas comment agir face à des réponses qui dépassent les capacités de lecture et de jugement d'un utilisateur sur une grande banque de données. Une recherche qui donne une vingtaine de documents dans une petite banque de données (par exemple 10 000 documents) , donnera de l'ordre de 2000 documents sur une banque de données de 1 000 000 de

documents (à rapporter aux 5 à 8 millions de documents des grandes banques de données diffusées sur les serveurs commerciaux). Le changement de taille des banques de données impose des changements de nature linguistique (les termes discriminants et les combinaisons de termes ne sont plus les mêmes) et corrélativement des changements dans la nature même du processus de recherche documentaire.

. évaluer les documents pertinents non extraits à partir de variations autour du vocabulaire choisi, autour des combinaisons booléennes, et éventuellement à partir d'échantillonnage de la banque de données. Ces trois méthodes combinées peuvent permettre une approche de la valeur de R_{pert}, cardinal de l'ensemble des documents pertinents non extraits. Un des problèmes est que nous ne pouvons savoir si cette valeur approchée l'est par défaut (reste-t-il encore des documents pertinents ?) ou par excès (avons-nous utilisé une partie plus significative de la banque de données ?).

Cette difficulté à évaluer le taux de rappel est souvent tournée en utilisant depuis de très nombreuses années (bientôt 20 ans !) les même banques de données tests. Ils s'agit de fichiers connus, fermés, associés à des questions déjà évaluées. Le problème rencontré avec ces fichiers est qu'ils semblent en dis connexion totale avec les systèmes actuels. Prenons les données statistiques sur quelques uns de ces fichiers :

Nom	Nombre de documents	Nombre de termes par document	Nombre de questions	Nombre de termes par question
CRANFIELD	1400	28,7		
WASWANI	11429	20,3	93	6,6
INSPEC-TEST	17070	33,4	65	25,8
UKCIS	27361	6,7	93	7,3

Sur les serveurs professionnels *(Questel, Dialog...)* il y a plusieurs centaines de banques de données ayant plus de 1 million de documents. Par ailleurs, le nombre moyen de termes d'une question est largement inférieur à celui des banques-tests, et atteint rarement les 25,8 termes de moyenne de la base

test-INSPEC. Cette remarque ne condamne pas pour autant l'ensemble des recherches pour évaluer le taux de couverture et la précision des systèmes documentaires, elle tend à les remettre à une place plus limitée. Ce point de vue était déjà exprimé il y a 10 ans par Karen Sparck-Jones dans une revue des diverses études couverture/précision menées jusqu'alors : *"nous n'avons aucune information sur le taux de couverture réel des systèmes en ligne commerciaux, ni des données réelles sur la majeure partie des schémas étudiés par les chercheurs",* [cité par BLA90, p. 84].

Une étude menée par Maron et Blair sur un système en texte intégral de plus grande envergure (40 000 documents utilisant le logiciel STAIRS) dans des conditions draconiennes d'évaluation (l'étude a duré plus de 6 mois, a utilisé deux chercheurs, six personnes, et a coûté environ un demi million de dollars) montre que dès que l'on aborde des systèmes d'une taille correspondant à la réalité commerciale, les résultats sont largement insatisfaisants. Les conclusions principales de cette étude ([BLA85], [BLA90], [BLA90b], et [SAL86] pour des remarques critiques) sont les suivantes :

. en dépit de la croyance des utilisateurs sur l'exhaustivité de leur recherche (évaluée subjectivement à 75 %), le taux de couverture réel est beaucoup plus faible (en réalité 20 %). Il n'y avait pas de limite au nombre de questions posées par les utilisateurs, les documents étaient retrouvés, après autant d'itérations et de reformulations de la recherche que souhaitées.

. il n'y a pas de différence significative entre les taux obtenus par différents chercheurs

. il n'y a pas de différence significative entre la recherche menée par l'utilisateur ou par un intermédiaire compétent (ici entre les hommes de loi concernés ou leurs assistants). Les problèmes sont avant tout inhérents au système utilisé (ici un système basé sur l'indexation en texte intégral).

. les utilisateurs ne pouvaient prévoir qu'une faible partie des mots et des expressions qui leur permettraient de retrouver les documents pertinents, mais qui seraient absents des documents non pertinents.

. la variabilité des mots et des expressions que les auteurs d'un document (traité par le système en texte intégral) emploient pour désigner le même objet est à la fois extraordinaire et imprévisible.

2 - La formulation de la question par l'utilisateur

L'analyse des systèmes d'information requiert de mettre en parallèle deux éléments contradictoires :

- la simplicité pour l'utilisateur
- la couverture et la précision des réponses.

La construction des systèmes d'information implique de prendre en charge toute une série d'opérations que les anciennes méthodes de l'informatique faisaient reposer sur l'utilisateur, et plus généralement, dans le cas des systèmes documentaires, sur l'intermédiaire en information. Concevoir un système d'information efficace implique la conception d'une interface utilisateur convivial, qui puisse remplir trois fonctions :

- permettre à l'utilisateur de formuler sa question le plus librement possible.
- utiliser le jugement de l'utilisateur comme moteur de l'interaction avec le système
- construire une interface de présentation des actions possibles et des documents extraits qui soit facilement acceptable et compréhensible par l'utilisateur.

2.a • Convivialité

Même s'il a été souvent souligné que l'on ne s'assied pas devant un terminal pour raconter sa vie ([PUJ89]), il importe de permettre à l'utilisateur de formuler sa requête dans les termes qui lui viennent à l'esprit spontanément.

Watt [WATT68] défini la convivialité *(habitability)* d'un système d'information comme la facilité offerte à l'utilisateur de rester à l'intérieur du

langage d'un système en exprimant ses besoins sous les formes qu'il désire. Dans ce cadre il distingue la convivialité conceptuelle et la convivialité fonctionnelle :

- la convivialité conceptuelle renvoie au domaine d'application du système. Elle définit pour l'utilisateur les objets et les actions que le système peut traiter, ce qui conduit l'utilisateur à ne poser des questions que sur les domaines pour lesquels le système possède des informations. On conçoit qu'il y a là une difficulté fondamentale, antérieure même à l'interaction homme système, et qui concerne le choix des outils adéquats pour des tâches données. Dans l'idée générale de l'utilisateur face à la documentation, il s'agit toujours de la même tâche et toujours des mêmes méthodes : espérer des réponses, qui doivent bien avoir été écrites quelque part. Nous avons appris à distinguer l'utilisation d'une pince et d'un tournevis, même si les deux instruments servent à bricoler. Il faut que le système documentaire permette à l'utilisateur d'apprendre son fonctionnement pour arriver au même résultat avec les outils documentaires. Les "interfaces coopératives" doivent permettre d'appliquer la convivialité conceptuelle en indiquant à l'utilisateur les raisons qui empêchent le système de répondre à sa question, et éventuellement de l'orienter vers des systèmes plus adéquats. Peu de systèmes, même expérimentaux, ne possèdent des qualités de convivialité conceptuelle suffisantes.

- la convivialité fonctionnelle définit les contraintes sur ce qui peut être accepté par le système pour comprendre la question de l'utilisateur. Alors que la convivialité conceptuelle détermine ce qui peut être demandé au système, la convivialité fonctionnelle s'intéresse à comment cela peut être demandé au système. L'objectif de créer des interfaces en "formulation libre" vise à développer la convivialité fonctionnelle des systèmes documentaires.

Pour préciser les opérations qui permettent de définir la convivialité fonctionnelle, Ogden [OGD87] poursuit le travail de Watt en définissant :

- la convivialité syntaxique qui correspond aux capacités du système à admettre des paraphrases de chaque requête. La formulation libre des questions peut engendrer un grand nombre de phrases valides pour une même question. Les capacités d'analyse syntaxique du système sont importantes pour faciliter

l'interrogation. Développer la convivialité syntaxique vise à obtenir les mêmes réponses pour diverses formulations de la même question.

- la convivialité lexicale traduit la capacité du système à connaître les mots utilisés dans les requêtes. Un bon système devrait pouvoir apprendre de nouveaux mots en les insérant dans un réseau de termes (synonymes, termes associés...), éventuellement en demandant à l'utilisateur des précisions sur les mots qu'il emploie. Les outils documentaires (thésaurus, réseaux sémantiques...) s'il sont utilisés en ligne au cours de la recherche, sont les outils principaux de la convivialité lexicale.

2.b - Les limites des systèmes commerciaux

Les systèmes documentaires actuels ne correspondent guère à ces critères de convivialité adaptés aux besoins d'information. Même si on les pare des attributs modernes de la convivialité (souris, menus déroulants, couleurs...comme dans le cas des D.O.C.), les systèmes documentaires commerciaux restent profondément marqués par deux limites déterminantes :

. la formulation des questions s'appuie sur les seuls termes présents dans la question. Même avec une syntaxe plus élégante que les langages de commande, comme dans les systèmes à base de menus hiérarchisé (accès vidéotex ou versions grand public des banques de données :*BRS After Dark, Dialog Bussiness connection, Questel entreprises...*), la formulation des questions reste soumise à la compréhension des méthodes de la recherche par unitermes. La convivialité lexicale n'est jamais au rendez-vous dans les systèmes commerciaux. La convivialité syntaxique renvoie au problème encore plus aigu du modèle booléen.

. l'articulation logique entre les termes de la question doit être posée dans les limites de la formulation booléenne, éventuellement élargie aux opérateurs de proximité.

La logique booléenne est fort éloignée des modes de pensée de l'utilisateur. Pourtant, elle a su prouver sa capacité à résoudre des problèmes quand ils sont

bien formulés, par exemple, dans les langages de programmation de machines-outils, ou dans la définition de processus industriels. Le problème de l'utilisation de la logique booléenne et des trois connecteurs principaux (ET, OU, SAUF) dans le cadre des systèmes documentaires renvoie au fait que l'informatique documentaire traite du langage. Or les trois connecteurs booléens ont un sens différent en logique et dans la langue courante. Ceci est souvent la source de nombreux quiproquos dans la formulation des questions.

Ainsi, le connecteur ET est en général associé dans la langue quotidienne à une addition, une augmentation des arguments de recherche, alors qu'au contraire dans la logique booléenne, augmentant le nombre de contraintes, il tend à réduire le nombre de documents extraits.

En français, la conjonction OU exprime très souvent une alternative, alors que l'opérateur booléen fait l'inverse en mêlant plusieurs termes, qui ne sont souvent que des synonymes partiels.

Cette opposition relative est souvent source de confusion. On voit fréquemment des utilisateurs étonnés de ne pas avoir assez de réponses proposer d'ajouter un terme (opérateur ET) à leur requête, ne comprenant pas que cela ne ferait qu'empirer la situation. Par exemple, l'expression *"fruits et légumes"* du langage quotidien devra se traduire par *"fruits* OU *légumes"* en formulation booléenne.

Ces problèmes naissent du fait que la question d'un utilisateur est formulée comme une proposition langagière, alors que son traitement utilise l'expression constituée au seul titre de chaîne de caractères. Dans une formulation linguistique les termes utilisés sont considérés comme des représentants d'un signifié. Les synonymes sont de ce fait intégrés par l'utilisateur comme étant déjà inclus dans sa formulation. Les termes associés, ou les généralisations, ou encore les termes spécialisés représentant des formes particulières d'un concept sont déjà présents dans l'expression de ce concept par un mot du langage. Mais la recherche documentaire n'utilise le langage qu'à titre accessoire. La comparaison entre les documents et la question se situe au niveau des chaînes de caractères. La question *"informatique* ET *documentation"* ne correspond pas du tout à la recherche de documents traitant de l'informatique et de la

documentation, comme une approche linguistique pourrait nous le faire croire, mais simplement à la recherche de documents qui contiennent conjointement les deux chaînes de caractères *"informatique"* et *"documentation"*. La formulation *"informatique documentaire"* sera ignorée, de même que *"recherche documentaire informatisée", "documentation informatisée", "systèmes documentaires informatisés»..... Déjà dangereux* avec les connecteurs booléens ET et OU, cette confusion entre signifiant et chaîne de caractères devient caricaturale avec l'opérateur booléen SAUF. On cherche à éliminer d'un ensemble de documents non pas les documents qui traitent d'un sujet donné, mais ceux qui contiennent la chaîne de caractères correspondante, même si justement cette chaîne est là pour dire que le document ne traite pas de ce sujet.

Cette distinction entre la formulation linguistique et la formulation booléenne d'équation entre chaînes de caractères n'est pas souvent comprise par les utilisateurs car les signes qui portent la requête sont les mêmes : les mots du langage. La recherche documentaire ne donne jamais de meilleurs résultats qu'en employant des formulations codées (numéros de registres, codes de classification...) qui sont justement inventés pour que le signifié n'existe qu'en tant que chaîne de symboles, et que le signifiant ne soit en aucune mesure rapporté à des éléments linguistiques habituels. La définition de langages documentaires, utilisant les termes du langage dans un sens unique, articulant selon une syntaxe propre les diverses facettes donnant le sens d'un document, et limitant le vocabulaire à certains termes, correspond à une volonté de réduire le langage à des chaînes de caractères. Mais, à l'instar des connecteurs booléens eux-mêmes, les termes du langage documentaire subissent les effets en retour de l'utilisation de termes du langage : l'utilisateur les considère comme des signifiants linguistiques et s'attend donc à ce qu'ils expriment tous les sens qu'il donne à ce signifiant linguistique, alors qu'ils ne sont utilisés que dans le cadre d'une formulation précise, en général hiérarchisée, mais sans que les relations de hiérarchies ne soient directement prises en compte dans le système.

Pour réduire ces difficultés, on a souvent proposé de *"former les utilisateurs"*. L'expérience de la formation des utilisateurs montre que ces derniers n'utilisent réellement cet apprentissage qu'afin de se familiariser avec les systèmes, et mieux savoir ce qu'ils peuvent demander à un intermédiaire en information [REI85].

Il y a dans l'interrogation d'un système documentaire complémentarité entre trois connaissances ([LEC87a], [DES85]):

- la connaissance du vocabulaire de la banque de données : méthodes d'indexation, thésaurus, vocabulaire contrôlé ou libre...
- la connaissance de la structure de la banque de données, et la place de chaque banque de données dans l'ensemble des services documentaires (couverture, spécialité, qualité...)
- la connaissance du langage de requête des serveurs.

On conçoit aisément que jamais un utilisateur ne voudra maîtriser ces trois techniques pour satisfaire des besoins ponctuels. On conçoit aussi que la maîtrise du problème posé par l'utilisation linguistique des termes des requêtes pour désigner des concepts et l'utilisation de chaînes de caractères pour effectuer la recherche documentaire soit le produit d'un apprentissage long et difficile.

Seule une approche qui tende à améliorer la convivialité fonctionnelle des systèmes documentaires reste crédible. La métonymie traditionnelle, qui est d'autant plus forte qu'elle s'applique à des termes linguistiques pour lesquelles elle est habituelle, en assimilant chaîne et sens doit trouver un écho dans le système. Celui-ci doit accepter les formulations des questions comme étant des expressions linguistiques et s'efforcer de les traiter comme telles, c'est à dire utiliser toutes les ressources sémantiques contenues dans les termes de la question (utilisation de thésaurus, d'analyseurs linguistiques...). C'est un des enjeux de la recherche en sciences de l'information.

2.c - Utiliser le jugement de l'utilisateur

Une autre méthode, qui peut être appliquée en complément, consiste à considérer la question de l'utilisateur comme un simple symptôme de ses besoins documentaires. Les résultats de la première recherche que la requête de l'utilisateur permet d'effectuer sont alors soumis à son jugement. Les documents retenus deviennent, au travers de leur représentation dans le système, des questions plus précises, et qui correspondent mieux aux besoins documentaires de l'utilisateur que la première formulation.

Dans une première interaction avec un système documentaire, il faut permettre à l'utilisateur de formuler sa question de la manière la plus libre possible. Mais il importe dans la conception du système de ne pas considérer que cette formulation représente à elle seule le besoin documentaire de l'utilisateur. Il ne s'agit que d'un moyen pour lui d'exprimer sa demande face à un système qu'il ne connaît pas. L'interprétation de la question par le système est un pas très important, parce qu'il conditionne la suite de l'interaction. Mais on ne peut estimer que l'on a compris le besoin de l'utilisateur parce que l'on a interprété sa question.

Cette remarque renvoie à la distinction entre les systèmes documentaires et les systèmes de gestion de données. La gestion de données a ce caractère particulier que la réponse à une question est univoque. Comprendre la question *"Je veux la liste de tous les employés du Sud-Ouest qui gagnent moins de 150 000 F."* est un travail de linguistique informatique difficile, mais la réponse est claire et ne supporte pas d'ambiguïté. En revanche, comprendre la requête *"Je veux des documents sur la salmonelle dans l'industrie des œufs"* n'est qu'un premier pas. Il faudrait ensuite savoir si l'utilisateur cherche des bilans épidémiologiques, des méthodes de traitement, des normes sanitaires, les conséquences de tel ou tel traitement sur les qualités gustatives ou la conservation....Lui-même ne sait peut-être pas, avant d'avoir vu ce que le système est capable de lui proposer, dans quelle voie il va préciser sa demande.

Le meilleur moyen pour que le système obtienne des informations sur le besoin de l'utilisateur, tel qu'il va évoluer après sa première question est d'utiliser le jugement qu'il porte sur les premiers documents fournis. Un utilisateur ne reformule que très rarement de lui-même une question. Quand il s'y résoud, la méthode la plus courante consiste à ajouter des termes, à préciser à partir de la première formulation. Or, si cela apparaît efficace dans certains cas, c'est souvent très restrictif, soit parce qu'un premier lot de réponses était déjà trop petit, soit parce qu'on passe alors à côté de tous les documents qui seraient pertinents avec une autre formulation. Cette incapacité à modifier, à "paraphraser", la première question est une caractéristique psychologique qui dépasse de loin le seul cadre de la recherche documentaire. Les psychologues Tversky et Kahneman (cités par [BLA90a] p. 15) considèrent que lorsqu'une personne doit évaluer une valeur inconnue, elle commence par estimer une

valeur initiale, et procède ensuite par corrections autour de cette valeur initiale, corrections qui s'avèrent en général insuffisantes. Ils appellent ce phénomène 'l'ancrage"*(anchorage)*. En recherche documentaire, l'utilisateur, dans sa première formulation, établit lui aussi une estimation, celle concernant la qualité des termes qu'il utilise à décrire effectivement son besoin documentaire dans le système. Il construit ainsi un "ensemble d'ancrage" *(anchor set),* dont il aura le plus grand mal à s'éloigner.

Si l'on admet, comme la pratique des intermédiaires en information le confirme chaque jour, l'existence de ce phénomène d'ancrage, il convient de travailler à ce que le système se charge lui-même de la reformulation des questions et propose de nouveaux résultats à l'utilisateur qui vont dans le sens des premiers jugements établis. Le jugement de pertinence est l'instrument de cette reformulation. Il revient à considérer les documents eux-mêmes comme une question, et permettre la recherche des documents les plus proches de ceux qui sont valorisés.

Cette notion d'un jugement de pertinence de l'utilisateur sur les résultats qui lui sont proposés est une pierre angulaire du développement futur des systèmes d'information. Elle prend sa source d'une analyse des documents eux-mêmes, et du réseau d'intertextualité qui s'établit entre eux.

En ce sens, le jugement de pertinence dans une recherche documentaire est différent de la méthode "Q.B.E." *(Query By Example* - Recherche par l'exemple) utilisée par les Systèmes de Gestion de Données. Dans le modèle QBE [ZL077], c'est la structure de la base de données qui est présentée à l'utilisateur, afin que celui-ci intègre dans le cadre prédéfini un exemple des résultats qu'il souhaite obtenir. Il y a derrière l'acceptabilité de la métaphore du QBE (au sens où toute interface utilisateur est une métaphore d'une application destinée à rendre sensible à l'utilisateur les fonctions du système) l'idée que le cadre de définition des données est suffisant pour indiquer les objectifs d'une recherche. Comme précédemment, on peut constater que ceci est vrai en grande partie dans les systèmes gérant des données, alors qu'on ne peut guère en tenir compte dans les systèmes documentaires. Par exemple, proposer une grille à remplir pour définir des critères de recherche dans un SGBD est cohérent avec la structure des informations (à chaque champ une information spécifique : Nom, Fonction, Lieu

de travail, Salaire... dans un système d'entreprise) et avec les objectifs d'une recherche (e.g. *"constituer une liste classée alphabétiquement des employés du centre de Lyon ayant un salaire brut mensuel inférieur à 10 000 F.").* En revanche, chaque champ d'un document informatique (par exemple le titre, les descripteurs contrôlés, le texte du résumé ou du document lui-même, et dans une moindre mesure auteurs, affiliation, source documentaire) n'est jamais qu'une piste d'accès permettant d'extraire de la banque de données les documents qui, d'une manière souvent difficile à définir, servent à l'utilisateur pour circonscrire une question. Les attributs sont des représentants des documents, avec toutes les incertitudes que cela représente, et non les définitions de l'objet, comme dans les systèmes de gestion de données.

Le concept de reformulation s'appuie sur des relations de connexité entre les documents. Ceux-ci sont reliés entre eux par divers chemins, qui assurent des liaisons par des similitudes d'ordre documentaires, lexicales et sémantiques.

- aspect documentaire : Les publications d'un même auteur ont des chances d'intéresser (ou de ne pas intéresser) celui qui favorise (ou rejette) un document extrait. De même, les publications d'un même laboratoire de recherche sont souvent centrées sur des sujets connexes. Les périodiques, notamment dans le domaine de la recherche scientifique sont centrés sur des sujets particuliers, qui font cohabiter des éléments d'information ayant des éléments de proximité plus ou moins forts selon le type de périodique. Le réseau des citations croisées entre documents permet lui aussi de tisser des liens. Les documents occupant une même place dans un plan de classement sont aussi proches du point de vue de leur contenu. Les méthodes bibliométriques permettent de mieux connaître ces liens entre documents.

- aspect lexical : Les termes utilisés, soit directement dans le texte (méthodes en "texte intégral"), soit dans ses parties significatives (titre, résumé), soit dans les termes d'indexation choisis par des spécialistes, peuvent être considérés comme autant d'éléments d'une nouvelle question permettant de retrouver les documents les plus proches des documents jugés pertinents (voir [TRI89] pour l'utilisation en grandeur réelle sur le serveur *DowJones* de ce procédé, et [SAL88b] pour des critiques).

- aspect sémantique : l'analyse des termes d'un document, en permettant d'en extraire des éléments de sens (indexation automatique) permet aussi de regrouper les documents en fonction de leur contenu, même si cela n'est pas directement explicité. L'utilisation de thésaurus (relations générique/spécifique, et relations d'association), de classifications documentaires et de dictionnaires intégrant des aspects sémantiques (réseaux sémantiques, analyseurs linguistiques, langages de représentation [ZAR90], [RAU89]) sont des instruments permettant de définir des liens tissés directement au niveau sémantique entre les documents.

Admettre la possibilité de sélectionner dans un premier lot de documents ceux qui correspondent pleinement aux souhaits de l'utilisateur, c'est ouvrir une série de pistes pour retrouver des documents liés. Toutefois, on ne peut prendre ces pistes pour certitudes. Il est nécessaire de pondérer les éléments d'information fournis par l'analyse du document, et il importe de faire intervenir des éléments extérieurs, comme par exemple la fatigue de l'utilisateur (après combien de documents pertinents souhaite-t-il continuer) ou le taux de rejet (quel pourcentage de documents non pertinents est acceptable par l'utilisateur).

La reformulation des questions peut aussi passer par la possibilité offerte à l'utilisateur de naviguer dans le système documentaire à partir des documents extraits par sa première question. On trouve plusieurs méthodes utilisant les informations contenues dans les documents pour organiser en réseau la banque de données :

- la réalisation d'agrégats *(clusters)* de documents ayant de nombreux points en commun. On définit alors une distance entre documents, par exemple en fonction du nombre de termes communs entre deux documents (coefficient de Dice, ou de Jaccard), éventuellement corrigée en utilisant des pondérations entre les termes, soit pondérations d'origine statistique (un terme fréquent est considéré comme moins important qu'un terme rare ; le nombre occurrences d'un terme dans un document doit tenir compte de la longueur de ce document,...), soit pondérations d'origine combinatoire (les citations croisées), soit pondérations déterminées pragmatiquement (on affecte une importance différente aux noms d'auteurs, au titre du périodique, aux descripteurs, aux mots du titre, ou du résumé...).

- la cartographie des interactions entre chercheurs obtenue notamment par l'analyse des citations. Deux documents seront alors d'autant plus proches qu'ils seront cités en commun. Ce procédé permet de repérer des connexions entre les secteurs scientifiques, même s'il faut garder présent à l'esprit les limites de la méthode repérées dans le premier chapitre. Utilisée comme instrument d'indexation, cette méthode reste cependant plus fiable, ou du moins pas plus aléatoire qu'une autre, comme nous le montrerons dans le chapitre suivant.

- la constitution de réseaux de mots ou d'expressions à partir de l'analyse statistique des documents. Des relations sont tissées entre les termes eux-mêmes du fait de leur présence conjointe dans plusieurs documents apparentés. L'intérêt de cette démarche est d'utiliser ces relations entre les mots pour offrir des pistes à la reformulation par l'utilisateur, à partir des mots qu'il a utilisé dans sa question. Le jugement de pertinence ne s'applique alors plus seulement aux documents, mais aux termes de la recherche, ce qui peut permettre de résoudre des cas où le nombre de documents extraits en réponse à une première question reste trop élevé pour permettre à l'utilisateur de les balayer sans ressentir rapidement les phénomènes de dégoût et de fatigue repérés ci-dessus. Cette méthode permet aussi de réaliser automatiquement les prémisses d'un thésaurus par analyse du fonds documentaire.

3 - Interface de manipulation des informations

L'interface utilisateur permet de présenter les capacités du système de façon à le rendre

- acceptable : compréhension par l'utilisateur des tâches à accomplir pour réaliser l'objectif qu'il s'est fixé
- efficace : permettre de formuler au mieux les besoins dans le cadre des requêtes acceptées par le système.

Ces deux notions sont couplées, mais cependant distinctes. Même dans le cadre d'un système documentaire spécifique, limité (par exemple un modèle de

recherche booléenne basé sur des unitermes et n'acceptant pas les opérateurs de proximité), on peut définir plusieurs interfaces utilisateur qui auront des incidences sur la qualité des recherches documentaires, du point de vue de la réalisation des objectifs de l'utilisateur. Si les écrans sont lisibles, ne présentant que les informations directement utiles, l'impact sur les tâches cognitives mises en œuvre sera déterminant dans la compréhension par l'utilisateur des informations. De même, la maîtrise des possibilités offerte à l'utilisateur à chaque moment de sa recherche dépend de la clarté avec laquelle les fonctionnalités du système sont présentées.

La réalisation d'une interface utilisateur est un problème fondamental de la recherche en informatique et en ergonomie cognitive. Pour l'utilisateur, une application informatique se limite souvent à son interface.

Une bonne interface doit :

- minimiser la mémorisation technique
- améliorer la cohérence de l'application
- donner des retours d'information (au système et à l'utilisateur)
- assurer la tolérance aux erreurs.

3.a - Les styles de dialogues homme - système informatique

l'objectif du dialogue homme système est de présenter de manière souple les diverses actions possibles à un moment donné du cheminement dans une application. Cet aspect est pris en charge par le module de dialogue. En suivant Pierre Lévine et Jean-Charles Pomerol ([LEV89], p. 84 et suivantes), on peut repérer plusieurs styles de dialogues :

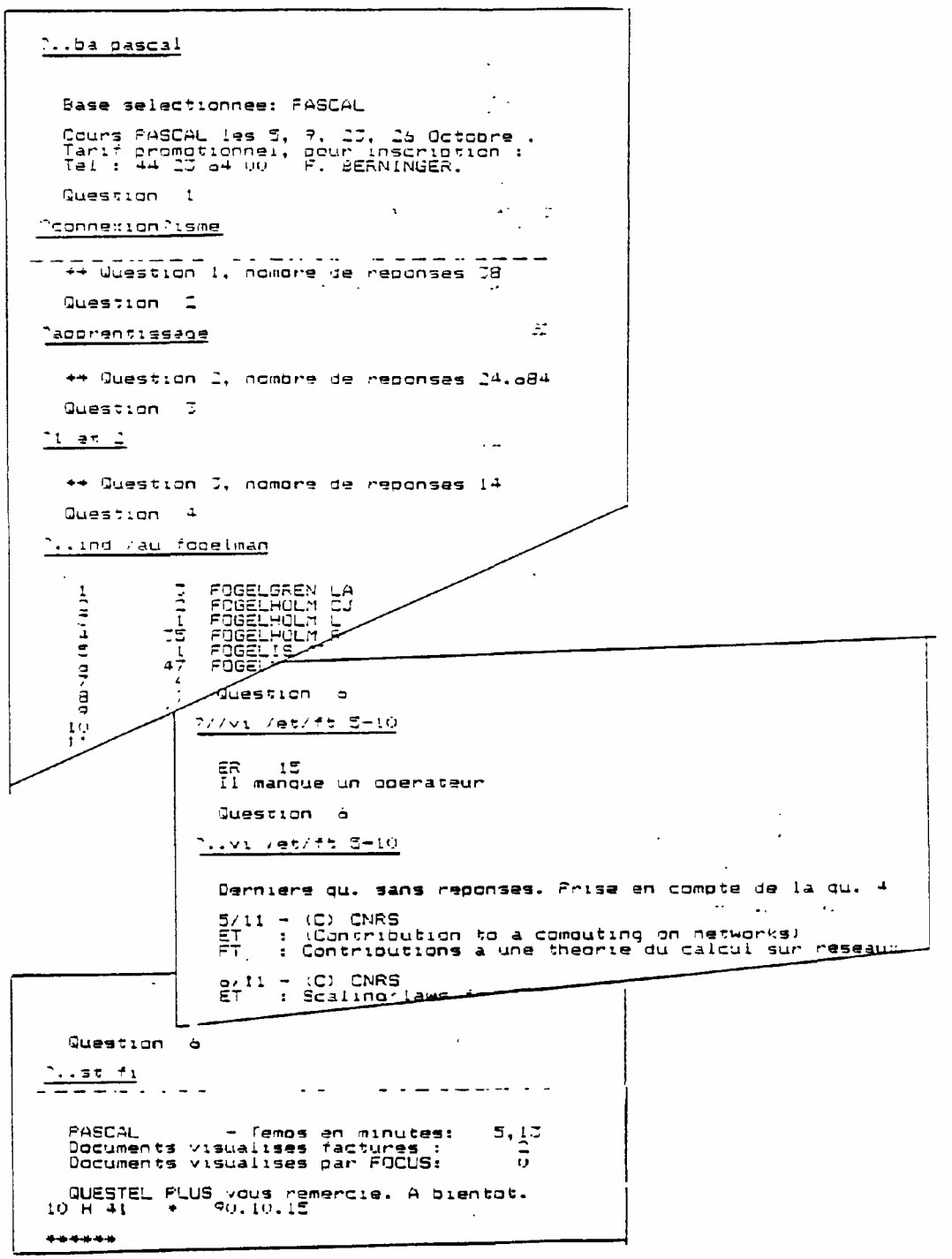
- le style "commandes"
- le style "question-réponse"
- le style "masque"
- le style "à la carte"
- le style "environnement".

- le style "commandes".

Il s'agit du style de dialogue le plus élémentaire, du point de vue des capacités de l'interface. L'interface par commande requiert un apprentissage,

souvent long et complexe. Cet apprentissage peut toutefois être réduit en nommant correctement les commandes, c'est-à-dire en autorisant des équivalents pour une même commande, et en tenant compte des habitudes des utilisateurs ([FUR83], [FUR87]).

Les serveurs professionnels de banques de données utilisent ce type de dialogue par "commandes". Voici un exemple d'une recherche documentaire sur *Questel.*



Même dans ce cadre limité du langage par commandes, on peut améliorer l'interface utilisateur :

. en présentant les commandes disponibles (exemple le système DARC),

. en autorisant des synonymes aux commandes (le serveur *Dialog* permet depuis peu de quitter le système par une liste d'expressions : logoff, logout, bye, off, end, quit... Il est le seul !)

. ou en assurant la traduction d'un jeu de commandes dans un autre.

Le pivot de transformation des langages de commandes dans le domaine documentaire est constitué par le langage C.C.L. *(Common Command Language),* projet de norme ISO N° DP 8777. Il est toutefois rarement implémenté par les serveurs, qui développent leur langage propre. Cependant, les divers langages de commandes des serveurs respectent des structures et des fonctions comparables entre elles, et comparables avec le C.C.L. Pour une analyse et une comparaison de l'ensemble des commandes des serveurs professionnels, on peut se reporter à [LEC89]. Certains anté-serveurs, comme *EasyNet,* implantent une version du langage normalisé CCL et s'en servent pour traduire les commandes d'un jeu à l'autre, ce qui permet de formuler les questions à partir d'un langage de commandes connu, quel que soit le serveur concerné [WILC88].

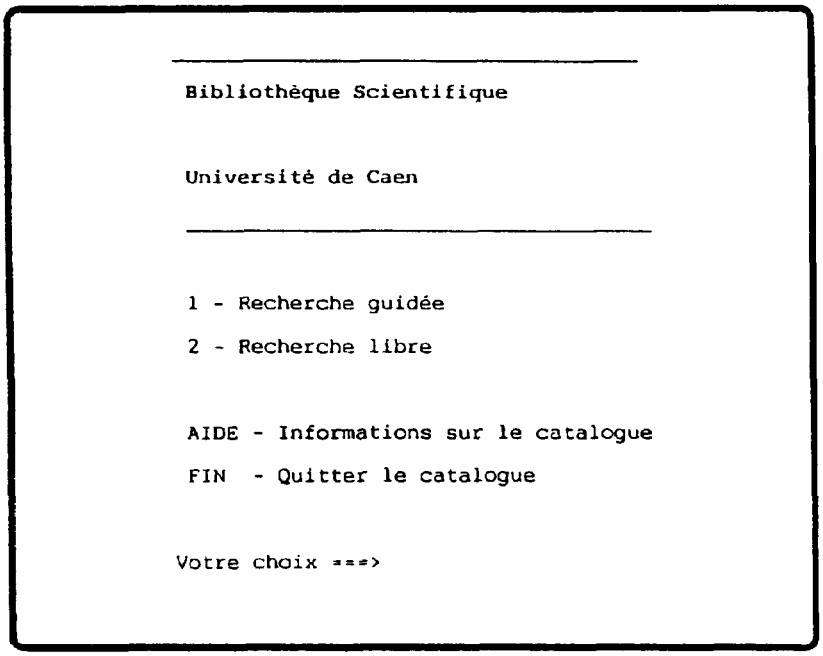
Le langage de commandes, parce qu'il constitue une propédeutique à toute interaction avec le système documentaire limite radicalement l'accès direct par les utilisateurs. L'apprentissage du langage de commandes rend encore plus difficile d'aborder ce qui constitue la véritable difficulté de la recherche documentaire : la formulation de la question.

. le style "question-réponse" :

L'utilisateur répond à des questions formulées par le système. Il est en général placé devant un menu présentant un choix limité. Les réponses acceptables sont précisées dans le menu. A la fin de ce cheminement de menu en menu, l'utilisateur peut poser la question qui motive sa recherche. Le style question-réponse est relativement facile à développer, mais ôte en général toute

marge de manœuvre à l'utilisateur. L'écriture même des questions est déterminante sur la compréhension des fonctionnalités du système.

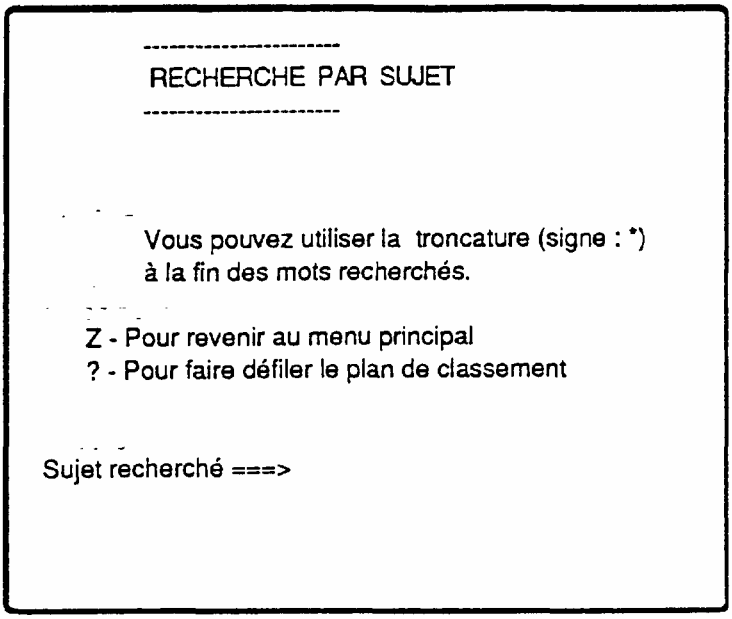
Ce type de dialogue est utilisé par tous les systèmes d'accès "grand public" aux banques de données professionnelles *(Questel-Entréprise, BRS AfterDark, EasyNet).* La plupart des logiciels documentaires disposent d'un langage associé permettant de construire des interfaces de ce type (par exemple *Logotel,* associé à *Texto,* ou *MNS* associé à *BRS-Micro).* Voici un exemple du style question-réponse rédigé en *Logotel,* pour le catalogue de la Bibliothèque Scientifique de l'Université de Caen.



Une des limites de ce type de dialogue est liée à la difficulté à formuler des questions qui soient toujours acceptables et compréhensibles par l'utilisateur, et dont les réponses soient facilement interprétées par le système. Prenons trois exemples :

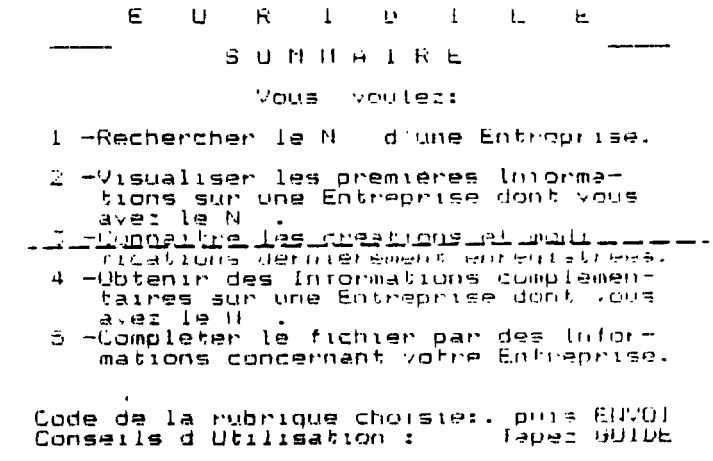
. Une société de conseil en vidéotex est chargée de mettre en place un système de réservation d'hôtels pour une grande chaîne. Une première interface est testée auprès d'un panel d'usagers. Une fois que l'utilisateur a choisi la ville, le lieu (près de la gare...), l'hôtel, l'interface propose : *"Pour réserver : "*. Et on retrouve alors une situation d'échec, l'utilisateur n'osant pas franchir le pas sans avoir la certitude de pouvoir se rétracter. Le même logiciel (qui permettait bien sûr de revenir sur sa décision), fonctionnant en remplaçant cette phrase par *"Choisir sa chambre : "* connaît alors le succès.

. Dans une ancienne version de l'interface utilisateur du catalogue de la Bibliothèque Scientifique de l'Université de Caen qui sera décrit plus loin, on retrouvait l'écran :



La majeure partie des utilisateurs choisissaient alors le module plan de classement (par ?), ou revenaient au départ, ne comprenant pas qu'ils pouvaient à cet endroit formuler leur question. Quand l'écran a été remplacé par un texte plus explicite (on a simplement ajouté la phrase : *"Ou posez directement votre question"),* le taux de succès a été largement amélioré... sans qu'une seule ligne du programme d'interaction vers le système d'information ait été modifiée.

. L'écran suivant est l'écran d'entrée du service 36-17 EURIDILE, produit par l'INPI, destiné à connaître les informations légales sur les entreprises françaises. Il faut souvent s'y reprendre à deux fois pour savoir quel choix correspond à la question évidente de la majeure partie des utilisateurs : écrire le nom d'une entreprise pour obtenir des informations.



Ces exemples montrent qu'un audit linguistique et ergonomique des systèmes d'information serait souvent nécessaire. On n'insistera jamais assez sur l'importance des formulations dans les interfaces pour l'efficacité des systèmes.

Les dialogues "question-réponse" conduisent à de fréquents changements d'écrans. Dans la conception de ce type de systèmes, il faut tenir compte du fait que chaque changement d'écran est comme un changement d'univers de référence, et implique une phase de prise de repères dans le nouvel écran, ce qui limite l'efficacité globale du système. Les écrans intermédiaires placés entre l'entrée dans le système et la visualisation des informations recherchées sont vécus comme des obstacles dont le contenu n'est pas pris en compte dans le cheminement intellectuel de la recherche [MIT89].

. Le style "masque"

Pour compenser les limites du style "question-réponse", on a souvent recours à un masque présentant d'emblée les diverses informations à fournir au système pour engager une action de recherche. L'exemple type est celui de l'annuaire électronique.

Le style "masque" sera d'autant plus efficace que la circulation entre les diverses parties du masque sera facile (utilisation d'une souris dans les interfaces graphiques, touches "suite" et "retour" du vidéotex...) et que l'on pourra distinguer la formulation de la question (circulation dans le masque) de l'envoi de la question au système.

On est toutefois, dans le domaine documentaire, confronté à un problème de complexité : l'utilisateur a tendance à remplir toutes les cases du masque qui lui est proposé. Si on lui propose ainsi une grille "Sujet, Auteur, Titre, Niveau, Date" (par exemple) il aura tendance à remplir toutes les hypothèses, ou à se sentir désarçonné s'il ne connaît pas une information (date de publication par exemple). Cette habitude, plus marquée à l'étranger, est moins sensible en France, en raison du succès de l'Annuaire Electronique.

On peut aussi proposer des dialogues emboîtés, présentés dans les seuls cas

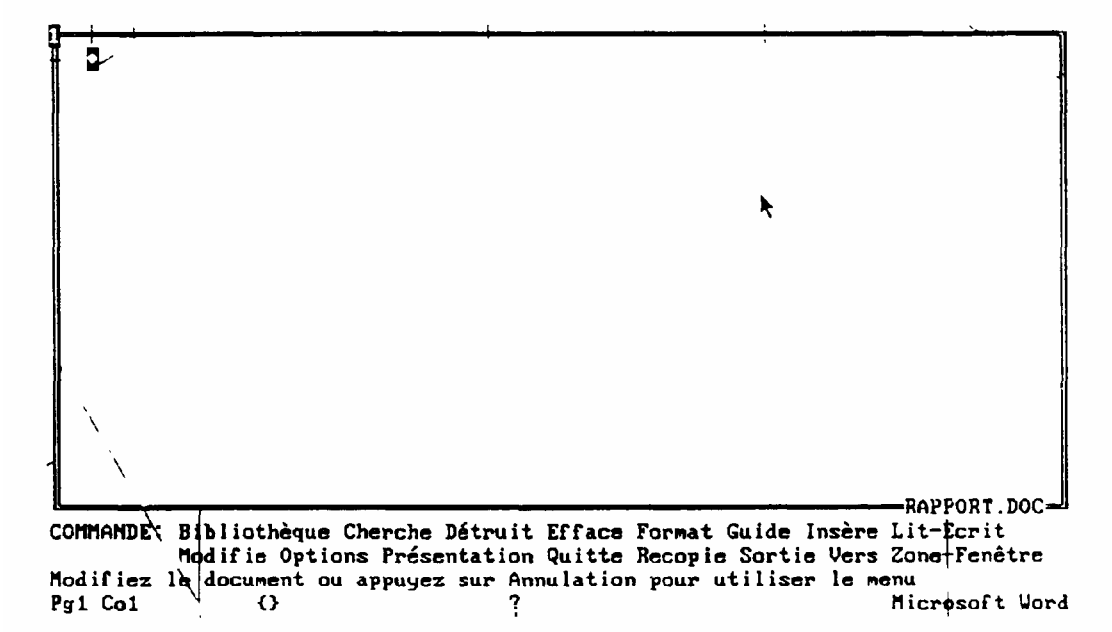
où cela est utile, par exemple par surabondance ou manque de réponses. Cette demande complémentaire s'inscrit alors dans une "boîte de dialogue" qui se surajoute au masque, sans le faire disparaître, ce qui permet de conserver les points de repère.

. Le style "à la carte"

Particulièrement adapté aux interfaces graphiques, le style "à la carte" permet à l'utilisateur de choisir le "menu" sur lequel il va travailler. Les diverses opérations possibles sont présentées en permanence à l'écran, l'utilisateur va choisir le cadre, puis l'action dont il a besoin.

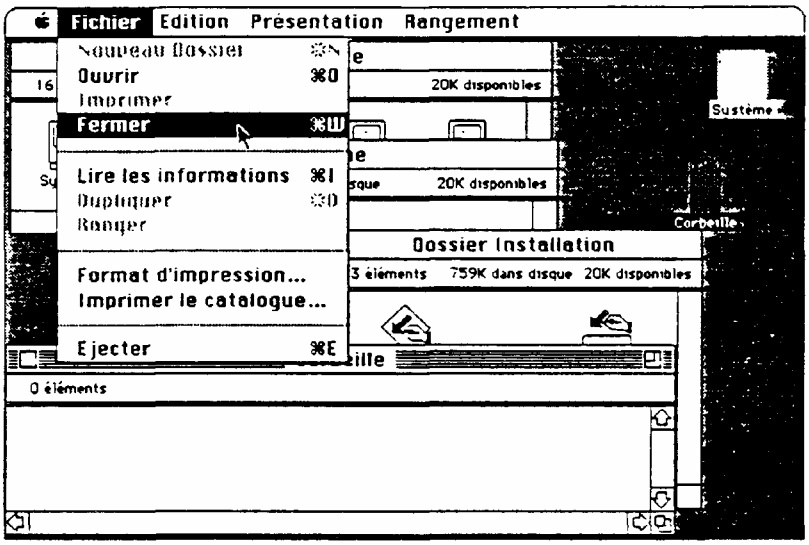
On distingue :

- les menus fixes : les actions sont repérées sur l'écran et une touche spéciale permet de basculer de l'application au menu. La circulation dans le menu se fait alors avec les flèches du clavier, la barre d'espace ou par la lettre de la commande mise en surbrillance. Le logiciel de traitement de texte *Word* est un bon exemple :



- les menus déroulants, pour lesquels l'activation d'un premier menu se traduit par l'ouverture d'une nouvelle liste de choix. Il faut éviter la multiplication des sous-menus, qui sont plutôt le signe d'applications de haut

niveau professionnel (par exemple en P.A.O. ou en C.A.O.), et limiter la liste des choix dans un menu. Les menus déroulants sont, par exemple, utilisés par le *Macintosh.*



- les "pop up" menus, qui apparaissent en surimpression à l'écran, en général associés à des fenêtres.

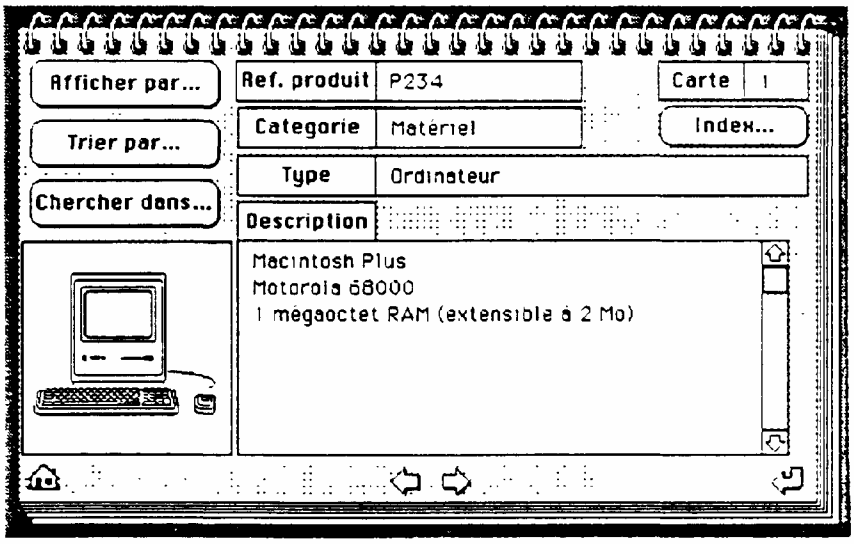
L'utilisation de ce style de "menu" laissé au choix de l'utilisateur commence à se généraliser dans le monde de la micro-informatique et des stations de travail. Dans le domaine documentaire, ce sont les applications sur Disque Optique Compact qui sont les plus friandes de ce type d'interaction. L'existence d'une interface graphique et d'un instrument de pointage (souris, trackball,...) est particulièrement bien adapté à ce style de dialogue. Le mode par menus laissés au choix de l'utilisateur prend bien en compte les nécessités d'un système documentaire de proposer des opérations multiples, notamment des opérations qui viennent en complément de la recherche documentaire : impression des résultats, conservation de certains documents (notion de "panier" où l'on entasse les documents sélectionnés), élimination d'autres, préparation de fichiers récupérant les informations, commande permettant de quitter le système, commandes permettant d'utiliser en parallèle d'autres applications (liens D.O.C./traitement de texte)...Les choix de l'utilisateur à chaque instant sont relativement élevés, et l'existence de menus cachés permet de gérer cette complexité sans encombrer les écrans.

La majeure partie des applications basées sur des menus déroulants offrent aussi des "raccourcis clavier", qui par combinaison de touches permettent aux "experts" ayant bien maîtrisé le système d'accélérer les actions.

On peut signaler aussi l'existence de menus en images, notamment dans les applications utilisant les vidéodisques. Le choix est alors réalisé par l'intermédiaire d'un écran tactile. Exemple : l'écran tactile du Musée Mémorial de Caen ou de la salle des dates du Musée d'Orsay (les icônes représentent les séquences filmées que l'utilisateur peut voir).

. Le style "environnement"

Dans le style environnement, les actions de l'utilisateur sont inscrites dans le cadre même de l'écran (par exemple l'écran comme métaphore d'un bureau du Macintosh), parfois "au travers" du document lui-même (hypertexte). Le style "environnement" fait appel à des icônes, qui représentent les actions à accomplir, et à la superposition de fenêtres, chacune regroupant une action ou un élément d'information. L'existence d'un choix, d'un croisement dans le chemin qui permet de parcourir les informations, peut aussi se traduire par une modification de la forme du curseur de pointage quand il passe au dessus d'une zone permettant de relancer l'action. Dans le cas du style environnement, les menus sont en quelque sorte introduits dans l'information elle-même *(Embedded menus :* [KOV86]). Le modèle hypertexte utilise pleinement le style environnement, comme dans l'écran *HyperCard* suivant :



3.b - La lisibilité des fonctionnalités

Une interface se caractérise par une métaphore de l'application. Il s'agit de proposer à l'utilisateur la manipulation d'objets informatiques qui représentent au mieux les objets informationnels de l'application. De ce point de vue, le document informatique peut se comporter comme un document papier : on pourra le ranger dans un dossier, le conserver, le dupliquer (équivalent de la photocopie), l'éliminer (corbeille à papier), comparer deux documents (partage de l'écran, multi fenêtrage)...Pour que la métaphore de l'interface soit cohérente, il importe qu'elle puisse être suivie jusqu'au bout. Par exemple, la métaphore du "bureau", qui définit le système d'exploitation du Macintosh, permet de récupérer les documents jetés à la "corbeille".

Pour l'utilisateur, la métaphore choisie doit aussi se maintenir de bout en bout de l'application. Devant une application informatique, on acquiert rapidement des "automatismes". Il est donc souhaitable que la même action se traduise par les mêmes effets. C'est cette logique de la métaphore qui fait par exemple le succès du Macintosh (un "air de famille" entre toutes les applications) et du Minitel (une même ergonomie des touches de fonctions pour toutes les applications). Rendre lisible les fonctionnalités d'une application informatique permet aussi d'en améliorer la cohérence.

Dans le domaine documentaire, il y aurait en ce domaine une piste de travail intéressante, qui consisterait à demander à des utilisateurs ce qu'ils aimeraient pouvoir réaliser à chaque étape de leur recherche. On a tendance en effet à considérer la recherche comme une activité linéaire : démarche tracée de la requête à la réception d'une liste de réponses. En réalité, pour que les applications documentaires soient plus en conformité avec les opérations intellectuelles en œuvre dans un travail documentaire, il faudrait tenir compte de plusieurs facteurs : lire, conserver, comparer.

- lire les informations renvoyées par le système et s'en servir pour l'interaction avec le système : équivalent à la discussion avec le bibliothécaire *("Cela m'intéresse, n'auriez-vous pas d'autres documents ?", "En fait, je recherche plus précisément...")*

Cela implique pour le concepteur de régler des problèmes de temps de connexion : trouver les méthodes informatiques pour ne pas encombrer l'ordinateur hôte pendant ces moments de réflexion (le temps de lire !), et trouver les méthodes de facturation qui rendent possible cette activité. Le choix du serveur de *l'Agence Spatiale Européenne* de facturer l'utilisation des banques de données sans tenir compte du temps de connexion (i.e. un forfait de connexion + un prix plus fort pour les références conservées) est un premier pas en ce sens. Malheureusement, il limite les recherches ponctuelles (une référence, une demande précise, qui nécessite peu de temps, et pour laquelle le forfait reste encore trop élevé). A l'opposé, le *type* de facturation de l'annuaire électronique, en assurant la gratuité d'une recherche rapide, tend à augmenter le nombre d'utilisations, même pour des besoins simples qu'un agenda pourrait facilement remplir.

- conserver facilement les documents électroniques dans un dossier personnalisé, que l'on pourrait aisément consulter au cours même de la recherche. Une recherche documentaire, c'est aussi une opération de sélection parmi des documents. *"L'information n'est que la matière première sur laquelle se branchent les systèmes de décision'* ([LEV89] p.49). L'information est un produit brut ou semi-fini, souligne Herbert Simon (cité par Levine [LEV89]), qui loin d'être rare est généralement trop abondant.

Réaliser une recherche documentaire, c'est avant tout synthétiser et filtrer l'information pour répondre à un but (une prise de décision) défini au moment de la recherche, et même parfois dans son cours lui même (par exemple dans le domaine scientifique). Il faut que l'utilisateur puisse juger de l'état d'avancement de son travail documentaire, notamment pour évaluer le bénéfice qu'il pourrait retirer à le poursuivre.

- permettre de comparer les documents extraits de la banque de données. Cette fonction a été particulièrement mise en avant dans le cas des banques d'images, mais elle peut être généralisée à toute la documentation. Les images, grâce à leur forme synthétique s'adressent directement au cerveau. Le temps de décodage et d'intégration est relativement court. Le choix d'une image répondant à un besoin déterminé (par exemple la sélection d'une photo de presse, ou des études en histoire de l'art) passe par une période de juxtaposition de

plusieurs images, avec des opérations de filtrage faisant émerger les images les plus adéquates ([LEC88a], p233-234).

Les imageurs documentaires ([HUD85], [HUD86]) remplissent ces fonctions. L'écran se comporte alors comme une table lumineuse sur lequel sont disposées les "imagettes", représentations réduites des images sélectionnées. Assurer la cohérence de la métaphore (écran = table lumineuse), c'est permettre de glisser des images de côté (élimination ou conservation pour un deuxième passage de sélection), de mettre côte à côte des images semblables pour les comparer, les évaluer. C'est pouvoir reprendre ensuite en format plein écran une ronde d'images à partir de la sélection sur les imagettes.

Nous avons l'habitude de voir des images sur écran, et la rapidité de perception fine des informations iconographiques facilite la mise au point d'interfaces permettant ces trois opérations (lire, conserver, comparer). Il est important de transporter ce savoir-faire dans la documentation textuelle, pour ouvrir la route à un nouveau type de lecture. La lecture électronique devient un enjeu culturel déterminant. Pour avoir des chances de succès, il faut que l'on puisse réaliser à l'écran le même *type* d'opérations que celles que l'on réalise avec des documents imprimés sur une table de travail : regrouper, lire aisément un document puis un autre, souligner, annoter, dupliquer, éliminer...De ce point de vue, les systèmes documentaires sont un excellent moyen de tester la validité des métaphores sur le travail intellectuel. Moins volumineux et en conséquences moins difficiles à traiter que les textes intégraux, les références ou les extraits de textes contenus dans les banques de données se prêtent au même type d'opérations que celles qui seront en œuvre dans la lecture électronique.

3.c - Les signes de l'échange

Les systèmes documentaires ne peuvent pas, nous venons de le voir, être considérés comme un simple mode de diffusion de l'information du système vers un lecteur qui prendrait séquentiellement les documents envoyés. Les processus mentaux engagés par l'activité documentaire sont plus complexes. Or souvent, les systèmes actuels se comportent de cette manière. Une fois l'information sélectionnée en réponse à une requête, elle est considérée comme pertinente et

envoyée à l'utilisateur qui ne pourra l'utiliser qu'ensuite, ayant quitté le système documentaire (lecture, sélection et comparaison sur le *"listing"* des références). Cette logique est aussi celle du réseau vidéotex (à l'origine, les imprimantes n'étaient pas prévues, le vidéotex étant un moyen de consultation des banques de données basé sur l'hypothèse : une question, une réponse). C'est aussi celle de certains anté-serveurs. Par exemple, le réseau *EasyNet* ne propose que vingt réponses à toute question, sans jugement de l'utilisateur.

Dans ce cadre, l'utilisateur interagit peu avec le système. Le retour de l'information est à sens unique : le système montre qu'il a bien enregistré la demande de l'utilisateur, et éventuellement indique qu'il ne veut pas prendre en compte certaines actions ("commande inconnue", "non adaptée", "fin de liste"...).

La longueur des traitements est un point important dans le dialogue homme système L'utilisateur ne sait pas si le système est bien en train de traiter sa question. Rares sont pourtant les interfaces qui indiquent par un message faisant prendre patience que le système travaille pour eux. Or l'utilisateur est lui-même en *"stand by"* quand l'ordinateur sur lequel il travaille est inactif. On attend la reprise du dialogue. Si le traitement est long, et en documentation informatisée un traitement devient long après seulement quelques secondes d'attente, un doute émerge : que fait le système ? Ma demande particulière est-elle prise en compte ?...Le style de travail actuel, accentué par les pratiques de facturation, ne permet pas de se détacher du traitement en cours, pour se consacrer à d'autres opérations (lecture des premières références, réflexion sur la recherche en cours...).

Dans les systèmes plus évolués, qui commencent à voir le jour, notamment au travers de l'utilisation des D.O.C., l'interaction est plus complexe, ce qui rend encore plus nécessaire de renvoyer à l'utilisateur un signe (passage en vidéo inverse, message, clignotement...) indiquant que le système a bien compris et pris en compte la demande ou le traitement voulus par l'utilisateur. Ainsi, permettre la sélection d'un document dans une liste pour le mettre en réserve dans un panier *(CD-Navigator,* développé par *ACT Informatique)* s'accompagne de la mise en valeur du document ainsi sélectionné (généralement mis en inverse vidéo) pour certifier à l'utilisateur la prise en compte de son choix. Dans les systèmes hypertexte *(Guide* sur *Macintosh),* le retour d'information se traduit par la

modification de la forme du curseur (la métaphore de la montre, qui tourne pour faire patienter).

Dans les systèmes à jugement de pertinence, le retour d'information devra être plus complexe : à la fois dirigé vers l'utilisateur pour lui confirmer ses choix, mais aussi vers le système pour relancer la dynamique de recherche, tout en donnant des informations à l'utilisateur sur les résultats de cette relance (nouveaux documents extraits). Dans ce cas, la métaphore adéquate est celle d'une navigation dans un espace documentaire. Elle reste cependant difficile à synthétiser de manière claire sur un écran.

3.d • La tolérance aux erreurs

Une interface homme système d'information a entre autres fonctions de gérer les erreurs de l'utilisateur. Dans le cadre documentaire, les erreurs les plus fréquentes sont de deux types [BORG86] :

. les problèmes mécaniques : manipulation du clavier, fautes de frappe, mauvaise maîtrise du langage de commande.
. les problèmes conceptuels qui renvoient à la difficulté de l'utilisateur à formuler sa question devant un système informatique.

1 - problèmes mécaniques

Plusieurs méthodes de conception de l'interface permettent de limiter les effets des erreurs mécaniques, notamment les erreurs de manipulation, liées aux commandes, et les fautes de frappe ou d'orthographe.

- *erreurs de manipulation.* Par exemple utilisation d'une mauvaise commande, envoi d'une question que l'on veut retirer...Une interface conviviale devrait permettre de revenir en arrière à tout moment (l'équivalent du bouton "annuler" du *Macintosh,* qui est présenté à chaque opération) et offrir la garantie que les informations sont conservées pour une lecture ultérieure (sauvegarde des informations automatique ou proposée à l'utilisateur).

- *erreurs de frappe ou d'orthographe.* Ce cas est plus complexe à traiter. Il implique en effet un choix sur "ce qu'a voulu exprimer l'utilisateur". En examinant une liste de demandes non abouties sur le catalogue de la Médiathèque de Valence, Pierre Le Loarer repérait les phénomènes suivants ([LEL89]) :

. Les utilisateurs utilisent souvent le *jeu de caractère réduit* (en majuscules non accentuées), mais cela est probablement lié à l'utilisation du minitel comme terminal. En sens inverse, d'autres systèmes ne traitent pas les caractères accentués (GRACE, catalogue de la Bibliothèque Universitaire de Bordeaux).

. les *fautes de frappe,* liées à la faible habitude des claviers, se divisent en quatre groupes :

- répétition de lettres : AANIMAL
- omission de lettre ou d'un espace : HISTIRE, SOCILOGIE
- remplacement d'une lettre par une touche proche sur le clavier, ou double frappe de deux touches proches : GENEALPOGIE, GYMNASYTIQUE, REGENCR
- inversion de lettres : DOUCMENTATION

. les *fautes d'orthographe* sont fréquentes : AMITIEE, DACTILOGRAPHIE.

L'ensemble des fautes (orthographe et frappe) privilégient quatre types d'erreurs ([TS086], [PUJ89]) :

. "omission" (O) : métode (th)
. "insertion" (I) : cabanne (n)
. "Substitution" (S) : Entécédent (An)
. "Transposition" (T) : Ihnèrent (nh)

Des études menées sur des textes dactylographiés montrent que ces quatre types d'erreurs couvrent environ 80 à 90 % des fautes lexicales, avec environ la moitié (40 %) pour les fautes par omission d'une lettre. Cette connaissance permet d'envisager des algorithmes de correction orthographique puissants, suivant plusieurs méthodes générales [PUJ89] :

. *méthodes combinatoires :* à partir de la chaîne de caractères erronée, on génère toutes les chaînes possibles qui pourraient donner la chaîne erronée par l'une de ces quatre opérations.

. *méthodes statistiques :* basées sur les fréquences d'apparition des digrammes ou des trigrammes (séquences de deux ou trois lettres) dans la langue considérée, ces algorithmes permettent surtout de corriger les mots longs.

. *méthode des alpha codes :* utilisée par le système SPIRIT, cette méthode se base sur la transposition des mots suivant une suite de lettres classées par ordre alphabétique (ALLIANCE -> AACEILLN). Un lexique des alpha codes permet de déterminer les mots candidats (en français, il y a en moyenne 1,05 mot candidat par alpha codes). A partir de alpha codes du mot erroné, on ajoute ou enlève une ou deux lettres et compare les mots - candidats obtenus avec la chaîne erronée, ce qui permet de sélectionner le terme le plus proche.

. *méthodes métriques :* on calcule une distance entre la chaîne erronée et les mots du lexique. En dessous d'une certaine valeur de cette distance, les mots sont candidats à la correction. La distance peut être le nombre d'opérations nécessaires pour passer d'une chaîne à l'autre, chaque opération étant pondérés par sa fréquence (statistiques d'erreur).

. *méthodes phonétiques :* Un fichier des représentations phonétiques des mots est conservé en mémoire. Cette méthode est utilisée par l'Annuaire électronique. Elle a l'inconvénient de ne pas corriger les fautes de frappe qui produisent des mots imprononçables.

2 - Les problèmes conceptuels

Reprenant la classification de Christine Borgman, Nathalie Mitev [MIT89] divise les problèmes conceptuels en quatre types :

. difficulté à exprimer la recherche à l'aide de critères d'interrogation précis, qui sont malheureusement nécessaires aux systèmes actuels

. difficulté à combiner les concepts selon la logique booléenne
. difficulté (et résistance) des utilisateurs à faire correspondre leurs propres termes au langage de la banque de données utilisée
. difficultés de positionnement de l'utilisateur, liées à la découverte d'un manque dans son savoir si le sujet lui est totalement ou partiellement inconnu.

Ces difficultés conceptuelles se traduisent souvent par des formulations qui pour être acceptables par un intermédiaire humain sont considérées comme des "erreurs" par le système. L'article de Pierre Le Loarer [LEL89] repère ainsi quelques exemples :

. l'utilisation de formulations respectant la syntaxe du français au lieu de celle des termes d'indexation : LES CADRES DANS L'ENTREPRISE, EXPEDITIONS DANS LE DESERT (au lieu de *"entreprises, encadrement"* ou *"désert, expéditions"* si tels sont les termes choisis dans le système).

. des formulations qui s'apparentent plus au conseil qu'à la démarche documentaire : LIVRES SPORTIFS, UN CONTE FACILE

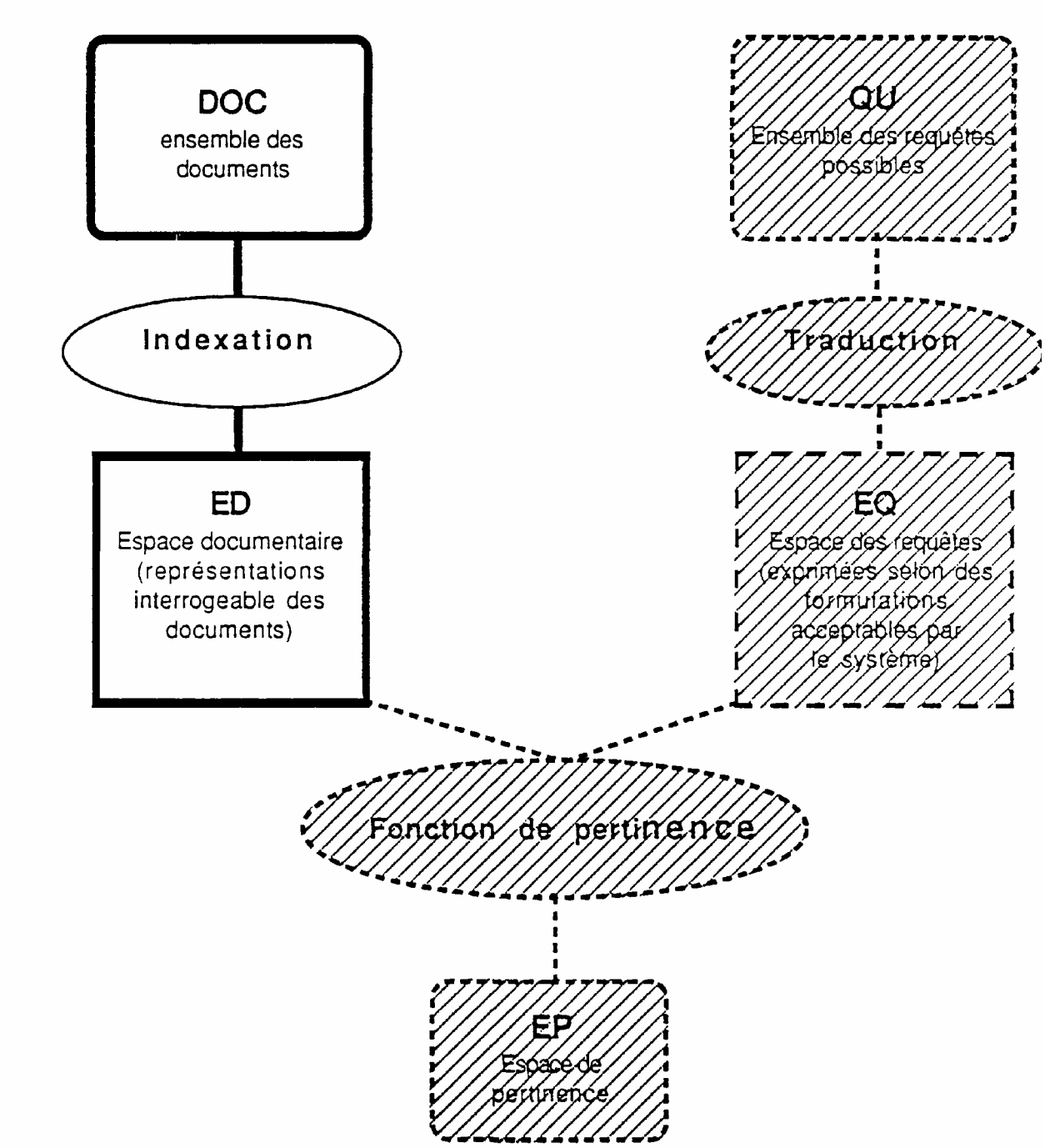
. une formulation traditionnelle des noms d'auteurs qui ne prend pas en compte l'ordre imposé par le système : ALBERT CAMUS, VICTOR HUGO (au lieu de *"Camus, Albert"* ou *"Hugo, Victor")*.

S'il est illusoire de considérer qu'une interface peut régler tous ces problèmes indépendamment de la conception du système documentaire lui-même, elle peut permettre de réduire les sources d'erreur par exemple en renvoyant à l'utilisateur les termes erronés (i.e. les termes inconnus par le système) pour qu'il corrige lui-même. Les fautes mécaniques sont inévitables. Il faut bien que l'utilisateur puisse formuler sa question (les méthodes de reconnaissance vocale sont étudiées, mais pas encore appliquées). Les fautes conceptuelles aussi. Elles doivent aussi être prises en compte au niveau de la transformation de la question de l'utilisateur en question acceptable par le système, et par la reformulation des questions.

III - f_i : DOC -> ED

La fonction d'indexation

Nous appellerons INDEXATION l'opération qui permet de passer d'un document à sa représentation manipulable par le système documentaire.



La fonction d'indexation a une double caractéristique :

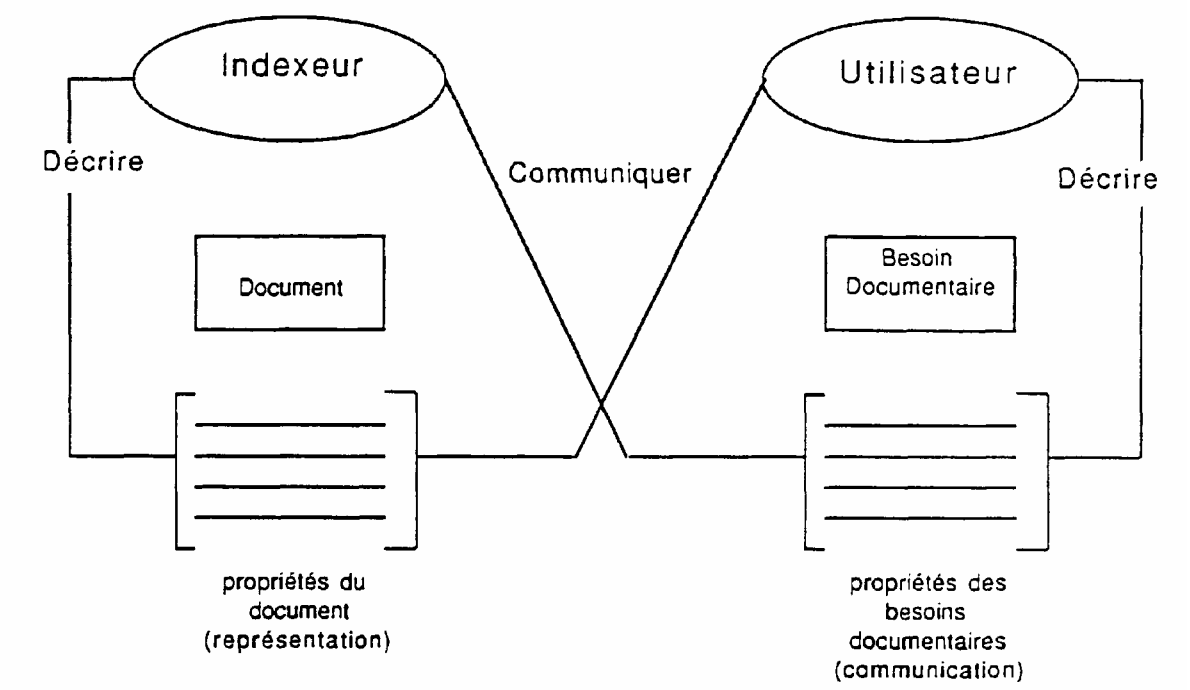
- elle dépend du document indexé : il s'agit de représenter le plus fidèlement possible son contenu, à la fois d'un point de vue synthétique (dans quel contexte se place le document) et d'un point de vue analytique (qu'est-ce qui le distingue des autres documents).

- elle dépend du système documentaire choisi, et notamment de sa convivialité, c'est-à-dire de la capacité du système documentaire à accepter et transformer des besoins documentaires exprimés par l'utilisateur en questions manipulables par le système. Par exemple, si l'on adopte une description codée des documents, cela n'a un sens que si l'utilisateur (ou l'intermédiaire, ou le système lui-même) est capable d'exprimer une requête suivant le même code.

On pourrait exprimer cette dualité sous une autre forme, en notant les deux utilisations de l'indexation d'un document

- *la représentation :* exprimer dans une grande économie de moyens des informations.
- *la communication :* choisir la représentation en fonction d'un utilisateur, et en fonction du canal de communication choisi.

On peut résumer cette approche par le schéma suivant, adapté de [THOM90]



1 - Une approche générale de l'indexation

L'indexation n'est pas une opération circonscrite au domaine documentaire. De nombreuses situations de la vie courante nous conduisent à "indexer" des informations, et à répondre à des questions formulées en utilisant des "descripteurs". On retrouve d'ailleurs dans ces situations toutes les caractéristiques de l'indexation dans les systèmes documentaires. Il semble utile de repérer dans ces relations informatives quotidiennes quelques éléments qui pourront nous guider dans l'étude de l'indexation dans le cadre des systèmes documentaires.

. exemple 1 : Il s'agit de décrire un ami à une tierce personne qui doit aller le chercher à votre place à l'arrivée d'un avion [BLA90, p. 172]. Nous savons décrire un individu pour le distinguer parmi une centaine de personnes, au moyen de critères-clés : taille, corpulence, couleur des cheveux, des yeux... Au pire, si deux ou trois personnes correspondent à votre description *("bruit documentaire"),* on peut toujours leur demander s'il s'agit bien de notre interlocuteur *("jugement de pertinence").*

La même situation devient beaucoup plus complexe s'il s'agit de décrire la même personne pour qu'on la retrouve parmi plusieurs milliers d'autres, disons parmi les spectateurs d'un match de football. Le nombre de grands moustachus de corpulence moyenne, blonds aux yeux clairs est tout de suite trop élevé pour qu'on puisse les interroger tous. Bien entendu, cette situation devient plus facile si le match de football se déroule au Cameroun, où les éléments de la description deviennent significatifs car rares. On retrouve là des questions relatives à la taille des systèmes documentaires d'une part, et aux connaissances pragmatiques qui figurent parmi les présupposés non explicites d'une recherche documentaire. On ne peut raisonner l'indexation (la description) en dehors de ces critères : quel système est interrogé ? Combien d'éléments peuvent comporter la même description dans le système ? La description est-elle discriminante dans le cadre choisi ?

On doit aussi remarquer au travers de cet exemple que la description est toujours soumise au regard particulier de celui qui décrit. La taille par exemple sera interprétée différemment si "l'indexeur" est lui-même petit ou grand

. exemple 2 : Les titres des journaux télévisés sont une forme d'indexation visant à la communication : faire passer en quelques termes le message qui sera développé ultérieurement. On retrouve alors le poids extraordinaire des connaissances pragmatiques indispensables pour la compréhension, mais qui ne sont jamais explicités dans le système. *"Ronan Pensée emporte le maillot jaune au sommet de l'Alpe d'Huez"* nous apparaît comme une phrase hautement significative, un condensé extrêmement bref (une description documentaire doit être courte, en moyenne une dizaine de termes dans les banques de données, voire un ou deux dans les catalogues de bibliothèques) et riche en termes précis (deux noms propres, une expression composée et un nom). Pourtant, nous ne comprenons cette phrase que dans la mesure où nous savons qu'il s'agit de cyclisme, que Ronan Pensée est un nom propre, le nom d'un coureur cycliste, que le maillot jaune est le signe distinctif du premier au classement général. Nous pouvons de plus ajouter que la course se déroule en montagne, et que l'arrivée est en haut d'une côte, que par ailleurs, pour peu que nous nous intéressions au cyclisme, nous savons être une des plus difficiles de l'épreuve. Quelle épreuve au fait, qui n'est jamais mentionnée dans le système ?

Ce poids des aspects de contexte dans la description documentaire est extrêmement difficile à gérer dans les systèmes documentaires dès qu'ils deviennent importants. Le contexte *"livre d'informatique"* est suffisant dans une bibliothèque municipale. Il devient inopérant dans une bibliothèque scientifique, et dérisoire dans une banque de données spécialisée. Dans la logique traditionnelle de l'indexation, on doit toujours partir du document lui-même, et le décrire pour lui-même. Cette opération engage pour l'indexeur toute une série d'opérations mentales destinées à positionner ce document dans le cadre général où il se situe, à déduire de sa lecture des informations propres, et éventuellement à rajouter des informations plus contextuelles. Il y a toujours un double jeu : agréger chaque document dans un contexte, et le distinguer par sa spécificité. C'est le propre de l'activité communicante (presse, publicité...) que de jouer sur le *"déjà su"* ou le *"déjà repéré"* et d'ajouter l'élément distinctif, de *"nouveauté"* ou de *"qualité"*. Or dans un système documentaire, l'arrière plan est inaccessible, non formulé dans une question d'utilisateur, et généralement pas explicité par le système lui-même (problèmes des plans de classement, problèmes d'un nouveau vocabulaire et des effets de "mode lexicale" [RIC90]).

. exemple 3 : Dans cette situation, deux amis discutent : " *Ils viennent de ressortir* Comme un torrent *de* Minnelli. *Sais-tu ce qu'en disent les critiques ?"* Telle que, la question appelle une recherche extensive de toutes les critiques du film de Minnelli, éventuellement depuis sa sortie en 1958. Or nous savons bien qu'en occurrence, il s'agit d'avoir un avis, et que la satisfaction de l'utilisateur est remplie par la mention de la lecture d'un ou deux articles récents, et surtout des jugements portés par le critique. Mais on ne peut exclure le besoin d'un utilisateur particulier de connaître tous ou du moins une large part des articles consacrés à ce film pour un travail spécifique.

Cette situation où la même formulation peut recouvrir plusieurs niveaux d'exigence et plusieurs types de besoins se retrouve fréquemment en recherche documentaire. L'utilisateur sera-t-il satisfait par quelques documents, par une large part, ou bien a-t-il besoin de connaître absolument tous les documents sur un sujet (par exemple dans le cas de recherches d'antériorité des brevets ou pour des procédures judiciaires) ? Dans notre exemple, les "mots-clés" restent les mêmes (titre du film et nom de l'auteur), ce qui implique d'ajouter d'autres critères distinctifs pour satisfaire en même temps les deux types de recherche (i.e. ne pas noyer l'utilisateur sous une masse de documents). Cette indexation qualitative est en général très difficile à obtenir. Ne serait-ce que pour des raisons juridiques : quel producteur de banque de données oserait affronter le courroux d'un auteur ou d'un éditeur en jugeant "faible", "inadapté" ou "répétitif le contenu d'un document ? Surtout s'il s'agit d'un texte vérifié et adopté par des pairs comme dans le cas des brevets ou des articles scientifiques.

. exemple 4 : Quand nous rangeons un objet, par exemple au retour des vacances, nous essayons de satisfaire deux critères : l'aspect pratique, qui veut que cet objet qui nous sera inutile pendant un an n'encombre pas, et l'aspect logique, qui veut que nous le retrouvions facilement quand nous en aurons besoin. De fait, nous attribuons une indexation de type classificatoire, qui permette de diminuer le *"bruit documentaire"* (un rangement efficace) en limitant le *"silence documentaire"* (les heures passées à rechercher un objet). Or combien de fois ne nous est-il arrivé de rechercher un objet en un tout autre endroit que celui où nous-même l'avions rangé ? L'indexation qui nous paraissait logique, lourde de signification, lors du rangement, ne correspondait plus du tout aux termes de

recherche, eux aussi totalement logiques, qui nous venaient à l'esprit pour retrouver cet objet.

Cet exemple souligne la part de subjectivité de l'activité d'indexation, qui fait que deux indexeurs, ou un indexeur et un utilisateur, n'utilisent pas les mêmes mots-clés pour décrire les mêmes choses. Que de plus ce sentiment confus d'accorder à un document une description adéquate varie énormément dans le temps. Ce qui était significatif et discriminant finit par ne plus l'être avec l'évolution du système documentaire (par exemple évolution du vocabulaire dans les domaines scientifiques) et avec l'évolution des types de demandes portées au système (ainsi, Noë Richter dans [RIC90] faisait remarquer que des descripteurs comme *"Réforme Haby",* significatifs en 1975 étaient devenus abscons en 1990 où le terme plus général *"Réforme du système scolaire, 1975"* aurait été plus adapté).

Ces exemples tirés de la vie quotidienne montrent que dans la discussion entre humains, la part du non-dit, de l'aspect pragmatique, ou de contexte, entre pour une part très importante dans le succès de la communication. Si l'on conçoit une relation verbale, il faut ajouter le poids des intonations, et, si les deux personnes sont en présence, celui des mimiques et des signes corporels divers de la communication. De plus, le dialogue entre humains permet de demander des précisions pour tout terme mal compris, ou ambigu. Ces caractéristiques disparaissent dans la plupart des systèmes documentaires (ils restent présents dans les bibliothèques, où l'utilisateur peut toujours demander au bibliothécaire *"un livre policier français moderne de qualité" ,* ce qui explique certainement le succès plus important des bibliothécaires que des systèmes informatisés). L'indexation pour les systèmes informatisés est donc une opération difficile qui doit prendre en compte les éléments d'arrière-plan, en évitant que leur explicitation rende le système trop lourd à manipuler (trop de documents ayant le même terme d'indexation).

Ces exemples montrent aussi que l'indexation est fondamentalement une opération inconsistante et aléatoire. Indexer, c'est attribuer à un document des descripteurs en fonction de critères implicites (aspect communication), avec une large dose d'incertitude (dans quelle mesure ce terme est-il discriminant ?) et un

coefficient de confiance dans l'opération fort limité (aurait-on adopté le même choix demain ?).

Si l'on rapporte les remarques faites ici dans le strict cadre de l'indexation documentaire, on retrouve de même les deux aspects permanents des activités d'indexation : la représentation et la communication. On retrouve aussi les problèmes relatifs au niveau de la description, à l'inconsistance et à la variabilité de l'indexation soulignés dans ces quelques exemples.

L'enseignement actuel de l'indexation ([RIC86], [RIC87], [RIC88], [RIC90], [ROY87], [MAN87]) distingue deux phases dans l'indexation : l'analyse qui *"est l'extraction des éléments caractéristiques du sujet"* et l'indexation *"qui est une sélection des mots-clés dégagés par l'analyse et leur transposition dans un langage structuré, codifié et normalisé"* [RIC88]. Il faut de plus tenir compte du lecteur dans le choix et l'attribution des descripteurs. Richard Roy dans [ROY87] citait l'exemple d'une recherche de documents sur l'holographie, constatant que le lycéen sera rebuté par l'article correspondant de l'*Encyclopedia Universalis* alors que le chercheur sera déçu par le dossier spécial de *Science et Vie.* Il faut aussi tenir compte des documents déjà intégrés au système documentaire, et des liens implicites (calculés par la réalisation d'agrégats) ou explicites (le jeu des citations entre documents) qui s'y sont tissés. Certains termes d'indexation deviennent inopérants car trop fréquents alors que d'autres sont souvent trop restrictifs car trop faiblement employés. Enfin, l'indexation n'est pas totalement distincte de la description matérielle du document (le catalogage), car des données de contexte utiles à la recherche documentaire sont aussi présentes à ce niveau (nom d'auteur, collection de livres, périodiques spécialisés, mots du titre...). En fait, il s'agit de concevoir l'indexation comme une opération visant à permettre, à partir de plusieurs méthodes critiquables et aléatoires en elles-mêmes, qu'un document entré dans le système soit retrouvé par l'utilisateur qui pourrait en avoir besoin. Nous allons voir ci-après les diverses méthodes utilisées, leurs avantages et leurs limites, notamment dans le cas des grandes banques de données, qui posent toujours un problème particulier.

Il faut auparavant souligner les problèmes de niveau pragmatique qui se posent lors d'une opération d'indexation. La connaissance du contexte pose un problème particulier en informatique documentaire. La recherche quotidienne

d'information passe souvent par des méthodes qui tiennent plus compte du contexte de production de cette information qu'elle ne procède de la recherche par des descripteurs attachés à l'information recherchée. Ainsi, dans une entreprise, rechercher un document sur "tel contrat" se traduit généralement par *"Va voir X. qui s'en occupait, il doit avoir un double"*. Les bibliothécaires connaissent bien ce type de demande : *"C'est un livre vert avec le titre en grosses lettres rondes, je l'ai vu chez Pivot"*. Sans demander à un système informatique de répondre à cet extrême (quoique le service vidéotex 3675 *Apos* soit tout à fait adapté, car classant les documents suivant au moins un aspect contextuel : la date de passage dans l'émission), il est important que le niveau pragmatique soit pris en compte dans l'indexation documentaire.

Ainsi, si l'on considère le système documentaire des avis techniques et des références des pièces d'un navire de guerre [BLA90, p. 188], les aspects pragmatiques dans la recherche d'information en cas de panne (e.g. l'endroit où la panne est arrivée, le système en cause,...) l'emportent sur les aspects descriptifs (le nom de la pièce défectueuse, e.g. une valve). Il existe certainement des centaines de valves semblables sur un bateau, ce qui rend la recherche par descripteurs irréaliste.

Cette insertion des éléments pragmatiques dans les systèmes documentaires n'est pas évidente. Elle implique de redéfinir les méthodes d'indexation et les traitements correspondants lors des recherches. Jusqu'à présent, les diverses méthodes que nous allons étudier sont souvent limitées aux données brutes que l'on peut extraire du document, par exemple les mots du document lui-même pour l'indexation en texte intégral. Or le propre des indications de contexte est de n'être pas toujours rappelées dans les documents eux-mêmes. L'activité classificatoire représente néanmoins un signe de cette volonté de replacer chaque élément d'information (niveau analytique) dans un contexte de point de vue (niveau synthétique). Toutefois, la capacité des systèmes documentaires informatisés à jouer sur ces deux niveaux reste encore élémentaire, et consiste le plus souvent à faire peser sur l'utilisateur ou l'intermédiaire en information la compréhension et l'utilisation de ce type d'informations quand elles sont contenues dans la banque de données.

2 - L'indexation manuelle

L'indexation manuelle reste aujourd'hui la forme la plus répandue d'indexation. Un indexeur, formé aux diverses techniques documentaires, attribue au document un certain nombre de *descripteurs.* Les descripteurs sont en général des formes nominales ou des codes. On peut distinguer :

- des mots-clés unitermes, formés d'un seul mot. Les expressions composées sont obtenues par conjonction de plusieurs unitermes. Ceci a l'inconvénient de mélanger des sens distincts : *"Droit* et *Travail"* retrouvera ainsi *"Droit du travail"* et *"Droit au travail"* [RIC88, p. 23].

- des descripteurs composés, constitués de syntagmes comportant deux ou trois termes. On trouve des expressions du type :

. nom adjectif : *Droit social*

. nom préposition nom : *Chemin de fer, Histoire de la musique*

. des termes possédant un trait d'union : *Libre-échange, Science-fiction.* Ces termes sont souvent difficiles à traiter. Par exemple on trouve les deux formes : *Agro-alimentaire* et *Agroalimentaire.*

. des termes avec rejet : *Boole, algèbre de* . Ces formes sont cependant adaptées aux catalogues manuels, en mettant en avant l'entrée de plus faible fréquence. L'indexation pour l'informatique documentaire peut s'en passer.

. des termes avec un qualificatif : *Mercure (métal), Mercure (Planète), Mercure (dieu)*

- des descripteurs structurés : sous une même entrée (dite *"vedette"),* on indique dans un descripteur structuré plusieurs informations. Ainsi, la norme Z 44-070 [AFN86] fait se succéder les descripteurs dans l'ordre suivant :

. tête de vedette, significative du sujet
. sous-vedette de point de vue
. sous-vedette de localisation géographique

. sous-vedette de localisation chronologique
. sous-vedette de forme (dictionnaire, bibliographie, congrès...)

La structure du descripteur se traduit par une syntaxe spécifique et des signes de ponctuation représentant les articulations logiques entre les termes. La ponctuation choisie dépend de la norme définissant la structure des vedettes.

- des codes numériques : les descripteurs peuvent être codés pour représenter simplement des notions conceptuelles difficiles à réduire à quelques mots. C'est par exemple le cas des *"concept codes"* de *Biological Abstracts* , des codes de produits de *Predicast,* ou des codes APE (France), SIC (Etats-unis) ou NACE (Europe) pour définir les secteurs économiques...

Dans cet exemple, extrait de *Biological Abstracts,* le *"code de concept"* permet de préciser l'ensemble des études couvertes par le code CC 13518, mais aussi de renvoyer certains sujets sur d'autres codes. Cet ensemble complexe d'informations se résume difficilement en un ou deux "mots-clés".

CC13518 DAIRY PRODUCTS

Frequencies Major (7230) Minor (1530)

Applications This code retrieves studies on milk and milk by-products, their processing, storage, composition and chemical properties.

Examples Studies on • milk • cream • ice cream • casein • caseinates • whey products • dried milk • cheese

Strategy Recommendations

- For dried milk used as an enrichment product, use appropriate keywords with this code and the *Synthetic, Supplemental and Enrichment Foods* code CC13534
- For microbiological fermentations in cheese, butter and fermented milks, use this code with appropriate keywords and the *Food and Beverage Fermentation* code CC39003.

- des indices de classification : les classifications ou les plans de classement permettent de situer un document dans la partie de la connaissance traitée par le système documentaire. Ils constituent souvent le seul repère contextuel. On trouve des classifications hiérarchiques (CDU, Classification Décimale Dewey, Classification soviétique BBK...), des classifications obtenues par juxtaposition d'indices (Library of Congress Classification) et une classification

par facettes (Colon Classification). Chaque classification permet de définir des *"indices",* qui sont un agencement de signes reliés par une syntaxe précise.

2.a - Règles et méthodes

Il existe de nombreuses règles d'indexation, l'objectif étant que deux indexeurs établissent la même indexation pour un même document. Nous verrons plus loin que cet objectif est parfaitement illusoire. Mais plus encore, pour que ce système ait des chances de fonctionner, il faudrait de plus que l'utilisateur connaisse et applique lui-même les mêmes règles d'indexation, afin de retrouver les documents qui l'intéressent.

On peut distinguer globalement trois types d'indexation manuelle :

. *une indexation libre,* où l'indexeur choisit lui-même les termes d'indexation, en s'efforçant de respecter les règles générales (nombre, organisation des informations...).

. *une indexation contrôlée,* dans laquelle les termes d'indexation sont obligatoirement choisis dans une liste pré-établie. On distingue alors les *"listes d'autorité",* classées alphabétiquement, qui ne gèrent que des relations de renvoi entre les termes, et les *"thésaurus"* qui définissent des relations sémantiques entre les termes (terme générique, spécifique, partie_de,...). L'informatique permettant de basculer d'un mode de présentation à l'autre (i.e. de l'ordre alphabétique des descripteurs à l'ordre sémantique des domaines de la connaissance) tend à rapprocher ces deux instruments.

. *une indexation par classification* du document dans un cadre prédéfini. La classification essaie de représenter le *"point de vue"* sous lequel est abordé le *"sujet"* défini par l'indexation littérale. Le point de vue est donné par la catégorie hiérarchique dans laquelle a été choisi l'indice de classification.

L'indexation par des termes littéraux permet de combiner les divers termes *a posteriori,* c'est-à-dire pendant l'interrogation du système. On parle d'indexation post-coordonnée. En sens inverse, le cadre des classifications est rigide et défini *a priori.* On parle alors d'indexation pré coordonnée [MAN87]. Les

divers outils documentaires (thésaurus, classifications, listes d'autorité) seront étudiés plus loin, afin de confronter leur utilisation pour l'indexation manuelle et pour l'indexation automatique.

Pour être utilisables par un système informatique, un document et ses données d'indexation sont associés dans une *"description structurée"*. Les informations sont réparties en champs. Chaque champ contient un élément d'information particulier. La polysémie du langage, le poids très important des connaissances pragmatiques nécessaires pour situer un document et ses divers qualificatifs ne permettent pas de regrouper toutes les informations dans un seul champ d'indexation. La recherche des textes de Victor Hugo (Auteur=Hugo, Victor) est différente de la recherche des biographies de Victor Hugo (Sujet=Hugo, Victor ET Biographie).

La structuration en champs est propre à la description informatique. Elle permet d'envisager autrement l'indexation en offrant de nombreux *"points d'accès"* aux documents. En ce sens, l'indexation des descriptions structurées se distingue du système des *"vedettes"* (vedette auteur, vedette matière) utilisé dans les catalogues sur fiches. Ce système, avec sa structure et sa syntaxe, correspond à une entrée unique pour un document. Il est remplacé par un accès plaçant l'intérêt sur l'un quelconque des points de la description documentaire (divers champs, divers éléments d'un champ, et combinaison de termes et de champs). Les mots du titre peuvent ainsi être considérés comme des moyens de retrouver un document particulier (utilisation du Titre comme une vedette) ou comme des indications de contenu (utilisation du titre des documentaires dans le lexique des sujets).

Dans la description structurée, on peut distinguer les *"champs interrogeables"* et les *"champs de visualisation"*. Cette distinction est particulièrement évidente pour les banques d'images, où l'interrogation se fait par l'intermédiaire des textes associés (même si des méthodes de jugement des images [HALI89] permettent de partir d'images connues pour reformuler la question, la reformulation est réalisée par le système à partir du niveau textuel). Malheureusement, bien des systèmes documentaires informatisés restent des héritiers de la pratique des vedettes. L'interrogation de ces fichiers documentaires est alors beaucoup moins souple. On parle dans ce cas de systèmes de recherche

" multicritères", par opposition à la recherche documentaire qui permet de travailler conjointement sur l'ensemble des champs de la description et qui autorise à la fois les restrictions (spécification d'un champ particulier) et les combinaisons booléennes entre propriétés des divers champs. La recherche multicritères, en descendance de la recherche par les "vedettes" ne permet que des intersections entre les diverses contraintes imposées sur les champs : il faut que l'ensemble des critères spécifiés soient remplis. De nombreux systèmes vidéotex, ou les catalogues de bibliothèques interrogeables en ligne fonctionnent sur le mode multicritères, qui reste le plus proche des habitudes et pratiques de la recherche dans les fichiers manuels. Ces habitudes ont pour conséquence de perpétuer dans le domaine de la documentation informatisée des règles d'indexation issues des pratiques du catalogue sur fiches. Même si certains commencent à vouloir intégrer les découvertes des systèmes documentaires dans les modes d'indexation des livres de bibliothèques ([BLAN89]), la majeure partie des bibliothécaires conservent la notion de vedettes structurées et l'indexation par un langage contrôlé ([RAM90]).

La majeure partie des articles sur la documentation informatisée considèrent un document comme représenté par une suite de descripteurs : *"soit D:=<t1, t2, t3,...t_k> un document du système..."*. Cela fait perdre de vue la nature complexe de la description documentaire. Un terme interrogeable est qualifié par le champ auquel il appartient, ce qui peut limiter les ambiguïtés. Ainsi, *"je cherche les articles sur l'isolation dans le domaine de l'électricité"* renvoie à *"sujet=isolation et domaine électricité"* mais reste différent de *"sujet=isolation et électricité"* qui ne prend pas en compte les différentes formes lexicales (notamment l'utilisation de la forme adjectif) que peut prendre la mention du domaine dans l'indexation sujet *("isolation électrique"* par exemple). Cette notation n'est valable que si l'on précise bien que le document lui-même, tel qu'il existe pour l'utilisateur, reste aussi complètement indépendant de sa description, c'est-à-dire de la manière dont le système voit ce même document. En fait, le document est accessible à l'utilisateur, alors que pour le système documentaire, il se réduit à un ensemble de descripteurs, chaque descripteur étant le représentant d'une qualité particulière (liée au champ dans lequel il apparaît). La représentation d'un document par une suite de descripteurs reste correcte si par *"Terme T_j",* on entend *"Terme T_j présent dans le champ X"*. On peut aussi supposer qu'un terme donné ne peut être présent que dans un seul champ

particulier [HALI89], mais cela semble imposer des contraintes inutiles et difficiles à tenir (par exemple, les mots du titre peuvent aussi être présents dans les descripteurs contrôlés).

Une fois acquise cette notion de description structurée, il subsiste de nombreux problèmes dans l'écriture même des termes d'indexation. Un ensemble de règles strictes s'imposent alors aux indexeurs pour homogénéiser l'emploi de certaines formes des mots. C'est par exemple le choix de règles imposant aux descripteurs d'être écrits au singulier ou au pluriel. Il existe par exemple deux normes françaises d'indexation matière établissant des règles pour rapporter chaque terme à un nombre précis. La norme Z 44-070 [AFN86] impose l'usage du singulier, qui est la forme habituelle des entrées dans les dictionnaires. A l'inverse, la norme Z 47-200 [AFN85] sur les listes d'autorité prône l'utilisation du pluriel. Ces deux règles souffrent bien entendu des exceptions suivant l'usage (la norme Z 44-070 accepte les termes mathématiques au pluriel) ou suivant le sens (la liste d'autorité RAMEAU, qui respecte la norme Z 47-200, accepte néanmoins Conscience ou Enseignement au singulier) [RIC88, p.30-31]. La contradiction entre les normes est un aspect montrant la complexité des choix d'indexation. De plus, l'application de règles strictes peut conduire à des aberrations. PASCAL, banque de données de l'INIST, qui choisit d'indexer au singulier, utilise ainsi le descripteur *Banque donnée,* que bien entendu aucun utilisateur ne pensera à orthographier ainsi. Or le choix d'une écriture des termes d'indexation n'est pas un vain exercice scolaire. Il détermine les capacités à retrouver des documents dans le système. Du moins, les partisans d'une indexation contrôlée clament-ils bien fort ce précepte, destiné à favoriser l'échange de données, toutes indexées sous la même forme. L'utilisateur doit alors apprendre les règles d'indexation pour retrouver les documents, une condition rarement remplie, ce qui explique de nombreux échecs des catalogues de bibliothèques en ligne.

Certains éléments d'information trouvent difficilement une forme normalisée unique. Par exemple la datation dans les banques de données d'objets muséographiques [LEC88]. On trouve dans ce cas plusieurs types de dates : des dates exactes (1789), des intervalles de dates (1852-1864) ou des mentions plus générales (premier quart du XVII^{ème} siècle). Les notations toponymiques (appellation ancienne ou actuelle, utilisation des termes en français ou en langue

d'origine...) ou géographiques (Orne (département) ou Orne (rivière)) posent aussi des problèmes particuliers. Cela renforce la nécessaire distinction entre systèmes documentaires et systèmes de données : même les "données" sont relativement mal définies dans les systèmes documentaires, et c'est au système de gérer ces incertitudes et ambiguïtés.

2.b - Une activité inconsistante

Plus généralement, on peut s'interroger sur le type d'activité à l'œuvre dans l'indexation humaine. Indexer un document, c'est lui associer un certain nombre de descripteurs afin de permettre à un utilisateur de le retrouver. Indexer n'est pas un processus de condensation (offrir une nouvelle forme plus réduite au document), mais un processus de communication. C'est même un processus réversible. En effet, un utilisateur, placé devant la tâche de retrouver un certain document dans le système, doit se poser la question : comment aurais-je indexé ce document ? Une situation difficile, comme l'indique Yves Courrier dans [COU76] *"deux indexeurs choisiront pour le même document très peu de descripteurs identiques, parfois moins de 50 %. Ce sera aussi le cas pour un même indexeur à deux périodes différentes."*.

Gérard Salton, dans [SAL86], reprenant les conclusions d'une étude de Cleverdon pour *l'Agence Spatiale Européenne ,* indique de même :

. si deux personnes ou deux groupes de personnes construisent un thésaurus sur un certain sujet, seulement 60 % des termes d'indexation seront communs aux deux thésaurus.

. si deux indexeurs confirmés indexent un même document avec un thésaurus donné, seuls 30 % des termes d'indexation seront communs entre les deux ensembles de descripteurs affectés.

. si deux intermédiaires réalisent la même recherche sur la même banque de données, seuls 40 % des résultats seront communs aux deux lots extraits.

. si deux scientifiques ou ingénieurs doivent juger la pertinence d'un ensemble de documents, leur zone de recouvrement sera seulement de 60 %.

Une étude massive de Furnas et al. ([FUR83], [FUR87]) reste aussi significative sur la grande variabilité de l'indexation (ou de la nominalisation). Plusieurs domaines très différents sont couverts par l'étude :

. les verbes principaux choisis par des secrétaires pour nommer des commandes de base d'un traitement de texte (effacer, remplacer...)

. le premier mot choisi par 337 étudiants pour décrire une sélection de 50 objets communs (les étudiants devaient choisir une alternative à certains mots qui leur étaient proposés *("love", "motorcycle", "Newsweek"...')*.

. le premier mot choisi par 30 agents immobiliers du New Jersey pour classer 64 annonces.

. Les premiers mots-clés choisis pour indexer 188 recettes de cuisine par 8 experts et 16 novices.

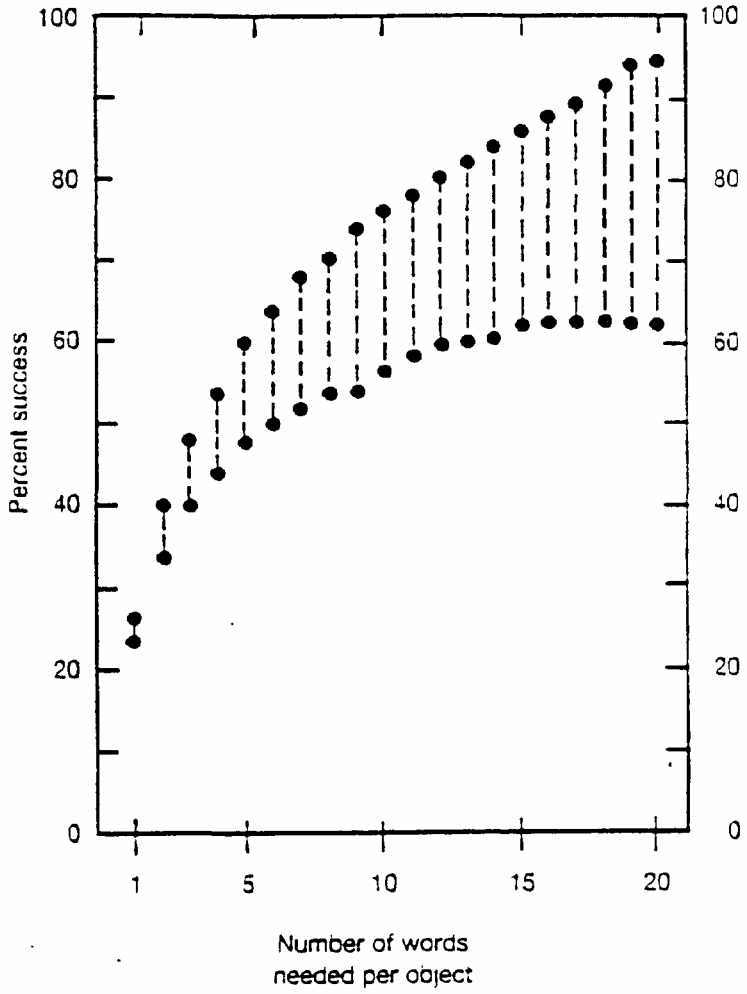
Le tableau suivant définit la probabilité de voir deux personnes appliquer le même terme au même objet :

Traitement de texte	0,07
Objets communs	0,12
Annonces	0,14
Recettes	0,18

L'analyse d'un si faible recouvrement entre les termes conduit à proposer d'utiliser plusieurs équivalents *("aliases")* pour un même objet afin de suivre la variabilité dans les comportements individuels. Le tableau ci-dessous indique l'amélioration du taux de succès d'une recherche (en supposant que le terme de la recherche est celui qui est utilisé par un des individus pour indexer) en fonction du nombre de termes équivalents pris en compte lors de l'indexation.

Outre une grande incertitude statistique, qui explique les différences entre les hypothèses optimistes et pessimistes sur le graphique, on admettra que le nombre de termes doit être relativement élevé pour des chances de succès qui

peuvent dans le pire des cas ne pas dépasser 60 % avec 20 termes équivalents. Or l'exemple de Furnas et al. ne concerne que quelques objets documentaires. Dans les grandes banques de données, multiplier le nombre de termes entraîne très vite un effet de "bruit" documentaire car des termes moins spécifiques peuvent concerner des objets documentaires très différents.



Toutes les autres études qui ont pu être menées concernant les capacités (ou plutôt l'incapacité) des individus à nommer de la même manière les mêmes objets confirment ce type de résultats. On pourra regarder ainsi [BR078], qui réalise une expérience similaire avec des personnes décrivant des objets destinés à tenir lieu de cadeaux. On s'aperçoit que non seulement la variation entre les individus est très grande, mais de plus que les termes affectés par un même individu en deux périodes différentes se recouvrent très peu.

Cette variabilité de l'indexation humaine conduit de nombreuses personnes à se poser le problème de faire réaliser ce travail directement par l'ordinateur.

3 - L'indexation automatique : remarques générales

L'indexation automatique consiste à utiliser les documents tels qu'ils sont entrés dans le système documentaire pour en extraire un jeu de descripteurs utiles pour la recherche ultérieure. Depuis des années des espoirs très importants sont mis dans l'indexation automatique. David Blair [BLA90, p. 49-50] distingue trois éléments qui tendent à favoriser l'indexation automatique par rapport à l'indexation manuelle dans les organisations (entreprises, bibliothèques...) :

. *les difficultés de l'indexation* elle-même, son inconsistance relevée plus haut.

. *les problèmes organisationnels* posés par l'existence d'une industrie d'indexeurs comme cela est nécessaire pour constituer une grande banque de données. Le statut des indexeurs est mal positionné, donc ardu à gérer. Leur tâche demande une large part de connaissances et d'initiative individuelle, mais reste cependant une activité de production.

. *les problèmes de coût :* le salaire des indexeurs est élevé, en fonction des compétences intellectuelles, mais le travail est lent, et crée même parfois (souvent ?) des goulots d'étranglement dans la production d'un système documentaire. Ces coûts sont considérés supérieurs aux coûts machine et donc favorables aux méthodes automatiques.

On peut penser que ces arguments seraient intéressants si les systèmes d'indexation automatique étaient performants au même titre que les systèmes d'indexation humaine. Quand une organisation décide de se doter d'un système documentaire, celui-ci est souvent appelé à devenir de taille importante et à occuper une place déterminante dans la structure de travail de l'entreprise. Par exemple la documentation technique d'une centrale nucléaire comporte plusieurs dizaines de milliers de documents et pour des raisons facilement compréhensibles constitue un enjeu fondamental pour l'entreprise. Une entreprise se doit donc de juger son système documentaire à la qualité des services rendus, ou du moins se doit d'établir la balance entre la qualité du service et son coût. Malheureusement, comme nous l'avons souligné, il n'existe pas de moyen vraiment fiable pour juger

les performances d'un système documentaire. Jouer l'indexation automatique contre l'indexation manuelle est dès lors un pari fondé sur des sentiments, des *a priori.*

On peut aussi considérer l'indexation automatique comme une aide à la productivité des indexeurs, permettant d'obtenir plus rapidement une représentation plus efficace d'un document au moment de son insertion dans le système documentaire. Cet axe de travail nous semble particulièrement productif, car il organise la jonction entre la puissance de calcul de l'ordinateur et la capacité à juger de façon pragmatique (i.e. en ayant des informations qui ne sont pas strictement contenues dans le document) de l'indexeur. Cette méthode a de plus l'avantage de pouvoir s'étendre au delà du texte vers d'autres types de documents qui devraient prendre une place de plus en plus importante (images, graphiques, plans, descriptifs techniques multimédias...). Dans ce cas, l'indexation manuelle reste la base de tout travail documentaire. Elle peut cependant être relayée par des méthodes de regroupement des images qui permettent de peaufiner les descriptions [HALI89].

On peut distinguer deux orientations en indexation automatique :

- se contenter de la forme de surface du document : indexation en *"texte intégral"* ou utilisation de méthodes statistiques sur les formes du texte.

- se baser sur les similitudes entre documents ou sur les réseaux de citations afin de décrire un document par un autre (i.e. faire hériter un document de termes d'indexation produits par ou pour un autre).

- travailler sur le sens du document et sur les réseaux d'association sémantique entre les termes d'indexation choisis.

On peut concevoir des méthodes d'indexation automatique qui utilisent des techniques élaborées pour chacun de ces aspects. Par exemple en mêlant des méthodes statistiques travaillant sur l'état de surface du document et des thésaurus ou des réseaux sémantiques travaillant sur le sens. De même, si l'indexation manuelle se réduit à l'affectation de mots-clés à des documents,

l'indexation automatique, grâce à la puissance de calcul des ordinateurs actuels peut jouer en plus sur la pondération des informations contenues dans les descripteurs. Cette utilisation de la pondération permet de mieux adapter les représentations aux besoins documentaires, mais aussi de tenir compte, dans l'indexation de chaque document particulier, d'éléments permettant de situer ce document relativement à tous les autres au sein de la banque de données considérée.

Un des objectifs de l'indexation est d'obtenir une normalisation des termes destinés à représenter les documents. C'est par exemple le but des règles d'indexation manuelle, ou des listes de descripteurs contrôlés. Dans le cas de l'indexation automatique, cette normalisation du vocabulaire passe par deux méthodes [PUJ89] :

. le système dispose d'un dictionnaire rassemblant toutes les formes possibles des entités lexicales. Celui-ci peut devenir très important : le système SPIRIT ([DEB89], [RAD88]) comporte ainsi un dictionnaire morphologique de plus de 450 000 entrées en français.

. le lexique ne renferme que les formes canoniques des termes (comme un dictionnaire classique) auxquelles on associe des règles de flexion. Le mot est reconnu à partir de sa racine et d'une terminaison possible, connue par la règle attachée à la forme de base conservée dans le lexique. Cette méthode est décrite en détail dans [LAI82, p.70-71]. Par exemple, la forme ROUGIE sera décomposée en

- ROUG	Base verbale
- I	marque du participe passé
- E	désinence (marque du féminin)

La liste des formes de base du dictionnaire sera construite en fonction des termes présents dans l'échantillon linguistique considéré. Ainsi, l'analyse de la forme ASSOCIATIONS se fera directement à partir de la racine ASSOCIATION et de la marque S du pluriel dans le cas où les seules formes présentes dans le texte (ou la banque de données) seraient ASSOCIATION et ASSOCIATIONS. Si en revanche des formes verbales dérivées du verbe ASSOCIER sont aussi présentes (ASSOCIE, ASSOCIAIENT...), le même terme ASSOCIATIONS sera rapporté à

la base la plus générale (ASSOCI), complétée par le suffixe ATION et la marque S du pluriel.

On peut penser que les deux méthodes de normalisation des termes ne sont pas incompatibles. La méthode basée sur l'existence d'un large dictionnaire comportant toutes les formes est rendue crédible par l'expérience. En effet, les lexiques de base des grandes banques de données en texte intégral, qui permettent des recherches très rapides, par exemple dans les dépêches de 1' AFP par le logiciel *Questel+,* sont très proches de ces grands dictionnaires. On peut penser que l'on augmenterait le taux de couverture des recherches en normalisant tous les termes, c'est-à-dire en renvoyant chaque forme présente dans le texte à une forme canonique déterminée pour l'indexation [RAD88]. Cette normalisation devra employer des méthodes d'analyse morphologique basées sur les diverses flexions à partir de formes de base, au moins pour établir la première projection du dictionnaire des formes sur le dictionnaire des termes choisis, même si l'analyse morphologique des mots n'est pas répétée à chaque fois, pour l'indexation de nouveaux documents.

Ces méthodes de normalisation ne sont cependant pas exemptes de problèmes. Elles sont utiles pour augmenter le taux de couverture, mais peuvent nuire à la précision. Ce n'est pas toujours un choix judicieux, notamment dans les très grandes banques de données. On trouve souvent dans la littérature spécialisée anglo-saxonne une méthode de normalisation lexicale par suppression des suffixes (réduire les mots à leur radical). Cette méthode peut engendrer un bruit documentaire très important. Quel est l'intérêt de réduire COMMUNISME à COMMUN ? ANALYSE et ANALYTIQUE à ANALY ? Il faut donc se garder d'appliquer une méthode de normalisation trop mécanique, qui ne tiendrait pas compte de certaines règles sémantiques. C'est au niveau de cette supervision du processus de normalisation que l'emploi d'un large dictionnaire des formes peut s'avérer utile, notamment si chaque normalisation d'un mot nouveau doit être validée par un administrateur du dictionnaire. Ces critiques à la méthode d'élimination automatique des suffixes sont confortées par des expériences menées sur des banques de données de taille importante qui sont beaucoup plus sensibles au bruit documentaire [HAR90, p. 108].

4 - L'indexation en texte intégral

L'indexation en texte intégral est certainement la méthode d'indexation automatique la plus répandue. Elle s'applique dans de très nombreuses banques de données diffusées dans les systèmes commerciaux. Les dépêches des agences de presse, les versions électroniques des journaux et des revues, les lettres d'information, les textes de loi et la jurisprudence, les documents bureautiques... sont traités en texte intégral pour être mis rapidement en diffusion. Les grands serveurs adoptent même depuis quelques temps la mise en service en continu de nouvelles informations textuelles [IWR90b] en utilisant des algorithmes performants de mise à jour des index.

L'indexation en texte intégral considère que tous les mots non grammaticaux d'un texte sont les meilleurs descripteurs de ce texte. Les descripteurs du document sont alors tous les mots lexicaux (noms, verbes, adjectifs...) présents dans le texte. Ils apparaissent sous la forme qu'ils revêtent dans ce texte précis, sans aucun processus de normalisation.

Pour indexer en texte intégral, le système repère d'abord chaque mot du texte. Après élimination des signes de ponctuation, un mot est la chaîne de caractères comprise entre deux espaces. Pour l'indexation, cette chaîne de caractères est transformée en lettres majuscules, à l'exception de certaines banques de données spécialisées pour lesquelles la présence des signes diacritiques est déterminante (par exemple FRANTEX, banque de données de textes littéraires français, utilisée pour des travaux linguistiques). Un anti-dictionnaire permet d'éliminer les termes ayant peu de valeur documentaire, notamment les mots grammaticaux. Jacques Virbel [VIR88] précise qu'en *"français, les mots outils (articles, pronoms, conjonctions, prépositions...) représentent 50 % de n'importe quel texte, l'autre moitié étant constituée par des mots pleins (substantifs, verbes, adjectifs, adverbes). On peut noter que dans le dictionnaire cette proportion est tout autre, les mots outils ne représentant que 0,5 % du lexique total"*. On comprend dès lors la nécessité d'alléger les index de moitié en ne traitant qu'un infime pourcentage des mots de la langue dans l'anti-dictionnaire.

Cette hypothèse d'élimination des mots outils est liée avec la restriction des

descripteurs aux unitermes qui sont les seuls termes versés dans l'index. Les mots outils sont lourds de sens dans la langue en ce qu'ils permettent d'articuler les mots dans des expressions au sens différent de la simple juxtaposition des unitermes *("coffre à bois', "coffre en bois").* L'anti-dictionnaire comporte aussi parfois des substantifs ayant un sens trop large (analyse, transformation, essai...). Il faut cependant faire attention à ne pas trop allonger la liste des mots éliminés. *"Analyse"* sera inutile dans une bibliothèque de lecture publique car équivalent à *"le sujet de ce livre est ...*", mais restera important en mathématiques.

On peut repérer plusieurs avantages à l'indexation en texte intégral. Elle est rapide, surtout quand elle s'appuie sur des documents existant déjà en version électronique (documents bureautiques, articles de presse récupérés à partir des bandes de photocomposition, dépêches d'agence produites directement sous forme électronique...) ou dont la saisie est effectuée par des personnels peu qualifiés (textes de lois saisis d'après les journaux officiels dans les usines à dactylographier d'Asie du Sud-est). Elle est peu coûteuse. En s'appuyant sur les textes eux-mêmes, elle permet de proposer divers équivalents des termes (synonymes, formes dérivées...) et diverses formes de leur organisation (ordre de succession). Un auteur de document ayant généralement utilisé les expressions propres à la spécialité de ce document (le jargon), l'indexation en texte intégral permettra une meilleure prise en compte de l'évolution du vocabulaire.

Mais elle est aussi lourde de limites. L'indexation en texte intégral fait porter sur l'utilisateur le fardeau de la description documentaire. S'il désire obtenir un bon taux de couverture pour sa recherche, l'utilisateur doit imaginer toutes les formes possibles que l'expression qu'il recherche pourrait prendre dans un texte. Or les termes de vocabulaire et les constructions de phrases utilisés pour désigner un même sujet sont très nombreux. Devant ce large spectre d'expression, il est impossible de deviner les termes et les expressions adéquats. Ethel Méaudre [MEAU88] cite ainsi les diverses formes prise dans la jurisprudence en matière de divorce pour indiquer qu'une femme "entretient mal son foyer" : *"mauvaise tenue du ménage", "ménage négligé", négligences ménagères", "négligence de la femme dans son intérieur".* Retrouver toutes ces formes tient de l'exploit sportif. D'autant que chacun des unitermes pris séparément ne peut guère aider à obtenir le meilleur taux de couverture.

"Ménage" peut recouvrir bien d'autres éléments que la tenue du domicile *("Femme de ménage", "ménage des bureaux"...),* et *"négligence"* reste encore plus vague et ambigu. En fait, on peut statuer que *"la variabilité dans les mots et dans les expressions que les auteurs des documents contenus dans un système utilisent pour parler du même sujet est extraordinaire et imprévisible"* [BLA90, p. 109].

L'indexation en texte intégral tend à disperser le vocabulaire. De ce point de vue, la grande facilité du langage à utiliser plusieurs synonymes ou termes équivalents pour le même sujet permet à un utilisateur d'obtenir fréquemment *"au moins"* quelques documents intéressant son sujet. Il s'agit là de l'opération inverse de celle qui est proposée plus haut : à partir de n'importe quel terme de l'utilisateur, on retrouvera un document l'utilisant. C'est cette capacité à fournir des réponses à presque toutes les questions qui fait le succès des systèmes en texte intégral. Mais trouver des documents ne signifie pas avoir trouvé tous les documents, ni même les documents les plus intéressants. Dans une expérience de grande envergure menée par Blair et Maron, et depuis largement discutée dans la communauté des chercheurs en sciences de l'information ([BLA85], BLA90], [BLA90b], [SAL86]), des utilisateurs pensaient avoir obtenu environ 75 % des documents pertinents contenus dans un système documentaire juridique utilisant le logiciel STAIRS, alors qu'une recherche plus poussée montrait qu'ils avaient extrait seulement 20 % des documents répondant à leur requête.

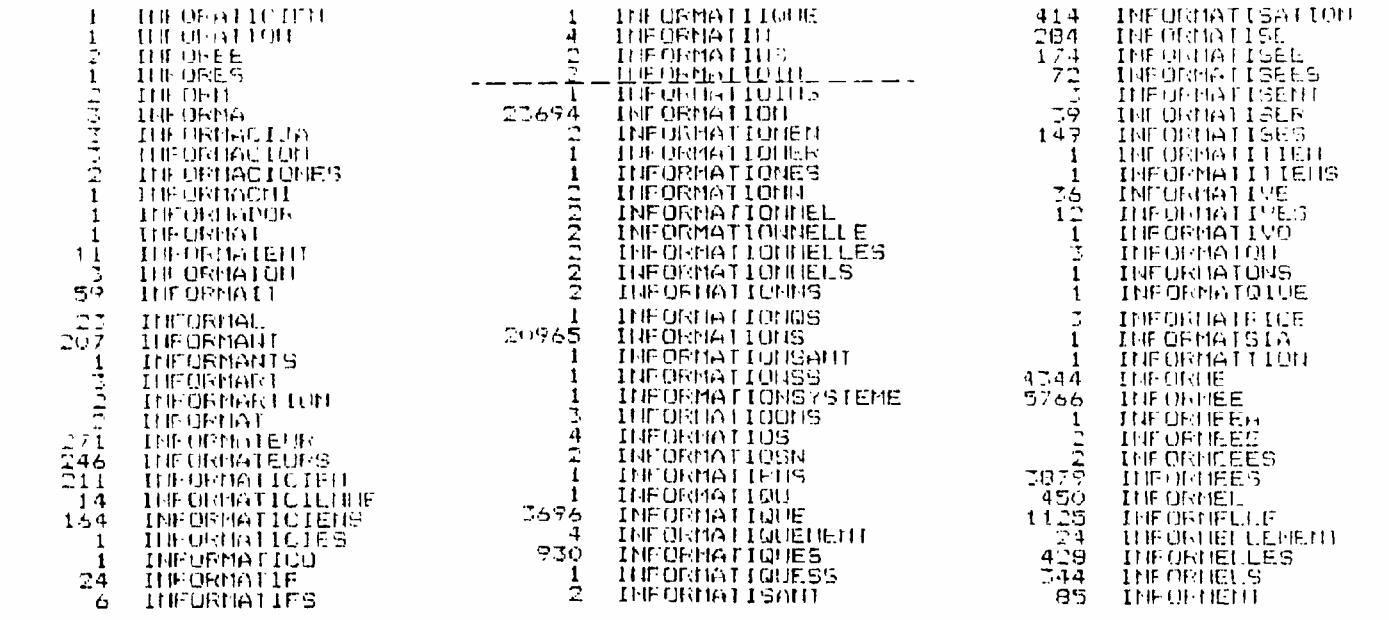
L'utilisateur d'un système documentaire basé sur l'indexation en texte intégral est confronté à une gageure : trouver tous les termes ou toutes les combinaisons de termes qui peuvent couvrir l'objet de sa recherche, mais surtout se limiter à ceux qui ne concernent que sa recherche, à l'exclusion de tout autre domaine. Il lui faut donc éliminer les textes qui contiennent des combinaisons des termes choisis mais ne concernent pas le sujet désiré. Cette capacité varie énormément avec la taille de la banque de données considérée ou suivant le type de question. La recherche sur une banque de données indexée en texte intégral sera d'autant plus performante que le sujet traité correspondra à des mots précis, inévitables, et qui ne sont pas contenus dans un trop grand nombre de documents. Ce type de recherche correspond par exemple au titre d'un livre ou d'un film dans une banque de données de presse, ou bien au nom d'une personne (s'il est peu fréquemment utilisé) ou d'un lieu géographique (encore qu'il sera plus facile de

rechercher Pyongyang que Paris dans la banque de données des dépêches de l'AFP). En revanche, si le sujet s'exprime par des termes plus généraux, le danger de bruit documentaire devient très important.

L'indexation en texte intégral part du postulat que chaque texte contient les informations utiles à sa compréhension. Or, nous avons souligné plus haut combien le poids des connaissances pragmatiques est fondamental dans la compréhension des phénomènes linguistiques. Il est de nombreux cas où l'on ne peut déduire de quelques mots d'un texte des informations sur le domaine traité par le document, sur son orientation. En sens inverse, de nombreux textes partent de l'hypothèse que le destinataire (lecteur d'un journal, récipiendaire d'une lettre, lecteur d'un manuel technique...) sait de quoi parle le texte. Les anglo-saxons distinguent ainsi *"meaning"* qui représente le contenu du texte de *"about"* ou *"aboutness"* qui caractérise le cadre général de ce dont parle le texte.

En plus de s'appuyer sur des bases théoriques discutables (les aspects de contexte ne sont justement pas souvent rappelés dans le texte lui-même), l'indexation en texte intégral, basée sur les formes de surface des textes hérite de toutes les imperfections.

Si l'on examine le lexique de la banque de données AGRA, des dépêches générales de l'AFP, à la date du premier juin 1990, on peut pointer quelques problèmes soulevés par l'absence de normalisation des termes dans l'indexation en texte intégral.



On constate par exemple que le concept d'informatisation est représenté par les termes suivants :

INFORMATISANT	2 doc.
INFORMATISATION	414 doc.
INFORMATISE	284 doc.
INFORMATISEE	174 doc.
INFORMATISEES	72 doc.
INFORMATISENT	3 doc.
INFORMATISER	39 doc.
INFORMATISES	145 doc.

Encore avons-nous la chance que la recherche utilisant une troncature INFORMATIS+, qui considère tous les termes dont la terminaison s'appuie sur la chaîne de caractère INFORMATIS, soient tous relatifs au verbe informatiser sans entraîner de bruit supplémentaire. L'utilisation des troncatures est souvent plus hasardeuse, et de toute façon rarement comprise par un utilisateur qui n'a pas reçu une formation spécifique.

Dans ce même lexique on trouve 120 termes entre INFOR et INFORUM, dont 44 termes sont visiblement des fautes de frappe (INFORATICIEN, INFORATION, INFORMATIOINS, INFORMATIONN...), soit un encombrement du lexique de 37 %, sans compter les termes très rares que l'on ne sait s'il faut considérer comme des fautes de frappe, des noms propres (INFORMATIN et INFORMATINS, INFORMART, INFORUM, INFORPAZ) ou des mots étrangers (INFORMACNI).

On peut conclure ce chapitre concernant l'indexation en texte intégral en constatant que si l'indexation manuelle est inconsistante et variable, l'indexation en texte intégral ne vient pas combler cette lacune. Dans les deux cas, le succès des recherches dépend de la capacité d'un utilisateur à prendre une décision (choix des termes de sa question, jugement des documents extraits, décision d'arrêter la recherche) dans un environnement peu fiable. Les difficultés de ces choix sont encadrées par deux éléments :

. la taille des banques de données qui permet d'obtenir souvent *"au moins"* quelques résultats satisfaisants

. et la confusion qui existe chez l'utilisateur entre le taux de

couverture d'une recherche (a-t-on extrait tous les documents pertinents ?) et le taux de précision (la part des documents pertinents parmi ceux qui sont extraits). Le succès des banques de données en texte intégral est à chercher dans ces phénomènes subjectifs.

Si l'on veut vraiment concurrencer l'indexation manuelle du point de vue de la qualité de la description, on doit donc envisager des méthodes d'indexation automatique plus perfectionnées. En effet, nombre des limites de l'indexation en texte intégral peuvent être dépassées par l'application conjointe d'autres techniques d'indexation automatique qui s'appuient elles aussi sur les formes de surface des textes.

5 • L'indexation par des méthodes statistiques

L'indexation en texte intégral, mais aussi l'indexation libre (indexation manuelle réalisée sans vocabulaire contrôlé), accordent la même importance à tous les termes. Un descripteur est présent ou non. On ne peut pas préciser qu'un document traite d'un sujet dans une certaine mesure (principalement, accessoirement, marginalement...), ni que le terme d'indexation choisi n'est pas spécifique de ce seul document, mais est affecté à de nombreux documents dans la banque de données.

Cette limite découle du fait que l'indexation est centrée sur le document, sans prendre en considération la banque de données dans laquelle il doit prendre place. L'utilité des termes d'indexation n'est pourtant pas une simple fonction du document. Elle varie fortement selon le système documentaire, plus précisément selon le nombre de documents et la dimension du domaine couvert. Un terme peut-être spécifique dans un système (par exemple POLOGNE dans une bibliothèque municipale française) mais trop général dans un autre (POLOGNE dans un système documentaire qui regrouperait les documents en français traitant de ce pays).

Pour résoudre cette contradiction, les méthodes statistiques partent de deux principes [SAL81] :

1. il existe une relation entre la fréquence d'un terme à l'intérieur

d'un document (ou de ses parties significatives - titre, résumé,...) et son importance pour la représentation de ce document

2. il existe une relation inversement proportionnelle entre l'importance d'un terme pour l'indexation d'un document, et le nombre total de documents contenant ce terme dans l'ensemble de la banque de données. Un terme rare, décrivant peu de documents sera privilégié par rapport à un ternie général se retrouvant dans un grand nombre de documents.

5.a - La pondération des descripteurs

Ces deux règles se traduisent par l'adoption d'un facteur d'importance interne au document (dit *tf,* soit *term frequency)* et d'un facteur d'importance externe (dit *idf, inverse document frequency).* Ces deux facteurs conduisent à pondérer les termes d'indexation.

Pour un terme donné T_j et un document D_i on définit :

. tf_{ij} = Fréquence du terme T_j dans le document D_i

. idf_j= 1/Nombre de documents possédant le terme T_j

Cela permet de définir un poids w_{ij} du terme T_j pour le document D_i

. $W_{ij} = tf_{ij} * idf_j$

La représentation d'un document par ses *n* termes d'indexation devient ainsi :

$D_i := <T_l,w_{il}; T_2'w_{i2} ; ...T_n,w_{in}>$

Il n'est pas toujours facile de connaître le nombre occurrences d'un terme à l'intérieur d'un document. On peut alors se contenter de la règle de l'inverse de la fréquence dans l'ensemble de la banque de données qui constitue une bonne approximation. Un terme est d'autant plus utile qu'il correspond à peu de documents dans la banque de données. Il existe plusieurs variantes permettant de calculer la fonction idf_j. L'utilisation d'une fonction logarithmique permet d'accentuer les différences de poids pour les termes de faible fréquence et tend à assimiler le poids des termes de fréquence élevée. En faisant intervenir la valeur N du nombre total de documents contenus dans la banque de données on peut

obtenir une fonction idf du type : $idf_i = - \log \frac{n}{N}$

Si le nombre de documents contenus dans la banque de données est très élevé, bien plus important que la fréquence du terme le plus utilisé, on considérera que N est alors égal à la fréquence du terme le plus élevé.

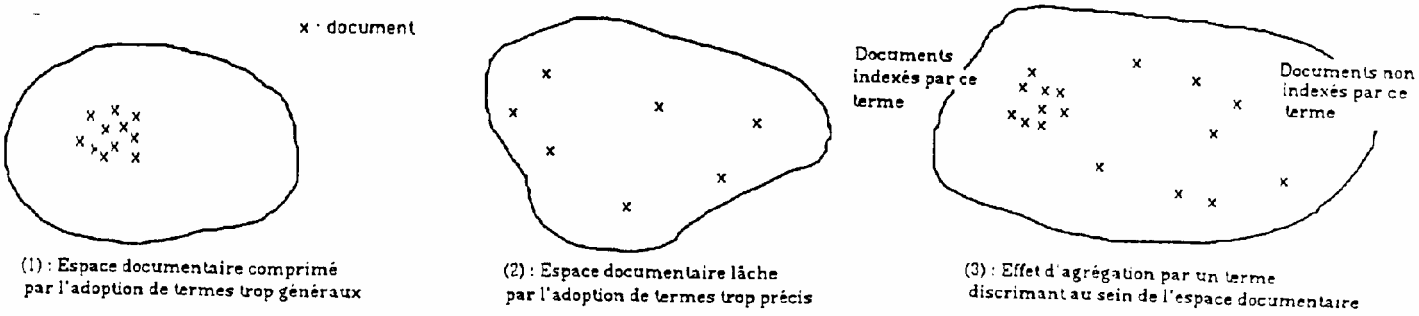
5.b - Valeur de discrimination des descripteurs

Cette approche d'une pondération marquée par la fonction *idf ne* tient pas compte du faible pouvoir de recherche d'un terme très rare. Celui-ci se trouve sur valorisé, alors que sa probabilité d'être utilisé comme question est faible, au moins aussi faible que sa probabilité d'être présent parmi les documents. Il convient alors de définir la qualité de discrimination d'un terme dans une banque de données ([SAL76], [SAL83], [BIR89]).

- Un terme très fréquent tend à agglomérer les documents autour de lui (i.e. de nombreux documents sont extraits par ce terme). Il est utile pour améliorer le taux de couverture, mais perd en spécificité.

- Un terme très rare est affecté à peu de documents et tend à rejeter les documents qu'il décrit lors d'une recherche effectuée avec des termes moins précis. Il est utile pour améliorer la précision d'une recherche, mais diminue les capacités à retrouver le document.

- Les termes de fréquence moyenne tendent à regrouper une partie des documents qui sont relativement semblables.



(1) : Espace documentaire comprimé par l'adoption de termes trop généraux

(2) : Espace documentaire lâche par l'adoption de termes trop précis

(3) : Effet d'agrégation par un terme discrimant au sein de l'espace documentaire

On peut définir dans l'espace documentaire ED une *densité,* qui est fonction de la similitude entre les divers documents de la banque de données. Une faible densité correspond à des documents épars, donc décrits par des termes d'indexation à faible fréquence, alors qu'une forte densité correspond à des documents rapprochés, donc à une indexation peu discriminante par des termes généraux. Le calcul de cette densité est donné par [CAN87].

Soit la matrice M de taille *m* * *n* constituée de *n* lignes déterminées par les *n* termes d'indexation utilisés dans l'ensemble de la banque de données et *m* colonnes correspondant aux *m* documents présents dans le système, chaque élément wj: de la matrice est donc est le poids du terme T: pour le document Dj.

	D_1	D_2	...	D_k	...	D_m
T_1	W_{11}	W_{21}		W_{k1}		W_{m1}
T_2	W_{12}	W_{22}		W_{k2}		W_{m2}
T_1	W_{1i}	WE		W_{ki}		W_{mi}
T_n	W1n	W_{2n}		Wkn		W_{mn}

La similitude entre deux documents $S(D_i\ D_j)$ est définie par une mesure, en général le cosinus de l'angle constitué par les vecteurs D_i et D_j dans l'espace vectoriel à *n* dimension constitué par les *n* termes d'indexation. Soit :

$$S(D_i, D_j) = \frac{\sum_{k=1}^{n} w_{ik} \cdot w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2 \cdot \sum_{k=1}^{n} w_{jk}^2}}$$

La densité Q de l'espace documentaire ED est la moyenne de la similitude calculée entre tous les documents pris deux à deux :

$$Q = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \left(\sum_{j=i+1}^{m} S(D_i, D_j) \right)$$

On peut obtenir un résultat semblable pour un temps de calcul plus faible en définissant un barycentre de l'ensemble des documents

$$G := < T_j g_j \text{ tel que } g_j = \frac{1}{m} \sum_{i=1}^{m} w_{ij} >$$

G serait un "document" dont le poids de chacun des termes d'indexation serait la moyenne des poids affectés pour l'ensemble des documents à chaque terme d'indexation.

La mesure de densité de l'espace documentaire se réduit alors à la moyenne de la similitude entre chaque document et ce barycentre :

$$Q = \frac{1}{m} \sum_{i=1}^{m} S(D_i, G)$$

La densité de l'espace documentaire varie en fonction de l'ajout ou du retrait d'un terme d'indexation. L'activité d'un terme d'indexation dans la banque de données dépend de la variation que son retrait produit sur la densité. On calcule ainsi une nouvelle densité Q_j après retrait du terme T_j :

$$Q_j = \frac{1}{m} \sum_{i=1}^{m} S(D_i^j, G_j)$$

où : D_i^j est le document D_i sans le terme T_j

$$D_i^j - = < T_1, w_{i1} : T_2, w_{i2} : .... : T_{j-1}.w_{i(j-1)} : T_{J+1}, W_{I(J+1)} : .... : T_n, w_{in} >$$

G_j est le barycentre après élimination du terme T_j

$$G_j = < T_1, g_1 : T_2, g_2 : ... T_{j-1}, g_{j-1} : T_{j+1}, g_{j+1} : ... : T_n, g_n >$$

La différence Qj - Q marque les changements provoqués par le retrait du terme T_j. On l'appelle la valeur de discrimination du terme T_j :
$VDT_j = Q_j - Q$

Cette valeur de discrimination a trois propriétés :

. si $VDT_j > 0$ le terme T_j est très discriminant, c'est-à-dire que sa présence fait diminuer la densité, tendant à éloigner les documents les uns des autres.

. si $VDT_j \approx 0$ le terme est indifférent, équitablement réparti.

. si $VDT_j < 0$ le terme T_j est peu discriminant. C'est un terme général qui est affecté à beaucoup de documents.

Ce calcul permet de mieux connaître l'activité de chacun des termes d'indexation dans la banque de données. On peut aussi s'en servir pour moduler les poids des différents termes d'indexation affectés à chaque document D_i en fonction de la valeur discriminante de chaque terme.

La description du document D_i par ses termes d'indexation devient :

$D_i := < T_l,(w_{il} * TDV_l) ;.... T_j(w_{ij} * TDV_j) ; ... T_n,(w_{in} * TDV_n) >$

Connaître la valeur de discrimination donne des indications pour recomposer les termes d'indexation selon deux méthodes :

- grouper les termes de faible fréquence qui sont très discriminants ($VDT_j > 0$) en termes ayant une valeur de discrimination proche de zéro. Pour Salton [SAL88], cette opération est l'équivalent statistique de la construction d'un thésaurus. On peut définir les termes à regrouper selon plusieurs manières : soit intellectuellement à partir de listes classées selon VDT_j, soit statistiquement en effectuant des regroupements locaux (i.e. dans le même champ sémantique défini par les termes généraux qui regroupent les documents malgré l'existence des termes très discriminants).

- constituer des descripteurs composés à partir des termes ayant une faible valeur discriminante (termes trop généraux dont $VDT_j < 0$). Un calcul de co-occurrence des termes permet de prévoir ces regroupements, même si la liste des termes composés doit être revue pour assurer la cohérence.

En amont de l'ensemble de ces méthodes, Salton [SAL81] propose d'éliminer les suffixes pour ne conserver que les racines des termes. Nous avons indiqué plus haut les limites de cette méthode.

5.c - Un guide pour l'indexation statistique

Gérard Salton [SAL81] propose une méthode statistique pour l'indexation automatique :

1. Identifier les mots dans le document

2. Utiliser une liste de mots-outils pour éliminer les termes fonctionnels.

3. Etablir une normalisation lexicale. La méthode la plus simple, mais parfois inefficace, est l'élimination automatique des suffixes. On peut aussi envisager d'utiliser un grand dictionnaire rapportant chaque terme à une forme normalisée choisie.

4. Calculer la valeur de discrimination de chaque terme d'indexation VDT_j pour $j=1$ à n, n étant le nombre total de termes d'indexation dans la banque de données

5. Grouper les termes de faible fréquence avec $VDT_j > 0$. On peut utiliser un thésaurus pour cette opération

6. Constituer des descripteurs composés à partir des termes peu discriminants ($VDT_j < 0$)

7. Calculer le poids de chaque terme d'indexation Tj dans le document Di par : $w_{ij} = (tf_j * idf_j)*VDT_j$

8. Indexer chaque document avec les termes originaux, les termes de groupage et les descripteurs composés, chaque descripteur étant affecté de son poids.

Cette méthode est cohérente, mais on peut toutefois se demander si elle n'est pas trop gourmande en temps de calcul pour pouvoir être effectuée sur une grande banque de données de plusieurs centaines de milliers de documents. Elle a de plus l'inconvénient d'être statique et non dynamique, c'est-à-dire d'entraîner la répétition de l'ensemble des calculs à chaque introduction de nouveaux documents ou de nouveaux termes dans la banque de données.

6 - L'indexation par les citations

A la suite de Kwok [KW085], on peut considérer qu'un document est écrit pour répondre à une *"question mentale"* Q_i. L'auteur répond à cette *"question mentale"* qu'il se pose à lui-même de deux manières :

. en rédigeant son texte (apport personnel)

. en citant d'autres documents qui sont relatifs à cette question.

En proposant des références à d'autres articles l'auteur effectue une forme d'indexation du document par d'autres documents qui sont pertinents pour cette question Q_i.

Dans cette hypothèse, un document D_i est indexé par :

- un ensemble de termes d'indexation T_j
- un ensemble de documents indexants (les documents cités) C_i.

La similitude entre deux documents peut donc être calculée à partir des termes d'indexation (cf. supra) mais aussi à partir des relations de co-citation. Les citations influent sur les formes d'agrégation des documents dans la banque de données au même titre que les termes d'indexation. On peut distinguer trois types de co-citation :

- un document D_i citant un document D_j
- les documents D_i et D_j sont cités conjointement par un troisième
- deux documents D_i et D_j citant un même troisième document.

Les limites de cette méthode tiennent aux difficultés propres à l'analyse des citations qui ont été soulignées dans le chapitre concernant la bibliométrie. Il faut souligner aussi que l'indexation par les citations n'est exploitable que dans le cas des articles scientifiques, des brevets (brevets cités) ou des objets muséographiques (présence dans les mêmes catalogues d'exposition). Les articles généralistes, les dépêches d'agence, les images...échappent à ce domaine : même si la citation (hommage) existe, elle n'est pas clairement revendiquée.

Les limites de l'analyse des citations ont des conséquences sociales et politiques importantes si l'on accorde une trop grande confiance à ce type d'indicateur. Ces mêmes limites ont toutefois une portée moindre si l'on considère les citations comme un élément d'une technique d'indexation des documents.

Nous avons vu plus haut que l'affectation de mots-clés à un document est, elle aussi, une activité aléatoire, souffrant de biais importants, généralement inconsistante, fortement dépendante des compétences personnelles des indexeurs... En ce sens, l'indexation par les citations ne fait que reprendre sous une autre forme les faiblesses générales de toute activité d'indexation. Les citations deviennent alors un indicateur du positionnement d'un document dans l'ensemble documentaire considéré par le système. Les citations prises en compte permettent de situer des liens entre les divers documents et de faire apparaître des regroupements entre documents, qui ont une probabilité importante de fonctionner en délimitant des secteurs d'intérêt équivalent en regard d'une même recherche.

Cette utilisation des citations pour agréger des documents en vue d'améliorer le taux de couverture lors des recherches n'est cependant pas exempte de limites. Par exemple, les articles fortement cités tendent à agréger des documents qui par de nombreux autres aspects sont différents. Ainsi, l'immense majorité des articles de Sciences de l'Information citent les ouvrages de base de Salton *(Modem Information Retrieval)* et de Van Rijsbergen *(Information retrieval),* même si les préoccupations des articles s'éloignent fortement du point de vue d'une recherche particulière. L'utilisation des citations comme élément d'indexation doit, à l'instar de l'étude d'une valeur discriminante des termes d'indexation, être pondérée à la fois par l'inverse de la fréquence de citation et par l'effet d'une citation particulière sur la densité de l'espace documentaire. Or l'affectation de citations est plus dispersée que celle de termes d'indexation (le nombre moyen de citations est en général supérieur au nombre moyen de termes d'indexation). Cela peut entraîner des temps de calcul trop importants, la dimension *n* de la matrice correspondant aux termes d'indexation (ici à l'ensemble des citations possibles dans un domaine donné) devenant très grande.

On peut aussi contester l'utilisation des citations comme termes d'indexation en faisant remarquer que chaque citation ne se rapporte pas à l'ensemble du texte citant, mais à un paragraphe particulier, et qu'en sens contraire, la citation ne concerne parfois qu'une partie du document cité. Le lien de citation s'apparente donc à une structuration de type hypertexte de l'information : le nœud d'appel est une partie du document citant, et le nœud appelé est une partie d'un autre document cité. Or l'indexation telle qu'elle est

envisagée dans ce chapitre concerne une autre forme de structuration des informations où un document est retrouvé pour lui-même, considéré comme un bloc d'information, enjeu de la recherche documentaire. Comme nous le verrons plus loin, on peut imaginer et souhaiter des systèmes qui permettent de lire une citation à partir d'un document citant (cf. le projet *Xanadu).* Mais on ne doit pas imposer ce lien comme déterminant dans le positionnement d'un document dans l'ensemble documentaire. La pratique quotidienne de lecture des documents scientifiques permet de concevoir cette remarque : un lecteur ne demande pas à voir tous les documents cités par un article qui est pertinent pour sa recherche, mais au contraire sélectionne ceux qui correspondent à l'objectif qui était le sien en lisant tel ou tel document.

Un apport particulier de l'indexation par les citations est aussi à souligner. On peut utiliser les termes d'indexation des documents cités, ou au moins les mots du titre de ces documents cités, pour établir une indexation du document citant par des descripteurs textuels. C'est la méthode dite *KeyWords Plus* de la banque de données *Scisearch,* dans sa version sur disquettes [GAR90]. Les occurrences de termes dans les titres des articles cités, en suivant les règles générales repérées par les techniques bibliométriques permettent d'établir un *"cœur"* concentrant l'information sur le document citant et d'une *"traîne"* indiquant les spécialisations. L'indexation de l'article citant se trouve ainsi renforcée par les termes apparaissant plusieurs fois dans les titres (ou éventuellement de manière plus générale dans les termes d'indexation) des documents cités. Eugène Garfield donne ainsi l'exemple de l'article *"Environmental effects of air pollution in Britain"* de S.J. Woodin [W0089] que l'application de la méthode *KeyWord Plus* permet de doter des descripteurs : Nitrogen déposition, soil acidification, forest décline, acid rain, growth, végétation, streams, sulfur, qui n'apparaissent pas directement dans le titre lui-même. Cette méthode est rendue réaliste car la banque de données *Scisearch,* qui permet de retrouver les articles cités, couvre la littérature scientifique depuis de nombreuses années, et possède donc les informations nécessaires pour établir les fréquences des mots dans les titres des articles cités.

7 - Les techniques d'agrégation des documents

L'hypothèse qui justifie les techniques d'agrégation *(clustering}* considère que des documents ayant des descriptions proches sont susceptibles d'être pertinents de manière conjointe pour une même requête. Cette notion intuitive n'est pas sans une longue tradition expérimentale. Ainsi les bibliothèques, depuis l'utilisation de classifications et le développement du libre accès, en permettant à l'utilisateur de butiner autour de son thème d'intérêt, ont-elles réalisé une forme d'agrégation de documents. Des livres ayant un contenu semblable sont placés en des endroits proches, ce qui tend à favoriser le taux de couverture dans une recherche sur les rayons.

7.a • Hypothèses de travail

Une fois les documents d'une banque de données regroupés en agrégats, on peut distinguer plusieurs types d'utilisation de ce travail lors de la recherche documentaire :

- chaque agrégat est représenté par son barycentre, document fictif synthétisant les divers documents regroupés. La recherche et le classement de pertinence formelle se fait d'abord au niveau des agrégats. Cette méthode tend à favoriser le taux de couverture, mais peut être préjudiciable à la précision.

- les agrégats sont utilisés en complément d'une recherche précise portant sur des documents, notamment après un jugement de pertinence, pour définir des liens dynamiques entre les documents. La sélection d'un document conduit le système à proposer à l'utilisateur les autres membres de son agrégat. Un document peut toutefois appartenir à plusieurs agrégats, et on se retrouve alors devant une hypothèse de navigation (vers quel agrégat le système doit-il se diriger ?). Cette méthode est fortement liée au concept d'hypertexte dynamique.

- les agrégats permettent de définir une classification propre à la banque de données à partir des documents déjà présents dans le système documentaire. On peut ainsi définir une technique d'agrégation hiérarchique, qui regroupe dans une deuxième phase les différents agrégats selon une arborescence, dont on peut définir la hauteur (mais sans maîtriser la taille de

chaque agrégat) ou au contraire le grain (chaque agrégat est de taille limitée mais la hauteur de la classification est directement dépendante du nombre total de documents insérés dans la banque de données).

Globalement, les méthodes d'agrégation tendent à favoriser le taux de couverture des recherches, souvent aux dépens de la précision. On distingue deux types de méthodes d'agrégation :

. les méthodes hiérarchiques
. les méthodes non hiérarchiques ou itératives.

Les méthodes d'agrégation hiérarchique consistent en $h+1$ opérations de regroupement pour passer de l'ensemble disjoint de tous les documents C_0 à la racine d'un arbre comportant toute la banque de données C_h . Les méthodes non hiérarchiques divisent l'ensemble de documents en une partition d'un seul niveau, éventuellement avec recouvrement des agrégats (un document appartenant à différents agrégats) [SOK76]. La même méthode non hiérarchique peut ensuite être appliquée aux divers agrégats constitués pour obtenir une classification. Les méthodes non hiérarchiques utilisent moins de temps de calcul, même si globalement les opérations d'agrégation sont fortes consommatrices, notamment pour les très grandes banques de données.

Van Rijsbergen ([VRI75], repris par [FAL85]) définit trois critères de qualité pour une opération d'agrégation :

. la méthode doit être stable quand la taille de la banque de données augmente (i.e. la partition ne doit pas changer radicalement avec l'insertion de nouveaux documents)

. de légères erreurs dans la description des documents ne doivent provoquer que de légers changements dans la classification obtenue

. la méthode doit être indépendante de l'ordre d'examen des documents.

A ces critères de qualité, il faut ajouter des critères quantitatifs, notamment le temps de calcul nécessaire pour réaliser l'agrégation.

Malheureusement, ces deux exigences sont contradictoires. Les méthodes hiérarchiques sont fortes consommatrices car elles reposent sur le calcul d'une matrice de similitude entre documents. Cette matrice doit être recalculée à chaque insertion de nouveaux documents afin que la classification soit une image de l'ensemble de la banque de données. Il ne semble pas que cela soit compatible avec un système dynamique comme un système documentaire, pour lequel les opérations d'insertion sont déterminantes. A l'inverse, les méthodes non hiérarchiques (ou méthodes itératives) sont basées sur l'incorporation de nouveaux documents à la banque de données déjà organisée en agrégats. Moins consommatrices en temps de calcul et plus adaptées aux systèmes d'information, elles cependant plus dépendantes de l'ordre initial dans lequel les documents sont examinés. En regardant ce phénomène à long terme (sur plusieurs années), cela revient à apporter un poids très important à l'histoire linguistique des termes d'indexation, tendant à limiter la nouveauté lexicale et à sur valoriser des termes utilisés à un moment donné, même s'ils sont tombés en désuétude.

Pour classifier les documents on utilise un coefficient de similitude entre documents, dont le calcul varie selon les méthodes. En général, il faut tenir compte de l'ensemble des termes d'indexation et effectuer un calcul de sommation des relations entre ces termes, chacun d'eux étant éventuellement pondéré, et établir une moyenne pour définir le coefficient de similitude.

Si, afin de simplifier l'exposé, on estime que les documents sont représentés par la liste de leurs descripteurs :
$D_i := < T_k, w_{ik} >$ et $Dj := < T_k, W_{jk} >$

On peut définir plusieurs calculs de similitude entre deux documents, calculés en fonction du nombre de termes en commun.

Le plus simple des coefficients de similitude est donné par le nombre de termes possédés en commun par les deux documents :

$S(D_i, D_j) = | D_i \text{ inter } D_j |$ ($|X|$ représente le cardinal de l'ensemble X.)

Ce coefficient est très sensible à la taille des documents (i.e. au nombre de termes d'indexation). Plusieurs calculs de similitude permettent de normaliser le

résultat de façon à éliminer ce facteur. On trouve ainsi [VRI75] :

- le coefficient de Dice : $S(D_i, D_j) = 2 \frac{|D_i \text{ inter } D_j|}{|D_i| + |D_j|}$

- le coefficient de Jaccard : $S(D_i, D_j) = \frac{|D_i \text{ inter } D_j|}{|D_i| \text{ union } |D_j|}$

- le coefficient du cosinus : $S(D_i, D_j) = \frac{|D_i \text{ inter } D_j|}{|D_i| \times |D_j|}$

- le coefficient de recouvrement : $S(D_i, D_j) = \frac{|D_i \text{ inter } D_j|}{\min(|D_i|, |D_j|)}$

Ces coefficients peuvent être calculés de façon à tenir compte du poids des termes d'indexation. Dans ce cas le coefficient de Dice, par exemple, entre deux documents D_i et D_j est donné par :

$$S(D_i, D_j) = 2 \times \frac{\sum_{k=1}^{n} w_{ik} \cdot w_{jk}}{\sum_{k=1}^{n} w_{ik} + \sum_{k=1}^{n} w_{jk}}$$

avec w_{ik} (w_{jk}) étant le poids du terme T_k pour le document D_i (D_j).

7.b - Méthodes d'agrégation hiérarchiques

La matrice de similitude utile pour les agrégations hiérarchique est une matrice de taille *n X n (n* correspond au nombre de documents). Elle est de la forme suivante :

$$\begin{array}{cccccc}
1 & S(D_1, D_2) & \cdots & S(D_1, D_i) & \cdots & S(D_1, D_n) \\
S(D_2, D_1) & 1 & \cdots & S(D_2, D_i) & \cdots & S(D_2, D_n) \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
S(D_i, D_1) & S(D_i, D_2) & \cdots & 1 & \cdots & S(D_i, D_n) \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
S(D_j, D_1) & S(D_j, D_2) & \cdots & S(D_j, D_i) & \cdots & S(D_j, D_n) \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
S(D_n, D_1) & S(D_n, D_2) & \cdots & S(D_n, D_i) & \cdots & 1
\end{array}$$

Deux documents dont le coefficient de similitude dépasse une certaine valeur sont considérés liés par un arc. Un agrégat est l'ensemble des documents liés entre eux par un coefficient de similitude supérieur à cette valeur.

On calcule ensuite le coefficient de similitude entre deux agrégats, pour définir une hiérarchie d'agrégats.

Supposons que l'on nomme AG_i l'agrégat constitué des documents D_i (respectivement AG_j constitué des documents D_j). On distingue trois types de méthodes pour obtenir ce coefficient de similitude entre agrégats ([CROU90], [GRI86]) :

- la méthode du lien simple ou du plus proche voisin *(single link method,* ou *nearest neighbour)* [WIL84]. Deux agrégats sont liés par le maximum de similitude entre deux quelconques de leurs composants (i.e. les documents les plus proches définissent le lien entre les agrégats).

$S(AG_i, AG_j) = \max [S(D_i, D_j)]$, Di appartient à AG_i et D_j appartient à AG_j.

La hiérarchie obtenue par cette méthode est lâche et les classes sont plus étendues que par les deux autres méthodes.

- la méthode du lien complet *(complète link method)* prend au contraire en compte le minimum de similitude entre deux quelconques des documents appartenant à chaque agrégat (i.e. le maximum de "distance").

$S(AG_i, AG_j) = \min [S(D_i, D_j)]$, Di appartient à AG_i et D_j appartient à AG_j.

Cette méthode tend à réduire la taille des agrégats, surtout si l'on est exigeant sur la valeur minimale de liaison entre agrégats à un niveau hiérarchique donné.

- la méthode du lien moyen *(average link method,* ou *méthode de Ward)* prend en compte la moyenne des similitudes calculées entre deux documents quelconques appartenant à chaque agrégat. Les agrégats obtenus sont plus resserrés que par la méthode du plus proche voisin, mais moins que par la méthode du lien complet.

$$S(AG_i, AG_j) = \frac{1}{|AG_i| + |AG_j|} \sum_{D_i \text{ dans } AG_i} \sum_{D_j \text{ dans } AG_j} S(D_i, D_j)$$

Deux caractéristiques permettent de moduler la réalisation d'agrégats :

- la valeur du seuil imposé pour que la similitude de deux documents puisse créer un agrégat.
- le nombre maximum de documents (ou d'agrégats) que peut contenir un agrégat.

Les méthodes d'agrégation hiérarchiques sont grandes consommatrices de temps de calcul, en général proportionnel au carré du nombre *n* de documents concernés : $O(n^2)$. De nombreuses études sont réalisées pour améliorer les algorithmes d'agrégation hiérarchique ([CAN84], [VOO86], [CAN89]), notamment en utilisant des ordinateurs parallèles ([RAS88], [RAS89]).

7.c • Méthodes d'agrégation non • hiérarchiques (ou itératives)

Les méthodes non hiérarchiques s'appuient sur des calculs réalisés au moment de l'introduction d'un nouveau document dans la banque de données. Les agrégats déjà présents peuvent être recomposés par cette introduction, ce qui entraîne un nouveau calcul de similitude (d'où l'appellation de méthodes itératives).

On distingue deux méthodes d'agrégation non hiérarchique [RAS87] :

- méthode en une passe qui opère suivant l'algorithme suivant :

1. le premier document est caractéristique de l'agrégat 1

2. calculer le coefficient de similitude du document suivant avec chacun des agrégats existants (un agrégat est représenté par son barycentre)

3. si le coefficient de similitude est supérieur à une valeur pré déterminée S_{max}, ajouter ce document à l'agrégat (ou les agrégats) correspondant(s) et recalculer le représentant (barycentre) pour cet (ces) agrégat(s). Si le coefficient de similitude calculé avec tous les agrégats existants reste inférieur à S_{max}, considérer que le document vient initialiser un nouvel agrégat. Recommencer l'étape 2 tant qu'il reste des documents.

Cette méthode est simple et efficace, mais le résultat reste fortement dépendant de l'ordre dans lequel les documents sont examinés, les premiers agrégats constitués regroupant plus de documents que les derniers.

- méthode de réallocation dans laquelle les agrégats sont modifiés jusqu'à ce qu'une forme stable soit obtenue. On utilise l'algorithme suivant :

1. Sélectionner *m* représentants d'agrégats (soit selon une méthode aléatoire, soit éventuellement par une technique d'agrégation en une passe comme ci-dessus, les représentants étant alors les barycentres des agrégats constitués)

2. Assigner chaque document à l'agrégat dont il est le plus proche et recalculer le barycentre de cet agrégat (si un document est déjà considéré comme la graine de l'agrégat il restera affecté à son propre agrégat)

3. Si un nombre de documents supérieur à une valeur donnée a changé d'agrégat (la classification est encore peu structurée) recommencer l'étape 2. Sinon considérer qu'un état stable est atteint.(on peut aussi limiter le nombre de passages par l'étape 2 pour être certain que l'algorithme converge).

Ces deux méthodes requièrent de rechercher pour chaque document l'agrégat dont il est le plus proche. Il semble que algorithme le plus efficace pour ce travail soit d'utiliser le fichier inverse pour limiter le nombre de calculs de similitude nécessaires ([LUC88]). Un document n'ayant aucun terme en commun avec le représentant d'un agrégat ne peut y être affecté. Le calcul de similitude pour le document Di s'effectue alors selon l'algorithme [PERR83] :

1. pour le premier terme T_j du document D_i rechercher les représentants d'agrégats qui comportent ce terme et attacher à chacun un coefficient de similitude valant le poids du terme T_j dans le représentant de l'agrégat multiplié par le poids du terme T_j dans le document D_i. Ce poids est obtenu comme précédemment en tenant compte de la fréquence d'apparition de T_j dans D_i et de l'inverse de la fréquence de T_j dans l'ensemble de la banque de données, afin de donner une importance prioritaire aux termes rares conjoints.

2. pour tous les autres termes T_k du document D_i rechercher les représentants d'agrégats qui comportent ce terme. Si un agrégat est déjà repéré, ajouter le poids pour ce nouveau terme T_k (i.e. Poids dans l'agrégat multiplié par poids dans le document et par l'inverse de sa fréquence dans la banque de données). Sinon ajouter un nouveau coefficient de similitude pour l'agrégat considéré. Répéter l'étape 2 tant que tous les termes du document Di n'ont pas été examinés.

3. l'agrégat le plus proche du document est celui dont le coefficient de similitude est le plus élevé.

Les méthodes non - hiérarchiques permettent d'ajouter des documents au système sans devoir recalculer l'ensemble de la matrice de similitude comme dans le cas des agrégations hiérarchiques. Elles restent cependant elles aussi gourmandes en temps de calcul. On peut cependant améliorer l'algorithme d'introduction d'un document en classant les termes Tj par ordre de poids décroissant, et en arrêtant de prendre en compte de nouveaux documents de la banque de données si le coefficient de similitude qui serait obtenu est inférieur à la valeur S_{max} requise.

Les techniques d'agrégation de documents ont une place particulière dans les techniques d'indexation en ce qu'elles concernent les équilibres qui s'instaurent dans l'ensemble de la banque de données. Un document est alors décrit en fonction de la place qu'il occupe dans l'ensemble des documents.

8 - L'indexation automatique par des méthodes sémantiques

Les recherches en traitement automatique de la langue sont souvent focalisées sur l'indexation de documents. Il s'agit alors d'obtenir :

- une représentation du document dans un langage de description qui prenne en compte les relations sémantiques entre les termes (i.e. un langage de description du SENS d'un document). Ce type de langage utilise par exemple les indicateurs de rôle qui permettent d'associer à chaque terme la fonction qu'il

occupe (acteur, objet...) ou un trait sémantique qui le caractérise (animé/inanimé, concret/abstrait...), ainsi qu'un certain nombre d'opérateurs sémantiques qui peuvent lui être associés *("permet-de-déduire", "dépend-de",..).* Cet objectif d'une représentation interne du contenu des documents est adapté à la production automatique de résumés ou à la réalisation de systèmes de *"question-réponse",* qui offrent à l'utilisateur non pas une liste de documents pertinents, mais directement l'information souhaitée, telle qu'elle a été extraite des documents par le système. SCISOR et RESEDA sont les systèmes de ce type les plus connus. On trouvera des indications sur cette méthode dans [ZAR88], [ZAR90] et [RAU89].

- une liste de descripteurs adaptés au document considéré. Ces descripteurs ne sont pas issus des formes de surface du texte comme dans les méthodes statistiques, mais recomposés à partir de connaissances du système concernant les liens sémantiques entre plusieurs termes, et sur les capacités d'analyse linguistique du système. Il s'agit de faire réaliser par le système le travail d'un indexeur humain. C'est de ce type de travail que nous traiterons dans ce chapitre, car il reste en phase avec le type de système documentaire qui nous concerne dans cette thèse.

L'indexation automatique s'appuie toujours sur une analyse lexicale des termes contenus dans les documents, suivie d'une normalisation lexicale. Cependant, pour produire des descripteurs plus adéquats, on doit pouvoir rapporter les diverses expressions normalisées issues de cette analyse à des formes sémantiques uniques. Par exemple [DEW89, p. 137], DETECTER sera considéré comme équivalent à DETECTION ; SUISSE rapporté à CONFEDERATION HELVETIQUE... Ce travail s'appuie sur des outils documentaires qui sont des versions informatisées des outils traditionnels des indexeurs humains, notamment les listes d'autorité, les thésaurus et les classifications.

8.a - Les outils documentaires

Les outils documentaires aident l'indexeur à choisir les termes d'indexation les plus adéquats au sein d'une liste de termes prédéfinis. Cette formule a pour avantage appréciable d'uniformiser l'écriture des termes

d'indexation. Si des relations de hiérarchie sont tissées entre les termes, l'outil documentaire permet en outre d'interroger le système avec une meilleure couverture. Cependant, les outils documentaires sont très lourds à mettre en place et à maintenir. Ils ont aussi tendance à éliminer le "jargon" spécialisé, qui est souvent le meilleur moyen de décrire et retrouver des documents pour la partie des utilisateurs qui est directement concernée. En fait, les outils documentaires sont conçus pour favoriser l'aspect *"description"* de l'activité d'indexation. L'aspect *"communication"* est rendu plus difficile par l'utilisation d'un vocabulaire contrôlé : il est nécessaire que l'utilisateur connaisse ce vocabulaire et l'emploie dans les termes de la recherche. Enfin, l'utilisation d'un outil documentaire ne garantit pas contre les aspects inconsistants et variables de l'activité d'indexation ([SAL86]).

D'un point de vue informatique, il est nécessaire que les termes du vocabulaire contrôlé et les liens qui les unissent soient intégrés dans le système. Gérard Salton définit un thésaurus comme un instrument pour améliorer le taux de couverture en définissant des synonymes ou des quasi-synonymes (calculés statistiquement) pour les termes de recherche [SAL76] ou en regroupant en classes (méthodes d'agrégation) les termes d'indexation pour obtenir des termes de degré supérieur [SAL83]. Cette définition reste à notre avis trop lâche. Elle ne rapproche pas l'instrument informatique de l'habitude des documentalistes, pour qui un thésaurus est un instrument beaucoup plus puissant, notamment parce qu'il introduit des relations hiérarchiques de type sémantique. L'histoire des thésaurus est relativement récente, et les premières conceptions de ce type d'instrument pour aider à la recherche documentaire datent de 1947 (cf. [ROBE84]). Nous nous contenterons de définir les concepts généraux qui président à l'établissement des outils documentaires.

Thésaurus

Un thésaurus est un instrument permettant de situer l'environnement sémantique d'un terme d'indexation. On peut définir trois types de relations permettant de qualifier un terme T_j (uniterme ou expression composée) dans le langage contrôlé :

. Une relation d'appartenance qui établit les liens entre termes du

vocabulaire contrôlé et un certain nombre de termes dits "rejetés", qui sont des termes connus, mais qui n'ont pas été sélectionnés dans le vocabulaire contrôlé. Il s'agit d'une relation de synonymie peut être considérée dans les deux sens :

- *"terme préférentiel",* qui renvoie le terme rejeté au terme Tj du langage contrôlé. C'est un instrument permettant d'améliorer le taux de rappel des recherches documentaires.

- *"employé pour",* qui est la relation inverse. En donnant la liste des termes rejetés considérés comme synonymes du terme Tj, on définit un espace sémantique plus large pour ce terme qui permet à l'indexeur de mieux juger de l'adéquation du terme au document.

On peut considérer que plusieurs relations appartiennent à ce même champ d'action de la synonymie, considérée comme une relation d'appartenance définissant les frontières entre les termes appartenant au langage contrôlé et les autres termes connus par le système :

- *"terme abrégé"* ,et sa relation réciproque *"abréviation pour".* Cette relation est particulièrement utile dans les domaines généralistes, surtout si plusieurs expressions peuvent conduire à la même abréviation (CGT : "Compagnie Générale Transatlantique", "Confédération Générale du Travail").

- *"équivalence de langage"* qui donne la ou les traductions du terme T_i dans les langages couverts par le thésaurus. Cette relation est souvent difficile à définir, compte tenu des difficultés à traduire terme à terme les expressions d'une langue dans une autre. Ce n'est pas toujours une relation symétrique.

- *"note"* est une relation qui permet de préciser les conditions d'emploi du terme T_i dans le cadre du vocabulaire contrôlé. Cette relation donne de même un environnement sémantique appréciable pour l'indexeur humain, en attendant une possible codification et donc utilisation lors de l'indexation automatique.

. une relation d'association permet de lister les termes qui peuvent être pertinents conjointement avec le terme Tj. On doit limiter le degré de

transitivité de cette relation d'association, par exemple en définissant un coefficient d'affaiblissement, afin d'éviter qu'elle perdre rapidement son sens. La relation d'association peut être considérée comme une suggestion pour définir une relation d'équivalence sémantique entre les termes. Elle est souvent employée dans ce sens pour la construction automatique de thésaurus ([GUN89]).

. une relation d'ordre permet de définir des termes génériques à partir des termes d'indexation. Si l'on considère une relation d'équivalence sémantique entre les termes appartenant au langage contrôlé, cette relation d'ordre est la relation qui associe un terme d'indexation à sa classe d'équivalence. La relation d'ordre n'existe qu'entre les termes du langage contrôlé (termes préférentiels).

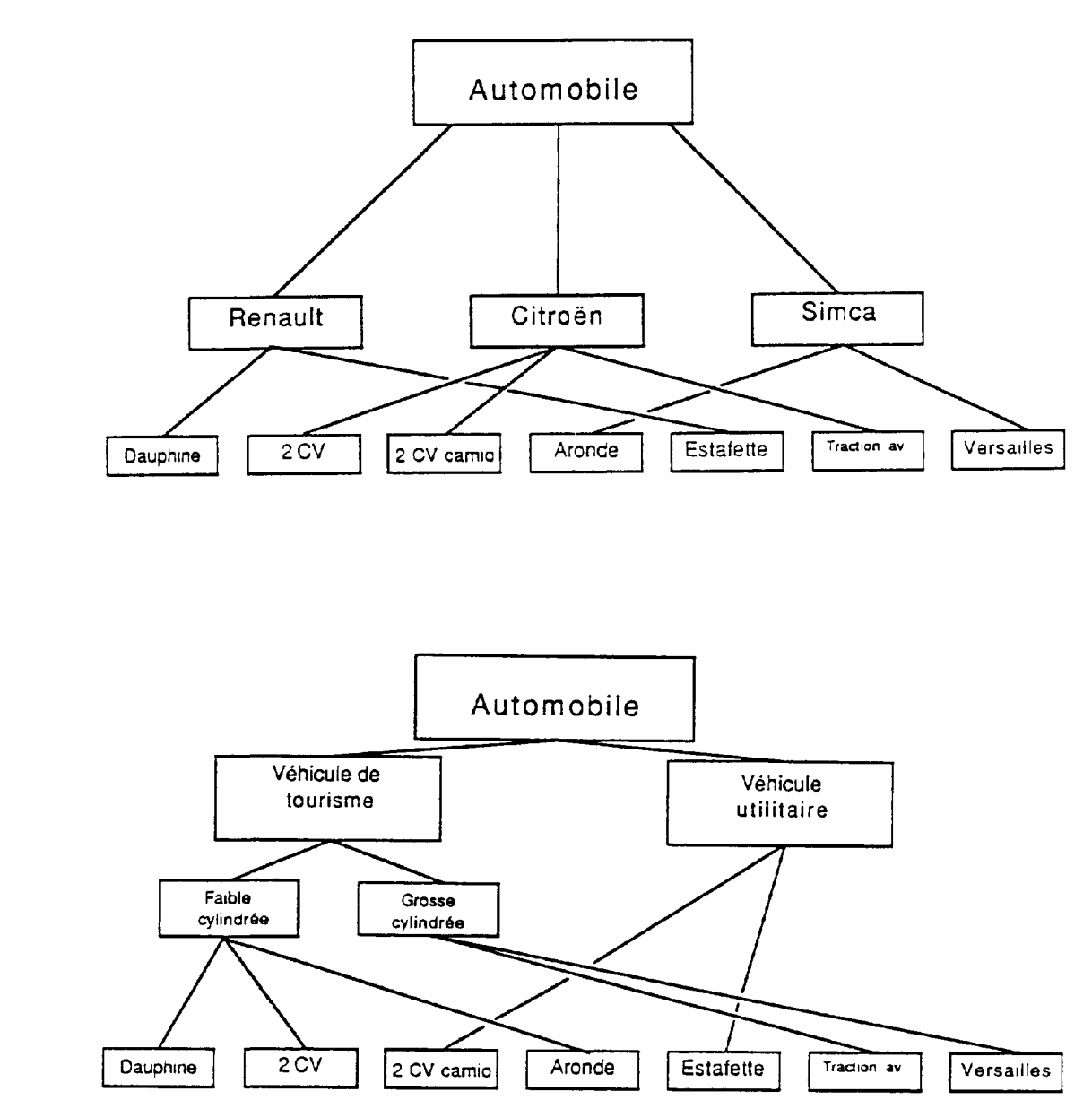
- un *"terme générique" (broder terme)* est le représentant de la classe d'équivalence d'un certain nombre de termes dans le champ sémantique considéré. Il est obtenu par un procédé d'agrégation hiérarchique des termes d'indexation. La relation de similitude est en général une relation sémantique, mais peut être obtenue par des méthodes statistiques (occurrence, utilisation des termes de la question de l'utilisateur...).

- un *"terme spécifique"* est un élément de la classe d'équivalence d'un terme Tj pour la relation sémantique considérée.

Pour que le thésaurus joue pleinement son rôle, il est nécessaire que les classes d'équivalence contiennent des termes dont les fréquences d'utilisation sont comparables [SAL83]. Il faut aussi que les classes elles-mêmes conduisent à un nombre de documents comparable.

La relation d'ordre permet de diviser le thésaurus en plusieurs micro thésaurus, spécifiques d'un domaine précis. Au sein d'un micro thésaurus, on peut définir des relations d'équivalence différentes, qui produisent plusieurs relations d'ordre parallèles. On parle alors d'un thésaurus poly-hiérarchique. Ce type de thésaurus est plus difficile encore à mettre en place et à maintenir. Le principal thésaurus poly-hiérarchique est constitué par le MeSH *(Médical Subject Heading),* créé par la *National Library of Medicine,* pour interroger la banque de données *MEDLINE.* On peut donner un exemple de

thésaurus poly-hiérarchique en explicitant deux relations qui existent entre plusieurs termes d'un vocabulaire consacré aux automobiles :



Avec les thésaurus poly-hiérarchiques, on se rapproche de la notion d'un réseau sémantique, qui est de définition plus complexe, mais aussi plus riche de possibilités. Avec les thésaurus poly-hiérarchiques, il convient de préciser le long de quelle relation d'équivalence il est nécessaire d'élargir ou de préciser les requêtes. Il convient donc de trouver des moyens d'étiqueter les relations pour que l'utilisateur choisisse le type de relation qui l'intéresse.

Réseaux sémantiques

Les réseaux sémantiques peuvent être considérés comme une extension de la notion de thésaurus. Dans un réseau sémantique, les termes d'indexation sont les nœuds d'un graphe, et les arcs sont porteurs d'une relation sémantique. On définit souvent les relations :

. *"est_un",* qui relie un terme à son ou ses génériques
. *"sorte_de",* qui est plus spécifique (un chat est une "*'sorte_de"* félin)
. *"partie_de"* est spécifique des classes d'objets complexes (coude est *partie_de* bras)
. *"synonyme_de"*

Dans un réseau sémantique, les objets héritent des propriétés de la classe à laquelle ils appartiennent, en fonction de la relation sémantique. Par exemple, le chat étant une *sorte_de* félin possède des griffes rétractiles. La propriété "griffes rétractiles" n'a pas besoin d'être directement attachée à chat, mais plutôt à félin, qui est un animal vivipare parce qu'il est une *sorte_de* mammifère...

Classifications documentaires

Les classifications documentaires se distinguent des classifications d'objets en ce qu'elles organisent dans un système méthodique les différents domaines de la connaissance [MAN87]. Les documents sont classés en fonction de leur sujet, et plus précisément en fonction du point de vue porté sur le sujet, ce point de vue s'exprimant par la partie de la classification dans laquelle le document est porté. Ainsi, le sujet *artisanat* peut correspondre à trois points de vue, selon qu'on le rapporte à la *sociologie* , à *l'économie* ou à *l'expression artistique.*

Les classifications documentaires servent en général à classer des livres. Elles remplissent la double fonction de système de classement et de système d'indexation. On parle d'ailleurs en général d'un "plan de classement" pour décrire un instrument de classification. Cette double préoccupation est présente dès les premières classifications bibliographiques *(Système des libraires de Paris* - 1804). Elle est accentuée par les grandes classifications universelles rédigées par des bibliothécaires *(Classification Décimale Dewey, Classification Décimale*

Universelle, Library of Congress Classification...).

Cette double fonction peut être remplie parce que les indices de classification sont représentés par une suite de symboles dont le sens n'est donné qu'au sein de la classification. La notation de l'indice représente donc le contenu sémantique du document en même temps qu'elle permet la succession des documents sur des étagères.

On distingue trois types de classifications :

. les classifications hiérarchiques *(Dewey. C.D.U., Classification Internationale des Brevets...)* considèrent que le savoir qu'elles traitent est constitué d'éléments emboîtés. La notation de ces classifications attribue à chaque symbole le soin de préciser le niveau de hiérarchie des connaissances. On trouve ainsi en Classification Dewey :

5	Sciences
51	Mathématiques
512	Algèbre
513	Arithmétique

Les classifications hiérarchiques sont difficiles à maintenir, notamment parce qu'elles considèrent que toute nouveauté doit s'intégrer dans un chapitre déjà défini des connaissances. Les nouvelles notions se traduisent alors par un allongement souvent rédhibitoire des indices.

. les classifications non hiérarchiques tendent à regrouper les connaissances connexes sans que chacune soit dépendante d'une autre. La Classification de la *Library of Congress,* élaborée à partir des livres pour les commodités du rangement est un exemple typique de classification non hiérarchique Par exemple, dans un domaine défini, les indices représentent d'abord les livres généraux (et non pas les "sujets" généraux) puis les monographies particulières.

Dans des domaines plus restreints, on peut considérer que le plan de classement de *Chemical Abstracts* ou le système des *Concept-codes* de *Biological Abstracts* sont des classifications non hiérarchiques

L'insertion de nouveaux éléments ou la division de classes est plus facile dans les classifications non hiérarchiques, surtout si des espaces ont été laissés vides dans la notation en prévision de nouvelles découvertes, ou de l'émergence de nouveaux points de vue.

. les classifications à facettes partent du principe que chaque sujet peut se décomposer en éléments appartenant à des espaces sémantiques prédéfinis. L'indice est construit par la juxtaposition (moyennant une syntaxe souvent complexe) des indices obtenus lors des diverses "projections" sur ces facettes.

La principale classification à facettes est la *Colon Classification* élaborée par Raganathan. Les facettes choisies (Personnalité, Matière, Energie, Espace, Temps) sont difficiles à comprendre pour les occidentaux, ce qui en limite certainement la portée en dehors de l'Inde. Pourtant, avec l'apport de l'informatique, on peut penser que les systèmes à facettes gardent un grand avenir, et que toutes les ressources d'expressivité de cette approche n'ont pas encore été cernées. L'informatique permet en effet de traiter séparément, et l'un après l'autre tous les aspects d'un document en se plaçant successivement dans chacun des champs sémantiques nécessaires pour le décrire. Le modèle QUID, qui sera développé dans la troisième partie s'inspire des méthodes de classification à facettes.

8.b **- L'extraction des descripteurs**

Alors que l'indexation en texte intégral ou l'indexation par des méthodes statistiques se contente d'extraire des descripteurs à partir des formes de surface du texte, et de les normaliser dans le meilleur des cas, l'indexation par des méthodes sémantiques cherche à définir les *notions* qui sont développées dans le texte. Cet objectif rend nécessaire la présence d'un outil documentaire du type de ceux présentés ci-dessus, qui offre une liste des notions connues par le système. Mais comme l'indexation doit procéder à une analyse linguistique des phrases du texte, il faut de plus posséder un dictionnaire linguistique.

L'analyse lexicale ne doit pas se fixer comme seul objectif de repérer et normaliser les mots présents dans le texte. Il faut au contraire reconnaître des

expressions sous des formes différentes, mais éliminer les termes utilisés dans un sens métaphorique. Par exemple [ERL87], il faut reconnaître le descripteur MISE A JOUR DE L'INFORMATION à partir de la forme *mettre à jour l'information.* Mais il faut éviter que l'expression *faire tache d'huile* n'induise le descripteur OLEAGINEUX.

L'analyse syntaxique doit permettre d'obtenir le même descripteur quelle que soit la paraphrase présente dans le texte. Ainsi, [ERL87] les phrases :

. élection du président de la République
. élection présidentielle
. élection au suffrage universel du président de la République
. scrutin présidentiel
. les Français vont élire un Président en 1988
. le chef de l'état a été élu avec 52 % des suffrages

doivent donner le même descripteur. En revanche, la phrase *"l'élection du Parlement Européen a été l'occasion pour le Président de la République de rappeler son attachement..."* ne doit pas se traduire par ce même descripteur, ce qui serait difficile avec des méthodes statistiques.

L'analyse syntaxique doit permettre de définir des parties homogènes dans les phrases. Les descripteurs seront retrouvés dans ces éléments syntaxiques. Il faut aussi opérer une nominalisation des verbes *(les prix augmentent -> augmentation des prix).*

Les expressions ainsi constituées doivent être comparées avec les termes descripteurs de l'outil documentaire informatique retenu. Cette comparaison ne peut se baser sur la simple similitude entre les expressions trouvées à partir du texte et les entrées du thésaurus. Plusieurs déformations de l'expression linguistique retrouvée par l'analyse seront réalisées jusqu'à retrouver une forme compatible avec une forme connue de l'outil documentaire. Cette forme, ou la forme synonyme si elle est *"terme rejeté",* représentera alors le descripteur choisi. Les entrées de l'outil documentaire pour cette comparaison peuvent être différentes du terme retenu, mais aussi prendre en compte des termes définis par les relations de synonymie ou d'association, jusqu'à retrouver le terme descripteur le plus adéquat.

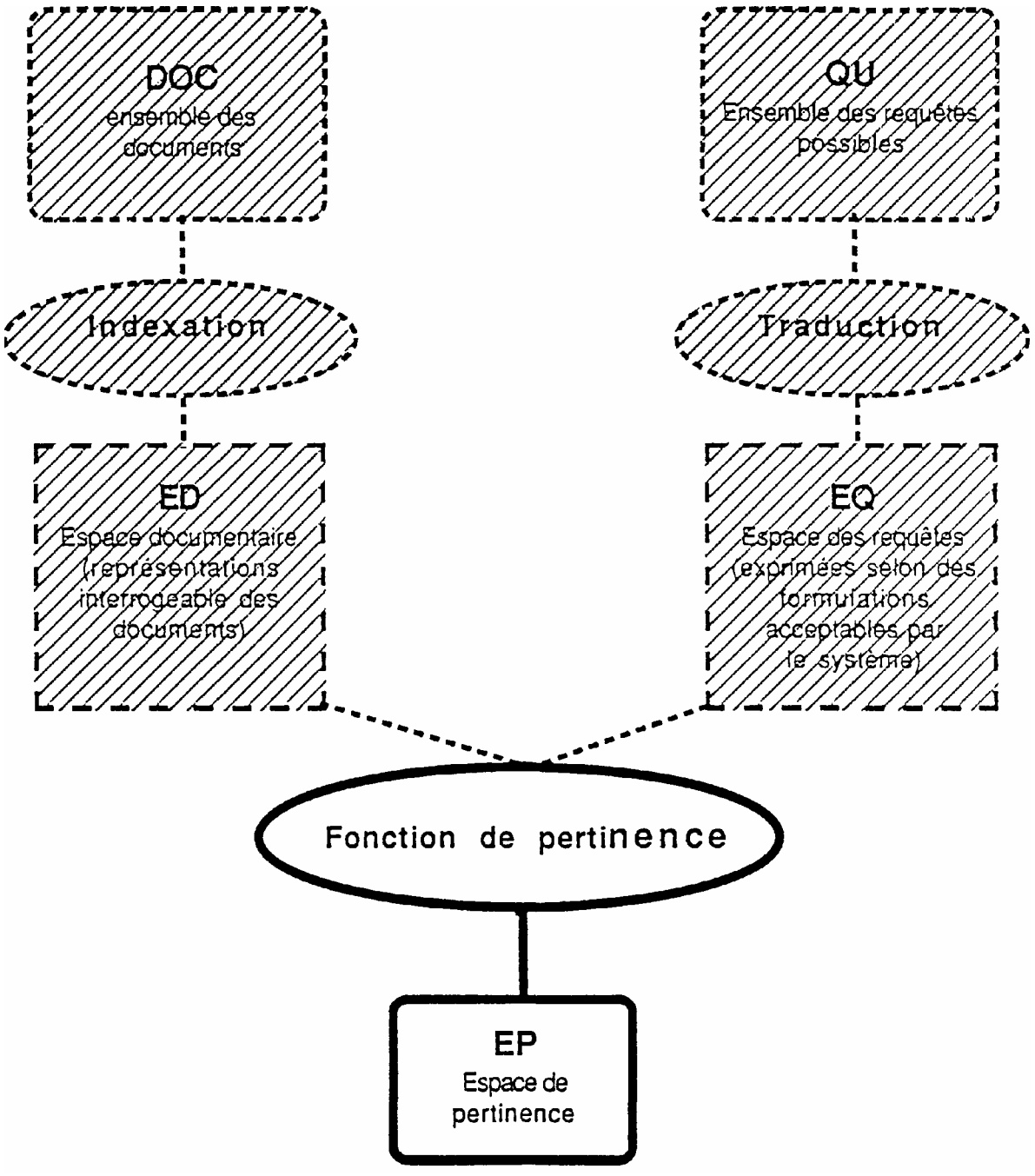
Ces processus d'analyse sémantique sont encore aléatoires, notamment parce qu'il existe peu de thésaurus informatisés, et parce que les méthodes d'analyse lexicale et syntaxique doivent encore se développer. Parce que le langage est une connaissance qui s'apprend par la pratique et non pas par les définitions comme l'enseigne Wittgenstein, il importe de trouver des méthodes informatiques évolutives, qui pourront constituer les outils d'analyse et les faire évoluer en relation avec les expériences concrètes d'indexation automatique. En ce sens, les méthodes sémantiques doivent dans un premier temps être utilisées sous contrôle et vérification d'un indexeur humain. Il s'agit alors de méthode d'aide à l'indexation.

Mais si l'on admet ce principe, ces méthodes d'aide à l'indexation peuvent aussi intégrer des apports des méthodes statistiques, des techniques d'analyse des citations et des méthodes d'agrégation de documents. On aboutit alors à un instrument puissant qui permet d'indexer un document en le traitant pour lui même (aspect description) et en tenant compte de son insertion dans la banque de données (aspect communication). Les faiblesses de l'indexation humaine seront certainement compensées, notamment par l'effet statistique, tout en gardant le pouvoir d'analyse des êtres humains, et en permettant au système d'apprendre, pour mieux remplir son rôle. Si l'on voulait adopter une image, avec un système d'aide à l'indexation qui regrouperait les méthodes d'indexation développées dans ce chapitre, les documents entrés dans la banque de données seraient représentés comme une boule dotée de tentacules (ses premiers termes d'indexation repérés par analyse du document lui-même). En se liant à d'autres documents grâce à ses tentacules (techniques d'agrégation, méthodes statistiques, analyse des citations) ce document pourrait enrichir son indexation de nouveaux termes, tout en pondérant les descripteurs déjà retenus. Ce processus serait sous contrôle de l'indexeur, ce qui garantirait une meilleure qualité au travail, mais aussi augmenterait la productivité de l'indexation. Ce qui est vrai pour un utilisateur, à savoir qu'il est plus facile de choisir dans une liste un descripteur que de l'imaginer, l'est aussi pour un indexeur. Plutôt que de s'orienter vers des systèmes entièrement automatiques aux résultats aléatoires, cette démarche de systèmes d'aide à l'indexation semble plus productive. Il reste à définir les outils informatiques adaptés à ce type de travail, notamment en terme de fonctionnalités, d'ergonomie et bien entendu de temps de traitement.

IV - fq : ED x EQ -> EP

La recherche documentaire

Une fois acquise la description de documents dans le cadre d'une banque de données particulière, il convient d'étudier les divers modèles de recherche documentaire.



On peut définir plusieurs modèles élémentaires de la recherche documentaire, ajoutant pièce à pièce des éléments à une construction théorique qui partirait du système le plus simple jusqu'au plus complexe. Le niveau élémentaire correspond à une question limitée à un terme seul (une *"vedette"}*. Les documents décrits par cette vedette sont extraits si la vedette, et elle seule, est posée en requête. Ce système est celui en vigueur dans la plupart des bibliothèques fonctionnant avec un fichier manuel, et correspond aux premières versions des systèmes de gestion de bibliothèques (même si certains subsistent encore, cf. la première version implantée en France en 1985 de LIBS 100 le logiciel de CLSI), David Blair [BLA90] distingue ainsi 13 niveaux élémentaires.

Malheureusement, cette approche ne tient pas compte de la réalité synthétique des modèles existants, et surtout des modèles à l'état de prototype de recherche. Nous choisirons donc de présenter les systèmes par une approche plus globale des alternatives qui se présentent au chercheur en informatique documentaire. Les systèmes décrits sont alors eux-mêmes composés de plusieurs "services élémentaires", par exemple un modèle booléen intégrant les opérateurs de proximité et la mise en ligne de thésaurus.

On peut ainsi considérer les modèles suivants :

- le modèle booléen
- le modèle vectoriel
- le modèle probabiliste
- le modèle hypertexte
- le modèle connexionniste

Avant d'examiner ces différents modèles, il convient de décrire le fonctionnement théorique du moteur de recherche, instrument permettant de trouver les documents répondant à une requête, ou du moins, dans une optique plus générale, d'évaluer le degré de pertinence d'un document par rapport à une requête.

1 - Le fonctionnement du moteur de recherche

On distingue globalement trois méthodes de recherche dans un fichier documentaire :

- une recherche séquentielle : l'ordinateur examine séquentiellement tout le fichier pour retrouver les occurrences du (ou des) termes de la requête. Compte tenu de la grande taille des banques de données, ce type de recherche est purement théorique, même s'il peut garder une efficacité sur un fichier restreint (éventuellement en complément de méthodes permettant de n'examiner qu'une faible partie de la banque de données).

- une recherche sur des fichiers inverses : à partir de l'examen des documents de la banque de données, le système crée un index en sortant dans l'ordre alphabétique une liste des termes d'indexation (en tenant compte des remarques précédentes sur ces termes d'indexation, qui peuvent être des noms d'auteur, des descripteurs, des unitermes issus de l'analyse du texte intégral, des citations, des codes de classification...tout type d'information qui sert à décrire un document dans le cadre du système documentaire). A chaque terme d'indexation est associé une liste des identificateurs des documents correspondants, et éventuellement des informations complémentaires (localisation, pondération...). La recherche d'un terme s'effectue alors dans ce fichier inverse, qui renvoie à l'adresse du document.

- une recherche sur fichier de signatures : chaque document est représenté par une signature qui est un vecteur binaire de longueur fixe. La recherche s'effectue en comparant le vecteur binaire de la question (obtenu par la même fonction signante) et celui du document. Cette méthode connaît un regain d'intérêt car elle peut plus facilement être traitée en parallèle sur des machines comportant de nombreux processeurs (64 000 processeurs pour la *Connection Machine* de *Thinking Machine Corp.* [WAL87] ; 16 000 pour la *Distributed Array Processor (D.A.P.)* de *ICL* [CAR88])

1.a - Recherche par fichier inverse

Le fichier inverse permet une entrée facile à des informations relatives aux termes d'indexation. Les termes d'indexation sont classés par ordre alphabétique, ce qui permet de retrouver rapidement les informations en recherchant par dichotomie dans la liste. Des algorithmes sont développés régulièrement qui permettent de retrouver encore plus rapidement dans la liste le terme considéré.

La principale information affectée à un terme est l'identificateur des documents décrits par ce terme. Si une requête est composée de plusieurs termes, le système réalise une liste fusionnée *(merged list)*. Gérard Salton [SAL88b] donne l'exemple simplifié suivant :

Term A : {D_2, D_{15}, D_{23}, D_{89}, D_{123}, D_{140}, D_{148},...}
Tenu B : {$D_{5>}$ D_{10}, D_{15}, D_{23}, $D_{50>}$ D_{90}, D_{123}, D_{190},...}
Liste fusionnée : {D_2, D_5, D_{10}, D_{15}, D_{15}, D_{23}, D_{50}, D_{89}, D_{90}, D_{123}, D_{123}, D_{140}, D_{148}, D_{190},...}

. La présence conjointe des deux termes A, B (correspondant à la conjonction booléenne) est représentée par les éléments présents deux fois dans la liste fusionnée :
Documents retrouvés pour (A *et* B) : {D_{15}, D_{123})

. La présence de l'un des deux termes au moins (correspondant à la disjonction booléenne) est représentée par l'ensemble de la liste, les éléments en double n'étant considérés qu'une seule fois :
Documents pour (A ou B) : {$D_{2>}$ D_5, D_{10}, D_{15}, D_{23}, $D_{50>}$ $D_{89>}$ D_{90}, D_{123}, D_{140}, D_{148}, D_{190},...}

La réalisation des listes fusionnées est une opération coûteuse en temps de calcul, surtout si l'on se trouve en présence de listes longues (termes ayant une grande fréquence) ou de nombreuses listes (questions composées de nombreux termes). Certains essaient en conséquence de construire des machines spécialisées (machines de bases de données) capables de réaliser ce type de traitement très rapidement. Le traitement en parallèle des listes est aussi envisageable.

Le fichier inverse peut contenir d'autres informations que la mention de l'identifiant des documents décrits par un terme. Pour permettre à l'utilisateur de juger l'efficacité de sa recherche, il faut aussi que le fichier inverse contienne des informations sur le nombre occurrences d'un terme dans la banque de données et sur le nombre de documents extraits. Pour faciliter la recherche sur un texte, il est aussi utile de disposer d'opérateurs de proximité, permettant de spécifier dans la question la distance entre deux mots. Les informations de position, qui sont alors indispensables, sont intégrées dans le fichier inverse.

Typiquement, le fichier inverse du logiciel de recherche en texte intégral BRS est constitué de la façon suivante [SIN89] :

. *un fichier "dictionnaire"* dont l'entrée est composée par un mot (terme d'indexation), lié aux informations suivantes : nombre occurrences du mot, nombre de documents qui contiennent ce mot, pointeur vers un second fichier de positionnement.

. *un fichier de positionnement* qui pour chaque terme possède les informations : numéro du document, nom du champ contenant le terme, numéro de phrase dans ce champ, numéro du mot à l'intérieur de la phrase.

Le fonctionnement quotidien de ce type de système comportant de nombreuses informations en sus de l'identifiant du document, permet d'envisager d'autres types d'extensions qui donneraient d'autres indications, tout en conservant des temps d'accès du même ordre de grandeur. Par exemple, le fichier inverse pourrait recevoir, comme dans le système SMART ([SAL76],[SAL88b]), des indications de pondération, calculées selon une des formules évoquées dans le chapitre sur l'indexation. La structure en deux fichiers reliés par un pointeur du logiciel *BRS* (issus des recherches sur le logiciel *STAIRS* d'IBM) permet dans le même ordre d'idées, d'envisager que le fichier dictionnaire puisse renvoyer à des formes normalisées des entités lexicales, qui alors contiendraient les informations de positionnement. Les formes INFORMATISER, INFORMATISATION, INFORMATISENT... seraient alors traitées de la même manière et la demande d'une des formes correspondrait à la recherche de toutes les formes associées.

Le fichier inverse peut de même contenir une liste des unitermes employés dans la description des documents, mais aussi des expressions composées (comme le fichier inverse des descripteurs de la banque de données *PASCAL),* voire un double traitement pour les mots possédant un trait d'union *(agro-alimentaire* et *agroalimentaire,* ou *science-fiction* et *science fiction).* Limiter une recherche à un champ d'information donné peut se réaliser par l'ouverture de plusieurs fichiers inverses (choix du logiciel documentaire *Texto)* ou grâce à l'indication de champ contenue dans le fichier de positionnement (choix de *BRS).*

Même si les fichiers inverses sont employés depuis de nombreuses années, on peut encore imaginer de nouveaux modes d'utilisation des index qui permettraient d'améliorer les systèmes documentaires ([PERR83], [SAL86]).

1.b • Recherche par fichier de signatures

Les signatures de textes connaissent aujourd'hui un grand développement, notamment dans le domaine de la bureautique ([CRO88]). Les idées à l'origine des fichiers signés ont été développées pour gérer des dictionnaires orthographiques. Elles sont aujourd'hui plus largement appliquées. Certaines applications documentaires commencent à être diffusées commercialement comme le système *DowQuest* du serveur *Dow Jones,* qui propose l'accès au texte intégral de plus de 175 périodiques économiques et d'information *(Wall Street Journal, Fortune, Washington Post,...).* L'appel vers l'utilisation de fichiers signés est provoqué par le développement des machines parallèles, capables de traiter en même temps plusieurs milliers de documents (64 000 documents pour la *Connection Machine* utilisée pour *DowQuest)* ([TRI89])

Une signature de texte est un vecteur binaire de longueur fixe *k.* A l'origine, chaque case du vecteur binaire est mise à 0. Un mot est représenté par le passage à 1 de une ou plusieurs cases. Les différents mots d'un texte sont ainsi représentés sur le même vecteur binaire.

Une formule simple pour obtenir une signature est la suivante [POG87] : Supposons le vocabulaire du système composé de S termes, et la signature de longueur *k.* Les mots du vocabulaire sont classés dans une liste alphabétique. Le

$i^{ème}$ terme sera mis en correspondance avec la case de numéro ci de la signature par la formule :

$C_i = | i * k/S |$ où $| x |$ est premier entier supérieur à x.

Par exemple, pour obtenir la signature de longueur 1024 cases de l'expression RECHERCHE DOCUMENTAIRE à partir d'un vocabulaire de 10240 termes, on recherche dans un fichier dictionnaire fonctionnant comme une table de fonction de hashing les deux termes : RECHERCHE (e.g. 897^{eme} dans la liste) et DOCUMENTAIRE (e.g. 6324^{eme} dans la liste). Les cases |897*1024/10240| =90 et 16324*1024/10240| =633 du vecteur binaire seront "allumées" (on utilise le terme "allumer" comme image pour exprimer que la valeur de la, ou des, case(s) concernée(s) passe à 1).

Comme on a réduit la taille du vocabulaire d'un facteur 10 (un très faible facteur en regard des besoins de réduction des banques de données réelles), on retrouvera 10 termes qui "allumeront" la même case du vecteur. Il existe donc une large probabilité d'erreur, qui a pour conséquence que l'on peut retrouver un document qui ne correspond pas à la demande *(false hit).* Si cette éventualité est grande, ce qui dépend de la méthode d'élaboration des signatures, la recherche sur fichier signé ne peut servir que de première étape avant l'examen mot par mot (séquentiel) d'un plus faible nombre de documents (ceux qui ont franchi cette étape de la recherche par signature avec succès). Ce procédé en deux temps est par exemple déjà utilisé, sous une forme différente mais similaire, dans la recherche structurale en chimie. Le logiciel DARC ne réalise la recherche de position "atome par atome" qu'après avoir réduit la taille du fichier par une recherche plus générale de groupes d'atomes. Même avec une stratégie en deux temps, cette méthode peut s'avérer rapide si l'on utilise des systèmes multiprocesseurs ([CRI88], [WIL88], [WIL89]).

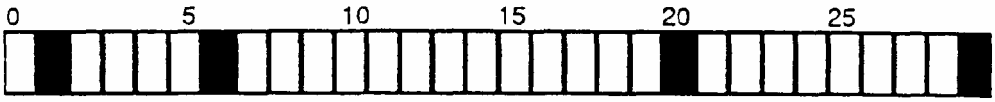
Une autre méthode est employée par les concepteurs de l'application documentaire sur la *Connection Machine.* Elle est basée sur une méthode décrite dès 1949 par Mooers [cité par FAL85] et connue sous l'appellation *superimposed coding.* Soit *k* la longueur de la signature. Pour insérer un terme dans cette signature, on allume *i* cases du vecteur. Une table de fonction de hashing permet de connaître les positions des *i* cases à partir du mot. La signature est utilisée pour placer *w* mots sur la même signature. Si le document dépasse *w* mots

significatifs, plusieurs signatures sont liées pour un même document, car la méthode est très sensible aux variations sur le nombre de mots. Les valeurs généralement utilisées sont :

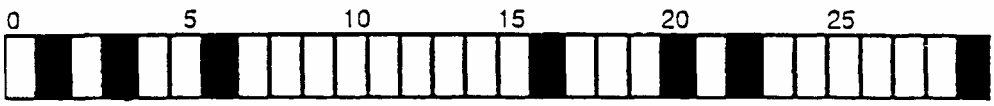
. k = 512 ou 1024 (longueur de la signature),
. i varie entre 10 et 30 (nombre de cases allumées par mot)
. w est compris entre 15 et 30 (nombre de mots par signature).

Un exemple extrait de [STA86b] permettra de mieux saisir le fonctionnement de ce codage.

Considérons une signature de longueur *k=30,* pour laquelle chaque terme allume *i=4* cases. La lecture de la table de correspondance nous indique pour le terme *"chimie"* que hash(chimie)=(1,6,29,20), ce qui se traduit par la signature :



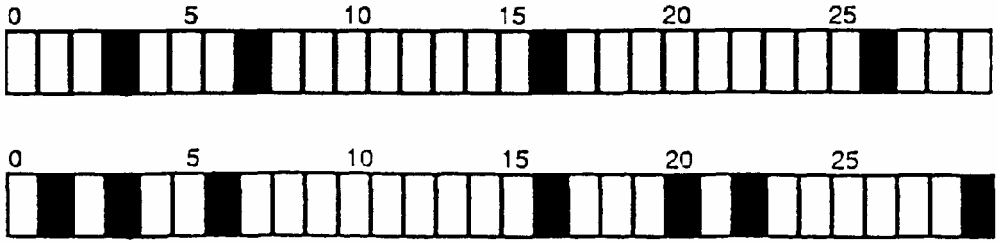
Si le document contient aussi le terme *"biologie"* tel que hash(biologie)=(16,22,29,3), la signature devient :



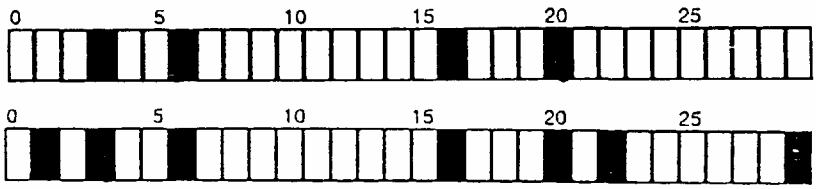
La case 29 qui était déjà allumée ne change pas.

Pour rechercher un mot, on établit sa signature de la même manière, en regardant la même table de correspondance, et l'on compare la signature obtenue avec les signatures des documents. Si les quatre cases sont allumées, le terme est considéré comme PRESENT.

Le terme *"physique"* tel que hash(physique)=(3,7,16,26) sera considéré comme ABSENT car les cases 7 et 26 ne sont pas allumées.



En revanche, si l'on recherche le terme *"mathématiques"* dont le codage est donné par : hash (mathématique) = (3,6,16,20), celui-ci sera considéré comme PRESENT, alors qu'il ne fait pas partie des deux termes *"biologie"* et *"chimie"* qui sont représentés dans la signature. Il s'agit d'une "erreur inévitable" *(false hit)*. Les valeurs respectives de *k, i,* et *w* permettent de minimiser le nombre d'erreurs inévitables. Un calcul probabiliste détaillé dans [STA86b] montre que le facteur principal est le nombre de mots encodés dans une signature. Le tableau suivant décrit les probabilités d'une fausse réponse en fonction de *w :*



w = Words per table
i = Bits set for each word, fixed at 10
k = Bits per table, fixed at 512
P_{false} = Probability of a false positive being returned when probing one table for one word

w	P_{false}
5	4.91×10^{-11}
10	3.12×10^{-8}
15	1.13×10^{-6}
20	1.26×10^{-5}
25	7.46×10^{-5}
30	2.96×10^{-4}

Avec cette méthode, telle qu'elle est utilisée par le serveur *Dow Jones,* les questions sont posées comme étant des "questions simples", c'est à dire obtenues par juxtaposition de termes, indépendamment de l'utilisation des connecteurs booléens. La phrase correspondant à la question est signée selon la même méthode que les documents, c'est à dire que les termes la composant sont juxtaposés sur la même signature question. La comparaison entre la signature de la question et celles des documents est ensuite réalisée. Un des avantages de cette méthode est de pouvoir aisément poser des questions longues, typiquement des documents entiers dont on utilise la signature pour retrouver les documents les plus proches.

Avec une structure de fichier signé, les problèmes de reconstruction des index sont minimisés, l'encombrement du fichier global (fichier documentaire + fichiers inverses) est réduit et les ajouts de documents sont plus aisés. Cette méthode devrait s'adapter de ce fait à la diffusion de banques de données sur Disques Optiques Compacts [C0089].

2 - Le modèle booléen

Le modèle booléen est largement le plus répandu en informatique documentaire. La majeure partie des systèmes commerciaux utilisent ce modèle, éventuellement élargi pour traiter le texte intégral. Il sert de base de référence à tous les nouveaux modèles. On peut même penser que des améliorations notables sont encore possibles à partir du modèle booléen de base. Notamment en utilisant des anté-serveurs qui peuvent implanter un système de jugement de pertinence et une visualisation ordonnée des documents retrouvés au dessus d'un serveur agissant suivant un modèle booléen([RADE88], [MOR82]).

2.a - Les opérateurs booléens

Le modèle booléen de base se définit par :

- *une indexation* des documents par un ou plusieurs descripteurs non pondérés. Les documents sont divisés en différents champs ce qui permet de d'indiquer qu'un descripteur a une fonction pré définie (auteur, sujet, classification, localisation géographique...).

- *une formulation des requêtes* qui respecte la logique booléenne. Celle-ci se base sur trois connecteurs :

. la conjonction (connecteur ET) exige que les termes (mots ou expressions composées selon les choix suivis au niveau de l'indexation) soient présents simultanément dans la description d'un document,
exp : IRAK et (PRODUCTION PETROLIERE)

. la disjonction (connecteur OU) exige que l'un au moins des termes soit présent dans la description des documents retrouvés.
exp. : LAIT ou (PRODUITS LAITIERS) retrouvera les documents indexés par LAIT comme ceux indexés par PRODUITS LAITIERS.

. la négation (connecteur SAUF) permet d'éliminer les documents possédant un terme particulier.
exp. : PVC sauf EMBALLAGES

Les diverses opérations peuvent être combinées dans une "équation de recherche".

exp. : ((TRAITEMENT DE TEXTE) et (MACINTOSH ou IBM-PC ou MS-DOS)) sauf (WORD ou MACWRITE)

- *les documents sont retrouvés* ou non suivant la présence ou l'absence des termes utilisés dans l'équation de recherche.

Ces règles de base sont élargies pour pouvoir traiter le texte intégral en ajoutant des *"opérateurs de proximité"* qui permettent de spécifier la fenêtre de texte dans laquelle les termes doivent être présents. Dans l'hypothèse de traitement du texte intégral, les expressions composées sont éclatées au moment de l'indexation en descripteurs unitermes. La reconstitution des expressions est dirigée par l'utilisateur lorsqu'il formule sa question grâce aux opérateurs de proximité.

On distingue en général les opérateurs de proximité :

. distance ordonnée entre termes (opérateurs av, w, adj selon les systèmes) : les termes doivent apparaître dans le texte suivant l'ordre de la question, éventuellement séparés par un nombre de mots précisé par l'utilisateur :

exp. : (DEUXIEME av TOUR) 2av (ELECTIONS av LEGISLATIVES) -> retrouvera la phrase *"le deuxième tour des dernières élections législatives...."* mais ignorera la formulation *"les résultats des élections législatives après le deuxième tour..."*

. distance entre termes (opérateur n) permet de résoudre ce problème en ne précisant pas l'ordre dans lequel les termes doivent apparaître.
exp. : (DEUXIEME av TOUR) 2n (ELECTIONS av LEGISLATIVES).
A nouveau la phrase *"le deuxième tour vient de sanctionner la victoire du parti arrivé en tête des élections législatives"* ne sera pas retrouvée.

. termes présents dans la même phrase (phr, same...) est

l'opérateur de proximité suivant :

exp. : (DEUXIEME av TOUR) phr (ELECTIONS av LEGISLATIVES).

Souvent cela ne suffit pas, par exemple pour le texte : *"Nous connaissons maintenant les résultats des élections législatives. A la suite du deuxième tour..."*

. termes présents dans le même paragraphe (prg, with,...) permet la recherche dans une fenêtre de texte encore plus large.

Les opérateurs de proximité sont adaptés au traitement du texte intégral des documents. Cependant, les exemples ci-dessus viennent montrer que le choix d'un opérateur, et plus généralement le choix d'une équation de recherche reste difficile. On peut penser qu'un utilisateur ne peut imaginer les diverses tournures que peut prendre une même idée dans des textes différents. Ce problème n'est pas totalement absent des systèmes basés sur une indexation manuelle. On trouve souvent des explications sur la logiques booléenne qui assimilent présence d'un concept dans un texte ou dans un document et utilisation des mots pour décrire ce concept. Il faut rester prudent sur cet amalgame du mot et du sens. Une équation de recherche booléenne permet de retrouver les documents décrits par les mêmes termes, dans la même combinaison. Plus précisément même, de retrouver les documents décrits par des chaînes de caractères exactement semblables aux chaînes de caractères posées dans la question (éventuellement avec des troncatures ou des masques, mais qui doivent être précisés dans la question). A aucun moment le modèle booléen ne permet de généralisation au-delà de cette stricte définition. Si le même concept est décrit par d'autres termes, ou par les mêmes termes combinés différemment, le document ne sera pas extrait.

2.b • Avantages et limites du modèle booléen

Le modèle booléen a encore de beaux jours devant lui. Belkin et Croft ([BEL87], repris par [DAC90a]) en donnent les raisons suivantes :

- les investissements réalisés dans les systèmes commerciaux fonctionnant sur ce modèle sont considérables, et un changement radical n'est guère envisageable actuellement.

- les autres techniques n'ont pas encore été testées dans des environnements opérationnels, en grandeur réelle. A contrario, malgré les nombreuses critiques que l'on peut lui porter, le modèle booléen a subi avec succès l'épreuve du feu. Il a su s'adapter au texte intégral, se baser sur des fichiers inverses toujours plus rapidement mis à jour, permettant l'intégration en continu de nouveaux documents dans les grandes banques de données (agences de presse...).

- les résultats obtenus par les techniques alternatives, même au niveau expérimental ne sont pas suffisamment supérieurs pour justifier le changement.

- la structure des requêtes booléennes permet à l'utilisateur d'exprimer, s'il arrive à la maîtriser (ou s'il arrive à s'attacher les services d'un intermédiaire compétent), les éléments principaux de sa recherche et les articulations logiques essentielles.

Pourtant, les critiques au modèle booléen sont nombreuses, qui traitent notamment de sa rigidité :

- la formulation des requêtes est complexe. Le sens booléen des connecteurs ET, OU, SAUF est différent du sens qu'ils ont généralement dans la langue quotidienne. Par exemple, l'expression *"fruits et légumes"* doit s'exprimer *"fruits OU légumes"* en booléen. Ce passage entre deux utilisations des mêmes mots outils, qui plus est de mots outils largement utilisés dans la langue quotidienne, est difficile. La logique booléenne serait bien plus facilement assimilable par l'utilisateur si elle n'était pas utilisée pour traiter des expressions linguistiques. C'est par exemple le cas dans les systèmes gérant les données.

- de nombreux textes pertinents ne sont pas retrouvés car leur description ne correspond qu'approximativement à la requête. Notamment, une requête du type : *"T1 et T2 et T3"* ne donnera des résultats que si les trois termes sont présents, indépendamment de l'importance relative pour l'utilisateur des trois termes. Un document contenant deux des trois termes sera rejeté de la même manière qu'un document n'en contenant aucun.

- l'importance relative des termes à l'intérieur de la requête ou à l'intérieur du texte n'est pas prise en compte. Les résultats ne sont pas ordonnés en fonction de leur degré de pertinence pour la question de l'utilisateur. Ainsi, une requête disjonctive : "T_1 *ou* T_2 *ou* T_3" extraira les documents qui contiennent au moins l'un des trois termes. Si un document contient les trois termes, ou deux des trois termes, il ne sera pas favorisé et présenté en premier à l'utilisateur.

- la comparaison entre la requête et les documents doit faire appel à des représentations utilisant le même vocabulaire. Ce point est toutefois tempéré par l'utilisation (encore rare) de thésaurus en ligne.

L'enjeu aujourd'hui des recherches en sciences de l'information est de passer du stade des critiques portées au modèle booléen à la résolution des défis qu'il pose aux utilisateurs. Les autres modèles que nous allons décrire maintenant sont efficaces en environnement de recherche. Doivent-ils être étendus tels quels ou doivent-ils servir de référence pour perfectionner le modèle booléen ? On peut penser qu'il est possible d'intégrer sur une base booléenne des thésaurus en ligne, des systèmes permettant de formuler librement les questions, de classer les documents par ordre de pertinence... Plus généralement, il semble possible de créer des systèmes documentaires comme une couche au dessus d'un modèle booléen, éventuellement au prix de légères modifications dans le fonctionnement des fichiers inverses, qui pourront alors intégrer d'autres informations que l'identifiant des documents décrits par un terme, comme nous l'avons souligné précédemment.

3 • Le modèle vectoriel

Le modèle vectoriel a été impulsé par le développement du projet SMART de Gérard Salton dans les années 60 ([SAL761, [SAL83]). Il s'agit de considérer un document comme un vecteur dans un espace vectoriel dont un référentiel est donné par l'ensemble des termes d'indexation, en admettant que tous ces termes sont deux à deux indépendants.

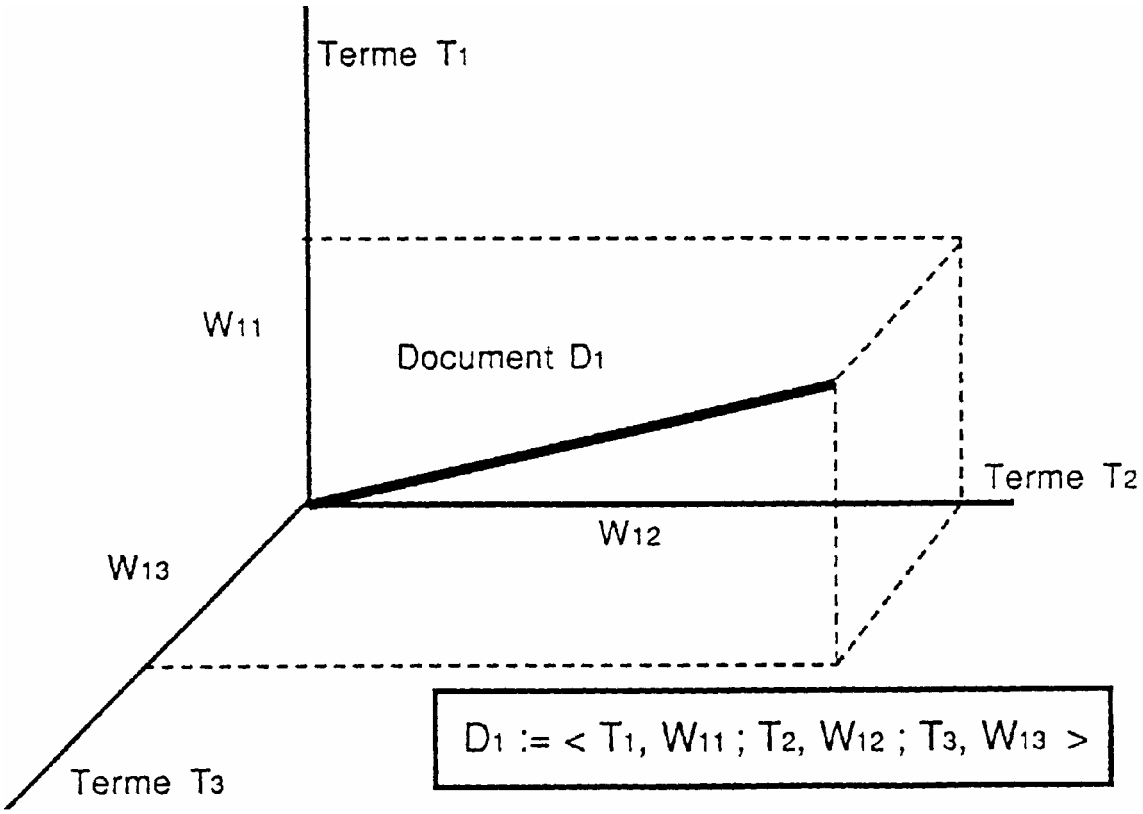
3.a - Documents et questions dans un espace vectoriel

Si les termes nécessaires pour décrire l'ensemble des documents du système sont T_1, T_2,....T_n ils forment un référentiel et chaque document est un vecteur représenté par une combinaison linéaire de ces termes :

$$D_i = d_{i1} T_1 + d_{i2} T_2 + ... + d_{ij} T_j + ... + d_{in} T_n.$$

On peut considérer que la majeure partie des facteurs d_{ij} est égale à zéro, car un document n'est représenté que par un nombre restreint de termes d'indexation. Les coefficients d_{ij} non nuls sont les poids du terme T_j pour le document D_i, calculés par l'une ou l'autre des méthodes décrites dans le chapitre sur l'indexation. Un document est donc un vecteur défini à partir de l'origine, ou de façon plus claire un point de l'espace euclidien défini par le référentiel $\{T_i\}$.

Prenons un exemple avec trois termes d'indexation.



On définit une mesure sur cet espace vectoriel, par exemple en considérant le produit scalaire de deux vecteurs. On estime que les vecteurs représentant les termes d'indexation sont orthogonaux et normalisés c'est-à-dire que les termes sont indépendants les uns des autres et ont tous la même importance dans le processus d'indexation. Dans ce cas, seul le poids d'un terme pour un document précis est pris en compte, même si ce poids est calculé en fonction de la fréquence générale du terme. Une mesure simple de la distance entre deux vecteurs est le cosinus de l'angle séparant ces deux vecteurs. Deux vecteurs orthogonaux (i.e. n'ayant aucun ternie en commun) donneront une valeur nulle du cosinus, alors que deux vecteurs confondus (i.e. ayant exactement les mêmes termes d'indexation avec le même poids) donneront une valeur de 1.

Une question Q est définie elle aussi comme un vecteur dans le même espace vectoriel :

$$Q= q_1T_1 + q_2T_2 + ...+ q_jT_j + ... + q_nT_n$$

La fonction de pertinence d'un document en regard de la question Q sera le cosinus de l'angle formé par la question et chaque document. Les documents sont alors retrouvés en fonction de leur proximité avec la question, même s'il ne correspondent pas exactement aux termes de celle-ci.

$$f_r(Q, D_i) = \mathrm{Cos}(Q, D_i) = \frac{\sum_{j=1}^{n} q_j \, d_{ij}}{\sqrt{\sum_{j=1}^{n} q_j^2 \sum_{j=1}^{n} d_{ij}^2}}$$

Les documents D_i sont classés en fonction de cette mesure. Cette possibilité de retrouver des documents de manière approchée est un des apports principaux du modèle vectoriel.

3.b - Modèle vectoriel et jugement de pertinence

Le modèle vectoriel a été le premier à pouvoir en pratique utiliser le jugement de pertinence pour reformuler les questions. A partir du moment où

documents et questions sont représentés de la même manière dans le même espace vectoriel et que leur distance est calculée à partir de cette représentation, un vecteur "document" et un vecteur "question" sont équivalent pour faire fonctionner le système. Dès lors, le choix de pertinence établi sur des "documents" vient enrichir la question "graine" posée à l'origine. Le vecteur de la requête se déplace pour se rapprocher des vecteurs représentant les documents pertinents et s'éloigner des vecteurs représentant les documents non pertinents. Le cosinus de l'angle se rapproche alors de 1 pour les documents les plus proches des premiers documents pertinents repérés.

Toutefois, afin de garder un poids prépondérant à la question originale, le modèle vectoriel permet de reformuler la question Q en fonction des documents pertinents (Pi) et non pertinents (Ni) :

$$Q = Q + a*\Sigma P_i - B*\Sigma NP_i$$

soit :

$$Q = \sum_{j=1}^{n} \left(q_j + \sum_{D_i \text{ pert}} a \,.\, d_{ij} - \sum_{D_i \text{ non-pert}} ß \,.\, d_{ij} \right) T_j$$

On peut cependant douter de l'efficacité de la prise en compte des documents non pertinents. Ceux-ci représentent en fait un nombre de documents très proche du nombre total de documents de la banque de données. Les valeurs de a et B permettent de moduler dans ce sens la prise en compte des deux types de documents lors du jugement de pertinence.

3.c - Avantages et limites du modèle vectoriel

Le modèle vectoriel est un élément fondamental de la recherche en science de l'information. Il a donné une base théorique solide (la théorie de l'algèbre linéaire) à ces recherches. Il n'est toutefois pas expurgé d'incohérences [RAG86]. En particulier, l'affirmation de l'indépendance des termes d'indexation,

notamment de leur répartition entre les documents pertinents et ceux qui ne le sont pas, est certainement très abusive.

Dans un espace vectoriel, si deux termes ne sont pas indépendants, il existe une combinaison linéaire permettant de décrire l'un en fonction de l'autre. On peut alors réduire la dimension du référentiel en éliminant tous les termes qui peuvent se réduire à une combinaison linéaire dans une base de l'espace vectoriel composé par un sous-ensemble des termes d'indexation. C'est intuitivement le travail que l'on fait en calculant les valeurs d'une matrice de occurrence des termes.

Cette démarche, mathématiquement cohérente, devrait alors nous conduire à considérer qu'un certain nombre de termes d'indexation sont inutiles, car pouvant se réduire à une combinaison d'autres termes. Une remarque qui est analogue à la logique de recherche de la Valeur de Discrimination d'un Terme (VDTj). Malheureusement, le monde des documents, donc celui du langage ne peut pas se réduire facilement à des termes "de base". Au contraire, c'est le flou, l'ambigu, l'indéterminé et la redondance qui fait aussi la force du langage. Il en va de même de l'indexation dès lors qu'on la considère comme un processus de communication.

De même, un document étant une combinaison linéaire de termes, on peut concevoir qu'il existerait une base de l'espace vectoriel formée d'un sous-ensemble des documents incorporés dans le système. C'est le choix qui sous-tend les méthodes d'agrégation, basées sur la matrice de similitude entre documents.

Ces deux approches, théoriquement correctes dans le cadre d'un espace vectoriel, sont contradictoires entre elles dans la pratique. Tant qu'il n'existe pas une base unique et minimale de description, une matrice de relation entre termes est indispensable pour conclure théoriquement à une matrice de similitude entre documents. Cette matrice doit pouvoir être obtenue en dehors même de la matrice termes documents. C'est par exemple ce qui est fait quand on définit un thésaurus : les liens sémantiques sont rapportés à des valeurs de pondération (dans un domaine de la connaissance donné) et ces valeurs sont utilisées ensuite dans la description des documents et éventuellement dans le calcul de pertinence lors de la recherche (cf. le système *Topic* de la société *Verity Inc* [VER89]).

Enfin, le modèle vectoriel ne tient pas assez compte de l'évolution du vocabulaire d'indexation. Un nouveau terme d'indexation ne naît pas par génération spontanée. Il était implicite dans certains documents avant même d'exister formellement. Par exemple, des documents sur l'hypertexte existaient bien avant que ce terme lui-même soit employé, qui plus est employé comme terme d'indexation (par exemple [KOV86] ou même [BUS45]). L'utilisation de ce terme, du strict point de vue du modèle est dès lors inutile, puisqu'un calcul pourrait obtenir la combinaison linéaire de termes qui serait équivalente à ce nouveau terme, notamment en la calculant à partir des documents qui auraient été décrits par le terme "hypertexte" si celui-ci avait été admis (i.e. les documents de l'agrégat dans lequel est inséré le nouveau document qui pour la première fois aurait pu être indexé par "hypertexte") . Or la pratique documentaire étant avant tout une pratique linguistique, l'invention lexicale et la modification sémantique des termes existant est un processus inéluctable, et surtout profitable.

Le modèle vectoriel reste cependant une notation pratique, car fortement corrélée avec des expériences intuitives. Nous connaissons tous des espaces vectoriels et avons appris à les manipuler (construction, habitat, circulation...). Il faut cependant concevoir les moyens de passer de cet espace aux dimensions rigides (d'un point de vue linguistique) à un espace topologique plus souple, où les termes et les documents seraient définis comme étant des suites convergentes de termes ou d'autres documents, ce qui laisse plus de place à l'improvisation linguistique. Une autre hypothèse d'évolution du modèle vectoriel est de considérer clairement le référentiel comme étant une base, c'est-à-dire un ensemble fermé et minimal dans lequel tous les documents et tous les termes d'indexation seraient décrits. Ce dernier point est étudié au sein de l'équipe de recherche en sciences de l'information de l'Université de Caen, et sera développé dans la troisième partie de cette thèse.

4 - Le modèle probabiliste

Quand un utilisateur est confronté à un système documentaire, il cherche à en extraire les documents pertinents pour sa question, et seulement ceux-ci. Or en général, il n'y a pas de relation stricte et connue entre les propriétés d'un

document (i.e. son indexation, prise au sens large) et sa pertinence. En revanche, il existe des *probabilités* que certaines propriétés d'un document rendent ce document pertinent pour une requête. Cette réflexion constitue la base pour la construction de modèles probabilistes de la recherche documentaire.

4.a - Recherche documentaire et prise de décision

En indexant manuellement un document, un indexeur juge la probabilité qu'un terme pris comme descripteur sera efficace pour que le document considéré puisse répondre à certains besoins. De même, en établissant son équation de recherche, un utilisateur juge que certains termes ont des chances importantes d'avoir été utilisés pour décrire des documents qui seraient pertinents pour son besoin documentaire. Il est possible que cette double assertion probabiliste fonctionne correctement (i.e., les termes de la question ont bien été utilisés comme termes d'indexation pour les documents adéquats). Toutefois, avec l'augmentation de la taille des banques de données et avec l'augmentation parallèle du nombre de termes d'indexation, d'autres phénomènes viennent contrecarrer ces premières affirmations : tous les termes ne sont pas équivalents, certains termes retrouvent un nombre trop important de documents pour qu'un utilisateur puisse les balayer, certains documents pertinents ne sont pas indexés par les termes jugés les plus probables par l'utilisateur (i.e. les termes de sa question)...

Une des méthodes probabilistes pour dépasser ce problème consiste à classer les documents par ordre de probabilité de pertinence décroissante *(Probability Ranking Principle - PRP)* ([MARO88], [RADE88]). Ce principe, qui est plus une heuristique qu'une loi, considère que l'efficacité globale d'un système documentaire, à partir des données dont il dispose (i.e. choix du type d'indexation et question posée), est accrue si les documents extraits par une requête sont classés en fonction de leur probabilité décroissante de répondre à cette requête, cette probabilité étant évaluée à partir des données dont le système dispose.

Une question découpe la banque de données en deux ensembles : les documents pertinents et ceux qui ne le sont pas, ce qui définit deux distributions de probabilités : $P1 = P (D_i \text{ pertinent} \mid D_i \text{ est représenté par } \{w_{ij}\ T_j\})$

$P2 = P (D_i \text{ non pertinent} \mid D_j \text{ est représenté par } \{w_{ij}\ T_j\})$

Ce sont les probabilités *a posteriori* de pertinence des documents.

Le théorème de Bayes permet de décomposer ces probabilités [RADE88]

$$P1 = \frac{P(D_i \text{ représenté par } \{ w_{ij}\ T_j \} \mid D_i \text{ pertinent }) * P(D_i \text{ pertinent })}{P(D_i \text{ représenté par } \{ w_{ij}\ T_j \})}$$

où :

. *P (D_i représenté par $\{w_{ij}\ T_j\}$ | Di pertinent)* est la probabilité qu'un document représenté par $\{w_{ij}\ T_j\}$ soit pertinent. Il s'agit de la vraisemblance de pertinence de D_i en fonction de sa représentation.

. *P (D_i pertinent)* est la probabilité *a priori* de trouver des documents pertinents

. *P (Di représenté par $\{w_{ij}\ T_j\}$* est la probabilité qu'un document soit représenté par $\{w_{ij}\ T_j\}$

Intervenir sur les probabilités *a priori* et sur les probabilités de représentation revient à intervenir sur la probabilité de sélectionner des documents pertinents lors d'une requête. Le théorème de Bayes permet donc de définir un critère opérationnel (un instrument de prise de décision pour le système).

4.b - Les trois modèles probabilités

Les deux premiers modèles probabilistes, représentés dans la littérature par modèle 1 et 2 jouent chacun sur l'une de ces deux probabilités.

Le modèle 1. proposé en 1960 par Maron et Kuhns [MAR060] définit la probabilité de pertinence comme étant le nombre de fois qu'un ensemble d'utilisateurs posant un terme Tj dans leur question jugera un document Di pertinent. C'est un modèle basé sur une vision probabiliste de l'indexation. Le

suivre revient à faire varier le poids des termes d'indexation dans la description du document, ce poids étant la valeur probabiliste qu'un document décrit par tel terme sera pertinent pour une question utilisant ce terme.

Le modèle 2. proposé en 1976 par Robertson et Sparck-Jones [ROB76], définit la probabilité de pertinence pour une question d'un ensemble de documents décrits par le terme Tj. Ce modèle implique une conception pondérée de la question, dans laquelle le poids de chaque terme sera la probabilité ci-dessus qu'un document ayant ce terme dans sa description soit en fait pertinent. Ce modèle 2 permet d'utiliser les méthodes probabilistes même avec des systèmes utilisant une indexation binaire (présence ou absence d'un terme dans la description) et semble utilisable pour construire des anté-serveurs [MAR088].

Ces deux modèles ont été réunis en 1982 par Robertson, Maron et Cooper [ROB82] dans un modèle généralisé, connu sous l'appellation modèle 3. Dans ce modèle, les termes d'indexation et les termes de la question sont pondérés.

Les modèles probabilistes permettent aussi d'implanter un système à jugement de pertinence. Les hypothèses d'origine étant modifiées par le jugement de l'utilisateur (i.e. la probabilité *a posteriori* de pertinence). Le modèle probabiliste permet aussi d'utiliser les découvertes de la théorie de la décision sous contrainte (risque conditionnel).

Une des limites du modèle probabiliste est la difficulté à évaluer certaines des probabilités. Les informations sur les documents sont en effet bien plus nombreuses que celles sur les questions et la conception d'un espace documentaire est plus aisée à concevoir et à faire évoluer qu'une distribution de probabilité pilotée par les questions. En particulier, le modèle probabiliste entraîne un nombre de calculs importants réalisés au moment de la question. Enfin, le modèle probabiliste, afin d'établir des probabilités crédibles part de l'hypothèse que les termes d'indexation sont indépendants les uns des autres, et équitablement répartis entre documents pertinents et non pertinents pendant la phase de jugement de pertinence. Or l'évidence sémantique vient bouleverser ce schéma : il y a une raison de l'ordre du sens pour que les termes ne soient pas indépendants.

5- Le modèle hypertexte

Le modèle hypertexte propose un mode particulier d'accès à l'information : la navigation. On peut en effet distinguer plusieurs types de recherche d'information :

- la *lecture séquentielle,* qui correspond au texte écrit, à l'image animée, au son...

- la *formulation de requêtes,* utilisée dans les banques de données ou les opérations de guichet (demande à un bibliothécaire, informations concernant la vie quotidienne - banque, assurance...)

- le *"butinage"* d'informations *(browsing),* qui fonctionne par association d'idées ou par approfondissement autour d'un point focal. Le "butinage" (ou "feuilletage" ou "flânerie") est l'opération typique de la recherche dans un dictionnaire ou une encyclopédie : à partir d'un point d'entrée, suivre les divers renvois (approfondissement, connaître le sens des mots employés dans une première définition...). Le "butinage" correspond aussi à l'attitude d'un lecteur devant les rayons d'une bibliothèque en libre accès.

Ces modèles de recherche d'information ne sont pas exclusifs les uns des autres. C'est le sens du modèle hypertexte de construire des systèmes d'information qui intègrent ces trois modes de recherche dans une opération générale de navigation.

Dans un réseau hypertexte, l'information est décomposée en blocs élémentaires (les *nœuds),* qui sont reliés entre eux par des liens qui autorisent le passage direct d'un nœud à l'autre. Cette notion de lien est l'essence même des systèmes hypertextes. Elle existait déjà dans d'autres formes d'organisation des informations. Par exemple, il existe des liens explicites dans les textes scientifiques (les citations), dans les textes juridiques (renvoi à des textes de loi, de décret ou de jurisprudence), dans les encyclopédies (renvois à d'autres articles) et des liens implicites dans certains types de documents qui renvoient à des illustrations à partir du texte (livres d'art, programmes de télévision...). Cependant, pour être qualifié d'hypertexte, le système doit fonctionner avec des

liens à l'action immédiate. Invoquer un lien doit provoquer instantanément l'action correspondante à ce type de lien.

5.a - Les quatre types d'outils hypertextes

C'est l'informatique qui permet vraiment la réalisation d'hypertextes. Pourtant, la première vision d'un système fonctionnant suivant ce mode associatif, le *memex* de Vannevar Bush [BUS45], fut imaginée sur la base de microfilms. A la fin de la guerre, Vannevar Bush, conseiller du président Roosevelt pour les affaires scientifiques, s'interrogeait sur le futur développement de la science, après les intenses efforts qui avaient été fournis par les chercheurs pour assurer le triomphe sur les nazis. Il pronostiquait l'avènement d'une époque où la maîtrise de l'information scientifique deviendrait un enjeu déterminant. Pour assurer une meilleure diffusion de la documentation, il établit le projet d'une machine permettant de tisser des liens analogiques entre les multiples documents scientifiques. Le *memex* qu'il décrivit dans son articles "As *we may think"* était une bibliothèque portable, basée sur la technologie des microfilms, permettant à l'utilisateur de noter ses propres commentaires, et de tisser ses propres chemins *(traits)* entre des éléments d'information. Il était certes conscient des problèmes technologiques à résoudre pour rendre son projet réalisable, mais entendait proposer une voie de recherche documentaire qui s'apparente à ce qu'il imaginait être fonctionnement intuitif de la mémoire humaine.

C'est avec l'informatique que les premiers éléments de réponse au projet de Bush purent voir le jour. Dans les années 60, Douglas Englebart a développé l'idée d'un ordinateur traitant des données symboliques, offrant les résultats sur des écrans, et laissant la décision à l'utilisateur dans une sorte de dialogue où l'ordinateur servait à opérer une "augmentation" du potentiel intellectuel de son utilisateur humain. A une époque où l'on nommait encore ces machines des "calculateurs", il faut reconnaître une longueur d'avance à cette vision. Engelbart développa ainsi le premier système hypertexte, NLS *(oNLine System),* en donnant une place prépondérante à sa dernière invention, la souris, qui permettait d'utiliser l'écran lui-même comme instrument du dialogue homme système. NLS, aujourd'hui commercialisé sous le nom *d'Augment,* accorde une place

particulière au travail en commun de plusieurs utilisateurs concourant à la rédaction d'un document (écriture, échanges, critiques).

A la même époque, Ted Nelson inventa le terme hypertexte, pour désigner un projet qui regrouperait toute la littérature sur un domaine, qui permettrait de circuler entre les textes par des liens associatifs, qui autoriserait des annotations par les différents "lecteurs", annotations accessibles à volonté, et qui assurerait une trace des diverses versions d'un document. Pour Nelson, chaque document s'inscrit dans le contexte de tous les autres et entretient des rapports explicites (la citation) ou implicites (analogie) avec une certain nombre d'entre eux. Son projet *Xanadu* d'une bibliothèque universelle se poursuit aujourd'hui. Ted Nelson, dans ce cadre, attache une importance fondamentale aux questions de droit d'auteur, à la fois pour préserver l'intégrité d'une œuvre (les annotations, renvois, critiques et citations ont un statut différent du document original) et pour assurer un retour à son auteur (copyright et reversements financiers en proportion de l'utilisation de son texte) [NELS88].

Ces trois projets dessinent une image générale de l'hypertexte :

- lecture active (annotation, liens analogiques)
- travail collectif autour d'un document ou d'un ensemble de documents
- cheminement personnel dont on peut éventuellement garder la trace (et même la faire partager par d'autres).
- découpage du texte en plusieurs "éléments" *(chunks,* gros morceaux), les liens permettant de passer instantanément d'un élément à un autre.

Ces concepts ont été utilisés pour d'autres types de projets. Jeff Conklin [CON87] distingue quatre types d'applications hypertextes utilisant les principes ci-dessus :

- les bibliothèques universelles *(macro-literary Systems)* pour lesquelles les liens entre documents et les liens de documents à commentaires (critiques, annotations,...) sont pris en compte dans le système. Les trois exemples cités ci-dessus en sont les principaux représentants.

- les outils pour la résolution de problèmes *(problem exploration tools)* permettent de prendre note des divers éléments déstructurés qui forment l'environnement d'un problème à résoudre. En autorisant l'utilisateur à tisser des liens entre ces éléments, ces outils permettent de dégager, dans le cours du travail d'élaboration lui-même, une cohérence et un projet.

Une version élémentaire de ces outils pour la résolution de problèmes est constituée par les "organisateurs de plans" *(outliners* ou *outline processors}* destinés à la rédaction de documents. Par exemple le "mode plan" de *Word IV*, ou le logiciel *More* sur *Macintosh.*

Les outils pour la résolution de problèmes sont plus particulièrement adaptés aux "problèmes faiblement structurés" *(wicked problems* de Horst Rittel, cité par [CON87]). On désigne ainsi un type de problème qui ne se conçoit qu'en fonction des réponses qu'on lui apporte. Il n'y a pas dans ce cas de succession rigide et organisée "problème -> réponse", mais une démarche qui permet de définir et préciser le problème en fonction du type de réponse que l'on peut lui apporter à un moment donné. Ainsi, de nombreuses tâches de prise de décision sont des problèmes faiblement structurés, car elles ne comportent pas en elles-mêmes de règles d'arrêt (i.e. règles permettant de dire que le problème est résolu). Ces tâches font dépendre l'arrêt ou la poursuite du processus de contraintes extérieures au problème (par exemple le manque de temps, d'argent ou même de patience). Les "problèmes faiblement structurés" ne reçoivent pas des solutions "justes" ou "fausses", mais seulement des solutions qui ont des degrés d'efficacité. Dans ce cas, le travail en collaboration et la capacité de chacun des participants de lier au système déjà en place ses informations, ses solutions et les jugements qu'il porte sur les apports des autres participants, sont des éléments de "résolution" déterminants. Le problème est alors organisé comme un hypertexte.

Les outils hypertextes destinés à la résolution de problèmes sont particulièrement adaptés à l'écriture collective de logiciels et à l'analyse de situations. Le représentant principal est *gIBIS* de MCC ([BEG88], [CON89]). Un des premiers systèmes opérationnels de ce type est le système d'aide à la décision *ZOG,* développé à l'université Carnegie-Mellon, qui est embarqué à bord du porte-avion nucléaire *USS Carl Vinson.* Ce système est aujourd'hui distribué sous *le nom de KMS* ([AKS88]).

- les systèmes de feuilletage ou de butinage d'information *(browsing Systems)* sont des systèmes permettant de circuler entre des éléments d'information, de les annoter, d'en extraire des parties. Ils concernent des domaines du savoir restreints et spécifiques ([SCA89]). Ils sont avant tout destinés à la consultation par le public. Dès lors, la qualité de leur interface utilisateur et les facilités d'utilisation sont déterminantes dans leur conception.

On retrouve ce type de systèmes dans les applications sur Disque Optique Compacts (par exemple le dictionnaire *Zyzomys* de *ACT-Informatique* et *Hachette)* ou sur vidéodisque (bornes interactives, disques du Musée du Louvre utilisant un *Macintosh* pour le pilotage...). L'Enseignement Assisté par Ordinateur est un domaine d'utilisation riche de perspectives. La documentation technique bénéficie aussi largement de ces outils (service des pièces détachées de *Renault* avec le logiciel *Hyperdoc).*

- les outils de réalisation d'hypertextes *(général hypertext technology)* qui sont à la fois des instruments pour gérer les données introduites dans l'hypertexte et pour construire les liens entre les éléments d'information, mais aussi des outils de rédaction des informations et d'intégration multimédia.

Le plus connu des outils de ce type offerts à la disposition des créateurs d'hypertextes est le logiciel *HyperCard* développé par Bill Atkinson pour le *Macintosh.* Il se compose d'un éditeur de liens, d'une boîte à outils pour la réalisation des écrans (dessins, fond de l'écran, texte, typographie...) et d'un langage de programmation associé *(HyperTalk)* permettant de personnaliser les applications.

Plus orienté vers la recherche dans le domaine de l'hypertexte, on trouve le système *NoteCards* développé au *PARC* de *Xerox (Palo Alto Research Center,* centre d'étude mythique d'où sont issues la majeure partie des innovations de l'informatique depuis 15 ans). *NoteCards* fonctionne sur des ordinateurs *UNIX* (stations de travail *Sun)* ([HAL88], [IRI89]).

Guide développé par Pete Brown à l'Université de Kent (G.B.) est diffusé par la société *Owl* pour *PC* et *Macintosh* ([BROW86], [BROW88], [DAC90b]). *Hyperdoc* de *GSI,* qui permet de gérer un nombre important de nœuds d'information sur

PC, est plus particulièrement destiné à la documentation technique, en intégrant des écrans à très haute résolution capables de présenter des plans ou des dessins industriels ([PAL89]). *CD-Navigator* est développé par *ACT-Informatique* pour la réalisation d'hypertextes sur D.O.C. et comporte une partie de recherche booléenne sur chaînes de caractères.

Ces quatre grands types d'applications permettent de cerner le concept d'hypertexte. Celui-ci reste toutefois dans une phase de développement, et il est encore difficile d'en concevoir une définition stricte et définitive.

5.b - La structure des hypertextes

Une structure hypertexte se compose principalement de trois éléments :

- une collection de *nœuds*. Les nœuds sont de taille et de structure variable. Ils peuvent contenir toute sorte d'informations : texte, graphiques, images fixes, images animées, logiciels, son...Cela conduit certains à parler "d'hypermédia", mais pour notre propos, nous conserverons le terme d'hypertexte, en laissant ouvert le type d'information contenu dans les nœuds. Il est évident qu'en fonction du type d'information, la grammaire du système est différente. Par exemple, il faut une fonction d'arrêt et de retour au nœud précédent à chaque fois que l'on a activé un nœud contenant de l'image animée, et éventuellement une fonction d'interruption ou d'arrêt sur image laissant le contrôle de la lecture d'information à l'utilisateur. Mais la notion d'hypermédia, en offrant une généralisation de surface (média au lieu de texte) ne permet pas plus de distinguer entre les grammaires diverses qu'il convient d'associer à chaque type d'information : texte, image fixe ou animée, son, graphiques ou logiciels.

- un réseau de *liens* permettant de naviguer d'un nœud à l'autre très rapidement. Cette capacité à invoquer des liens pour butiner entre les éléments d'information est la caractéristique principale d'un hypertexte. On ne peut toutefois parler de lien que s'il s'agit d'un appel direct (un pointage de souris ou la frappe de une ou deux touches de clavier), provoquant une réponse immédiate du système. Il existe plusieurs sortes de liens, qui définisent une syntaxe des hypertextes.

- un interface permettant d'invoquer des liens directement.

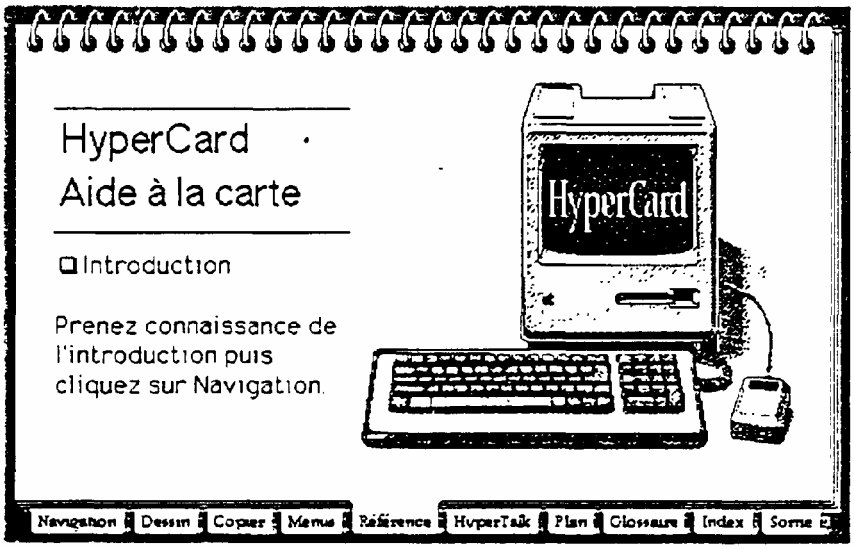
L'invocation peut se faire à partir de la lecture du contenu d'un nœud. On parle alors de "boutons" qui sont

. soit des points de l'écran dessinés à cet effet, représentés éventuellement par des icônes.

. soit des mots ou des expressions du texte (des parties soulignées ou grisées dans une image) qui sont mis en valeur (typographie, encadrement...),

. soit des zones de l'écran (ou de la fenêtre active dans le cas de systèmes multifenêtres) qui sont invisibles à la lecture, mais provoquent une modification du signe représentant le pointeur indiquant ainsi le passage sur un bouton.

L'invocation d'un lien peut aussi être guidée par une carte générale *(browser)* représentant les nœuds et les liens présents dans l'ensemble de l'hypertexte, chaque nœud de la carte étant sensible à l'action d'un pointeur.

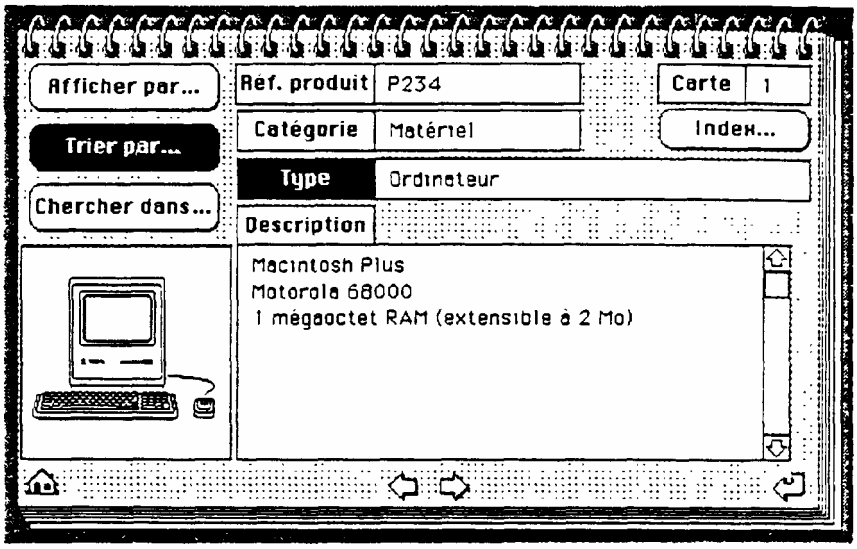


1 - les nœuds de l'hypertexte

Les nœuds d'un hypertexte se définissent d'abord par le type de contenu informationnel (texte, image, son, logiciel...).

Un nœud peut être représenté par une carte, c'est-à-dire par un contenant à la taille pré définie. Les cartes doivent alors contenir une unité syntaxique de l'hypertexte (un nœud) et une unité sémantique répondant à la division en éléments d'information. La taille limitée des cartes rend cette opération complexe. Pour contrer ce problème, on peut définir des nœuds de texte, dont la taille de visualisation (unité syntaxique correspondant à un écran ou à une fenêtre) est-indépendante de la taille du texte (unité sémantique). On fait ensuite défiler le texte (ou l'image...) par un ascenseur, ou tout autre moyen traditionnel.

Certains nœuds correspondent à des informations structurées, par exemple un carnet d'adresse ou un répertoire bibliographique. Dans ce cas, la carte correspondante est pré définie par un "fond de carte" commun qui permet la saisie et la visualisation des informations.



On peut aussi déterminer une typologie des nœuds, et en regard une typologie des liens qui appellent et qui partent de ces nœuds typés ([DAC90b], [BEG88]). On trouve ainsi des nœuds spécifiques des annotations d'auteur, des commentaires d'utilisateurs, et divers types de nœuds répondant aux buts spécifiques poursuivis par l'hypertexte. Par exemple si l'hypertexte s'inscrit dans

un système d'aide à la décision, on peut concevoir des nœuds pour enregistrer les les faits, d'autres pour les contraintes, d'autres encore pour les décisions prises ou proposées...Chaque type de nœud ne pourra s'insérer dans l'hypertexte que par un certain type de lien, tissé avec un certain type de nœud, ces précisions étant définies par le concepteur .

On peut aussi regrouper des nœuds élémentaires dans des nœuds composés pour obtenir une meilleure cohérence de l'hypertexte. Les nœuds composés fonctionnent ensemble et sont traités comme un seul nœud du point de vue offert par les autres nœuds. Ces nœuds composés permettent de réaliser des cartes de navigation plus lisibles.

2 - le réseau des liens

Le réseau des liens forme l'ossature sensible d'un hypertexte. On peut distinguer deux grands types de liens :

. les liens organisationnels permettent d'organiser les divers éléments d'information suivant des schémas traditionnels. En général, ces liens sont appelés par des boutons spécifiques, qui restent présents sur tous les écrans : *"aller au début", "nœud suivant", "nœud précédent"*. Les tables des matières permettent de structurer les éléments d'un hypertexte comme le serait un texte imprimé. Ce type de lien, parce qu'il correspond à une pratique multi centenaire reste le plus employé.

. les liens associatifs sont l'apport nouveau de l'hypertexte. Ils sont en général intégrés à l'intérieur même du contenu du nœud en cours de lecture. On peut à ce niveau distinguer ([DAC90b], [CON87]) :

. *les liens d'annotation,* qui permettent d'afficher une note dans une fenêtre spécifique de l'écran. Plusieurs stratégies sont employées pour lier la note au texte, en fonction de la taille de la note : ouverture d'une fenêtre permanente, avec un "ascenseur" pour lire tout le texte de la note, ou bien ouverture temporaire d'une petite fenêtre à l'endroit même du bouton appelant, qui se referme dès que l'on lâche la souris. On peut aussi concevoir des liens

d'annotation dont l'invocation appelle une information d'un autre type, par exemple une image d'illustration, un graphique, un document sonore...

. *les liens d'inclusion,* qui étendent le texte de référence sur un point précis. Un nouveau texte est inséré à l'endroit du bouton. Cette relation est semblable aux passages en petits caractères de certains livres, que l'on peut sauter en première lecture, mais qui offrent des précisions importantes. Ce lien est aussi utile pour avoir des indications sur le contenu d'un chapitre dans une table des matières. Ce lien peut éventuellement être doté de critères de confidentialité (comme dans les documents structurés - normes O.D.A. par exemple) ou de critères d'accessibilité (certaines informations ne sont accessibles qu'aux utilisateurs ayant obtenu un certain niveau de connaissance dans des applications d'enseignement assisté par ordinateur). L'existence d'un lien d'inclusion ne se conçoit que dans des systèmes ne limitant pas le contenu d'un nœud à la taille d'un écran. On parle alors de systèmes orientés vers le texte, à l'opposé de systèmes orientés vers les cartes pour lesquels la taille du nœud est établie avant la détermination de son contenu.

. *les liens de référence* qui permettent de passer d'un nœud à l'autre. En général, le point de départ est un bouton du nœud d'origine, et le point d'arrivé est constitué par un autre nœud. On peut toutefois concevoir que l'arrivée soit seulement une partie d'un autre nœud, et que l'appel soit une région entière du texte d'origine. Cette extension du lien de référence est difficile à maintenir, notamment si les contenus des nœuds de départ et d'arrivé sont modifiés. L'image première du lien de référence est celui de la citation bibliographique, mais on peut aussi le retrouver dans le "jeu du dictionnaire", qui permet d'appeler la définition des mots présents dans une définition donnée.

Plus difficile à concrétiser est la constitution de liens strictement analogiques, qui feraient correspondre le contenu global de deux nœuds. La proximité sémantique de deux nœuds est alors

. déterminée au moment de la conception par un "auteur" d'hypertexte,

. calculée par des méthodes d'agrégation de nœuds similaires à celles évoquées plus haut pour les documents,

. tracée par les utilisateurs au fur et à mesure de la construction de

cheminements particuliers qui sont conservés pour les utilisateurs futurs (cf. le *memex* de Vannevar Bush, ou le projet *Xanadu).*

Il n'existe pas encore de typologie établie des liens et des nœuds hypertextes, ni de spécification des actions provoquées par tel ou tel type de lien, ni de définition des attributs qui peuvent être associés à un lien (critères de confidentialité, accessibilité, changements typographiques...). Une telle normalisation syntaxique devrait permettre un meilleur échange entre les expériences, la possibilité d'accès aux hypertextes par des matériels hétérogènes et assurer une réutilisation des divers hypertextes constitués. Cette normalisation tendrait à dégager l'hypertexte, comme nouveau produit informationnel, du logiciel qui a présidé à sa création. La diffusion gratuite et massive du logiciel *HyperCard* a provoqué la création et la diffusion de *stackware,* piles de cartes organisées en hypertexte. La pérennité de ce type de travail n'est pas assurée, ni le transfert sur d'autres systèmes informatiques, ni même la possibilité d'y accéder à distance avec des outils généralistes de consultation.

Cette normalisation de la syntaxe des hypertextes ne peut cependant pas précéder la recherche d'une rhétorique spéciale à ce type d'organisation des informations [DAC90b]. Le texte écrit, linéaire, sait depuis des siècles indiquer par toute une série de connecteurs linguistiques, de formulations, ou même de construction de phrases, les diverses parties et intentions d'un texte. Le lecteur est guidé par l'auteur qui indique une annotation, un exemple, une digression, un résumé, un retour en arrière...Le discours oral, par l'apport supplémentaire de l'intonation, des pauses et du rythme sait mieux encore faire comprendre les articulations entre les éléments d'information. Enfin, le discours audio-visuel, bien qu'il se cherche encore, connaît des procédés, reconnus par tous, destinés à guider le spectateur : voix off, plans rapprochés, alternance champs/contre-champ, insertion de plages contemplatives entre deux interviews... Il faut qu'émergé un consensus du même type sur la construction d'hypertextes. Les artifices rhétoriques propres à l'hypertexte sont encore largement de l'ordre de l'idée *a priori* sur l'utilisation possible d'un stock d'information. En ce sens, les liens traditionnels d'organisation et de hiérarchisation du texte restent les liens dominants. Il y a dans ce domaine une piste de recherche productive, réunissant informaticiens, linguistes, spécialistes de l'audio-visuel, journalistes, psychologues ou praticiens et théoriciens de

l'éducation. Il semble en effet nécessaire de partir d'une analyse des méthodes d'apprentissage, et des méthodes de recherches d'information dans un univers peu structuré pour concevoir des méthodes hypertextes efficaces.

Les recherches portant sur cette compréhension des problèmes d'organisation des informations en hypertextes induisent des interrogations plus épistémologiques sur ce mode de diffusion de l'information [DOL88]. Le choix de découpage en noeuds, la définition des liens, et éventuellement des attributs d'accès afférents sont de la responsabilité de l'auteur, et de ce fait introduisent des biais subjectifs. Acceptée dans le texte "linéaire", cette influence de l'auteur sur le produit d'information est loin d'être reconnue dans le domaine de l'hypertexte, ses promoteurs aspirant à la "neutralité" en laissant à l'utilisateur le choix d'organiser sa lecture. Or, l'organisation de l'hypertexte, loin d'être un mode "neutre" permettant de proposer à chaque lecteur l'ensemble des informations, en lui laissant la liberté totale de lire ce qui l'intéresse, comporte aussi des présupposés (pourquoi telle information est-elle associée à un noeud ?), qui peuvent conduire à occulter des parties de l'information ou induire une lecture "idéologique" de certaines liaisons (pourquoi ce nœud est-il relié à tel autre ?). Ces aspects de la construction d'un hypertexte sont d'autant plus forts que nous ne connaissons pas bien la grammaire des hypertextes et les modes de lecture des utilisateurs.

5.c - Navigation et butinage

La navigation dans l'information, le butinage, sont des activités quotidiennes (lecture du journal, fréquentation des lieux publics, lecture des panneaux indicateurs routiers ou des cartes routières...) qui restent pourtant rebelles à la modélisation. Dans le domaine plus spécifique des hypertextes, on peut penser que la navigation procède de trois options :

- rechercher un mot-clé (chaîne de caractères, descripteurs ou équation de recherche booléenne) dans les nœuds d'information. Ce mode de recherche s'apparente aux modes de requête des banques de données. La capacité à retrouver et à classer les nœuds d'information en fonction de leur pertinence pour une question d'utilisateur est similaire aux problèmes évoqués avec les autres modèles documentaires. L'indexation des nœuds se pose de la même

manière, et les hypothèses allant de l'indexation en texte intégral à l'indexation par mots-clés, en passant par une pondération des descripteurs, sont retenues.

- invoquer un lien à partir d'un nœud. Présentés à l'écran, les liens permettent de suivre les informations associées (hiérarchiquement ou analogiquement) au nœud en cours de consultation. L'utilisateur décide ou non d'invoquer un lien et choisit donc son parcours de lecture. C'est l'aspect butinage de l'information. Le chemin ainsi défini par un utilisateur particulier est conservé dans un historique qui peut être affiché à l'écran pour permettre un retour en arrière.

- utiliser une carte générale du contenu de l'hypertexte *(browser)* pour situer un nœud et connaître les autres nœuds associés et le type de liens existant entre eux. Chaque nœud de la carte générale est représenté par un mot-clé, un icône ou une représentation en réduction (par exemple les imagettes d'un imageur documentaire).

Il semble utile de distinguer *navigation* et *butinage* qui sont les deux types d'activité permettant d'utiliser les potentialités particulières des hypertextes, c'est-à-dire la circulation suivant les liens ([MCA89], [FOS89]). On réservera le terme de *navigation* à la circulation utilisant une carte de navigation. Cette activité représente une action réfléchie et contrôlée à partir d'un projet général, ayant une destination particulière. Le *butinage* est obtenu par l'invocation directe des liens à partir des boutons. Il s'apparente plus à la flânerie au sein de l'univers informationnel, et constitue une activité cognitive plus difficile à modéliser. Par voie de conséquence, l'activité de butinage fait porter de plus lourdes responsabilités sur le concepteur du système s'il désire éviter que l'utilisateur ne soit *"perdu dans l'hyper-espace"*.

Ray McAleese [MCA89] distingue deux méthodes de butinage :

. un butinage spécifique, dans lequel l'utilisateur recherche des informations avec un but défini précisément

. un butinage thématique, qui correspond à un processus exploratoire, où la circulation dans les informations se fait avant de définir les frontières de la recherche.

L'enjeu du modèle hypertexte est de favoriser cette recherche exploratoire, répondant à un but qui n'est pas encore précis dans l'esprit de l'utilisateur. Un hypertexte doit être structuré afin de faciliter le butinage, mais doit aussi permettre de filtrer l'information en fonction des buts d'un utilisateur, lui offrir des instruments pour planifier sa recherche (des *"cartes de navigation")* et lui permettre de déterminer le niveau de détail dans l'information qui lui est nécessaire.

De ce point de vue, on peut distinguer cinq stratégies d'utilisateurs :

. le balayage *(scanning)* qui permet de couvrir un thème sans descendre dans les détails

. le butinage *(browsing)* où l'utilisateur poursuit un chemin jusqu'à la satisfaction de son besoin d'information

. la requête *(searching)* qui correspond à un but précis et bien défini

. l'exploration *(exploring)* permet de couvrir toute les perspectives de l'information recueillie

. le vagabondage *(wandering)* qui est une recherche d'information sans objectif défini et qui consiste en un cheminement non structure parmi les éléments d'information.

Les divers types d'interface choisis pour un hypertexte privilégient certaines de ces stratégies. Globalement, on peut distinguer deux grands types d'interfaces :

. *l'interface syntaxique* qui propose un langage intégré dans les éléments d'information. Cet interface est basé sur le repérage de"boutons" dans le contenu d'un nœud (texte, icônes, "points chauds"...) (e.g. HyperCard)

. *l'interface graphique* qui propose une représentation générale du contenu de l'hypertexte en indiquant les nœuds et les liens (par exemple une table des matières ou une carte de navigation) (e.g. *NoteCard)*

McAleese distingue ainsi l'adaptation d'un type d'interface en fonction des stratégies de recherche :

	Interface syntaxique (représentation des chemins)	**Interface graphique** (représentation du terrain)
Balayage	**	**
Butinage	***	*
Requête	**	*
Exploration	*	**
Vagabondage	*	***

5.d - Avantages et limites des systèmes hypertextes

Les hypertextes permettent à chaque utilisateur de retrouver des informations à partir de modes d'approche différents. Le texte est en quelque sorte organisé de manière différente pour chacun des lecteurs. Ils permettent de plus d'engager une lecture active, en annotant les textes, en traçant des chemins particuliers. En revanche, les hypertextes posent plusieurs problèmes aux utilisateurs, que Carolyn Foss [FOSS89], nomme :

. le "problème des digressions enchâssées" *(the embedded disgressions problem)*

. le "problème du musée d'art" *(the Art Muséum problem).*

Dans le premier cas, l'utilisateur suit des chemins de traverse et finit par perdre le fil de sa recherche originale. Les buts qu'il s'était fixé peuvent se perdre au cours de ce voyage dans l'information. La seconde métaphore représente la situation d'une personne qui a vu de nombreux éléments d'information, mais qui finit par ne plus savoir les distinguer les uns des autres, et par ne plus savoir généraliser à partir de ces éléments d'information pour en faire un savoir cohérent. Ce problème est déjà reconnu par les analystes de l'audiovisuel qui décrivent la manière dont les médias audiovisuels provoquent une *"culture mosaïque"* où les éléments d'information sont certes vus et parfois même connus, mais ne constituent pas pour autant un savoir cohérent ([POR76]).

L'utilisateur d'un hypertexte peut se trouver perdu dans l'hyperespace, et ne plus savoir dans quelle direction prolonger ses recherches. C'est l'objectif de la "carte de navigation" *(browser)* de montrer l'environnement d'un nœud en le re-situant dans l'ensemble de l'hypertexte. Toutefois, on peut alors se trouver débordé par la masse d'information, d'autant qu'il faut représenter les nœuds aussi bien que les liens !

Deux méthodes permettent de circonscrire ce problème :

. établir un filtre qui ne propose sur la carte qu'un certain nombre de nœuds, reliés par des liens d'un certain type. Ces techniques de filtrage seront certainement favorisées par une meilleure connaissance d'une syntaxe des hypertextes, comme souligné plus haut.

. proposer une vue particulière de l'hypertexte, centrée sur le nœud en cours de lecture et qui privilégie les nœuds environnants (au sens où les liens directs "rapprochent" des nœuds) et tend à faire se confondre les nœuds plus distants. Cette présentation, dite *"fish eyes view",* par analogie à certaines photographies prises avec un objectif à très courte focale *(fish eye), a* été proposée par Furnas [FUR86].

L'objectif général d'un concepteur d'hypertexte doit être de lutter en permanence contre la désorientation de l'utilisateur, et contre les problèmes de désorganisation cognitive, qui font perdre le sens des objectifs de recherche d'information qui étaient à l'origine de la consultation. C'est aussi l'enjeu passionnant des recherches à venir dans ce domaine.

5.e - Hypertexte et banques de données

Le concept d'hypertexte est en lui-même un modèle d'organisation des informations. Il permet cependant de regarder différemment de nombreuses difficultés rencontrées dans l'utilisation des banques de données. Deux éléments des systèmes hypertextes se retrouvent dans les recherches sur les banques de données :

. l'existence de liens entre documents qui permettent de faire aisément passer d'un point à l'autre de l'hypertexte, ces passages étant

matérialisés par une carte de navigation. Cette approche est analogue à la réalisation d'agrégats dans les banques de données. La réalisation d'une cartographie de la science au travers des documents reprend une partie des problèmes de conception et de lisibilité des cartes de navigation des hypertextes.

. la capacité à utiliser les informations contenues dans un nœud (document) comme signal pour invoquer un autre élément d'information. On retrouve alors les concepts de jugement de pertinence et de reformulation dynamique des requêtes.

Dès lors, les recherches sur la syntaxe des hypertextes et sur la résolution des problèmes de désorientation et de désorganisation cognitive vont pouvoir nous aider à établir de nouveaux types d'interfaces pour les banques de données, en considérant un document comme un nœud d'un hypertexte, les relations de similitude (descripteurs communs, co-citation, relations sémantiques...) comme des liens. La requête primaire devient alors une voie d'entrée dans le réseau hypertexte. Les résultats d'une requête peuvent aussi être considérés comme des nœuds composites, ouvrant l'accès à certains documents particuliers et considérés comme équivalents du point de vue de la requête (des agrégats).

Considérer les banques de données sous cet aspect nous conduit à envisager une modification profonde de l'opération de recherche documentaire sous deux angles :

. il faut que le terminal de consultation dispose des attributs d'un système hypertexte, notamment un écran graphique permettant le multi fenêtrage et l'utilisation d'un instrument de désignation (en général une souris). La souris pourrait être utilisée pour passer en mode bascule d'une version courte (titre-auteur-descripteurs) à une version longue (document complet) par un "lien d'annotation". Les divers "objets cognitifs" actifs au cours d'une recherche documentaire (écran de requête, écran historique, réserve de documents, écran du jugement de pertinence...) seraient activés chacun dans une fenêtre du terminal de consultation, comme dans les systèmes hypertextes.

. il faut que la recherche documentaire devienne une lecture active, avec la possibilité pour un utilisateur d'annoter les références qu'il sélectionne (pourquoi est-elle conservée, rangement dans un dossier particulier...)

et la conservation d'un historique du chemin parcouru, et plus encore d'un historique des nœuds considérés comme satisfaisants pour le besoin documentaire de l'utilisateur.

Ces deux points ne renvoient pas seulement à des questions scientifiques, mais aussi à des questions organisationnelles et économiques. Actuellement, la consultation de banques de données est une opération en trois temps : établir la requête, récupérer les résultats, lire et exploiter les résultats, en général après avoir quitté le système documentaire. Cette articulation des opérations, contrainte par des motifs économiques (mode de facturation de la recherche documentaire) doit être complètement dépassée pour intégrer le modèle hypertexte et les banques de données. Dans ce domaine, les Disques Optiques Compacts, parce qu'ils bénéficient de l'interface graphique des micro-ordinateurs, et parce que leur utilisation n'est pas dépendante du temps passé, peuvent présenter une première étape, permettant l'étude et la mise au point de ce type de système. On peut alors d'autant plus regretter que les D.O.C. contenant des banques de données ne soient en général que des versions "en livre de poche" des banques de données en ligne.

Quelques idées émanant du modèle hypertexte ont été utilisées pour augmenter les capacités des systèmes traditionnels de recherche documentaire. Ainsi, le logiciel *TINman* (et son application aux catalogues de bibliothèques *TINlib)* de la société britannique *I.M.E. (Information Management & Engineering)* permet d'utiliser le résultat d'une première recherche documentaire pour pointer des mots apparaissant à l'écran (noms d'auteurs, mots-clés....). Cette opération relance la recherche en utilisant le terme repéré comme nouvel attribut. Le procédé est séduisant, si l'on considère qu'il est plus facile de reconnaître et sélectionner une information que de deviner ce que contient le système. Cependant, la désorganisation cognitive soulignée plus haut est alors très forte. Dans les systèmes utilisant le jugement de pertinence, un ensemble de caractéristiques, calculé sur plusieurs documents, est utilisé pour relancer la recherche. Ici, seuls les termes pointés sont pris en compte. La recherche devient alors beaucoup plus sensible aux limites de l'indexation, et la dispersion due aux ambiguïtés des termes employés est plus forte. Ce logiciel marque cependant une première prise en compte dans le domaine des catalogues de bibliothèque de la notion de boutons interactifs, notion propre aux systèmes hypertextes. Les limites soulignées montrent cependant que cet aspect des

hypertextes ne peut pas être isolé des réflexions globales sur le modèle de navigation d'un utilisateur dans l'information.

L'hypertexte ne peut être considéré comme un "gadget", dont on pourrait utiliser les aspects de surface (le *look and feel)* en délaissant la structuration profonde des informations et le concept de carte de navigation qui répondent à une nouvelle grammaire de l'information.

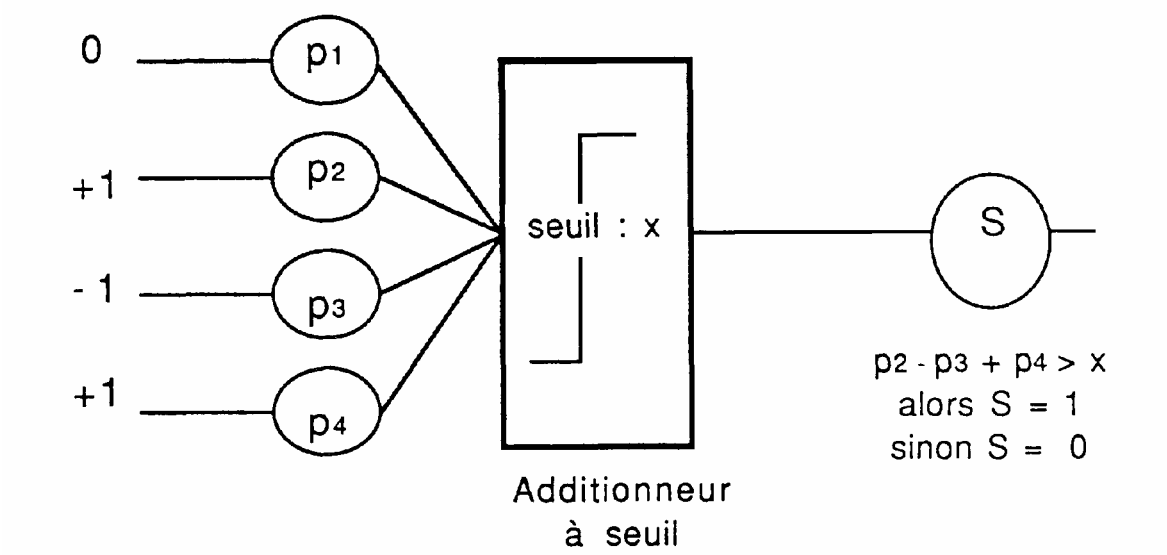
6 - Le modèle connexionniste

Le modèle connexionniste est né de réflexions sur le cerveau humain et ses capacités particulières de mémorisation. A l'opposé du paradigme de Von Neuman, qui sépare mémoire et programme, et qui s'interdit toute capacité de généralisation (on ne retrouve dans la mémoire que ce que l'on y a versé ou qui en est une conséquence directement calculable), les connexionnistes veulent trouver des modèles informatiques capables de généralisation, capables de retrouver des informations même tronquées, capables de fonctionner malgré des pannes localisées.à l'instar du cerveau humain.

Pourquoi reconnaissons-nous quelqu'un qui s'est laissé pousser la barbe ? Certainement pas parce que nous retrouvons dans un coin de notre cerveau une image que nous y aurions stockée puisque nous ne le connaissions que sans barbe. Au contraire, notre mémoire nous permet de reconnaître des éléments d'information, même si l'information est perturbée, distordue ou incomplète. Nous ne procédons certainement pas par "indexation" (une indexation d'un individu sans barbe) puis par recherche d'une information décrite par des éléments d'indexation. Il semblerait au contraire que l'information soit répartie entre de multiples neurones, sans qu'aucun d'eux ne soit affecté d'un sens précis, mais que seul l'ensemble permette un travail de mémoire. Cette structure est très compétente, notamment elle garantit une continuité de fonctionnement de notre cerveau malgré la perte régulière et massive de neurones. Elle permet aussi des généralisations à partir d'éléments incomplets (la susdite reconnaissance d'un individu avec ou sans barbe). Elle possède aussi ses revers, notamment un manque de fiabilité dans la recherche d'une information précise et les

glissements entre éléments mémorisés (le rêve). Mais même ces éléments sont une chance pour l'inventivité de notre pensée : nous pouvons obtenir des liens associatifs entre éléments d'information mis en mémoire.

Le connexionnisme, même s'il prend appui sur une conception de l'activité de la mémoire ne se veut pas pour autant une modélisation exacte du cerveau. Il s'agit plus modestement de se situer au niveau fonctionnel en utilisant des modèles d'automates interconnectés. Les premières idées furent développées par les cybernéticiens McCullogh et Pitts en 1943. Ils montrèrent alors que des fonctions logiques pouvaient être calculées par des automates simples constitués en réseau. Ces unités simples par analogie avec le cerveau furent désignées comme *"neurones formels"*. Les neurones sont capables de réaliser des "échanges synaptiques". Un "neurone formel" se comporte comme un additionneur à seuil de toutes les valeurs qui lui sont apportées par les autres neurones, chaque participation d'un neurone étant modulée par les *"poids synaptiques"*.



6.a - Les réseaux de neurones formels

Françoise Fogelman-Soulié [FOG87] considère qu'un neurone formel est un processeur élémentaire défini par :

. son état interne *a* (activité)

. des connexions avec d'autres automates ou l'environnement

. une fonction de transition *f* qui lui permet de calculer son état

interne *a* en fonction des signaux qu'il reçoit sur ses connexions : pour un neurone *i,* on a : $a_i=f(a_1, a_2,...,a_k,...a_n)$

où a_k représente l'activité du neurone *k,*

et *n* désigne le nombre de neurones connectés par des liaisons synaptiques au neurone *i.*

On considère de plus que l'apport d'un neurone *k* est modulée par "l'efficacité synaptique" de la liaison entre ce neurone et le neurone *i.* Cette efficacité est mesurée par un *"poids synaptique"* de la liaison : w_{ki}. On distingue des effets coopératifs (l'activité du neurone *k* fait augmenter l'activité du neurone *i* : w_{ki} est positif) et des effets inhibiteurs (l'activation du neurone *k* fait diminuer celle du neurone *i* : w_{ki} est négatif).

Le calcul de a_i devient :

$$a_i = f\left(\sum_{k=1}^{n} w_{ki}\, a_k\right)$$

La fonction utilisée permet "d'exciter" ou non un neurone en fonction des intrants. On peut concevoir une fonction à seuil (le neurone ne s'active qu'au delà d'une certaine valeur), une fonction à saturation (l'activité du neurone est linéaire entre deux valeurs données des intrants, mais fixe en dehors de ces valeurs) ou d'autres type de fonctions ayant en général des asymptotes.

1 - Le perceptron

Une des premières applications connexionnistes est le *perceptron* de Frank Rosenblatt ([MINS88]). Un perceptron est composé d'un ensemble d'automates regroupés en trois couches :

. *une "rétine"* qui reçoit les informations du monde extérieur. Elle est composée de *n* cellules dont l'activité x_k est 0 ou 1 *(k* est compris entre 1 et n).

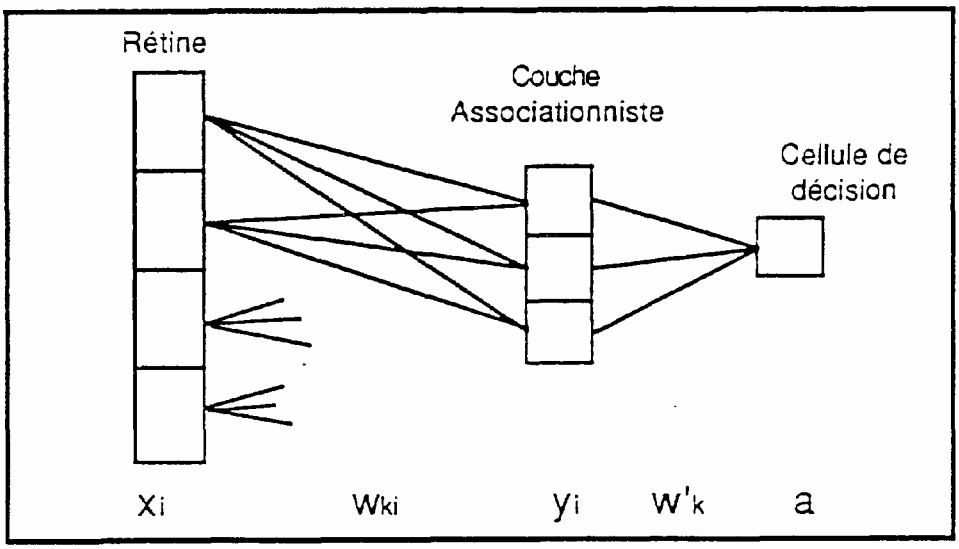
. *une couche associationniste,* ayant *p* cellules *(p<n).* Chaque cellule de la couche associationniste est reliée à toutes les cellules de la "rétine".

L'activité de la cellule *i* est : $$y_i = \sum_{k=1}^{n} w_{ki}\, x_k$$

Toutefois, certains poids *Wki* peuvent être fixés à 0, afin de créer des "masques" sur la rétine. Les poids de liaison entre la rétine et la couche associationniste sont définis par le concepteur du perceptron et ne peuvent être modifiés par apprentissage.

. *une couche de "décision",* composée d'une cellule reliée à toutes les cellules de la couche associationniste. Son activité est

$$a = f\left(\sum_{k=1}^{p} w'_k\ y_k\right)$$



Le perceptron est capable de reconnaître des objets déformés. Il peut ainsi reconnaître une lettre (la rétine est composée de points d'activité 1 ou 0 qui dessinent la lettre proposée) si elle lui a été apprise. La cellule de décision est excitée si la lettre présentée est bien celle apprise, ou reste éteinte dans le cas contraire (i.e. son activité reste inférieure à la valeur de décision). Les divers poids synaptiques w'_k , définissant les liens existant entre les cellules de la couche associationniste et la cellule de décision, sont adaptés par apprentissage. Cette notion d'apprentissage est une des caractéristiques fondamentales des systèmes de neurones formels. L'apprentissage consiste dans ce cas à présenter un jeu d'exemples, et à modifier les poids en fonction des réponses (correctes ou incorrectes) du perceptron.

En 1969, dans un ouvrage célèbre et récemment réédité ([MINS88]), Minsky et Papert ont sonné le glas du perceptron et des recherches connexionnistes en démontrant que le perceptron ne pouvait pas classifier des problèmes qui n'étaient pas linéairement séparables, en particulier la fonction logique "OU exclusif (XOR).

Le développement de nouvelles méthodes d'apprentissage ont cependant permis un renouveau des recherches connexionnistes dans les années 80. On distingue actuellement trois grand types de réseaux de neurones :

- les réseaux récurrents (réseaux de Hopfield)
- les réseaux en couches à apprentissage supervisé
- les réseaux en couches à apprentissage non supervisé (réseaux de Kohonen).

2 - Réseaux récurrents

Les recherches connexionnistes ont été encouragées par la publication en 1982 d'un article de John Hopfield [HOP82]. Celui-ci décrit un réseau de neurones dans lequel chaque neurone reçoit des informations de tous les autres et en envoie à tous les autres. Dans ce système de type coopératif, la décision est prise en plusieurs étapes, quand le système a atteint un état d'équilibre. On parle alors de *"réseaux récurrents",* car on ne peut connaître la valeur de l'activité d'une cellule sans connaître celle de toutes les autres. L'état stable, s'il existe, est obtenu par itérations successives du calcul d'activité de toutes les cellules en fonction des variations qui ont eu lieu lors de la "passe" précédente. Ces états stables sont dits *"attracteurs".* Le réseau apprend quand on lui présente un jeu de "prototypes", et que l'on minimise la fonction d'erreur entre une réponse connue et les réponses trouvées par le réseau. A la fin de l'apprentissage, chaque prototype différent est associé à un attracteur stable ce qui permet de définir une classification dans l'ensemble des prototypes. Si l'on présente un nouveau cas, le réseau se stabilisera suivant l'attracteur le plus proche de la réponse.

Un réseau de ce type a été mis au point à l'Ecole Supérieure de Physique et de Chimie Industrielle [PER88] pour reconnaître les dix chiffres dans le cadre d'un projet de lecture directe des codes postaux manuscrits. Le réseau est constitué de 600 neurones interconnectés. Il sait reconnaître 80 % des chiffres.

Dans 10 % des cas, il attribue une valeur erronée, et dans 10 % il se stabilise dans un état inconnu. Les réseaux de Hopfield ont introduit une reprise importante des recherches connexionnistes dans les années 80.

3 - Réseaux en couches à apprentissage supervisé.

Parallèlement aux réseaux récurrents, les chercheurs ont mis au point de nouvelles méthodes d'apprentissage pour des réseaux qui reprennent en l'élargissant la conception du perceptron. L'idée de base est de permettre une modification par apprentissage des poids synaptiques des liaisons existant entre la "rétine" et la couche associationniste. On obtient alors un modèle de réseaux en couches. On distingue les réseaux à apprentissage supervisé et les réseaux à apprentissage non supervisé, dits réseaux de Kohonen.

La méthode d'apprentissage supervisé la plus célèbre est dite méthode de *"rétro propagation du gradient"*. Elle consiste à évaluer la différence entre la décision prise par le réseau et la décision souhaitée et à contraindre une modification des poids synaptiques en utilisant la couche de décision comme point d'entrée. Dans ce cadre, la couche de décision peut comporter plusieurs neurones.

Si S_k est la sortie calculée par le réseau quand on présente l'exemple x_k , et y_k la sortie désirée, on peut déduire l'erreur sur chaque neurone *i* de la couche de sortie par $e^k_i = S^k_i - y^k_i$.. On modifie alors les poids des liaisons synaptiques en retranchant une valeur calculée à partir du gradient de la situation antérieure et de cette valeur de l'erreur.

Ce type de réseau est à la base du programme NETtalk de Terence Sejnowski [DUR87]. Ce programme convertit du texte anglais dactylographié en parole. Le réseau comporte 309 neurones organisés en trois couches. En entrée, une lettre est codée sur 29 cellules, le système examine 7 caractères en même temps. Cette couche d'entrée de 203 neurones est totalement connectée à une couche associationniste intermédiaire de 80 cellules, qui sont reliées à une couche de sortie de 26 cellules, représentant 50 phonèmes. Le réseau produit lui-même le codage des informations, grâce à un apprentissage de 12 heures (de temps machine) réalisé par comparaison entre un texte (lu par la couche d'entrée) et son

découpage en phonèmes (présenté à la couche de sortie). Un synthétiseur de voix permet la prononciation des phonèmes ainsi repérés. Si on présente, après l'apprentissage, un texte inconnu au réseau, le système prononce correctement 90 % des mots. Lors de l'apprentissage, la couche de neurones cachés a réalisé une classification des règles de prononciation pour chaque groupe de lettres en isolant les exceptions.

4 - Réseaux en couches à apprentissage non supervisé.

Les réseaux de Kohonen utilisent pour leur part une méthode d'apprentissage non supervisé. L'objectif est alors de conduire le réseau à formaliser une classification des prototypes présentés sans que celle-ci soit connue au préalable. Quand un exemple x^k est présenté au réseau, on détermine la cellule de sortie la plus sensible, et on modifie les poids des cellules de son voisinage. Ce modèle est une application de la propriété de "rétinotopie" des êtres humains, qui fait que les relations de voisinage sont conservées entre les neurones qui traitent en chaîne des informations dans diverses parties du cerveau (par exemple de la zone de la vision à la zone de compréhension de l'image). A la fin de l'apprentissage, le réseau s'est structuré lui-même de façon à ce que chaque cellule de la couche de sortie réponde sélectivement à une classe d'exemples (i.e. s'active plus que les autres).

En marge de ces divers types de réseaux, les recherches connexionnistes distinguent aussi deux grands modèles de représentation des informations :

- la représentation locale, attribue une information à chaque entité. C'est la représentation la plus aisément compréhensible. Elle correspond aux diverses expériences connexionnistes dans le domaine documentaire ([KW089], [BEL89], [VER090]). Un neurone particulier représente alors soit un terme d'indexation, soit un document. Les liens synaptiques sont pondérés en fonction des relations existant entre ces informations.

- la représentation distribuée, bien que plus difficile à mettre en œuvre considère que chaque information est représentée par un ensemble d'activation de

plusieurs neurones dans le réseau [HIN86]. Ce modèle permet d'utiliser un même neurone dans le cadre de la représentation de plusieurs éléments, et conduit à une meilleure optimisation de la mémoire et de la puissance de calcul. Il permet de favoriser la maintenance d'une information même si certains éléments d'information (i.e. neurones) sont hors d'usage. On obtient alors des mémoires adressables par le contenu. Les propriétés des modèles distribués sont plus proches des propriétés de la mémoire humaine. Cependant, ils ne constituent pas totalement une alternative au modèle local, celui-ci gardant une force pour les représentations de "haut niveau" (représentations symboliques). Le modèle distribué semble plus efficace dans les opérations de reconnaissance plus élémentaires.

Les représentations locales ou distribuées sont compatibles avec tous les types de réseaux connexionnistes Les relations entre le réseau et l'univers extérieur se situent au niveau symbolique (interprétation de la décision proposée par le réseau). Cette relation doit proposer une information décryptée, que cette information ait été codée sous une forme locale ou distribuée.

6.b - Une base de données connexionniste : le modèle des *Jets et des Sharks*

Le modèle *"des jets et des sharks"* de McClelland [MCC86] offre une bonne approche des modèles connexionnistes appliqués à la gestion de l'information. Il s'agit d'un système de gestion de données (les attributs sont précis et constituent l'objet des requêtes). Cependant, ce modèle nous apprend comment peut fonctionner un modèle documentaire connexionniste

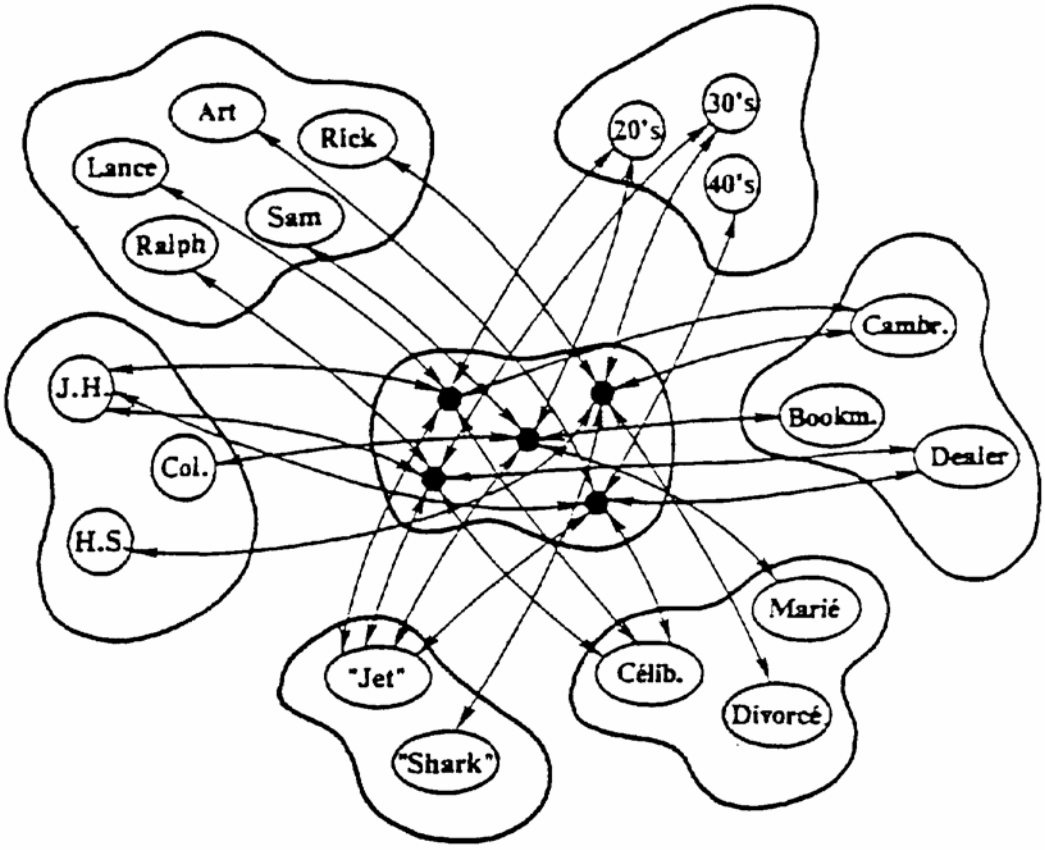
Considérons une banque de données répartie composée d'une trentaine d'individus malfamés comme on pourrait en rencontrer dans une grande ville. Ces individus sont décrits par certains attributs : nom, gang (les *"Jets"* et les *"Sharks"},* classe d'âge, niveau scolaire, situation familiale et profession. Cette banque de données peut être représentée comme dans les systèmes traditionnels par un tableau.

On peut aussi la considérer comme un réseau, dans lequel chaque individu

est représenté par une unité. Cette unité de représentation est liée aux attributs par des connexions mutuellement excitatrices. Les attributs sont eux-mêmes représentés par des unités, les connexions entre unités représentant un même attribut étant pour leur part mutuellement inhibitrices (on ne peut appartenir en même temps à plusieurs classes d'âge). Cette représentation correspond au schéma suivant :

LES JETS ET LES SHARKS *

Nom	Gang	Classe d'âge	Éduc.	Sit. fam.	Profession
Art	Jets	40	Collège	Célib.	Dealer
Al	Jets	30	Collège	Marié	Cambrioleur
Sam	Jets	20	Univers.	Célib.	Bookmaker
Clyde	Jets	40	Collège	Célib.	Bookmaker
Mike	Jets	30	Collège	Célib.	Bookmaker
Jim	Jets	20	Collège	Divorcé	Cambrioleur
Greg	Jets	20	Lycée	Marié	Dealer
John	Jets	20	Collège	Marié	Cambrioleur
Doug	Jets	30	Lycée	Célib.	Bookmaker
Lance	Jets	20	Collège	Marié	Cambrioleur
George	Jets	20	Collège	Divorcé	Cambrioleur
Pete	Jets	20	Lycée	Célib.	Bookmaker
Fred	Jets	20	Lycée	Célib.	Dealer
Gene	Jets	20	Univers.	Célib.	Dealer
Ralph	Jets	30	Collège	Célib.	Dealer
Phil	Sharks	30	Univers.	Marié	Dealer
Ike	Sharks	30	Collège	Célib.	Bookmaker
Nick	Sharks	30	Lycée		
Don	Sharks	30			
Ned	Sharks	30			
Karl	Sharks	40			
Ken	Sharks				
Earl	Sharks				
Rick	Sharks				
Ol	Shark				
Neal	S				
Dave					



La recherche dans un tel réseau se déroule par extension de l'activité *(spreading activation)* de l'unité correspondant à la requête sur l'ensemble du réseau. Si l'on désire connaître les propriétés d'un individu à partir d'un de ses attributs univoques, par exemple son nom (e.g. Lance), on active l'unité représentant cet individu dans l'ensemble des noms. Par extension de cette activité le long des liens synaptiques, la cellule attribuée à l'individu en question sera alors activée. Etant activée, elle répercutera cette information à toutes les cellules qui lui sont liées. Nous connaîtrons alors l'âge, la situation familiale la profession et le gang de Lance qui seront les valeurs "excitées" dans les ensembles correspondants.

Si l'on pose une requête moins spécifique, par exemple si l'on recherche le (ou les) individu(s) ayant entre 20 et 30 ans et appartenant au clan des sharks, on activera les entrées correspondantes dans les ensembles des attributs "âge" et "clan", et par extension de ces activations, nous "exciterons" le nom de Ken, qui en retour "excitera" les autres propriétés (niveau d'étude, profession...) de l'individu Kent.

Ce réseau offre quelques avantages sur les modèles traditionnels :

- une dégradation progressive des performances. Si nous posons une question qui n'admet pas de réponse exacte, l'élément qui répondra "le mieux' (i.e. ayant le plus grand nombre de points communs avec la question) sera "excité", mais bien entendu plus faiblement, entraînant un renforcement, lui aussi plus faible, mais cependant significatif, des caractéristiques vraisemblables de la réponse approchée. Ainsi, le système peut donner une bonne réponse, en soulignant qu'elle n'est pas la réponse idéale. Nous n'obtenons jamais de réponse nulle.

- une affectation de valeurs inconnues par relation de proximité. Si nous ne connaissons pas, au moment de l'alimentation de la base de données, la valeur d'un attribut pour un individu donné, le phénomène d'extension de l'activation autour des autres caractéristiques de cet individu tendra à allumer (plus faiblement) les individus qui lui sont le plus proches, et en retour, la valeur la plus élevée pour l'attribut inconnu sera, selon la plus forte vraisemblance, celle qui manque dans la description de cet individu. Vraisemblance ne veut pas dire

certitude, mais là encore, le système est capable de proposer des réponses, même approchées, dans une situation où des données sont manquantes.

- une généralisation spontanée. Supposons que l'on active le gang "Jets". L'extension de l'activation permettra d'établir une liste des noms des Jets. Les individus excités propageront l'excitation vers les autres attributs (âge, niveau d'étude, profession...), ce qui aura pour effet d'établir le portrait du *"Jet moyen"*. Le système aura généralisé des connaissances, sans que cette généralisation ait été stockée en un point quelconque de la mémoire.

Ces caractéristiques du modèle *"Jets-Sharks"* sont obtenues grâce à une autre forme d'organisation des mémoires. Dans ce modèle, il n'y a aucune représentation en mémoire des individus (i.e. de fiche documentaire regroupant les propriétés d'un individu). Au contraire, la définition d'un individu est fonction de la force des liaisons entretenues entre le point représentatif et les diverses valeurs des attributs. Cela a une conséquence sur les capacités d'apprentissage du modèle. Apprendre ne consiste plus à enregistrer de nouvelles configurations ou à établir des règles explicites (cas de l'Intelligence Artificielle), mais à modifier les forces de liaisons, pour que le système puisse réagir *comme s'il savait.*

6.c • Connexionnisme et systèmes documentaires

Les recherches connexionnistes en science de l'information sont aujourd'hui proches du modèle des Jets et des Sharks en ce qu'elles utilisent des représentations locales. Ces expériences offrent un attrait particulier car elles se rapprochent de ce qui est la nature même de l'information documentaire.

Un document n'est jamais un produit intellectuel isolé. Au contraire, il n'existe que par son environnement. Il tend à apporter des réponses, des contradictions ou à enrichir un débat en posant de nouvelles questions. Un document est un élément au sein d'un ensemble de documents, produit dans le "micro monde" intéressé par des thèmes similaires. La réalité d'un document n'est pas uniquement incarnée dans ce document lui-même, mais dans les "forces de liaisons", au sein du micro monde documentaire dans lequel il s'insère. Cette conception de la production documentaire a animé l'œuvre de

Manfred Kochen [SWA90]. Un de ses élèves, Richard Belew a réalisé, suivant cette conception, le système AIR *(Adaptive Information Retrieval),* qui est un des premiers modèles connexionnistes adaptés au domaine documentaire [BEL/89].

Le système AIR est constitué par un réseau d'environ 5000 cellules, correspondant à 1600 documents sur l'intelligence artificielle. Un nœud du réseau est affecté à chaque document. Lors de l'introduction du document, le système crée un nœud pour les auteurs ou ajoute une liaison si le nœud auteur existe déjà. Il fait de même avec les mots du titre qui sont considérés comme des descripteurs. Le poids des liens est fonction de l'inverse de la fréquence des termes. La somme des poids de toutes les liaisons sortant d'un nœud est maintenue constante (valeur 1). Il existe deux liens entre un descripteur et un document, représentant les deux directions que peut prendre l'extension de l'activation. Les poids associés peuvent prendre des valeurs différentes.

Une requête place de l'activité dans les nœuds correspondant aux termes choisis (ou aux auteurs, voire aux documents eux-mêmes qui peuvent être considérés comme une question). L'extension de cette activité va "exciter" un certain nombre de nœuds du réseau, correspondant à des documents, des termes ou des auteurs. Cette excitation est proportionnelle à la valeur des poids synaptiques associés à chaque liaison. L'ensemble des nœuds excités au dessus d'un certain seuil constitue la "réponse" à la question. Les nœuds sont affichés à l'écran, ainsi que les liens. L'affichage des réponses est dynamique, le premier nœud de type document activé de façon significative est placé au milieu de l'écran. Cela induit l'affichage des nœuds auteur et descripteurs de ce premier document. Il est ensuite rejoint par les autres nœuds au fur et à mesure que l'extension de l'activation leur fait atteindre un certain niveau d'excitation. L'interface distingue les nœuds de type document placés au centre de l'écran des nœuds descripteurs en haut et auteurs en bas.

Un procédé de jugement de pertinence permet d'activer plus spécifiquement certains des nœuds apparaissant à l'écran (une activation négative peut éventuellement être donnée, correspondant à la non-pertinence), ce qui conduit à une modification du réseau constituant la réponse. Ce processus permet d'obtenir de nouvelles réponses à une question. Il permet aussi au réseau AIR d'apprendre à chaque utilisation. Si par exemple un terme de la question n'est pas connu par le système, il est placé dans un nouveau nœud. La question est traitée sans ce terme, et le jugement de pertinence de l'utilisateur va constituer la première insertion de ce terme dans le réseau, en donnant un poids aux liens tissés entre ce terme et les documents pertinents. De même, l'extension de l'activation va tendre à modifier les poids des liaisons entre documents jugés ensemble pertinents (ou non pertinents), car la somme des poids est maintenue égale à un pour chaque neurone.

Cet apprentissage est rendu possible par l'assimilation entre l'activité et la pertinence. Un facteur de correction des poids est déterminé par la relation entre l'activation d'un neurone et la pertinence d'un autre.

L'extension de l'activation permet de concevoir des relations transitives entre nœuds. Si un nœud A est associé à un nœud B, et que celui-ci est associé à un nœud C, alors A et C sont associés mais avec un coefficient plus faible. Cette transitivité peut être illusoire dans la réalité (par exemple un auteur ayant travaillé sur plusieurs sujets différents), mais la capacité de AIR d'apprendre à partir de l'utilisation permet de supprimer les extensions abusives en s'appuyant sur le jugement de pertinence.

La fonction d'association entre deux nœuds est calculée en tenant compte de tous les chemins qui mène d'un nœud à l'autre (bien qu'une longueur maximale soit fixée pour éviter des calculs trop longs). Cela permet par exemple de rapprocher entre eux des mots-clés, ou des documents, ou encore des mots-clés et des documents. On retrouve la notion de recherche documentaire associative [YAK87].

Le système AIR conduit Richard Belew à concevoir un modèle wittgensteinnien du langage appliqué en informatique documentaire. Wittgenstein insiste sur le fait qu'un mot n'existe que parce qu'il s'insère dans un

acte de langage, et que nous n'apprenons le sens d'un mot qu'en expérimentant son usage, hors de toute construction préalable dans notre esprit d'un dictionnaire de définitions. Ces idées sont reprises dans AIR, en considérant qu'un mot n'est connu par le système qu'en fonction des utilisations qui en sont faites (i.e. utilisation dans les liens entre terme et document) et que l'expérimentation langagière (jugement de pertinence, transitivité contrôlée par les utilisateurs des relations associatives) constitue le moyen de créer et maintenir un réseau de signification entre les mots.

D'autres systèmes documentaires s'appuyant sur un modèle connexionniste ont été étudiés. Il existe une forte corrélation entre le modèle connexionniste et le modèle probabiliste [KW089]. Le modèle connexionniste est aussi utilisé dans une opération d'analyse des données pour obtenir une classification des thèmes efficaces dans un environnement local au sein d'un espace documentaire. Le travail de Alain Lelu [LELU86] vise à obtenir une proposition de pistes de reformulation guidant un utilisateur qui a déjà sélectionné un ensemble de documents (en occurrence des images) par analyse des représentations associées à ces documents et classification des termes en fonction de leur caractère opératoire (i.e. ne pas présenter des termes inutiles, ou synthétiser dans des termes plus généraux plusieurs expressions induisant une circulation semblable dans l'ensemble des documents). Pour cela, il utilise un réseau de type réseau de Kohonen, en réduisant le nombre de cellules de la couche de décision, ce qui tend à ne sélectionner que les composantes principales de la classification ([LELU89]).

L'extension de l'activation est l'élément central de l'application du modèle connexionniste à la recherche documentaire. Le modèle connexionniste se distingue cependant des réseaux sémantiques qui fait lui aussi usage de l'extension de l'activation.

Dans un réseau sémantique, les liens entre les termes sont étiquetés par un certain type de relation *(est_un, a_pour_partie, est_partie_de....* et un lien ne permet le passage de l'activation qu'en fonction du type de relation considéré dans une recherche particulière, et dans un seul sens. Au contraire, les liens dans les modèles connexionnistes sont sans signification particulière, seul le résultat d'ensemble se voit attribuer une quelconque importance. Le poids des liens n'est plus fixé par le concepteur comme dans un réseau sémantique ou un thésaurus,

mais reste capable d'évoluer quand le réseau apprend. Toutefois, on peut envisager la coopération entre les deux éléments, le réseau connexionniste permettant de souligner des relations de type probabiliste conduisant à la construction supervisée d'un réseau sémantique, et en sens inverse, les liens forts définis par le réseau sémantique influençant le calcul des poids de liaison du réseau connexionniste Cette conception est d'ailleurs en phase avec une certaine évolution des recherches connexionnistes, qui tendent à spécialiser l'utilisation des réseaux pour certaines applications dans le cadre de modèles symboliques basés sur l'explicitation de connaissance [KNI90].

deuxième partie

Réalisation d'un catalogue informatisé

I - Le choix d'une hypothèse de travail

La réalisation du catalogue informatisé de la bibliothèque scientifique de l'Université de Caen est intervenue entre mai 1989 et juillet 1990. Elle s'est déroulée en plusieurs étapes, et avait des objectifs très concrets et immédiats, en termes d'organisation de la bibliothèque, d'amélioration de l'accès aux références et aux ouvrages, d'économie de temps et de personnel. Le catalogue informatisé a donc été réalisé avec les moyens dont disposait l'université. Il ne saurait être conçu comme une application des recherches et des réflexions décrites plus haut. H s'agit au contraire d'une expérience pratique qui permet de soulever des problèmes intéressant la recherche, et de montrer quelques voies de résolution dans un environnement informatique et documentaire traditionnel.

La réalisation d'un catalogue informatisé de bibliothèque pose un grand nombre de problèmes, qui peuvent se diviser en plusieurs thèmes :

- des problèmes politiques : Une bibliothèque universitaire s'inscrit dans un environnement diversifié, depuis l'échelle locale (le Service Commun de la Documentation, l'Université...) jusqu'à l'échelle nationale (le réseau des bibliothèques universitaires et sa Sous Direction au Ministère de l'Education Nationale) et même internationale (les réseaux bibliographiques tels OCLC). Le choix de réaliser un catalogue informatisé se pose alors dans un ensemble de contraintes, à la fois en termes budgétaires, informatiques (structures de données, réseaux d'accès à l'information), bibliothéconomiques (précision du catalogage, points d'accès aux notices, plan de classement...), voire idéologiques (rapport entre l'informatisation et la démocratisation des bibliothèques). Ces questions doivent être débattues, car elles conditionnent les futurs développements de l'informatisation des bibliothèques.

- des problèmes économiques et organisationnels : Informatiser une bibliothèque renvoie à une gestion particulière des frais induits et à une réorganisation interne des tâches du personnel. On a trop souvent dit qu'informatiser coûtait très cher mais permettait d'économiser sur les tâches répétitives, et de rendre le personnel plus disponible au public. Il s'agit

d'envisager ces affirmations en regard de la pratique. A quelles conditions sont-elles vraies dans chaque condition concrète ? Quel nouveau mode de gestion du personnel permet l'informatisation ? Faut-il jouer la carte du réseau et de la récupération ou de l'achat des données bibliographiques ? Quelle est la politique suivie par l'Etat dans la mise en œuvre d'un réseau de fourniture de données bibliographiques ? Faut-il s'en remettre à des prestataires de service étrangers ?...

- des problèmes informatiques et de télécommunication : Informatiser, c'est choisir une architecture générale de système informatique, envisager les possibilités de diffuser le catalogue sur les divers réseaux de télécommunication. C'est choisir un logiciel, en fonction des besoins exprimés par la bibliothèque, mais aussi en fonction des contraintes financières, et en tenant compte des instruments disponibles sur place à un moment donné. Informatiser, c'est de même prévoir la disparition des matériels ou des logiciels du marché. On peut établir une durée de vie de trois à cinq ans pour les ordinateurs, et de cinq à dix ans pour les logiciels. Or les données bibliographiques ont une pérennité plus importante, difficile d'ailleurs à évaluer, le catalogage descriptif étant plus stable que l'indexation ou les données locales (numéros d'inventaire comptables, localisation, conditions de prêt...). La récupération des données, les formats d'échanges de données, l'ouverture relative des logiciels et des matériels... doivent être pris en compte dans les choix.

- des problèmes linguistiques et documentaires : Informatiser, c'est aussi choisir un mode d'écriture des données qui permette leur utilisation par une machine. Cela recouvre des problèmes d'indexation, des problèmes de formulation des questions par l'utilisateur, et des problèmes de définition d'une interface de recherche. De plus, il faut choisir la langue du système d'information. Dans un monde entièrement dominé par les anglo-saxons, cela revêt une grande importance. La langue du système se juge au niveau de l'interface utilisateur (présentation des écrans) mais aussi en fonction des possibilités d'écrire les données elles-mêmes en français (le système accepte-t-il les signes diacritiques du français, et les traite-t-il différemment lors de l'indexation et de la visualisation ?).

L'ensemble des questions soulevées dans la première partie trouvent à ce niveau des implications concrètes. Toutefois, un catalogue de bibliothèque est un

cas particulier de système documentaire car il est destiné à un public spécifique, placé dans une situation de recherche particulière. Cet aspect est plus sensible dans une bibliothèque universitaire, qui dessert beaucoup d'étudiants de premier cycle, désirant principalement des manuels d'enseignement ou des livres d'exercices.

- des problèmes bibliothéconomiques : Le catalogue d'une bibliothèque n'est qu'un instrument d'accès à des références de documents. H s'inscrit donc dans des choix de classement, de mode d'accès, dans une répartition des lieux (articulation entre bibliothèque centrale et bibliothèques de laboratoires ou d'instituts). Il s'inscrit aussi dans un mode de gestion des prêts, des inscriptions, des réservations... A ce titre, il convient d'envisager, au niveau de la représentation informatique, les divers aspects d'un livre dans une bibliothèque : commande, inscription au catalogue, situation à un moment donné, conditions particulières de prêt...

- des problèmes pédagogiques : Les besoins auxquels un catalogue de bibliothèque, ou plus généralement un système d'information d'une université, doivent répondre sont variables, et dépendent de choix suivis dans l'organisation des études (par exemple les étudiants doivent-ils rendre un mémoire ? Ont-ils des travaux spécifiques à accomplir en bibliothèque ?), dans l'organisation de la recherche (liens avec le prêt inter bibliothèques, avec le réseau informatique de la recherche de l'Université, et sa messagerie...) et dans l'enseignement des techniques d'accès à la documentation au sein du cursus universitaire.

L'ensemble de ces problèmes se synthétisent en deux types de "prise de décision" :

- quelle sera la place du catalogue informatisé dans la bibliothèque, en regard des autres pratiques bibliothéconomiques, et en regard des projets de développement des réseaux d'information dans l'université, la ville et la région.

- doit-on distinguer l'accès public au catalogue et l'informatisation des autres fonctions ? Peut-on développer un système documentaire sans s'insérer au préalable dans un réseau bibliographique ?

Les réponses que nous avons choisies pourront paraître iconoclastes à une génération de bibliothécaires élevés au lait du partage des corvées (le catalogage "partagé", puis "en réseau") et de l'intégration documentaire universelle (une super bibliothèque virtuelle permettant de localiser toute la richesse documentaire). Ces réponses sont néanmoins conjoncturelles. Elles permettent de pointer les directions vers lesquelles doivent évoluer les conceptions de l'informatisation des bibliothèques. Il convient en effet de briser des dogmes qui n'ont abouti jusqu'à présent qu'à limiter les avancées informatiques en France (et en langue française !) et à enrichir indûment des fournisseurs de logiciels souvent mal adaptés aux besoins du public.

1 - Présentation de la Bibliothèque Scientifique de l'Université de Caen

L'Université de Caen regroupe environ 20 000 étudiants, majoritairement dans les premiers cycles. C'est une université pluridisciplinaire, couvrant l'ensemble des domaines de recherche et de formation, ce qui n'est pas sans poser des problèmes documentaires.

La documentation à l'université est gérée par un *"Service Commun de la Documentation"* qui doit regrouper les bibliothèques d'instituts et de laboratoires avec la bibliothèque universitaire. Cependant, cette logique collective n'est pas encore entrée dans les faits, et le SCD correspond principalement à la B.U., qui est divisée en trois sections (Lettres-Droit, Médecine, Sciences) réparties en trois endroits différents du campus.

La bibliothèque scientifique propose environ 40 000 ouvrages, dont seulement 10 à 15 000 sont encore utiles aujourd'hui. Le budget d'achat de livres est d'environ 80 000 francs, soit moins de mille exemplaires par an. La collection de périodiques scientifiques est plus riche, avec 250 titres courants entrés par abonnements (750 000 francs de budget) et 300 titres entrés par dons et échanges. Après une longue période sans achats de livres, il a fallu reconstituer un fonds correspondant aux besoins des premier et second cycles (quelques ouvrages en de nombreux exemplaires), ce qui ne nous a pas permis d'acheter beaucoup d'ouvrages de niveau recherche. Dans ces conditions, la majeure partie des livres

spécialisés ou des revues sont répartis dans les laboratoires, souvent dans de petites bibliothèques propres à un service. Cet endettement de la documentation, propre aux universités françaises, pose bien entendu des problèmes de gestion des fonds. Le catalogue doit tenir compte de cet élément pour l'insertion de données de localisation et d'accessibilité.

Comme toute bibliothèque universitaire [MIQ89], la bibliothèque scientifique vit au régime de la pénurie de personnel et de moyens. Jusqu'en 1989, pour emprunter un ouvrage, l'étudiant devait passer par la consultation d'un catalogue sur fiches pour obtenir la cote du livre qui était rangé en magasin. L'opération de prêt a été informatisée en 1988 grâce au système sur micro-ordinateur MOBIBOP. Le système installé à la bibliothèque permet la gestion des prêts et des inscriptions à partir de deux terminaux. Il fonctionne sur un compatible AT (FORUM 286) sous le système d'exploitation PROLOGUE. L'enregistrement des prêts est réalisé dans des tables mettant en relation l'identifiant du lecteur et l'identifiant de l'ouvrage (lecture de codes à barre). Des contraintes particulières peuvent être ajoutées, soit sur certains ouvrages, soit sur certains types de lecteurs. Ce système est bien adapté à la taille et au rythme de fonctionnement de la bibliothèque (35 000 prêts par an).

En juin 1989 a été installé un système anti-vol, ce qui permettait d'envisager la mise en libre accès des ouvrages. Il convient de mesurer l'impact de cette transformation sur le fonctionnement de la bibliothèque au cours de l'année universitaire 1989-1990. Une enquête informelle auprès des usagers sur l'utilisation du catalogue informatisé a montré que pour leur grande majorité, ils ne l'avaient jamais utilisé, entièrement satisfaits par la mise en libre accès des livres, qui constituait pour eux la véritable innovation dans la bibliothèque, une remarque qui doit faire réfléchir ceux qui ne jugent les progrès des bibliothèques qu'à l'aune de l'introduction des nouvelles technologies de l'information.

C'est dans le cadre de cette opération de mise en libre accès des ouvrages qu'est né le projet de constituer un catalogue informatique. Proposer directement les ouvrages aux lecteurs nous oblige à les classer par sujets sur les rayons. Il faut donc choisir un mode de classement thématique et indiquer sur l'instrument de référence la localisation de chaque ouvrage. Il était impensable de modifier toutes les fiches du catalogue à chaque déplacement d'un livre (5 000 livres ont été

déplacés en un an du magasin à la salle de libre accès). Le catalogue informatique a été conçu pour servir de "filtre" entre un état antérieur, qui perdure et risque de durer encore, marqué par l'accès via le catalogue sur fiches à des livres conservés en magasin, avec un numéro d'inventaire attribué par ordre d'entrée dans la bibliothèque, et la nouvelle situation de libre accès pour les ouvrages les plus récents ou les plus utilisés, qui sont repérés par leur classement systématique.

Parallèlement à cette ouverture du libre accès, nous avons décidé d'appliquer les nouvelles directives préparées par la DBMIST (Direction des Bibliothèques, des Musées et de l'Information Scientifique et Technique) concernant l'organisation des collections dans les Bibliothèques Universitaires ([DBM88],[SANS88]). Cette circulaire prend acte de l'introduction de l'informatique documentaire, en accordant une place plus spécifique à l'indexation décimale : *"l'indexation n'a pas une perspective analytique, mais plutôt synthétique en vue d'un regroupement utile à la recherche"*. Ce choix se traduisait par l'adoption de critères de regroupement des ouvrages (utilisation des classifications pour répartir les livres sur les étagères) au lieu de critères de dispersion par une indexation trop fine, utilisant des indices développés. La circulaire conseillait de ne conserver que les trois premiers chiffres des indices des classifications. La proposition, liée à la possibilité d'obtenir des listes par domaines d'acquisition, était d'unifier la cote d'un ouvrage (indiquant sa localisation dans la bibliothèque) et son numéro d'inventaire (comptable), en donnant à ce numéro unique la forme :

***XXX-1005, b* où:**

. *XXX* est un indice de classification, regroupant dans les listes comme dans les rayons tous les ouvrages traitant d'un même domaine.

. *1005* est (par exemple) le numéro d'ordre dans l'indice XXX. Le choix d'une notation sur quatre chiffres permet de conserver une cote de longueur fixe (de 1001 à 9999) dans un même indice, ce qui permet de faire trier des listes par cote (équivalent à un registre d'inventaire) même en suivant le classement alphanumérique. Ce choix permet aussi de présenter toujours au bout de l'étage affecté à un indice particulier les ouvrages dont l'entrée dans la bibliothèque est la plus récente (numéro le plus élevé).

. *b* permet d'indiquer, si nécessaire, les différents exemplaires d'un même ouvrage. Cette distinction permet de traiter chaque ouvrage comme

une donnée univoque dans un système de gestion des stocks (inventaire, prêt, disparition...).

Cette proposition visait à trouver un système simple, aisément compréhensible, qui allie les besoins de regroupement des ouvrages et les nécessités de gestion exemplaire par exemplaire (comptabilité, prêt, récolement...). Après un an, notre expérience a montré la validité de ces choix.

Toutefois, pour pouvoir respecter la règle d'une limitation de l'indice à trois chiffres tout en prenant en compte les spécificités de notre fonds, nous n'avons pas voulu utiliser une classification existante (Classification Décimale Dewey, C.D.U.). Celles-ci sont trop encyclopédiques pour une bibliothèque spécialisée comme la nôtre. Il est par exemple impossible d'exprimer l'ensemble de l'informatique avec un indice à trois chiffres dans la Classification Décimale Dewey. Alors le langage PASCAL, l'intelligence artificielle ou le connexionnisme... Nous avons donc choisi de réaliser un plan de classement spécifique dont les lignes directrices sont exposées plus loin.

Pour informatiser en fonction d'un besoin conjoncturel comme celui décrit ci-dessus, nous étions contraints à des choix draconiens. Impossible d'attendre près de deux ans pour rejoindre un réseau comme SIBIL. Impossible d'attendre un choix plus général d'informatisation, concernant l'ensemble des services de l'université, et extensible aux divers systèmes gérant les bibliothèques municipales de l'agglomération. Indépendamment même des jugements que l'on peut porter sur les réseaux bibliographiques, il nous fallait agir rapidement, en nous appuyant sur les ressources de l'université. D'emblée, le choix d'un catalogue sur micro-ordinateur a été exclu, car dans un univers dispersé comme une université, on se doit de concevoir un catalogue directement accessible depuis un terminal minitel. Nous disposions par ailleurs d'un Centre de Calcul très coopératif, doté d'un ordinateur *VAX* sous *VMS* de *DEC (Digital Equipment Corporation)* et du logiciel *Texto + Logotel* de *Chemdata.* D'un certain point de vue, ces contraintes ont été bénéfiques, car elles nous ont permis de nous concentrer sur les problèmes d'accès par les utilisateurs, ce qui nous a offert une plus grande marge de manoeuvre pour tester et améliorer le système documentaire qu'un logiciel "clé en main". Il nous fallait cependant assurer une compatibilité éventuelle avec les divers systèmes existants, afin de préserver

l'avenir. Nous pensons que cette nécessité peut aussi être prise en compte par des logiciels documentaires généralistes.

La pénétration de logiciels documentaires du type de *Texto* ou *BRS* offre à notre avis, une meilleure garantie de pérennité des informations que les systèmes clos du monde des bibliothèques. Un problème subsiste toutefois concernant le choix de la description bibliographique (normes de catalogage) et les diverses possibilités d'échange de données. Nous traiterons plus loin le transcodage en format MARC, mais il faut préciser que certains logiciels spécifiques du monde des bibliothèques (notamment *Datatrek)* permettent de récupérer des informations venant de fichiers documentaires pour peu que ceux-ci puissent fournir des fichiers ASCII balisés. Ceci étant le cas de *Texto,* notre choix reste cohérent avec l'éventualité d'une insertion ultérieure dans un réseau national, ou plus prosaïquement avec le changement, au niveau local, de système documentaire, ce qui, d'évidence, interviendra dans les années qui viennent.

2 - Les systèmes intégrés de gestion de bibliothèque

Les systèmes intégrés veulent regrouper autour d'une base de données l'ensemble des opérations de gestion d'une bibliothèque :

- le catalogage et l'indexation des ouvrages
- l'accès en ligne au catalogue par les bibliothécaires
- l'accès public en ligne, éventuellement à partir du domicile
- la gestion des prêts, l'émission de lettres de rappel
- les commandes d'ouvrages et la gestion comptable
- le bulletinage des périodiques
- la collecte de statistiques sur le fonctionnement de la bibliothèque.

Qui oserait contester un tel programme ? Quel bibliothécaire n'a jamais souhaité une vérification automatique de ses commandes (ce livre est-il déjà présent ?), une édition complète et illustrée des statistiques (dans le quart d'heure, comme dans les publicités pour l'informatique) ? Qui ne souhaite connaître immédiatement la situation d'un ouvrage qu'il trouve dans le catalogue ?

Malheureusement, comme tout programme mirobolant, le système intégré

de gestion de bibliothèque a un coût important : coût financier, organisationnel et intellectuel (indexation et recherche documentaire). Pour permettre les nombreux accès simultanés à la base de données requis par l'ensemble des fonctions, il faut se doter d'un matériel imposant, que l'expérience montre cependant toujours sous-estime.

On doit aussi considérer la distinction entre les systèmes de gestion "stock et flux" et les systèmes de recherche documentaire. Les contraintes de la gestion des livres (commandes, catalogage, prêt) ne sont pas les mêmes que les contraintes de l'accès public au catalogue (indexation, interface utilisateur...). On retrouve là encore la distinction entre les systèmes gérant les données (le livre comme objet matériel, avec un prix, un fournisseur, un numéro d'identification et éventuellement un lien avec un identifiant dans le fichier des lecteurs) et les systèmes d'information, qui doivent porter leur attention sur d'autres aspects du livre (son contenu et les moyens de le décrire). Cette distinction a des conséquences sur l'interface de manipulation des informations. La convivialité d'un système documentaire se mesure à sa capacité à prendre en charge des éléments d'incertitude, tant dans la formulation des questions que dans l'indexation des ouvrages. A l'opposé, un interface de gestion des prêts se doit une rigueur exemplaire, construite sur une rigueur semblable de la description et de l'indexation documentaire (par exemple, le système des vedettes des catalogues sur fiches, ou le système des "clés auteur titre" de certains systèmes intégrés).

Nous avons fait le choix d'informatiser la bibliothèque en nous appuyant sur des systèmes éclatés, chacun remplissant des fonctions spécifiques, tout en réservant la possibilité d'échanger des données entre ces systèmes quand le besoin s'en ferait sentir et que les solutions techniques (informatiques et organisationnelles) seraient au rendez-vous. Car informatiser une fonction spécifique n'est pas toujours évident. Il convient dans ce cas d'accorder une indépendance à chaque tâche, pour éviter que les retards pris dans l'une d'entre elles (en général le catalogage des livres) ne viennent ralentir l'ensemble du processus. Ce choix est par ailleurs cohérent avec la logique générale de l'évolution de l'informatique, qui va dans le sens de systèmes spécifiques, capables de remplir au mieux les tâches qui leurs sont assignées, tout en conservant un aspect communicant pour échanger les données avec d'autres systèmes spécifiques. Par exemple, dans le domaine de la bureautique, la mode des

"logiciels intégrés" a été abandonnée au profit de logiciels de plus en plus performants dans leur tâche principale (traitement de texte, tableurs, logiciels de mise en page, gestionnaires de plans...) et organisés afin de récupérer ou d'exporter des données les uns vers les autres. Dans le même ordre d'idées, on observe la généralisation du concept *OSI (Open Systems Interconnexion - Interconnexion de Systèmes Ouverts)* dans l'ensemble des applications informatiques, surtout si elles possèdent des aspects télécommunicants. Un concept qui commence aussi à s'imposer dans les bibliothèques ([CAI89a], [CAI89b]).

A la bibliothèque scientifique, les diverses opérations quotidiennes sont assurées par des systèmes spécialisés. L'organisation des échanges entre ces systèmes est encore élémentaire, mais il convient avant d'envisager son extension que chacun d'entre eux fonctionne correctement, et plus encore que les pratiques professionnelles s'adaptent à l'informatisation de ces opérations, et cela pour l'ensemble du personnel. Ce qui implique un travail de formation et de diffusion de la pratique informatique que chacun sait long et difficile.

- La fonction de prêt est assurée par *MOBIBOP,* un logiciel spécifique, développé par la société *ISL* à la demande de la DBMIST, pour répondre aux besoins des bibliothèques universitaires. Pour ce système, un livre est représenté par un identifiant (un code à barre ou un numéro d'inventaire). Le logiciel MOBIBOP permet en plus de gérer des statistiques évoluées. Ce système était en place avant l'ouverture du catalogue. Il correspond exactement à l'idée d'un gestionnaire stock et flux. Les items sont repérés par des identifiants univoques et dépourvus de sens (codes à barre) et les circulations sont contrôlées au travers de règles propres à chaque couple id_lecteur et id_ouvrage (conditions de prêt, droits spécifiques de certains types de lecteurs ou restrictions attachées à certains types d'ouvrages). Malgré quelques difficultés techniques, ce logiciel nous donne pleinement satisfaction pour sa conception fonctionnelle.

- La fonction de commande des ouvrages peut être totalement indépendante du système documentaire, charge à son responsable de vérifier l'éventuelle présence des livres au catalogue. Une tâche que de toute façon les systèmes intégrés ne remplissent pas toujours. Par ailleurs, pour commander des livres, point n'est besoin d'une description complète. En général, le libraire se

débrouille parfaitement avec l'éditeur, le premier auteur et le titre. C'est du moins un des avantages à travailler avec des libraires, en plus de défendre la vente de livres dans sa propre ville. Du point de vue du système documentaire, il est souvent difficile de récupérer une information incomplète pour la transformer ensuite, lors de la réception de l'ouvrage, en catalogage définitif. En sens inverse, s'efforcer de retrouver la description exacte d'un livre avant sa commande est une tâche fastidieuse, qui fait perdre de vue l'essentiel d'une opération de commande : la satisfaction rapide d'un besoin exprimé, ou le suivi de la production éditoriale courante en fonction des critères spécifiques adaptés au public de la bibliothèque.

- La fonction de bulletinage des périodiques avait été, elle aussi, informatisée avant le choix d'ouvrir un catalogue électronique. Le logiciel avait été écrit en *dBase* sur un compatible AT (Normerel ATC12) mais quelques systèmes commerciaux remplissent les mêmes fonctions *(Oasis* diffusé par *Dawson)*. Plusieurs hypothèses ont été testées pour assurer une transmission régulière des informations de bulletinage dans le catalogue, aucune n'ayant pleinement donné satisfaction. Aujourd'hui, le catalogue ne comporte que l'état général de la collection (date de début et de fin, éventuellement la mention "abonnement en cours"). Les lacunes dans la collection et le bulletinage des derniers numéros ne sont mentionnées que dans le système de gestion des périodiques. La fonction de bulletinage est elle aussi une fonction de gestion : repérage des numéros manquants, vérification des retards, suivi des fournisseurs, partage du coût des abonnements entre partenaires... Elle est dès lors distincte des fonctions de catalogage et de recherche documentaire sur les titres des périodiques.

- L'accès public au catalogue constitue donc une opération indépendante, qui mérite d'être traitée en tant que telle. Cette distinction est d'ailleurs reconnue par les participants au réseau *SIBIL,* qui à l'instar de la bibliothèque de l'Université de Bordeaux ont choisi de récupérer les données de *SIBIL* pour confectionner en mode local un catalogue accessible au public *(GRACE).*

Les systèmes intégrés de gestion de bibliothèque sont bâtis pour permettre un accès par des spécialistes à des livres déterminés. Typiquement, on connaît le titre et l'auteur, et on veut savoir si le livre est présent. Dans la culture professionnelle, on distingue ainsi les systèmes de gestion de bibliothèque des

systèmes documentaires qui proposent des accès par descripteurs. Or cette distinction n'est plus de mise. Même dans une bibliothèque, on peut souhaiter obtenir des livres par leur sujet. Il s'agit d'ailleurs du cas général dans une bibliothèque scientifique. Cette recherche doit pouvoir s'effectuer sans devoir passer sous les fourches caudines d'une classification ou d'un langage documentaire trop structuré (ordonnancement des vedettes et des sous-vedettes). Or la majeure partie des systèmes intégrés de bibliothèques gèrent les informations de contenu comme des informations de gestion : le nombre de descripteurs est fixe, et l'interrogation doit respecter l'ordre des termes dans les vedettes. De ce point de vue, l'accès public au catalogue n'est qu'une couche de vernis passée sur un système dont la structure principale n'est pas adaptée à la recherche documentaire. On agrémente l'interface, éventuellement on se dote d'un transcodeur de protocole pour permettre l'accès vidéotex. Mais on ne peut pas prendre en charge les diverses questions linguistiques (indexation) ni les questions concernant la satisfaction d'un utilisateur (modèle de recherche) qui ont été évoquées dans la première partie. Les difficultés sont alors remplacées par le respect figé de la "normalisation" des vedettes matières, charge à l'utilisateur de s'adapter à ce type de codification.

On doit cependant remarquer des exceptions, c'est-à-dire des logiciels qui sont à l'inverse construits autour d'un module de recherche documentaire. Par exemple, *Datatrek,* dans la dernière version disponible sur des ordinateurs MS-DOS, gère les informations de catalogage comme une suite de phrases traitées en "texte intégral". Il autorise ainsi des recherches souples et efficaces et reste plus adaptatif que la majorité de ses concurrents. Ce logiciel permet ainsi, dans sa nouvelle version qui n'était pas disponible lors de la réalisation de notre catalogue, de définir une utilisation propre du format MARC. On peut alors utiliser les zones de résumé prévues dans ce format pour gérer des références bibliographiques. Toutefois, l'interface utilisateur reste encore marqué par les habitudes de la profession des bibliothécaires et n'est pas assez souple pour une utilisation directe par le public, notamment dès que l'on aborde la recherche par sujet. Ce logiciel marque cependant l'évolution à venir, d'une réunion entre les deux conceptions d'un catalogue de bibliothèque et d'une banque de données documentaire. La fin d'une guerre de prérogatives qui de toute façon est déjà inscrite dans le fonctionnement même des bibliothèques [MEL89]

3 • Les réseaux bibliographiques

La réalisation d'un catalogue de bibliothèque n'est pas, dans la culture professionnelle des bibliothécaires, une opération isolée, d'obédience strictement locale. Les bibliothécaires se vivent, à juste titre, comme les éléments d'un vaste réseau d'accès à l'information. Ce réseau est orchestré par deux axes de travail définis par la *Fédération Internationale des Associations de Bibliothécaires (FIAB - IFLA).*

- le *Contrôle Bibliographique Universel* (C.B.U.)
- l'Accès *Universel aux Publications* (UAP - *Universal Availability of Publications).*

Ces deux objectifs sont distincts, et ont des conséquences différentes sur la politique à mener dans le domaine des réseaux bibliographiques. Alors que le CBU rend indispensable la production normalisée de descriptions bibliographiques, l'UAP peut se contenter de descriptions moins complètes, mais doit s'appuyer sur des informations de localisation des documents dans les bibliothèques. Ces deux objectifs permettent de pointer les distinctions qui existent entre la notion de catalogue collectif et la notion de serveur bibliographique.

3.a • Le Contrôle Bibliographique Universel

Le *Contrôle Bibliographique Universel* (CBU) vise à organiser et assurer la connaissance mondiale de tout ce qui paraît dans le monde. Chaque pays est responsable de son application au travers de sa bibliographie nationale, sous la tutelle de l'organisme chargé de son élaboration (Bibliothèque Nationale en France).

Pour que le CBU puisse entrer en application à l'époque informatique, chaque pays doit pouvoir échanger les données sous une forme compatible, ce qui conduit à quatre types de normalisation [INT75] :

- la normalisation des descriptions bibliographiques : contenu des divers éléments de la notice bibliographique, écriture des vedettes, choix des points d'accès à la description bibliographique. Ce travail est réalisé par l'édition des recommandations ISBD *(International Standard Bibliographie Description)* qui

définissent le contenu des diverses zones de la description. Les diverses normes qui en découlent dans les pays (édictées par l'AFNOR en France) sont en général très proches les unes des autres, et compatibles moyennant quelques aménagements [DUS89].

- la normalisation d'identifiants internationaux. Ceux-ci permettent la désignation univoque des livres (numéro ISBN) et des publications en série (numéro ISSN). Cette pratique est aujourd'hui reconnue et généralisée. On considère par exemple que seulement 15 % des ouvrages publiés en France ne possèdent pas d'ISBN [MIN89]. Toutefois, des problèmes subsistent dans l'attribution, confiée aux éditeurs, de ces identifiants. En particulier, on constate l'existence de cas où le même ISBN est attribué à deux ouvrages différents. Certains éditeurs attribuent au contraire un nouvel ISBN pour une simple réimpression. Enfin se pose la question de l'identification des ouvrages en plusieurs volumes. Malgré cela, on peut considérer que l'ISBN est un moyen pratique et relativement fiable de retrouver un ouvrage publié après 1975. Même si une vérification s'impose, elle est rapide et légère. Ces identifiants sont structurés avec une clé de vérification, qui limite les possibilités d'erreur par faute de frappe.

- la normalisation des jeux de caractères, des translittérations de langues non romaines, des divers codes de pays, de langue... L'ISO *(International Standard Organisation)* se charge de ce travail : norme ISO-646 pour les caractères de base codés en 7 bits, et norme ISO-2022 pour l'obtention des autres caractères, notamment les signes diacritiques, par combinaison de codes à 7 bits. [INT75].

- l'élaboration d'un "format d'échange" pour que ces données soient lisibles par diverses machines de par le monde. Ce format d'échange est basé sur le travail de la Library of Congress (format MARC - *Machine Readable Catalogue)* et fait l'objet d'une norme internationale (ISO-2709). Il existe toutefois plusieurs types de formats MARC, et le format UNIMARC, privilégié par la Communauté Européenne, n'est toujours pas publié en France, avec des exemples adaptés à notre pays [MIN89].

3.b - L'Accès Universel aux Publications

L'Accès *Universel aux Publications (UAP - Universal Availability of Publications),* fut le thème du congrès de *l'IFLA* en 1973. Son objectif est *"de fournir à n'importe qui, n'importe où et n'importe quand, tout document publié dans le monde"* [BOS83]. Cet objectif est ambitieux, mais l'UAP fonctionne comme un parapluie sous lequel viennent s'abriter plusieurs programmes nationaux. Par exemple, en France, il concerne le fonctionnement du Prêt Inter Bibliothèques, ou la logique de répartition des acquisitions dans les bibliothèques universitaires au travers des CADIST, chargés d'acquérir pour une diffusion nationale toutes les publications concernant un domaine précis.

Au niveau européen, ce programme sert de guide à l'action de la BL-DSC *(British Library - Document Resource Center)* dont l'objectif est de devenir la principale bibliothèque de ressources à l'échelle européenne, voire mondiale. L'UAP guide aussi les volontés actuelles de tisser des réseaux de messageries entre les bibliothèques européennes pour faciliter la circulation des documents, tout en respectant les règles propres à chaque réseau ou en autorisant la connexion de matériels hétérogènes. Ces échanges sont rendus possibles par l'utilisation des protocoles *OSI (Open System Interconnexion),* qui permettent de distinguer les niveaux d'échange des applications informatiques. Leur application dans les bibliothèques est souhaitée et soutenue par la Commission des Communautés Européennes ([CAI89a], [CAI89b]).

Ces deux règles de l'UAP et du CBU sont indispensables pour la conduite des opérations nationales et internationales de développement des acquisitions documentaires et d'informatisation des bibliothèques. Il convient cependant d'en relativiser les implications dès que l'on s'adresse à des bibliothèques particulières, notamment à de petites bibliothèques.

La première remarque concerne le catalogage des ouvrages. L'AFNOR, suivant en cela l'exemple des bibliothèques américaines, a établi trois types de normes de catalogage, en proposant deux décalques moins complets que la norme intégrale. Le catalogage complet est long et difficile. Il s'agit d'un travail de spécialistes, qui doit être suivi par la Bibliothèque Nationale. Les bibliothèques

municipales et universitaires, du moins quand elles cataloguent leurs fonds courants, peuvent se contenter de notices ayant moins d'informations. La circulaire de la DBMIST [DBM88] recommande ainsi *"d'adopter une description bibliographique allégée. En l'état actuel de cette question, elles doivent normalement recourir aux fascicules de documentation AFNOR sur la notice moyenne ou minimale"*.

La seconde remarque concerne l'adhésion préalable à un réseau pour lancer une opération locale d'informatisation. Cette idée largement répandue parmi la profession reste d'application difficile. Par exemple, le réseau SIBIL, qui regroupe plusieurs bibliothèques universitaires demande près de deux ans pour former le personnel et installer les terminaux. Cette confusion entre l'appartenance à une structure réseau, répondant à un choix organisationnel, et la nécessité d'offrir un service à un réseau d'utilisateurs, qui est le véritable objectif du programme *d'Accès Universel aux Publications,* est une importante source de blocages. C'est du rôle de la déontologie professionnelle, et de l'action des pouvoirs publics que de promouvoir l'accès universel aux publications. Cette opération a besoin d'un catalogue collectif de localisation, dont la forme reste à définir. C'est une autre opération que de concevoir la constitution d'un catalogue local, éventuellement par achat de notices bibliographiques auprès d'un producteur d'informations bibliographiques. Cette distinction est à la base du nouveau *"Schéma directeur de l'information bibliographique"* que tentent de mettre en place les pouvoirs publics.

3.c • L'échange de données bibliographiques : le format MARC

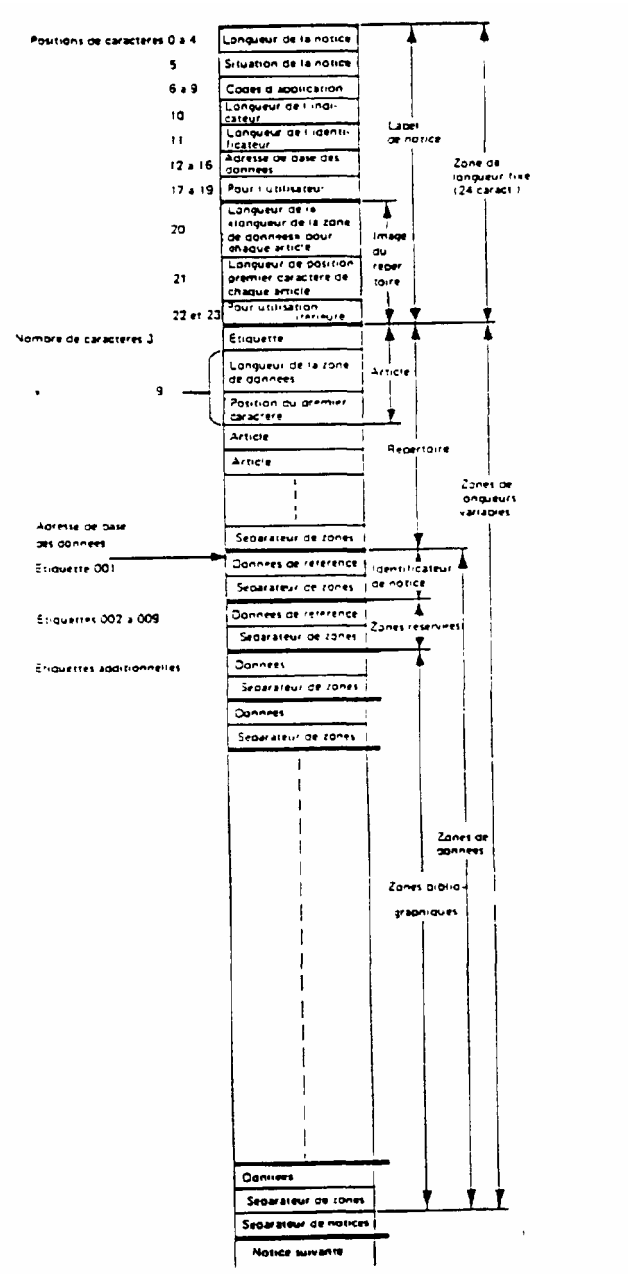
Le format MARC est un format d'échange des données, qu'il faut distinguer du format de traitement, propre à chaque système documentaire. Pour un échange en format MARC, les notices bibliographiques sont reportées sur un support magnétique, les unes après les autres. Les notices sont séparées par un code de fin de notice dit IS3 (ce code n'a pas d'équivalent graphique). Chaque notice est composée de trois parties :

. *le guide,* zone fixe de 24 caractères, chaque caractère représentant des informations sur l'enregistrement qui va suivre : longueur de l'enregistrement, statut de la notice (corrigé, nouveau, provisoire...), type de matériel, adresse du début du texte de la notice...)

. *le répertoire* indique les zones présentes dans la notice et le nombre de caractères de chaque zone. Chaque zone est représentée dans ce répertoire par 12 caractères. Les trois premiers indiquent l'étiquette de la zone considérée (par exemple 100 : vedette principale, personne physique). Les 4 suivants donnent la longueur de la zone (exprimée en caractères) et les 5 derniers expriment l'adresse du premier caractère de cette zone, calculée à partir du caractère 00000 représentant le début du texte de la notice (i.e. début de la zone 001).

. *le texte de la notice* est découpé en zones. Chaque zone est séparée de la suivante par le caractère IS2, qui n'a pas de représentation graphique (on indique cependant souvent ce caractère par le symbole : $)

La structure générale d'une notice en format MARC est donnée par le schéma suivant



Certaines zones du format MARC sont obligatoires et de longueur fixe : zone 001 - Numéro d'identification ; zone 004 - Liste des étiquettes corrigées ; zone 008 - qui définit sur 49 positions des informations générales sur la notice (date, niveau intellectuel, type de publication, sujet principal...). Les autres zones sont de longueur variable et peuvent être répétitives et divisées en sous zones, elles-mêmes répétitives. Les zones "bibliographiques" ont des étiquettes allant de 010 à 999 *("selon les besoins"* indique la norme ISO-2709).

Un système de gestion de bibliothèque devrait être capable de transformer les données de son format interne spécifique en format MARC et réciproquement. Il convient toutefois de s'entendre sur la signification apportée aux diverses étiquettes, et éventuellement de limiter le jeu d'étiquettes (i.e. de zones de description) en fonction de besoins spécifiques (une bibliothèque pour enfants n'a pas besoin de toutes les zones ; les références bibliographiques pour le dépouillement d'articles ne correspondent pas aux même critères que les ouvrages anciens...). La transparence ne joue guère dans ce domaine, et si chaque constructeur déclare traiter des données en format MARC, il y a souvent loin de la coupe aux lèvres.

Cette situation tend à minimiser le rôle du format MARC dans l'échange de données. Plusieurs normes de format MARC cohabitent, le format ayant le plus d'avenir (UNIMARC) n'étant même pas publié en France. De plus, chaque constructeur prenant ses libertés avec la norme, on aboutit à la nécessité de créer des interfaces spécifiques à chaque fois que l'on veut échanger des données en format MARC. Cette situation n'est cependant pas catastrophique. L'existence du format MARC sert de guide à la pratique de l'échange des données. En ce sens il est très utile car les divers formats MARC ne divergent que sur des points de détail, qui peuvent être traités dès lors qu'ils sont connus. La fixation sur une notion rigide du format d'échange a pour seule conséquence concrète de conduire les concepteurs de logiciels à ne pas publier leur version du format MARC.

3.d - La notion de catalogue collectif

La notion de catalogue collectif est souvent vécue comme intimement liée à l'existence d'un super réseau, où chaque bibliothèque partagerait le catalogage.

Or il ne s'agit là que d'une hypothèse pour l'existence d'un catalogue collectif. Le *"Schéma directeur de l'information bibliographique"* publié en juillet 1989 par le CESIA [MIN89] est très clair : *"Bien distinguer les concepts de catalogue collectif et d'information bibliographique".*

Un catalogue collectif est un instrument qui dépend de la pratique collective des bibliothèques. Il n'a de sens que si l'on accepte la mise à disposition des ouvrages à l'échelle de toute la zone d'influence de ce catalogue collectif (ville, région,...). Son objectif est donc inséparable du fonctionnement du "prêt inter bibliothèques". A ce titre, un catalogue collectif peut fonctionner avec des notices très dépouillées.

Un catalogue collectif ne peut pas non plus être confondu avec le catalogue propre d'un établissement, qui comporte des points d'accès par sujet suivant des hypothèses d'indexation qui sont adaptées à chaque public, et des informations locales. Il ne peut pas non plus être confondu avec les nécessaires ressources bibliographiques nationales, dont le caractère scientifique et la complétude sont des éléments primordiaux. Ce qui doit en revanche être envisagé, c'est la possibilité pour les bibliothèques de nourrir ce catalogue collectif à partir de leurs ressources propres. Le sens d'utilisation est celui de la coopération librement consentie entre les bibliothèques. L'exemple type est le fonctionnement du CCN *(Catalogue Collectif National des Périodiques),* alimenté par chaque bibliothèque participant au réseau, et particulièrement utile pour la localisation des périodiques, même si les notices n'ont pas la précision des notices bibliographiques respectant les normes de catalogage.

Le problème posé aujourd'hui concernant ce type de catalogue collectif est lié à la faible intervention de l'Etat, qui tendait, jusqu'à l'élaboration de ce schéma directeur, à confondre à l'intérieur du projet *Pancatalogue* ([PEN87]) cette tâche avec celle de la fourniture de notices bibliographiques.

Le catalogue collectif est un système raisonné, qui permet une sélection des ouvrages par ville ou par région, ou éventuellement selon les spécialités des bibliothèques participantes, et la disponibilité probable des ouvrages. On peut dans ce cadre envisager deux types de catalogue collectif :

. un modèle centralisé, décalqué sur SIBIL ou OCLC, dans lequel

chaque bibliothèque apporte ses informations de localisation à une vaste banque de données bibliographiques. Cette conception va de pair avec la notion de "catalogage partagé".

. un modèle réparti, dans lequel chaque bibliothèque conserve la liberté de définition de son catalogue en fonction de ses besoins propres, l'interrogation de localisation d'un ouvrage précis étant effectuée par un anté-serveur, qui recherche successivement, en fonctions de certaines heuristiques (proximité géographique, grand domaines de spécialité...), les divers catalogues pour proposer une localisation d'un ouvrage clairement identifié.

Pour des raisons d'efficacité, un panachage des deux types de catalogues collectifs (catalogue magnétique ou catalogue virtuel) doit être envisagé, les grands établissements versant leurs ressources dans une banque de données couvrant la majeure partie des besoins (le modèle de la BL-DSC, de l'INIST ou du réseau des CADIST), et les petits établissements étant sollicités par le système anté-serveur pour des cas particuliers (préférence d'une localisation géographique, ou absence du livre dans le catalogue de première instance) [PAQ89].

Cette solution laisse de plus une grande marge de manœuvre aux collectivités territoriales (villes, département, régions) pour promouvoir des opérations de coopération particulières entre les diverses institutions publiques ou privées de leur zone de compétence. A l'heure de la construction de l'Europe, où des bassins économico culturels se constituent par delà les frontières nationales, cette solution semble plus adaptée aux besoins et aux désirs des régions. Elle permet par exemple de définir des coopérations transfrontières à l'échelle du pourtour méditerranéen, voire de sous régions (Catalogne et Languedoc-Roussillon), ou à l'échelle de zones ayant une autonomie linguistique (Alsace, Euskadi) ou culturelle (Bretagne - Pays de Galles - Irlande).

3.e - La notion de serveur bibliographique

L'information bibliographique est un bien d'une tout autre nature que l'information de localisation. On définit l'information bibliographique comme *"tout ce qui est créé ou utilisé dans une bibliothèque en matière de descriptif*

normalisé de support documentaire" [MIN89]. L'information bibliographique est alors un bien marchand, que l'on peut créer, acheter ou vendre. Il est réutilisable si certaines normes sont respectées, et si les appareils de lecture sont adaptés à ces normes.

La situation de l'information bibliographique est semblable à celle de tous les produits culturels, notamment les produits audio-visuels. Les normes sont indispensables pour lire un disque ou une cassette de par le monde, mais elles peuvent changer, en fonction des intérêts et du poids économique des fabricants de matériels (passage du disque analogique au disque compact, nouveau codage D2MAC des images télévisuelles, passage à la Télévision Haute Définition...) [MIE89]. Il y a distinction entre le contenu culturel (l'information bibliographique) et le réseau de production/diffusion de ce bien.

Du point de vue d'une bibliothèque, il faut se placer suivant une logique consumériste : qui vend cette information ? A quel prix ? Est-elle adaptée à mes besoins ? Est-elle facilement ou au contraire difficilement récupérable dans mon système documentaire ? Est-il plus efficace, plus économique et plus satisfaisant scientifiquement d'acheter les notices ou de les créer ? Ces réflexions sont analogues à toutes celles qui sont mises en œuvre pour des achats d'autres biens culturels, notamment s'ils sont innovateurs (vidéodisques, disques compacts il y a quelques années...).

Une différence subsiste toutefois, mais devrait s'estomper dans les années qui viennent en proportion de la prise de conscience des coûts afférents : les bibliothèques ont la possibilité de créer elles-mêmes les notices. Le coût de ce travail reste cependant difficile à évaluer. Après une enquête dans les bibliothèques françaises, Geneviève Boisard [BOI89] estime que le coût d'une notice varie de 63 à 441 Francs, avec une moyenne de 151 Francs pour les bibliothèques de lecture publique et 257 Francs pour les bibliothèques universitaires. Ce coût varie très fortement entre les établissements, les petites équipes très surchargées cataloguant plus vite que les équipes de spécialistes. Il varie aussi en fonction du degré scientifique désiré : choix de la norme complète, précision de l'indexation matière... Dans tous les cas, le facteur principal dans l'évaluation du coût est le temps passé par le personnel, notamment par le personnel scientifique (conservateurs).

En regard de cette démarche consumériste, se situe une stratégie de l'offre. Or dans ce domaine, les choses avancent très lentement. On trouve actuellement différents acteurs proposant leurs services aux bibliothèques. Si l'on met de côté les sociétés de service qui proposent la conversion rétrospective de tous les types de catalogues de bibliothèques à partir de la version sur fiches (par exemple JOUVE), on trouve des producteurs publics et privés de notices bibliographiques informatisées.

Il règne une absence de définition des règles du jeu de ce marché :

. Quels droits de propriété littéraire et artistique peut revendiquer l'Etat sur les notices crées sous son autorité : les notices de la Bibliothèque Nationale, mais plus généralement celles de l'INIST ou des Bibliothèques Universitaires ?

. Quels droits ont les acheteurs de notices notamment s'il s'agit d'instances collectives (un serveur régional de données bibliographiques ou un réseau de coopération entre bibliothèques) ? Ces acheteurs peuvent-ils redistribuer les données ? Sous quelles formes ? A quelles conditions ?

. Quelles sont les garanties de pérennité des informations achetées ? Par exemple en cas de changement de matériel, une bibliothèque peut-elle transcoder les informations achetées auparavant ? Ou doit-elle à nouveau racheter les notices ?

En marge de ces problèmes juridiques, on est confronté à une situation où sont confondus les problèmes concernant la production des notices et ceux concernant leur diffusion technique. La diffusion des données regroupe deux services : la sélection des notices intéressant une bibliothèque particulière et le transcodage de ces références pour l'adapter au matériel disponible dans cette bibliothèque.

Le *Schéma directeur de l'information bibliographique* s'appuie sur une distinction stricte entre la production des notices et leur diffusion. En cela, il rompt radicalement avec la logique antérieure du "catalogage partagé". Le catalogage partagé est organisé en France par les réseaux LIBRA (Bibliothèques

Centrales de Prêt et quelques bibliothèques municipales) ou SIBIL (bibliothèques universitaires) et à l'échelle mondiale par OCLC. Il s'appuie sur deux principes qui sont à mon avis aujourd'hui caducs :

. une grande banque de données collective regroupe toutes les informations bibliographiques, et chaque bibliothèque ajoute des données de localisation.

. une bibliothèque qui ne trouve pas dans cette banque de données les informations qu'elle souhaite est chargée de nourrir le système avec ses propres notices. Il y a là une situation déséquilibrée. La récupération des notices est un acte commercial (achat), alors que la fourniture est un acte coopératif. Il s'agit de deux logiques différentes, qui impliquent différemment chaque entité.

Cette distinction entre les deux logiques est d'autant plus grave que le fournisseur de notices bibliographiques est situé hors des frontières, plus encore aux Etats-Unis comme c'est le cas du réseau OCLC. La dépendance culturelle qui en découle (l'indexation matière est principalement rédigée en anglais) est doublée de la fourniture à une société multinationale d'origine américaine d'une information de grande valeur (i.e. les descriptions bibliographiques des fonds rares, anciens et précieux des bibliothèques françaises). On peut même ajouter que cette situation est de type colonial, quand le pays le plus fort offre au plus faible le simple droit d'acheter ses produits, alors qu'il ne rétribue pas à sa juste valeur les biens qu'il récupère (achat d'une notice à 17 F., ristourne de 4 F. par notice crée !). Quand on pense que c'est l'Etat qui a signé cette soumission au nom des bibliothèques universitaires, on prend conscience du besoin de redéfinir les enjeux des réseaux d'information et leurs implications géopolitiques mondiales.

Par exemple, quelles garanties avons-nous que les prix des notices créées par les instances de l'Etat ne seront pas revendues à des prix très élevés à d'autres instances de l'Etat dans quelques mois ou quelques années ? Cette interrogation est une formule de rhétorique : d'ores et déjà les bibliothèques universitaires françaises , dépendant de l'Etat, achètent les notices crées par d'autres bibliothèques universitaires françaises dépendant elles aussi de l'Etat, et tout cela de manière incontrôlée, puisque le financement est assuré par l'Etat, sans implication réelle des bibliothèques acheteuses. Quelles garanties avons-nous que

la fourniture de notices ne cessera pas quand cela apparaîtra plus rentable (économiquement ou politiquement) à la société qui possède le fonds bibliographique enrichi par nos soins ? N'oublions jamais qu'il n'y a pas, à l'échelle économique mondiale de cadeaux au nom de l'amitié des peuples et des grands desseins culturels. N'oublions pas que les mesures de rétorsion économique existent dans les échanges commerciaux (par exemple dans le domaine agro-alimentaire entre l'Europe et les Etats-Unis).

En regard de ces incertitudes et de cette soumission à un producteur incontrôlable, la nouvelle structure proposée par le schéma directeur apparaît autrement plus logique et cohérente. Cette structure abandonne la notion de catalogage partagé pour celle de la diffusion par des organismes spécialisés des notices bibliographiques. On distingue alors le travail de production des notices de celui de leur commercialisation.

- la production des notices est assurée par l'Etat ou par des producteurs privés (Le *Cercle de la Librairie* en France). Les notices produites par l'Etat sont vendues à un organisme chargé de leur diffusion. Cette vente s'accompagne de deux limites : l'Etat lui-même ne distribuera pas ses notices parallèlement à cet organisme, et la cession des notices est libre de droits. Le prix de vente des notices par l'Etat est fonction de décisions politiques. Trop faible il obérerait la possibilité de concurrence avec les producteurs privés, trop élevé, l'Etat se trouverait dans la situation de racheter, par le biais des bibliothèques universitaires, à un prix très élevé des notices dont il finance la création par ailleurs (Bibliothèque Nationale, grands organismes documentaires).

- la diffusion des notices de l'Etat est prise en charge par un établissement national de type EPIC (Etablissement Public Industriel et Commercial), les producteurs privés étant libres de se placer aussi sur ce créneau *(Electre* principalement). Cet établissement national devrait gérer un serveur de données dont les missions seraient [MIN89] :

. organiser la mission de service public de diffusion des notices (collecte des données, actualité des données, administration de la base et des accès télécommunicants...)

. permettre une sélection aisée des notices intéressant une

bibliothèque particulière. Le système doit pouvoir fonctionner en ligne, en mode interactif, pour les vérifications de notices, notamment celles qui ne peuvent être recherchées par le simple numéro ISBN.

. offrir des services de conversion de format pour s'adapter aux divers systèmes informatiques de ses clients. Ce service viendrait bien entendu en sus de la vente des notices dans un format implicite (de préférence UNIMARC).

. assurer la promotion et la commercialisation des notices bibliographiques françaises, notamment dans des opérations internationales

. concevoir et commercialiser des sous-produits dérivés à partir des notices bibliographiques (Disques Optiques Compacts, catalogues thématiques...).

Agissant sur un marché, l'établissement public pourrait se charger de collecter d'autres types d'informations bibliographiques auprès de fournisseurs différents pour d'autres médias que le livre (dépouillement d'articles de périodiques, phonogrammes, vidéogrammes...). L'existence d'un tel prestataire de service devrait être à même de débloquer la situation, notamment en ce qui concerne la régularité de l'approvisionnement en notice en particulier pour les ouvrages récents (une grande faiblesse de la *Bibliothèque Nationale)* et le caractère scientifique du catalogage (une faiblesse *d'Electre,* qui pratique souvent le catalogage à partir des publicités d'éditeurs) [TES89].

Ces propositions, et surtout la volonté des pouvoirs publics de les voir aboutir, volonté marquée notamment par le lancement d'un appel d'offre durant l'été 1990 pour la création du serveur national, permettent de poser de façon beaucoup plus ouverte la question du réseau des bibliothèques. L'initiative est enfin redonnée aux établissements et aux collectivités locales, qui doivent faire leurs choix dans un monde ouvert (plusieurs propositions possibles) et dont l'avenir est assuré par le marché (peut-on acheter des notices et à quel prix ?) et non par des décisions administratives, qui prennent plus souvent la forme de projets nébuleux que de réalisations concrètes (le pancatalogue par exemple).

4 - L'architecture générale du système documentaire

4.a - Les critères du choix

L'informatisation de la bibliothèque scientifique de l'Université de Caen a été mise en œuvre avant que le projet de serveur national de données bibliographiques ne voit le jour. Nous avons donc choisi de ne pas récupérer de données bibliographiques. Ce choix correspondait avec la volonté d'agir rapidement, et au moindre coût. Il pouvait se justifier si l'opération d'informatisation était plus rapide que la maintenance du catalogue sur fiches. Dans cette hypothèse, comme nous fonctionnions sans aucun achat de matériel, le choix était de toute façon positif. L'expérience a montré que ce pari était justifié, puisque en un an, nous avons pu nourrir le catalogue avec tous les livres déplacés et mis en libre accès et toutes les nouveautés.

En juillet 1990, le catalogue comportait ainsi :

1276 Monographies
1374 Manuels d'enseignement
910 Livres d'exercices
34 Usuels
29 Rapports
305 Thèses
482 Mémoires d'étudiants
548 Titres de périodiques
1005 Articles issus des *Techniques de l'Ingénieur*
6846 Articles de revues de vulgarisation scientifique.

Ce catalogue a été constitué en plus du travail quotidien de la bibliothèque (achats, gestion, recherches documentaires informatisées, service du public, prêt inter bibliothèques, mise en place du libre accès), sans personnel supplémentaire. On peut donc estimer que le coût d'une notice, indépendamment même du fait qu'il était moins important que celui de l'entretien du catalogue manuel, est de moins de 40 Francs si l'on ne considère que les livres. Elle tombe à 15 Francs si l'on ajoute les articles de périodiques (considérant que nous avons bénéficié du don d'un fichier de 6 000 titres par la bibliothèque de l'Université de Pau). Bien entendu, des coûts de ce type sont associés au fait que nous ne suivons

pas les règles pointues du catalogage scientifique. Nous avons en effet adopté les normes minimales AFNOR [AFN85] et choisi une indexation libre, sans utiliser la liste d'autorité RAMEAU. En ce sens, ils faut relativiser ces économies dès lors que l'on voudrait étendre cette expérience. Il faut bien que quelqu'un se charge en France du catalogage scientifique. Mais dans notre cas, nos responsabilités en ce domaine sont très limitées : nous ne sommes pas CADIST, et nous ne possédons aucun fonds rare, ancien ou précieux. Au contraire, la majeure partie de nos ouvrages possèdent un numéro ISBN, ce qui devrait faciliter les échanges et les insertions futures dans des projets nationaux.

On pourrait résumer les conséquences de nos choix en deux rubriques :

Inconvénients :

- catalogage descriptif simplifié, principalement adapté aux besoins du public, mais isolé des projets nationaux.

- indexation matière par descripteurs libres, moins uniforme que l'indexation sur liste d'autorité. De ce point de vue, nous ne participons pas à l'amélioration et à la mise à jour de RAMEAU.

- le système ne sait pas encore lire des notices en format MARC et les incorporer. Il ne sait pas encore transformer le format de données interne en format MARC. Il s'agit là d'une amélioration qu'il faudra apporter rapidement. Si on se limite aux zones principales du format MARC, cela reste possible.

- l'utilisateur ne peut pas connaître la position d'un livre au moment où il interroge le catalogue. De même, il ne sait pas le nombre d'exemplaires possédés par la bibliothèque, ce qui lui donnerait au moins des indications générales. Ce dernier point devrait être adapté rapidement.

- il n'existe pas de thésaurus en ligne, ce qui limite la couverture des recherches documentaires. Il est envisagé cependant de réaliser une extension automatique des recherches de livres en utilisant les codes de classement.

- le système n'est pas finalisé comme un produit commercial, ce qui rend

nécessaire la présence d'une personne connaissant *Texto,* pour régler certains cas particuliers. De même, la portabilité n'est pas garantie par un contrat en bonne et due forme en cas de changement de matériel, même si elle reste possible par toute personne maîtrisant *Texto*. Il convient aussi de renforcer les programmes de transfert de données et les capacités du néophyte à intervenir dans le fichier pour corriger des données (moyennant la résolution de problèmes de sécurité des accès).

- l'accès vidéotex n'est pas encore établi. Seul l'accès par minitel 1B est possible. L'utilisation du réseau Télétel (36-14) est aujourd'hui impossible pour des problèmes techniques liés au Centre de Calcul.

Avantages

- le système a pu être mis en place rapidement, répondant aux besoins liés au passage en libre accès des ouvrages.

- le coût du système est dérisoire. Si l'on exclut le temps de travail du personnel, qui de toute façon est affecté à la bibliothèque pour gérer son fonctionnement quotidien, les seuls frais sont liés à l'utilisation de l'ordinateur du Centre de Calcul, c'est-à-dire correspondent à des transferts de charge à l'intérieur du budget de l'Université. La création du logiciel, le transfert des données et la mise au point des index a ainsi coûté à la bibliothèque scientifique 22 185 Francs en un an et demi (de janvier 1989 à septembre 1990).

A l'échelle globale de l'Université, cette somme ne correspondant qu'à de l'utilisation de matériel déjà disponible, ce catalogue n'a strictement rien coûté.

- le système peut aisément intégrer des informations provenant d'autres systèmes documentaires, comme l'a montré l'insertion de plus de 6 000 références d'articles provenant d'un fichier *dBase* produit à la bibliothèque de l'Université de Pau. Conçut sur la base d'un logiciel documentaire, il est adapté aux besoins de la communauté scientifique, qui peut envisager des échanges de données, notamment avec les banques de données bibliographiques.

- la recherche documentaire par le public est simple. Les chances de succès, pour peu que des documents répondant à la question soient présents, sont

assez grandes car le système est ciblé sur l'amélioration du taux de couverture aux dépens de la précision. Cela correspond aux besoins des étudiants qui fréquentent une bibliothèque scientifique, même si cela reste souvent trop général pour la recherche.

- les échanges avec tout système acceptant un format ASCII balisé sont possibles.

- le système possède des informations écrites en utilisant tous les caractères de la langue française, y compris les signes diacritiques. Les écrans sont de ce fait de lecture facile.

- l'accès réparti, à partir d'un minitel 1B, est immédiatement possible. Cela correspond à la vocation de notre bibliothèque d'être un centre de ressource documentaire scientifique et technique à l'échelle régionale.

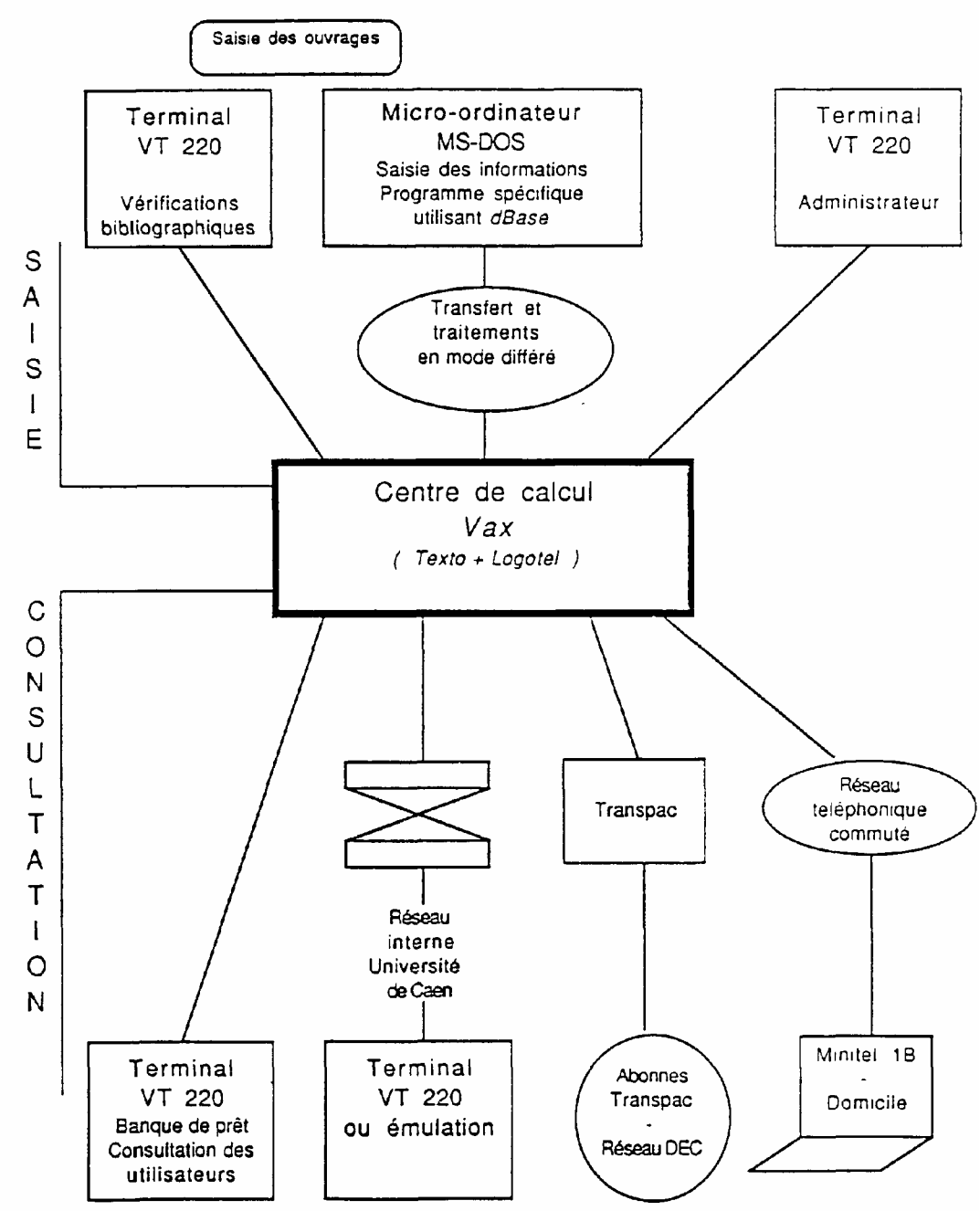
4.b • L'architecture informatique du système

Notre choix d'informatisation a voulu s'appuyer sur les instruments disponibles à l'intérieur de l'université. Ce choix nous a conduit à concevoir un système réparti, basé autour d'un ordinateur de puissance moyenne *(Vax* de *D.E.C. - Digital Equipment Corporation)* accessible par le réseau de télécommunication, aussi bien à l'intérieur de l'université qu'à partir d'un poste téléphonique ou d'une autre université appartenant au réseau DEC. Les logiciels *Texto* et *Logotel* étaient déjà implantés sur l'ordinateur, un disque dur spécifique étant réservé aux applications documentaires. Depuis plusieurs années le *Centre de Calcul de l'Université de Caen (C.C.U.C.)* héberge sous *Texto* des applications internes : la bibliothèque de l'Institut d'Histoire Ancienne, une banque de données de Thiochimie [MET88], le catalogue des produits chimiques de l'Université...

Le logiciel *Texto,* s'il est assez puissant et capable de gérer des banques de données importantes (près de 200 000 références dans le cas de *Téléthèses),* reste d'un abord ardu. L'interface utilisateur est marqué par sa conception déjà ancienne, notamment par l'utilisation du mode ligne à ligne. Même si des efforts ont été faits récemment pour réaliser un mode de saisie pleine page, on doit conclure à un échec, la saisie n'étant guère plus conviviale. L'utilisation

disproportionnée de l'unité centrale de l'ordinateur par ce mode de saisie rend de toute façon son utilisation prohibitive dans une architecture répartie. Nous avons donc choisi d'opérer la saisie des informations en mode local sur un micro-ordinateur compatible MS-DOS. Les fichiers sont versés en différé dans la banque de données. Cette opération est l'occasion de recomposer les informations et de compléter l'indexation en traitement par lot *(batch)*. Cette organisation est d'ailleurs utilisée par de nombreux concepteurs sous *Texto*. Par exemple la bibliothèque de l'Université de Metz, qui utilise un masque de saisie sur micro développé en Turbo-Pascal.

Cela nous conduit à l'architecture suivante :



Les modes d'accès en ligne pour l'interrogation du catalogue sont alors les suivants :

- terminal VT220+modem 9600 bits par seconde à l'intérieur de l'université. Le modem subvocal utilisé permet de conserver l'utilisation du poste téléphonique, ce qui a permis l'ouverture du service sans aucun frais d'installation. Il y a trois accès de ce type à l'intérieur de la bibliothèque scientifique. Un offert au public, et deux autres pour le travail interne. Tout possesseur d'un terminal ou d'une émulation de terminal VT sur micro-ordinateur peut utiliser une formule de ce type pour accéder au catalogue. Cela concerne les utilisateurs réguliers du Centre de Calcul, ou les membres du réseau interne de l'université, qui est en voie de développement et qui doit relier toutes les machines hétérogènes utilisées pour l'enseignement ou la recherche.

- Accès par le réseau DEC *(Digital Equipment Corporation)* des universités. Ce réseau permet de connecter tous les ordinateurs VAX situés dans les universités françaises. On peut donc interroger le catalogue depuis de nombreuses universités. Cela permet d'envisager des opérations de collaboration entre bibliothèques, notamment pour des opérations de dépouillement de périodiques.

- Accès depuis un terminal minitel 1B (minitel bi-standard, qui fonctionne alors en mode ASCII français, 80 colonnes).

Deux voies sont possibles :

. par le réseau téléphonique commuté, l'accès 31-45-55-56 offrant une entrée directe sur le VAX du CCUC.

. par le réseau vidéotex Télétel2 (36-14). Le terminal est cependant basculé en mode 80 colonnes, la version vidéotex réalisée par des étudiants de licence d'informatique [GHE90] n'ayant pas encore été finalisée et implantée sur l'ensemble de la banque de données. On peut d'ailleurs s'interroger sur la capacité du vidéotex à offrir un service de consultation de banque de données. Il y a peu d'informations sur l'écran, ce qui conduit à des manipulations fréquentes, et les caractères sont souvent peu lisibles dès que la page-écran est remplie. Le mode mixte (écran 80 colonnes, mais utilisation des touches de fonction vidéotex) est certainement plus adapté, et un développement de ce type est

envisagé. L'accès par le réseau Télétel reste cependant soumis à la résolution d'un problème d'écho sur le terminal. Le type de liaison actuel entre l'ordinateur, l'autocommutateur de l'université et le réseau Télétel conduit à annuler l'écho du PAVI. Dans les conditions actuelles, l'utilisateur ne peut pas voir lettre par lettre ce qu'il frappe au clavier et ne reçoit son propre message qu'après l'avoir envoyé à l'ordinateur (écho du VAX).

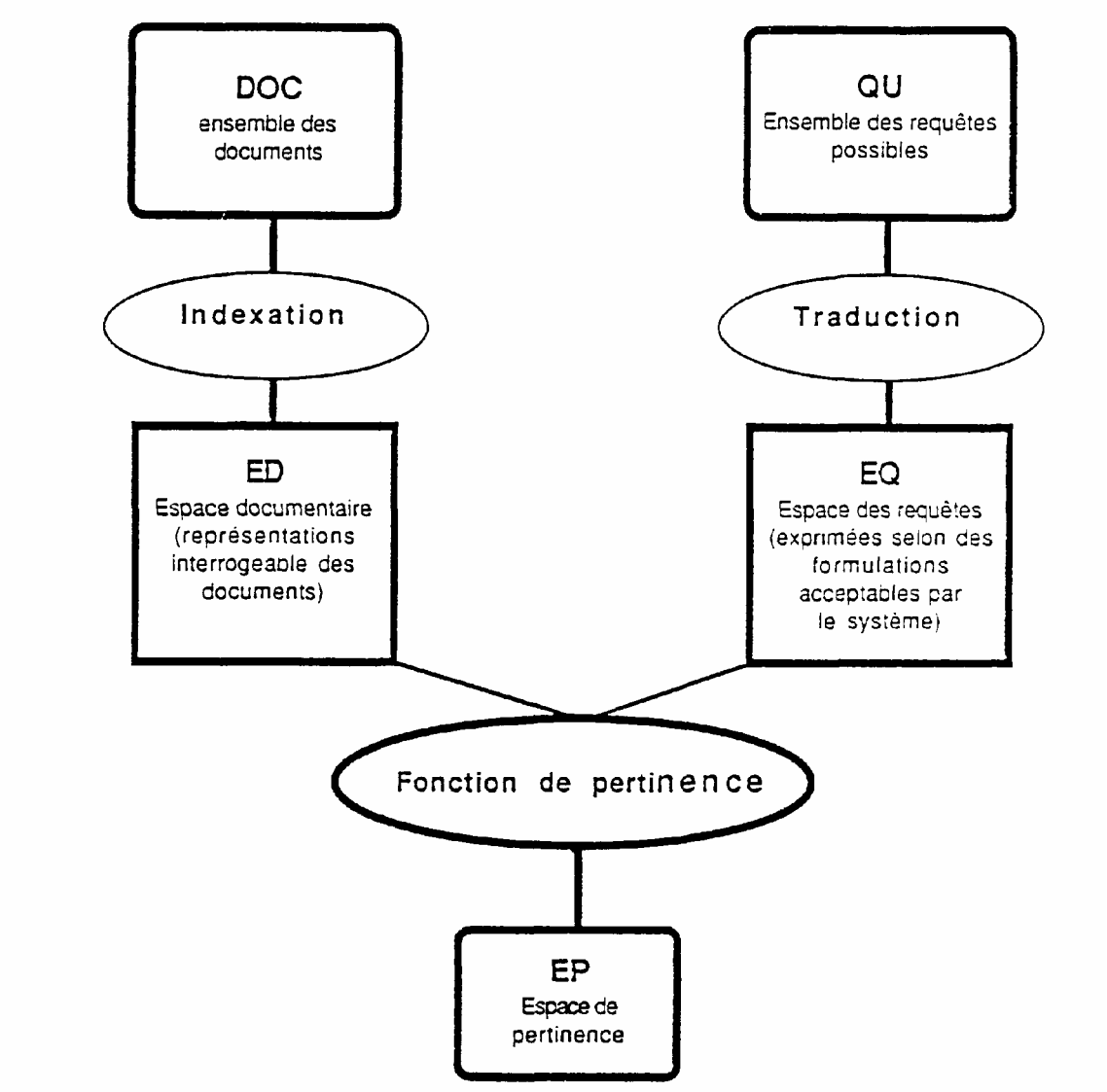
L'accès par terminal minitel 1B a été envisagé dès le début du projet. Il est en effet indispensable qu'un chercheur d'université, qui n'est pas toujours présent sur place, puisse connaître à tout moment les ressources de sa bibliothèque. La répartition géographique du campus dans la ville plaidait aussi pour ce type d'accès depuis les laboratoires, en l'attente de la finalisation du réseau de la recherche. Le type d'interface prévu pour l'interrogation de *Texto* par *Logotel* repose de toute façon sur le mode ligne à ligne, ce qui a facilité la conception simultanée de la version pour terminal et pour minitel 1B.

II - Présentation et mise en place du catalogue informatisé.

Reprenons la définition donnée en introduction d'un système documentaire :

$$SD := < DOC,\ QU, ED, EQ, EP, f_i, f_q, f_p >$$

Nous allons considérer sous cet angle chacune des propriétés du système documentaire constituant le catalogue de la bibliothèque scientifique de l'Université de Caen.



1 - Le système documentaire choisi

1 a • DOC, ensemble des documents.

Le catalogue contient des références bibliographiques, qui constituent les documents du point de vue du système. Il y a pour l'instant cinq types d'objets documentaires décrits par leurs références bibliographiques :

. des livres.

Les livres sont majoritairement ceux qui sont conservés à la bibliothèque scientifique. Le développement du système peut cependant intégrer des ouvrages appartenant à d'autres institutions de l'université mais confiés à la bibliothèque sous certaines conditions (prêt réservé...), ou des ouvrages localisés en divers endroits sur le campus (laboratoires...). La description des livres contient les indications traditionnelles du catalogage descriptif (auteur, titre, collection, éditeur, année de publication...) telles qu'elles sont définies par la norme Z 44-072 : *"Catalogage des monographies : description bibliographique minimale"* [AFN85]. Une grande partie des livres est composée de manuels d'enseignement, dont la description documentaire est difficile *("Cours de Physique - DEUG B").*

Voici la description d'un livre telle qu'elle apparaît à l'écran :

La Cellule : Biologie moléculaire

Darnell, James ; Lodish, Harvey ; Baltimore, David

Vigot, 1989. - XXXVI-1189 p. : fig. en noir et en coul.

I.S.B.N. : 2-89137-085-6

Ouvrage conservé à la **Bibliothèque scientifique**

Salle du Libre-Accès - Cote : 718 - 1006

Indice principal : 718 - Biologie moléculaire

. des titres de périodiques.

La référence représente alors l'état des collections de périodiques de la bibliothèque scientifique. La description est limitée au titre, aux dates de début et de fin de la collection, et à leur localisation dans la bibliothèque (salle, étage, travée).

Biochemistry and Cell Biology : Biochimie et Biologie Cellulaire

Périodique scientifique - Mensuel - I.S.S.N. : 0829-8211
Titre reçu par abonnement

Début de la collection :1986, Vol. 64 - Périodique en cours
Salle Pasteur - Cote : BIO 3 - Bas - Travée 5

Fait suite à : Canadian Journal of Biochemistry and Cell Biology (Issn : 0714-7511) en 1986

. des articles de revues de vulgarisation scientifique.

6000 articles sont ainsi décrits. Le fichier a été récupéré auprès de la bibliothèque de l'Université de Pau. La description est minimale : titre de l'article (quelques descripteurs sont éventuellement ajoutés entre parenthèses à la fin du titre) et source bibliographique (journal, année, mois). Toutefois, un essai de saisie plus complète des années 1989 et 1990 de certains titres est mis en place. Dans ce cas, un "résumé" est ajouté en prenant soit le résumé d'auteur, soit les principaux titres et intertitres de l'article. Ce travail n'est pas réalisé par un documentaliste, et ne prétend pas à l'exactitude de l'analyse. L'objectif est plutôt d'obtenir des indications de contenu au moindre coût.

Pour l'instant, nous avons privilégié les revues généralistes, afin d'offrir des instruments de travail aux étudiants pour la réalisation de mémoires. Les revues couvertes sont les suivantes :

- La Recherche
- Pour la Science
- Sciences et Avenir
- Science et Vie
- Microsystèmes
- Bulletin de la société linnéenne de Normandie.

Arc de Gibraltar, un volcanisme hors des normes.

de Larouzière, François Dominique

La Recherche. - 1989, Vol. 20, Num 214, Pages 1272-1275

Article de périodique

Salle du Libre-Accès - Cote : **BDP 1**

Comment convergent l'Afrique et l'Eurasie ? Une brèche dans le dogmatisme des sciences de la Terre ? Des laves considérées comme typiques. Quand la géophysique confirme les observations géologiques.

. le dépouillement des *"Techniques de l'ingénieur"*. Cette encyclopédie par fascicules est une mine de renseignements à la fois théoriques et pratiques. Elle reste largement sous-exploitée dans les bibliothèques universitaires, alors que les articles sont en général très synthétiques, écrits par des spécialistes du sujet, et peuvent constituer des dossiers de mise au point et des recueils de données et de formules de calculs de grande qualité. Il n'existe à notre connaissance aucun dépouillement de cette encyclopédie, et l'index imprimé reste très sommaire, et peu utilisable. Les fascicules traitent en général d'un sujet précis, qui s'intègre dans un volume plus large regroupant des sujets proches (bâtiment, informatique, métallurgie...).

Pour rester fidèles à notre logique de rendre un service au moindre coût (i.e. sans pouvoir malheureusement disposer de personnel supplémentaire), la

référence est reprise directement sur le fascicule, sans indexation supplémentaire. Elle est constituée du titre de l'article, et de son sommaire.

Systèmes d'irrigation

Clément, René/Galand, Alain

Les Techniques de l'Ingénieur - Volume :C5 - Fascicule :C5250-5252

Salle Linnée - Cote :

Structure d'un système d'irrigation. - Irrigation à la parcelle : irrigation par aspersion. - Irrigation à la parcelle : irrigation localisée. - Réseau collectif de distribution par canaux. - Réseau collectif de distribution par conduites sous pression. - Stations de pompage. - Ouvrages de transport

---

. les thèses et mémoires d'étudiants. Le catalogue décrit toutes les thèses soutenues à l'université et un certain nombre de mémoires d'un institut spécialisé dans le lait et la viande (ILVENUC).

Etude des possibilités d'action d'un test immunologique pour la détection de clostridium tyrobutyricum dans le lait et les fromages

Lorbat, Valérie

Univ. Caen, 1986. - 42 p.

Ouvrage conservé à la Bibliothèque scientifique

Ouvrage conservé en Magasin - Numéro : DEA-L-1986-1

Indice principal : 942 - Toxicologie alimentaire, contamination des aliments

Cet ensemble de documents est assez hétéroclite. Dans un premier temps, nous avons pensé proposer plusieurs fichiers séparés à l'utilisateur. L'expérience de la première version, qui allait dans ce sens a montré que l'utilisateur, dans sa pratique quotidienne, ne change pas de catalogue. Ainsi, alors que la majeure partie des recherches se fait par sujet, de nombreuses informations risquaient d'échapper à l'utilisateur qui n'aurait pas reconduit sa requête dans les autres fichiers. La version actuelle a donc regroupé l'ensemble des informations dans un même fichier. Bien entendu, cela provoque un bruit documentaire car nous n'avons pas de langage d'indexation fixe, et certains termes (analyse, thermodynamique, transport...) sont très généraux (présence dans les résumés ou les sommaires des articles), en même temps que spécifiques dans le cadre scientifique (titre de livres d'enseignement). Une solution éventuelle pourrait consister à décomposer automatiquement les requêtes qui extraient un grand nombre de références en plusieurs ensembles distincts selon le type de document. Sinon, pour les requêtes ayant peu d'occurrences, les utilisateurs semblent très satisfaits d'obtenir des articles de périodiques, ou des références qu'ils n'auraient pas pu trouver simplement en parcourant les rayons. De toutes façon, dans un système qui possède globalement peu de références (aujourd'hui 12 000 documents), il est préférable d'augmenter le taux de couverture, quitte à engendrer du bruit documentaire.

1.b - QU, l' ensemble des questions

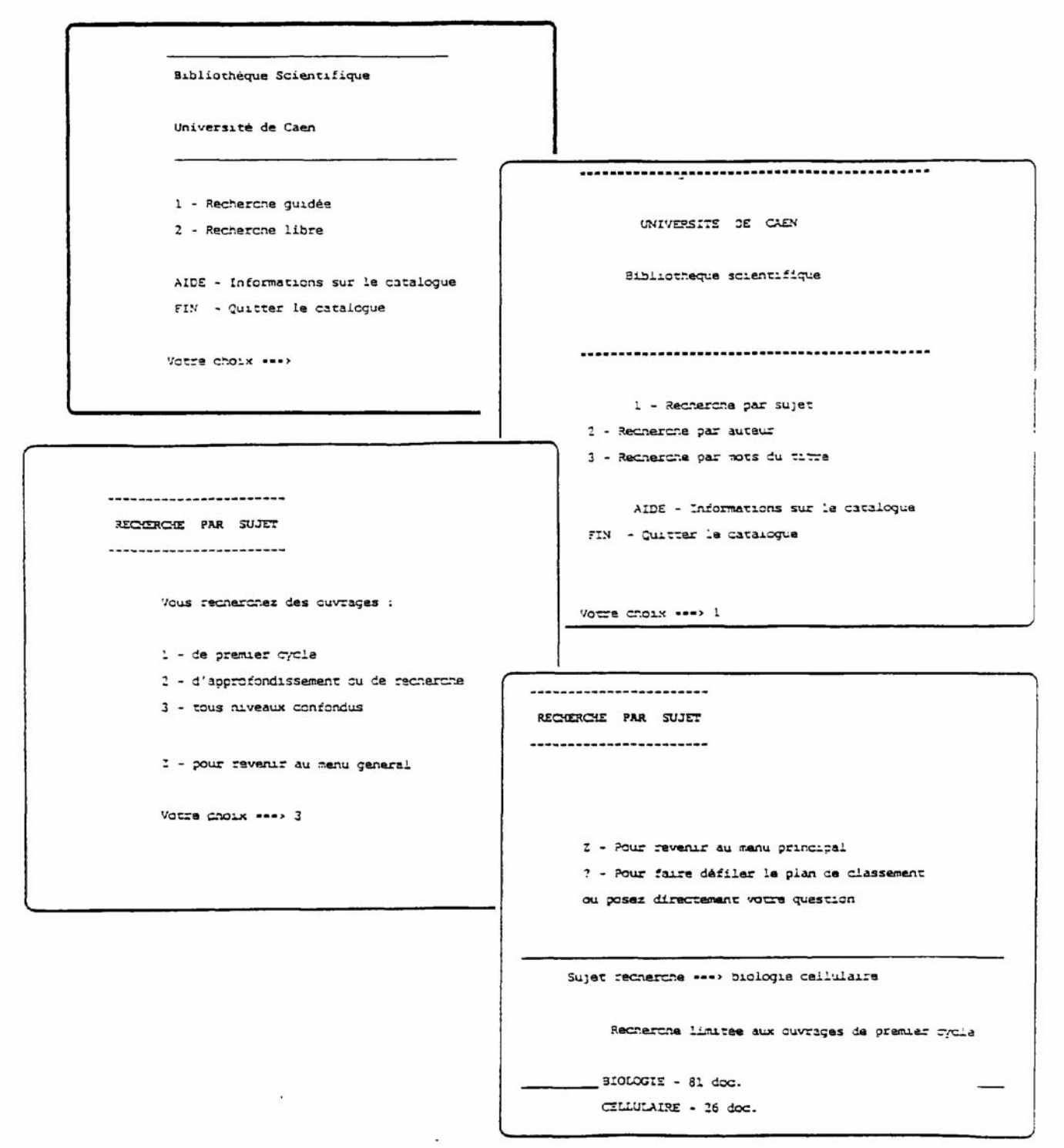
Plusieurs hypothèses ont été testées pour la formulation des questions par l'utilisateur. *Texto* fonctionne sur un mode booléen traditionnel, et les questions sont du type : C=A, où C est le nom d'un champ tel qu'il figure dans le fichier documentaire considéré et A le terme de la requête (ou l'équation booléenne entre les termes de la requête). Ce type de formulation, s'il est acceptable par un documentaliste formé, est très pénible et ardu pour un utilisateur. Non seulement il correspond à un utilisateur qui connaît la logique booléenne, mais de plus qui maîtrise le nom des divers champs et leur place dans le système (nom exact du champ, index associés, type de descripteurs...).

Au contraire, nous avons cherché à permettre à l'utilisateur de poser une question dans les termes qui lui viennent spontanément à l'esprit. Il nous semble

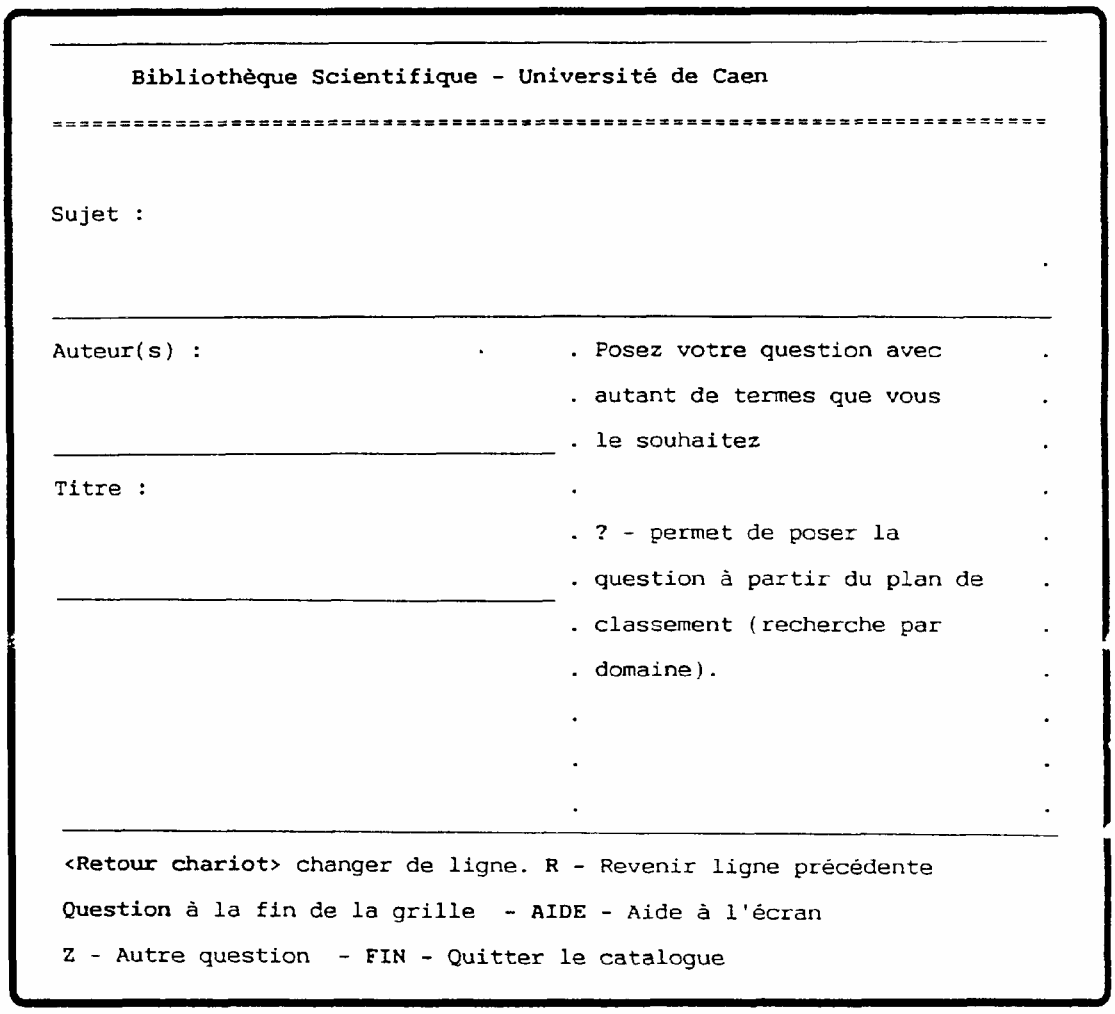
préférable d'obliger le système à se débrouiller avec les questions de l'utilisateur, plutôt que d'organiser des séances d'apprentissage qui se révéleront vite inutiles.

Il existe actuellement deux méthodes pour poser une question au système :

.*la recherche guidée* propose une succession de menus (questions à choix multiples) pour savoir dans quel champ mener la recherche.

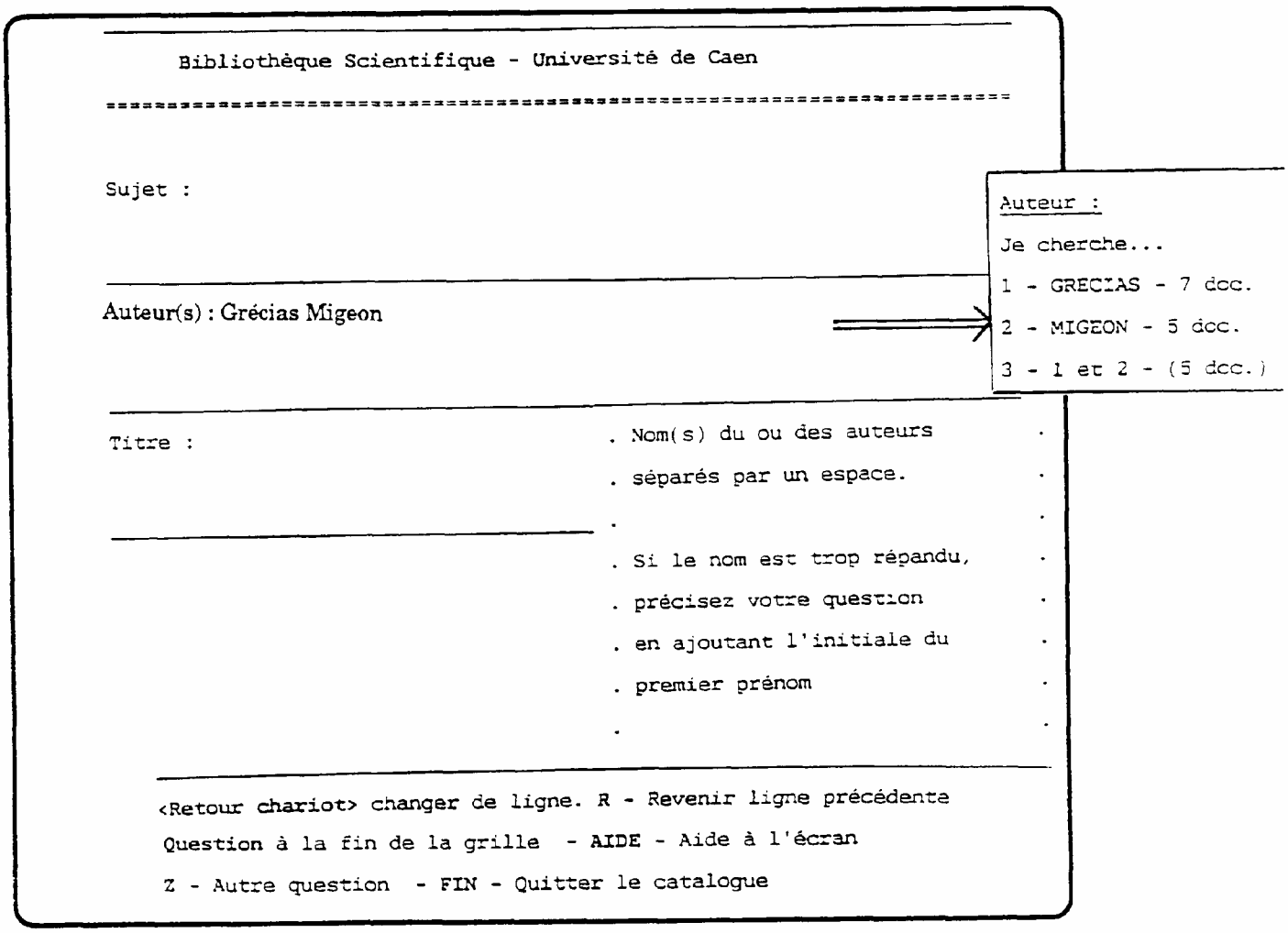


. *la recherche libre* propose une grille de saisie dans laquelle l'utilisateur se déplace avec les touches <*RC*> (<retour chariot>) ou *R*+<*RC*> (respectivement champ suivant et précédent).



Dans les deux cas, la formulation de la recherche par sujet ou par mots du titre est entièrement libre. On peut ainsi trouver une succession de mots *(technique analyse biologie cellulaire),* éventuellement séparés par des signes de ponctuation *(traitement, image),* des phrases (ou quasi phrases) complètes *(lichens comme bio indicateurs de contamination radioactive),* des mots reliés par des articles *(théorie des nombres)* ou par des traits d'union *(ferromagnétisme),* des expressions comportant des articles élidés *(systèmes d'information)...* De même, pour retrouver un titre précis, l'utilisateur peut indiquer le titre en entier ou seulement les mots significatifs. L'utilisateur peut saisir sa requête indifféremment en majuscules ou en minuscules, accentuées ou non.

La recherche par noms d'auteurs est plus contrainte : on ne donne que le nom de l'auteur, éventuellement une liste de noms séparés par des espaces ou des signes de ponctuation. La précision au niveau du prénom ne peut-être apportée que par l'initiale du prénom. En fait, cette limitation est acceptable car le fonds de la bibliothèque reste limité, et le bruit documentaire provoqué est acceptable. La formulation des noms de personne dans les systèmes documentaires reste un problème complexe. Une syntaxe rigide (du type Nom, Prénom) est souvent mal comprise, mais la reconnaissance de nom et du prénom par un système intelligent n'est guère envisageable (exp. Victor Hugo).



1.c - ED, l'espace documentaire et f_i, la fonction d'indexation

Les références bibliographiques sont indexées dans un espace documentaire relativement étendu (indexation libre). Il est composé de deux types de descripteurs :

1 - *des unitermes* en majuscules non accentuées.

Les unitermes sont obtenus en transformant en majuscules les mots significatifs pris parmi certains champs :

- les mots du titre, pour les livres comme pour les articles.

- les mots-clés qui sont éventuellement ajoutés lors de la saisie des livres. Cette indexation manuelle n'est pour l'instant réalisée que pour les ouvrages. Elle est libre, c'est-à-dire que nous n'utilisons pas de vocabulaire contrôlé, essentiellement pour des questions de coût (temps d'indexation). Les mots clés sont aussi constitués par la traduction en français des titres des ouvrages étrangers. L'indexation manuelle des livres est plus difficile que celle des articles, car les termes choisis sont tout de suite très généraux *(algèbre, circuits électriques...),* le livre couvrant un domaine plus large qu'un article. Si des mots clés sont ajoutés, ils sont saisis en typographie riche, sous la forme d'expressions habituelles du langage, avec les connecteurs linguistiques appropriés (traitement de texte, analyse d'images, transformée en Z,....). Le système se charge de retrouver les unitermes. Ce choix permet de conserver la richesse linguistique des expressions, qui sera peut-être exploitable par un système plus évolué.

- les mots automatiquement associés aux codes de classement pour les ouvrages (exemple : 358 -> Traitement du signal -> TRAITEMENT/SIGNAL)

- les mots du "résumé" pour les articles des revues ou des techniques de l'ingénieur.

Le choix de descripteurs unitermes entraîne le découpage des expressions composées utilisées dans l'indexation manuelle. Les expressions sont recomposées par l'utilisation de l'opérateur booléen ET. On peut ainsi retrouver *traitement d'images* (qui est le mot-clé donné par l'indexeur) par *TRAITEMENT*

et IMAGES. Bien entendu, cette solution peut entraîner un bruit documentaire *(les images intolérables des mauvais traitements infligés aux animaux de laboratoire...),* mais dans l'hypothèse d'une formulation libre confrontée à une petite banque de données (12 000 documents en septembre 1990), ce bruit semble préférable au silence qu'entraînerait le choix de descripteurs composés *(traitement d'images, traitement des images, traitement informatique de l'image....)*. De plus, ce choix est cohérent avec celui de ne pas indexer manuellement des documents (articles), et celui de pratiquer une indexation sans vocabulaire contrôlé, pour des raisons de coût.

2 - des codes de classification appartenant au plan de classement des ouvrages établi à la bibliothèque. La genèse et la structure de ce plan sont décrites plus loin. Les indices du plan de classement servent à l'indexation selon deux méthodes :

. *les indices eux-mêmes* sont portés dans l'index, ce qui permet la recherche par domaine directement en posant l'indice correspondant. Même si des champs distincts sont conservés pour les indices principaux et secondaires, ils sont reportés à un niveau équivalent dans l'index. Un système plus évolué pourra cependant utiliser ces deux champs pour une pondération de l'indexation.

. on verse aussi dans l'index *les mots qui définissent un indice* (e.g. 718 *->Biologie moléculaire)* et un certain nombre de mots associés à un indice (e.g. *480 - Connexionisme, réseaux de neurones* a pour mots associés *CONNEXIONNISME/RESEAUX/NEURONAUX)*. Ce travail est réalisé automatiquement lors du versement dans le catalogue d'un lot de nouveaux documents.

1.d • EQ, l'espace des questions et f_q la fonction de traduction

Cette définition de l'espace documentaire induit une définition de l'espace des questions manipulables par le système. Si les requêtes sont exprimées en formulation libre, indifféremment en caractères majuscules ou minuscules, accentués ou non, le système ne connaît que des Unitermes d'indexation en majuscule et des codes de classements. Les Unitermes doivent de plus être reliés par des opérateurs booléens.

La fonction f_q de traduction procède en décomposant la requête de l'utilisateur transformée en majuscules en unitermes, et implicitement recherche la conjonction de ces unitermes (opérateur ET). Le choix d'utiliser la conjonction répond au besoin de recomposer les expressions composées *(intelligence artificielle, biologie moléculaire,...).*

Plusieurs opérations sont réalisées lors de cette traduction de la question de l'utilisateur. L'algorithme correspondant à f_q est le suivant.

1 - transformer la requête en majuscules. La fonctions *majusc* de *Logotel* ne transforme pas les signes diacritiques. Il faut donc ajouter une routine qui recherche toutes les lettres pourvues de tels signes et les transforme en majuscules non accentuées. Cette routine est d'exécution relativement longue, car *Logotel* est un langage interprété. Toutefois, elle semble obligatoire si l'on veut offrir le maximum de liberté à l'utilisateur, notamment si on lui propose un clavier français. L'habitude de la frappe sur clavier se développe, qui conduit à des automatismes.

2 - décomposer le résultat en unitermes. On commence par éliminer les signes de ponctuation et les articles élidés (l', d', c', s'). Puis on remplace les espaces par le signe $. Un uniterme est alors défini comme la chaîne de caractères comprise entre deux $.

3 - Si la requête est composée d'un seul terme, on vérifie qu'il n'appartient pas à une liste pré définie de termes trop généraux. Cette liste est établie en fonction du cadre particulier d'une bibliothèque scientifique. La liste des termes traités est la suivante : INFO, INFORMATIQUE, MATH, MATHS, MATHEMATIQUE, MATHEMATIQUES, OPTIQUE, CHIMIE, GEOLOGIE, ELECTRICITE, ELECTROMAGNETISME, ELECTRONIQUE, ELECTROSTATIQUE, ALGEBRE, ANALYSE, MECANIQUE, PHYSIQUE, BIOLOGIE, GENETIQUE, THERMO, THERMODYNAMIQUE.

. Si la requête n'est pas dans cette liste on passe à l'étape 4.
. Autrement, on affiche le plan de classement à l'endroit correspondant au terme posé et on laisse l'utilisateur reformuler sa question en

indiquant un ou plusieurs codes de classification. Dans ce cas, la recherche ne peut trouver que des livres, car les autres types de documents ne sont pas indexés suivant le plan de classement. Cela n'est guère sensible, car l'utilisateur qui pose *biologie* ou *informatique,* voire même *info* ou *math,* à un catalogue cherche plutôt des ouvrages généraux.

4 - On élimine les unitermes qui correspondent à des mots outils. Le système ne fonctionne pas avec un anti-dictionnaire car l'ouverture de plusieurs fichiers sous *Texto-Logotel* est très longue. Seuls certains mots sont repérés en cascade : A, AUX, LE, LA, DANS, DE, DU, DES, LES, ET, OU, SAUF, SUR, UN, UNE, EN. On remarquera que les opérateurs booléens sont retirés à ce niveau, et ne sont donc pas traités par le système. On pourrait envisager un traitement dans certains cas simples (un seul booléen OU entre deux termes), mais une analyse morphosyntaxique permettant de déterminer le sens exact du connecteur (sens de la langue quotidienne contre sens en formulation booléenne) doit faire appel à une autre méthode de programmation.

5 - Prendre chaque uniterme successivement.

. si la dernière lettre est un S (uniterme du type AAS), on pose la question "AA ou AAS"

. autrement (uniterme du type AAA), on pose la question "AAA ou AAAS".

Cette prise en compte élémentaire du problème du nombre dans les requêtes n'est qu'une approximation. Elle est cependant efficace dans de nombreux cas. Les pluriels en S induisent une prononciation identique du terme, ce qui conduit l'utilisateur à oublier les diverses formes que peut prendre un mot. On pourrait améliorer cette opération en traitant d'autres cas simples (pluriels en X), mais les temps de traitement risqueraient de devenir trop longs pour un bénéfice réduit. Bien entendu, cette opération est transparente pour l'utilisateur, qui ne visionne le résultat qu'associé à l'écran avec le terme utilisé dans sa question.

6 - On effectue une conjonction booléenne entre les termes d'occurrence non nulle.

Cette fonction de traduction reste relativement sommaire. On est loin des traitements de la langue proposés par les systèmes d'intelligence artificielle. Mais elle s'appuie cependant sur quelques opérations de "bon sens", qui peuvent grandement aider l'utilisateur, notamment les pluriels en S, la suppression des articles, et particulièrement des articles élidés et le passage en majuscule de l'ensemble des caractères, une opération qu'oublient de nombreux systèmes travaillant avec *Texto*. De plus, comme le fonctionnement logique du système apparaît à l'écran, par l'affichage des unitermes posés et des réponses obtenues, on obtient un système qui renvoie à l'utilisateur des indications sur son fonctionnement, qui peuvent lui suggérer des perfectionnements dans sa conduite de recherche, sans pour autant se bloquer devant une formulation inattendue. Après quelques utilisations, l'usager sait qu'il peut poser plusieurs termes, que ses opérateurs disjonctifs (connecteur OU) sont éliminés, et seulement proposés dans une deuxième étape...

L'application de cette fonction de traduction au champ des auteurs est cependant particulière : on n'utilise pas, bien entendu, le module d'ajout du S, mais on conserve la décomposition de la requête en unitermes. Cela permet aisément de traiter le cas d'une question ayant deux noms d'auteurs, mais laisse passer quelques anomalies dans le cas de noms possédant un article. La requête AUTEUR=LA FONTAINE sera traitée sous la forme "LA et FONTAINE". A nouveau, ces limites sur la cohérence sémantique des recherches sur le champ auteur sont de moindre importance dans le cas d'une bibliothèque scientifique où la recherche par sujet prédomine, et où le nombre d'auteurs différents est relativement faible (peu d'ouvrages). Dans les autres cas, on pourrait toujours aisément passer à une saisie plus stricte des noms d'auteurs, éventuellement en séparant sur deux lignes la saisie du nom et du prénom.

1.e - EP, l'espace de pertinence et f_p, la fonction de pertinence

Nous nous situons totalement dans un cadre booléen, donc les valeurs de l'espace de pertinence sont limitées à l'ensemble {0,1}. Les documents apparaissent dans l'ordre chronologique inverse (le plus récent en tête), ce qui est la seule forme de classement des réponses.

Les opérations booléennes ne sont pas accessibles au premier niveau, car

elles rebuteraient un grand nombre d'utilisateurs. On peut toutefois composer des équations de recherche complexes, soit en donnant dans la question une liste de termes et en opérant ensuite les combinaisons booléennes à partir des numéros des réponses, soit dans le mode en "formulation libre" en ajoutant des précisions concernant d'autres champs (date de publication, type de document, niveau d'études et même auteur, titre ou sujet pour ajouter un nouveau terme) après obtention des premiers résultats.

2 - La réalisation d'un plan de classement adapté

L'élaboration du plan de classement et son utilisation dans le catalogue informatisé est un point nodal de la définition du système documentaire choisi. Les codes de classement permettent une indexation des livres, en même temps qu'ils définissent la place occupée par ce livre sur les étagères. Le plan de classement a été établi en fonction des ouvrages disponibles à la bibliothèque. Il ne s'agit aucunement d'une "classification de la connaissance", mais d'un outil pratique, adapté à un fonds particulier, et à une fréquentation particulière (principalement les étudiants des premier et second cycles scientifiques). Les indices sont composés de trois chiffres (e.g. *131, 480, 930...),* ce qui correspond à l'esprit de la circulaire de la DBMIST [DBM88].

Il y a eu deux versions radicalement différentes du plan de classement. La première mouture s'est voulue dérivée de la *Classification Décimale Dewey.* Nous avions étendu sur dix classes (premier chiffre de 0 à 9) ce qui était limité en deux classes (5 et 6) de la C.D.D. Des aménagements avaient été faits pour insérer l'informatique. L'articulation générale du plan de classement s'inspirait aussi du plan de classement de la banque de données PASCAL. Cette version donnait ainsi une structure hiérarchisée.

On trouvait ainsi la chimie décomposée suivant la classification suivante :

500 : Chimie-généralités
- **501** : Histoire de la chimie
- **502** : Traités généraux
- **503** : Recueils de données
- **505** : Manuels portant sur l'ensemble du programme
- **509** : Exercices portant sur l'ensemble du programme

510 : Structure de la matière

512 : Chimie descriptive
515 : Chimie quantique

520 : Chimie physique
521 : Solutions
522 : Thermodynamique chimique
523 : Electrochimie
524 : Cinétique chimique
525 : Réaction chimique, catalyse

530 : Chimie minérale
531 : Monographie concernant un corps particulier ou une famille de corps
535 : Minéralogie

540 : Chimie organique
541 : Monographies concernant un corps particulier, ou une famille de corps
545 : Polymères
546 : Organo-métalliques
547 : Produits naturels

550 : Chimie analytique

560 : Appareils et méthodes
561 : Chromatographie
562 : Analyse spectrochimique
563 : Résonance Magnétique Nucléaire

570 : Chimie industrielle.

Ces choix posaient plusieurs inconvénients :

. le plan de classement n'étant pas structuré comme un thésaurus, la recherche d'un indice ne se traduisait pas par la recherche de tous les indices de niveau inférieur, alors que la présentation emboîtée semblait intuitivement l'indiquer. En sens inverse, une recherche sur un indice de bas niveau (e.g. " *561 - chromatographie")* ne permettait pas de retrouver les ouvrages plus généraux incluant une partie sur la chromatographie, qui auraient alors été classés en *"560 - Appareils et méthodes").*

. l'indexation des ouvrages généraux n'était pas aisée. Il aurait alors été préférable de transformer la classification en taxinomie [HAL89], c'est à dire de réserver dans les niveaux inférieurs un indice pour les livres que l'on ne pouvait pas classer à ce niveau parmi les indices présents. Par exemple, dans l'indice *430, langages de programmation,* qui se subdivise en *431 - Langage PASCAL, 432* -...., il fallait prévoir un *439 - Autres langages de programmation.* Cette structure permet d'éviter les inconvénients d'une indexation d'ouvrages généraux *(Les langages de programmation)* ou couvrant un domaine qui n'était

pas prévu dans le plan *(Utiliser Modula-2)*. Cette opération aurait du être possible pour tous les niveaux. Or la structure hiérarchique n'est guère adaptée à ce type d'extension.

La pratique de l'indexation décimale [RIC87] veut que l'on réserve le code "tête de chapitre" pour les sujets généraux et les codes de niveau plus spécifique pour les précisions. Or cette habitude ne correspond pas aux pratiques des utilisateurs. Notre fonctionnement général d'élaboration des concepts est radicalement différent. Un concept est construit de façon ascendante en regroupant des connaissances diverses que nous considérons comme semblables (e.g. le concept de champignon est primitivement le moyen de désigner des plantes que nous reconnaissons comme ayant de nombreux caractères communs). Mais quand il existe, le concept agit en retour sur les distinctions et les classements que nous opérons dans l'ensemble des objets (le concept de champignon nous sert à désigner une classe d'objets, et en même temps sert de "service d'accueil" pour accepter un nouvel objet dans notre connaissance : même si nous ne connaissons pas ce nouvel objet, nous pouvons néanmoins lui attacher des critères car c'est un champignon). Cette distinction entre la notion intuitive de concept et les méthodes de l'indexation décimale utilisant des classifications hiérarchisées est source de très nombreuses confusions dans la recherche documentaire, qui correspondent d'ailleurs aux dilemmes de l'indexeur.

. la structure hiérarchique impose une extension en profondeur dès que l'on veut ajouter de nouvelles informations [MAN87]. Or, pour les raisons pratiques exposées plus haut, nous voulions limiter à 3 chiffres la longueur des indices. De plus, la structure hiérarchique occupe la même place dans la classification pour les domaines ayant peu d'ouvrages et peu de subdivisions et pour ceux correspondant à beaucoup d'ouvrages et de nombreuses divisions conceptuelles.

. La structure hiérarchique est peu évolutive. Un nouveau concept doit s'insérer dans un cadre prédéfini, être un sous-ensemble d'un concept déjà utilisé. Une classification hiérarchique est donc difficile à entretenir. Les nombreuses commissions qui président à la mise à jour de la Classification Décimale Dewey en sont la preuve, de même que l'absence de réactualisation, du moins en langue française, de la C.D.U. *(Classification Décimale Universelle)*. Or

notre choix de construire un plan de classement adapté à notre bibliothèque, c'est-à-dire modifiable avec l'introduction de nouveaux enseignements ou de nouveaux axes de recherche dans l'université, était contradictoire avec un tel entretien. Par exemple, le secteur agro-alimentaire, largement développé à Caen prend une place importante dans la classification, à l'opposé des secteurs techniques et des sciences de l'ingénieur, réduits à la portion congrue dans le fonds d'ouvrages comme dans le plan de classement. Bien entendu, cela peut changer.

En marge de ces critiques générales concernant les classifications hiérarchiques, nous avons fait circuler notre premier plan de classement auprès de plusieurs enseignants chercheurs pour obtenir des compléments dans leur discipline. Cette démarche nous a montré d'une part que le cadre hiérarchique d'origine (la Classification Dewey) était largement périmé en regard de la recherche, les têtes de chapitres de la classification ne correspondant pas à la manière dont se situaient les scientifiques, d'autre part que les principes d'une classification sont difficiles à faire comprendre à des non-spécialistes de la documentation. Les modifications qui nous ont été proposées consistaient en général à ajouter des indices en fonction des concepts utilisés dans la recherche, mais sans tenir compte des livres réellement présents dans la bibliothèque, et des besoins de regroupement des secteurs peu fournis.

Nous avons alors décidé de changer complètement de méthode de travail, en construisant un plan de classement non hiérarchique. Malheureusement, comme nous avions déjà coté un nombre important d'ouvrages, il n'a pas été possible de recommencer à zéro, et nous avons été obligés de tenir compte des premiers choix. La volonté de construire une classification non hiérarchique, obtenue par regroupements en grands domaines, puis par domaines plus restreints, en s'efforçant que deux indices de classification proches correspondent à deux concepts connexes a été celui qui a présidé à l'établissement de la *Library of Congress Classification* au siècle dernier [MAN87], C'est aussi une méthode qui s'apparente, en terme de démarche, aux techniques d'agrégation non hiérarchiques décrites dans la première partie. Elle est ainsi parée de bases théoriques solides, et la suite de notre expérience a montré que cette méthode était bien plus souple et évolutive que la logique d'un plan de classement hiérarchique.

Les indices sont regroupés de façon à ce que les deux premiers chiffres définissent un espace sémantique, et que le troisième chiffre serve de précision. Mais cette structure n'est pas obligatoire, ce qui permet de récupérer de la place pour insérer dans le plan de classement des domaines qui ne sont pas directement dépendant du domaine qui aurait les mêmes premiers chiffres. On s'efforce toutefois de conserver une proximité sémantique.

Par exemple, les sciences de la vie peuvent s'étendre sur les codes commençant par les chiffres 7, 8 et 9, évitant une extension en profondeur, alourdissant les indices. Néanmoins, les codes commençant par 7 touchent des problèmes plus théoriques, alors que ceux débutant par 8 sont axés sur l'étude des diverses espèces du vivant et qu'un code commençant par 9 correspond à la biologie appliquée.

De même, on trouve à proximité :

680 : Cartographie
685 : Géomorphologie
690 : Télédétection

qui pour être des concepts séparés restent cependant, dans le cadre de notre bibliothèque, des notions que l'aspect pratique tend à rapprocher.

Le plan étant établi manuellement, cette proximité est entièrement empirique, tenant compte des livres présents dans la bibliothèque, des enseignements assurés à l'université et des regroupements qui sont faits dans la définition des unités de valeur, et bien sûr des connaissances et inclinaisons personnelles du personnel de la bibliothèque. Par exemple, j'ai tenu à ce que les sciences de l'information et les ouvrages sur l'information interactive (indice 397) soient proches des ouvrages sur les télécommunications (indices 390-395) et de l'informatique (indices 4..) alors que de nombreuses classifications rapprochent ce secteur de recherche de la bibliothéconomie, dans le chapitre des "généralités". Cette subjectivité de la classification est accentuée par le faible nombre de personnes ayant élaboré ce plan de classement, et par les contraintes de temps, qui nous empêchaient d'hésiter ou de revenir en arrière.

Chaque ouvrage se voit attribué un *"indice principal",* qui servira à déterminer sa cote, et qui indiquera sa position sur les étagères. Bien entendu, cet indice principal suffit rarement à définir un livre. Il y a souvent des livres

frontière, ou des livres qui peuvent intéresser plusieurs disciplines. Par exemple, les "réseaux de neurones" intéressent à la fois les informaticiens, les physiciens et les neurophysiologistes. "L'apprentissage" est étudié en psychophysiologie et en intelligence artificielle. On accorde alors des *"indices secondaires",* qui permettent de dessiner les divers centres d'intérêt du livre. Les indices secondaires sont utiles pour l'indexation du catalogue informatisé. Chaque livre peut recevoir un *"indice principal"* et trois *"indices secondaires"*.

L'attribution des indices est une forme d'indexation manuelle, et de ce fait soumise aux aléas et aux incohérences mentionnées dans la première partie. L'expérience montre que des livres semblables, voire parfois le même titre traité à deux périodes différentes, sont souvent placés en deux endroits différents du plan. Une correction globale de la banque de données et de l'indexation a été faite lors du changement de classification, qui a permis de repérer certaines de ces anomalies. Lors de ce changement, nous avons éliminé les indices qui concernaient peu d'ouvrages, pour respecter la volonté synthétique de ce choix d'indexation et de classement. En sens inverse, avec le développement du système, il serait souhaitable de diviser certains indices ayant un trop grand nombre d'ouvrages. Il reste à établir une procédure rapide et fiable pour permettre aisément ces changements. Or c'est là une faiblesse de toute activité classificatoire, surtout si elle est doublée d'une activité de classement, que d'être très difficile à maintenir et à faire évoluer. Car même si on affirme s'en tenir à la réalité matérielle des ouvrages, les présupposés et les conceptions personnelles (ou multi personnelles dans le cas de classifications établies collectivement) jouent un très grand rôle dans l'élaboration d'une classification, et dès lors les divisions ne sont pas toujours évidentes.

Un autre problème rencontré avec ce plan de classement a été la difficulté à faire cohabiter les ouvrages d'enseignement, notamment de premier cycle, qui restent très généraux (correspondant souvent à plusieurs indices du point de vue intellectuel) et qui sont très nombreux (encombrement des étagères), avec des ouvrages de recherche, plus pointus et moins nombreux, donc facilement dissous aux yeux des utilisateurs dans la masse des livres disposés sur les rayons. La solution trouvée s'est appuyée sur le fait pragmatique que nous avions peu d'ouvrages de recherche, et qu'il était donc possible de créer des indices les regroupant par disciplines.

On trouve ainsi :

157 : Analyse, approfondissement et recherche
183 : Théorie des probabilités et des statistiques
546 : Chimie organique, approfondissement et recherche
797 : Physiologie et biochimie végétale, approfondissement et recherche.

Cette solution, pour être efficace car correspondant globalement aux demandes d'achats d'ouvrages émanant des laboratoires, n'en est pas pour autant satisfaisante car trop liée à une situation locale particulière. Cependant, un fonds spécialisé ne pourrait de toute façon pas se satisfaire d'un plan de classement aussi général, qui correspond à un fonds encyclopédique (dans le domaine scientifique) principalement axé sur la satisfaction des besoins des étudiants.

Certains [MAN87] proposent de s'appuyer sur des classifications dont le premier symbole est une lettre, ce qui autorise 26 domaines au lieu de 10 avec les chiffres. Dès lors, chaque domaine peut plus aisément être divisé en quelques classes significatives. Toutefois, le repérage dans un fonds dont les grands domaines, qui dans une université correspondent aux enseignements, sont dispersés est plus difficile. Les bibliothécaires ont l'expérience d'étudiants demandant "des ouvrages de maths", même à un catalogue informatisé ! L'aspect pratique doit donc être privilégié, surtout quand le nombre d'ouvrages reste faible, et facilement maîtrisable. Pour traiter un fonds plus encyclopédique (par exemple une bibliothèque de lettres et sciences humaines) il y aurait certainement de nombreux avantages à opérer à partir de 26 lettres.

A l'intérieur du système informatisé, le plan de classement est organisé comme un réseau sémantique élémentaire dans le cadre limité d'un fichier *Texto*. A un indice, qui sert de clé d'accès, est associé un texte de définition, des "mots associés", une note d'application pour guider l'indexeur, et d'autres indices qui traitent, selon deux niveaux de proximité, de domaines connexes.

Même si toutes ces informations ne sont pas exploitées aujourd'hui, on peut envisager de s'en servir pour des reformulations élémentaires (par exemple, pour "élargir" une requête en interrogeant sur les "indices associés" des indices les plus fréquemment utilisés dans le premier lot de réponses). Un développement

de ce type est prévu. Les limites de *Texto,* notamment l'impossibilité d'ouvrir deux fichiers en même temps et la lenteur de basculement lors de l'ouverture d'autres fichiers, sont aujourd'hui le principal obstacle à l'écriture de ces programmes.

3 • La réalisation informatique

3.a • Présentation de *Texto*

Le logiciel *Texto* est un logiciel documentaire réalisé par la société lyonnaise *Chemdata.* Il est certainement le logiciel documentaire le plus utilisé en France, avec près de 2 000 installations, sur divers matériels (micro-ordinateurs, mini et gros systèmes) et sous plusieurs systèmes d'exploitation (MS-DOS, VMS, VM, UNIX, GCOS,...).

Un document dans *Texto* est défini par trois critères :

. une clé d'accès, univoque, qui peut être numérique (avec numérotation manuelle ou automatique) ou alphanumérique.

. une division en champs, chaque champ étant défini par son *nom* et son *contenu.* Le contenu des champs est de longueur variable, ce qui réduit l'encombrement du disque à l'espace strictement nécessaire. Un champ sans contenu ne prend donc aucun espace mémoire. Toutefois, un fichier *Texto* reste un document ASCII, et n'est pas compressé.

. le contenu d'un champ est divisé en articles. Plusieurs attributs sont alors affectés à un même champ, l'index les considérant comme équivalents (une des distinctions fondamentales entre un logiciel documentaire et un SGBD). Le *séparateur d'articles* est laissé au choix de l'utilisateur. Nous avons choisi le "slash" (/). Un article est donc la chaîne de caractères contenue entre deux "slash". En fonction du choix de système documentaire développé plus haut, les articles sont les unitermes repérés pendant la phase d'indexation.

Les documents sont regroupés en fichiers. *Texto* ne permet d'ouvrir qu'un seul fichier à la fois, ce qui est une de ses faiblesses (à la différence, par

exemple, de *BRS).* Un fichier est exploitable au travers d'un *document de paramètres,* qui définit les noms des champs, le format de la clé d'accès, le séparateur d'articles, les index associés à des champs, leur mode de mise à jour, et éventuellement les fichiers chaînés associés à un fichier principal.
Voici un exemple simple de fichier *Texto.*

```
nom      .pdemo
general .6 1 . /
champs   .num AUTEUR TITRE REF DP MCL




num      .000001
AUTEUR   .Minski, M./Papert, S.
TITRE    .Perceptrons
REF      .MIT-Press
DP       .1987
MCL      .PERCEPTRON/CONNEXIONISME/INTELLIGENCE ARTIFICIELLE

num      .000002
AUTEUR   .Horfstäder, D.
TITRE    .Gödel, Escher et Bach : Les brins d'une guirlande éternelle
REF      .Interéditions
DP       .1987
MCL      .INTELLIGENCE ARTIFICIELLE/SYSTEMES FORMELS/LINGUISTIQUE
```

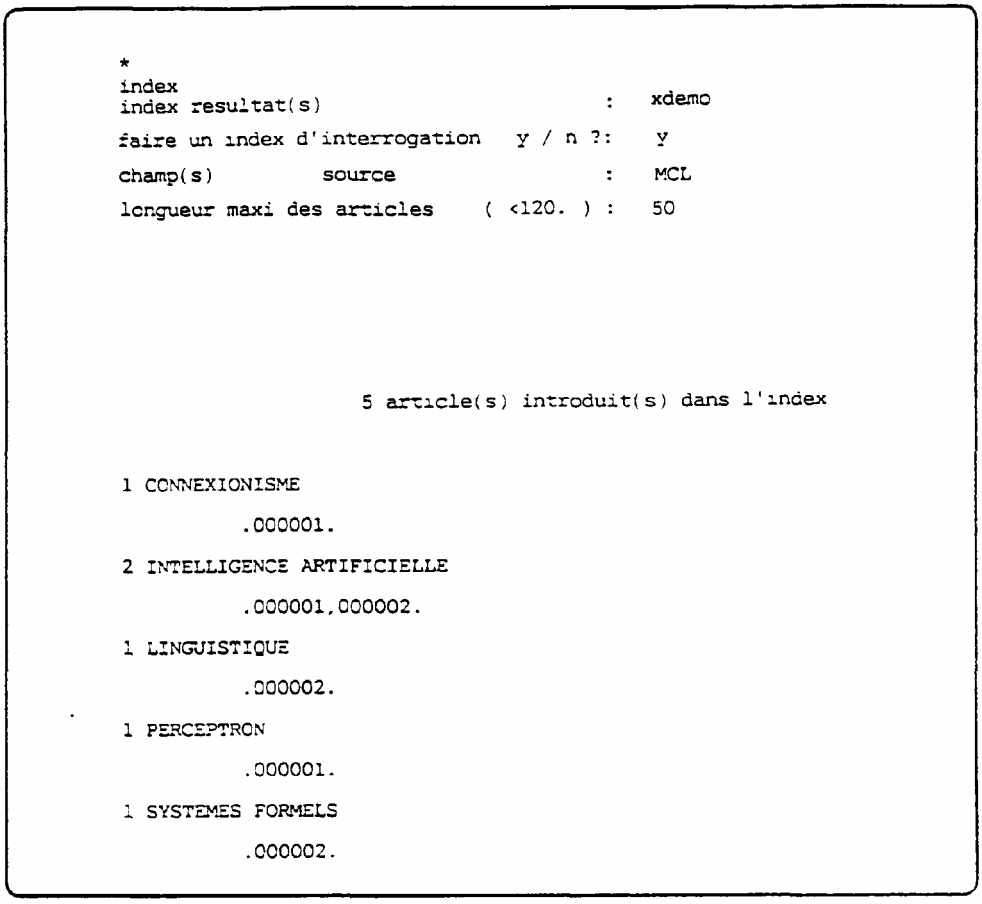
Plusieurs *documents de paramètres* peuvent être associés au même fichier documentaire, selon les objectifs de traitement désirés. Toutefois, l'ouverture du couple fichier+document de paramètre est relativement longue, surtout si de nombreux index sont associés au fichier. En pratique, il vaut mieux travailler avec un seul document de paramètres plus complet.

Voici le *Document de Paramètres* que nous avons utilisé.

```
nom      .pcatal
general .6 1 . /
champs   .num ECAUT ECTIT ECEDREF ECCOLL DP ANTIDP ISBN RECHISBN TH CGR NAT,
         .(x+)xnat PROP LOC LANG NIVEAU,(x+)xniv INDICE INDSEC BIB INV COTE
         .MAJAUT,(x+)xaut MAJTIT,(x+)xtit MAJEDC MAJCOLL MAJMCL,(x+)xbase
         .MAJCGR MAJIND NOMCLASS MCL AMS INSPEC PASCAL CAS RESUME JOURNAL MOIS
         .VOLUME FASC SUITE DEVIENT SUPPLEM MODENT TRAVEE
```

L'accès aux fichier peut être séquentiel, mais cette opération est trop longue au delà de quelques centaines de documents. On utilise alors un fichier inversé (index). L'index d'interrogation de *Texto* associe à chaque *article* la liste des *clés d'accès* des documents possédant cet article et le nombre de documents concernés.

Au fichier de démonstration présenté ci-dessus est associé l'index de mots-clés :

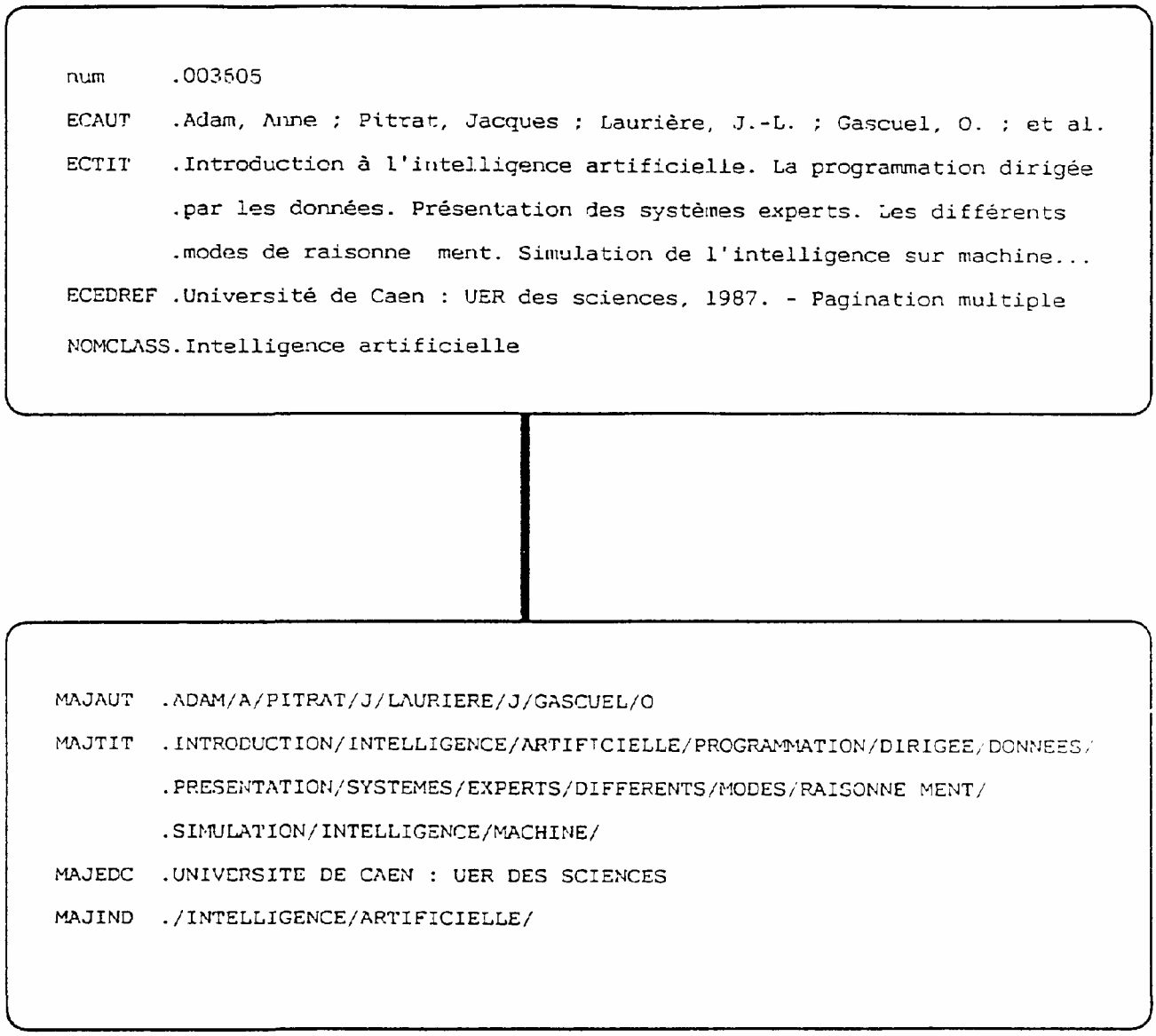


```
*
index
index resultat(s)                          :   xdemo
faire un index d'interrogation   y / n ?:   y
champ(s)          source                   :   MCL
longueur maxi des articles    ( <120. ) :   50

                    5 article(s) introduit(s) dans l'index

1 CONNEXIONISME
          .000001.
2 INTELLIGENCE ARTIFICIELLE
          .000001,000002.
1 LINGUISTIQUE
          .000002.
1 PERCEPTRON
          .000001.
1 SYSTEMES FORMELS
          .000002.
```

On remarquera que les descripteurs composés sont traités comme un seul "article" de *Texto*. Pour une indexation par unitermes, il aurait fallu auparavant remplacer les espaces présents dans le champ des mots-clés par un séparateur d'articles.

Texto n'est pas un logiciel permettant l'indexation en texte intégral. Il se borne à reconnaître les *articles* (au sens précis de *Texto)* comme étant la chaîne de caractères placée entre deux séparateurs d'articles, et traite les caractères tels qu'ils sont saisis. Il n'y a pas d'indexation automatique en majuscules, bien que cette possibilité devrait être mise au point dans la prochaine version du logiciel. Si l'on veut indexer les mots du titre d'un document, et en même temps conserver la

possibilité de visualiser ce titre en caractères riches (minuscules avec les signes diacritiques), il faut définir un champ spécifique pour l'interrogation, qui contiendra les mêmes mots, en majuscules. C'est lors du remplissage automatique de ce champ interrogeable que l'on élimine les mots outils, que l'on normalise l'emploi des traits d'union...



Les index de *Texto* sont soit spécifiques d'un champ donné, soit associés à plusieurs champs. On peut définir ainsi un index sur les mots du titre, mais aussi un index plus général qui regroupe les mots du titre et les mots-clés (l'équivalent du *Lexique de base* - *Basic Index* - des banques de données en ligne des serveurs commerciaux).

Enfin, *Texto* permet facilement des échanges de données avec d'autres logiciels. Le format d'échange de *Texto,* dit *ajout piloté,* est du type ASCII balisé. *Texto* peut accepter en entrée tout fichier texte du type :

Nom d'un champ - CHP1

Contenu de ce champ

Nom d'un champ - CHP2

Contenu de ce champ

...

// *(signe de changement de document)*

Nom d'un champ - CHP1

Contenu de ce champ

...

//

Dans notre application, les documents sont saisis sur micro-ordinateur. Le programme constitue un fichier de ce type, qui est intégré dans *Texto* en mode "ajout piloté". Voici un exemple de document en format ajout piloté avant le transfert dans *Texto :*

```
ECAUT
Lehninger, Albert L. ; Duquesne, Maurice (éd.) ; Duquesne, Janine (éd.)
ECTIT
Bioénergétique : Base moléculaire des transformations de l'énergie biologique
ECEDREF
Ediscience, 1969. - 238 p.
DP
1969
ANTIDP
131
NAT
LM
PROP
BS
LOC
LA
LANG
FRE
INDICE
718
BIB
BS
INV
8ϕ4007/8ϕ4008
COTE
718 - 1011
MAJAUT
LEHNINGER, ALBERT L./DUQUESNE, MAURICE (ED.)/DUQUESNE, JANINE (ED.)
MAJTIT
BIOENERGETIQUE/BASE MOLECULAIRE DES TRANSFORMATIONS DE L'ENERGIE BIOLOGIQUE
MAJEDC
EDISCIENCE
MAJMCL
ATP/ADENOSINE TRIPHOSPHATE/PHOTOSYNTHESE/BIOSYNTHESE
MCL
ATP/Adénosine triphosphate/Photosynthèse/Biosynthèse
//
ECAUT
Milsant, Francis
ECTIT
Automatismes à séquences : 3e éd.
ECEDREF
Eyrolles, 1979. - 238 p.
```

Texto peut facilement produire un fichier dans ce même format à partir d'un fichier structuré *Texto*. Cela permet les échanges entre diverses machines, et assure la pérennité des informations. De plus, *Texto* permet de fournir des fichiers au format texte (suite de caractères) selon les désirs de l'utilisateur. Cela permet de construire un fichier balisé de type MARC à partir d'un fichier *Texto*. Les données sont alors récupérables par tout système capable d'entrer du format MARC. Bien entendu, il convient de définir les balises MARC utilisées, et d'effectuer une association entre les champs du fichier *Texto* et les zones du fichier MARC. Par exemple, comme nous ne saisissons pas la "ville d'édition" du format ISBD *(International Standard Bibliographie Description)* car nous respectons la norme minimale AFNOR Z 44 - 072, nous ne pourrons la fournir en sortie. En revanche, les "auteurs secondaires" dans notre fichier ne sont pas représentés dans un champ particulier, alors qu'ils sont repérés par une zone MARC spécifique. Cependant, la mention de leur apport particulier est indiquée (ill., pref.,...). On peut donc concevoir une recherche de ces qualificatifs et placer ainsi la balise MARC correspondante.

La réalisation d'un convertisseur en format MARC de ce type est envisagée, notamment pour les échanges avec la bibliothèque municipale de Caen (si toutefois la société CLSI qui équipe cette bibliothèque veut bien nous donner la structure de *son* propre format MARC).

Texto distingue deux types de questions :

. la *question directe,* qui donne immédiatement les réponses à une équation booléenne,

. *l'interrogation composée* qui permet de combiner les résultats obtenus aux diverses étapes de recherche. C'est ce dernier mode qui est utilisé dans notre application.

La syntaxe des questions est de la forme C=A, où C représente le nom exact du champ interrogé, et A représente le terme, ou la combinaison booléenne de termes, de la recherche.

Voici le déroulement typique d'une recherche en mode *interrogation composée.* Même si cela reste transparent pour l'utilisateur, c'est ce type de questions que pose le système informatique. Les questions sont donc traduites dans des expressions connues de *Texto.*

```
*
ques

******  Interrogation  Composee  ******


?
MAJMCL=INTELLIGENCE et ARTIFICIELLE

  $1           94  reponse(s)  pour :   MAJMCL=INTELLIGENCE et ARTIFICIELLE


?
MAJMCL=SYSTEMES et EXPERTS
  $2           31  reponse(s)  pour :   MAJMCL=SYSTEMES et EXPERTS


?
MAJMCL=SYSTEME et EXPERT
  $3           16  reponse(s)  pour :   MAJMCL=SYSTEME et EXPERT


?
$2 ou $3

  $4           44  reponse(s)  pour :   $2 ou $3


?
$1 et $4

  $5            6  reponse(s)  pour :   $1 et $4


?
NAT=ART
  $6         6846  reponse(s)  pour :   NAT=ART


?
$5 sauf $6

  $7            2  reponse(s)  pour :   $5 sauf $6
```

3.b - Présentation de *Logotel*

Logotel est un langage de manipulation des données de *Texto*. Il s'agit d'un langage de programmation assez fruste, mais qui permet de lancer des commandes de *Texto,* notamment les opérations de question et de listage. L'objectif de *Logotel* est de permettre le développement d'interfaces utilisateur pour des fichiers *Texto.*

Le langage *Logotel* reste un langage de programmation de bas niveau. Il ne connaît que la structure de branchement *("si...aller tel sous-programme"),* ce qui rend complexe la définition de boucles (on est obligé d'accompagner toute opération d'une variable d'état). Les variables sont en nombre limité (200 variables) et uniquement alphanumériques. Cela augmente le nombre d'opérations nécessaires pour effectuer des comparaisons sur les chiffres. Par exemple, 20 est inférieur à 5 en mode alphanumérique. Il faut alors comparer 1020 et 1005 pour retrouver l'ordre numérique. Ces opérations supplémentaires dans un langage interprété restent grandes consommatrices de temps.

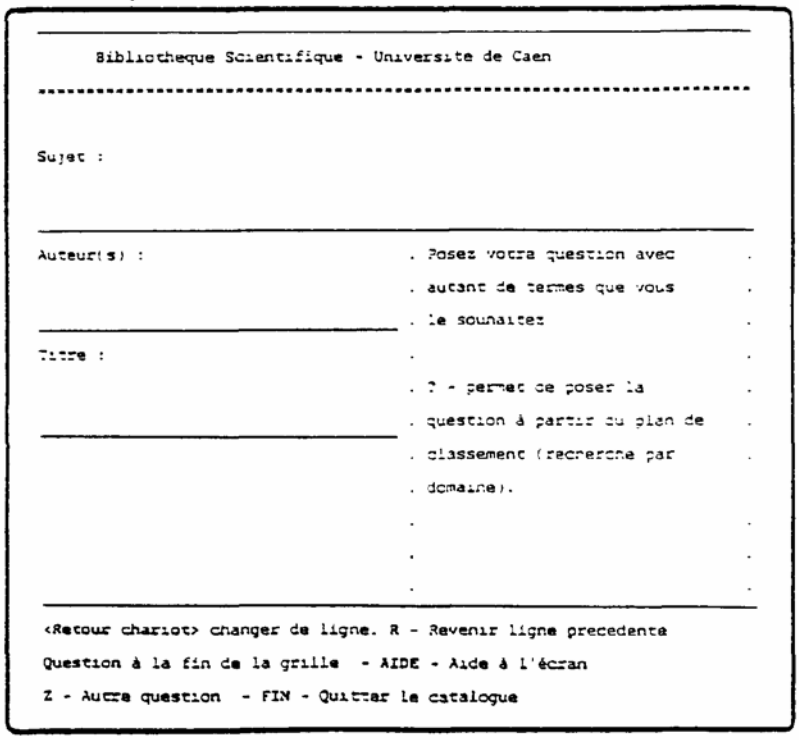
En revanche, *Logotel* permet d'indicer les variables en utilisant un "compteur" dans la définition de la variable. La variable générique MOT[X] définit ainsi les diverses variables MOT1, MOT2, MOT3...uniquement en incrémentant la valeur de X. Malheureusement, il faut en permanence conserver un état de la dernière opération car *Logotel* ne contient pas de commande pour la remise à zéro de toutes les variables.

Le langage *Logotel* est une couche logicielle ajoutée sur *Texto,* qui intervient principalement au niveau de la présentation des écrans, de la transformation des questions de l'utilisateur en question acceptable pas *Texto* (forme C=A) et sur la présentation des données. Le fonctionnement général du système d'information est déterminé auparavant par les choix de structuration et d'interrogation du fichier *Texto* (méthodes d'indexation, questions acceptables, contenu des informations...).

1 - La définition des écrans est réalisée avec l'ordre *Imprimer.* Tout le texte qui suit cette commande *Logotel* est affiché à l'écran. On peut utiliser dans le programme les séquences d'échappement correspondant au positionnement du

curseur ou pour baliser le début et la fin du passage en caractères de surbrillance (gras à l'impression) ou d'inverse vidéo. Les codes hexadécimaux correspondants sont introduits par antislash (\). Le programme suivant permet de définir l'écran de saisie des informations.

```
modifier CLS1 : \1B[2J
modifier CLS2 : \1B[H
modifier CLS : [CLS1][CLS2]
modifier DCG : \1B[1m
modifier FCG : \1B[0m
modifier TIRET : ____________________________________
modifier DBTIRET : ====================================


imprimer [CLS]
\1B[1;1H[TIRET][TIRET]
[DCG]      Bibliothèque Scientifique - Université de Caen  [FCG]
[DBTIRET][DBTIRET]

Sujet :

[TIRET][TIRET]
Auteur(s) :

[TIRET][TIRET]
Titre :

[TIRET][TIRET]
\1B[19;1H[TIRET][TIRET]
#1<Retour chariot>#0 changer de ligne. #1R#0 - Revenir ligne précédente
#1AIDE#0 - Aide à l'écran
#1Z#0 - Autre question  - #1FIN#0 - Quitter le catalogue

imprimer \1B[6;45H
\1B[7;45H_____________________________________
\1B[8;45H . Posez votre question avec       .
\1B[9;45H . autant de termes que vous       .
\1B[10;45H . le souhaitez                   .
\1B[11;45H .                                .
\1B[12;45H . #1?#0 - permet de poser la     .
\1B[13;45H . question à partir du plan de   .
\1B[14;45H . classement (recherche par      .
\1B[15;45H . domaine).                      .
\1B[16;45H .                                .
\1B[17;45H .
\1B[18;45H .
```



2 - La présentation des données est gérée par le mode de "lecture virtuelle". Les contenus des champs *Texto* peuvent être versés dans une variable *Logotel* par ce mode. La lecture virtuelle d'un document crée autant de variables qu'il y a de champs dans ce document, chaque variable portant le nom du champ. On utilise cette possibilité pour faire présenter les informations de *Texto* dans un format agréable. Le programme de la page suivante permet de définir la présentation de la référence complète d'un ouvrage. Le résultat est présenté ci-dessous.

```
1

_____________________________________________________________

Algorithms for graphics and image processing
Pavlidis, Theodosios
Computer science press, 1982. - XI-416 p. : 29 p. de pl. en noir


I.S.B.N. : 0-914894-65-X


Ouvrage conservé à la Bibliothèque scientifique
Salle du Libre-Accès - Cote : 465 - 1012
Emprunt à domicile réservé aux étudiants du DEA Instrumentation

Indice principal : 465 - Image et ordinateur, infographie

_____________________________________________________________
```

```
texte lvl [VM[NV]]
si ECCOLL=*
modifier COLL : Collection : [ECCOLL]
aller ,nat

etiquette nat
si NAT=C
modifier NT : Actes de congrés - [CGR]
aller ,isbn
si NAT=R
modifier NT : Rapport
aller ,isbn
si NAT=ART
modifier NT : Article de périodique
aller ,isbn
si NAT=T
modifier NT : [TH]
aller ,isbn
si NAT=B
modifier NT : Brochure
aller ,isbn
si NAT=U
modifier NT : Usuel ou ouvrage de référence. Ne peut être emprunté.
aller ,isbn
modifier NT :
aller ,isbn

etiquette isbn
si ISBN=*
modifier ISB : I.S.B.N. : [ISBN]
aller ,niveau

etiquette niveau
si NIVEAU=1
modifier NIV : Ouvrage principalement destiné aux étudiants de
modifier NIV : [NIV] [DCG]Premier cycle[FCG]
aller ,bib
modifier NIV :
aller ,bib

etiquette bib
si BIB=BS
modifier BIBLIO : Ouvrage conservé à la [DCG]Bibliothèque scientifique[FCG]
aller ,prop
si BIB=PHYS
modifier BIBLIO : Ouvrage conservé à la [DCG]Bibliothèque de Physique[FCG]
aller ,prop
modifier BIBLIO :
aller ,prop

etiquette prop
si PROP=INSTRUM
modifier PRP : Emprunt à domicile réservé aux étudiants du DEA Instrument
aller ,loc
si PROP=PHYSCORP
modifier PRP : Emprunt à domicile réservé aux étudiants du DEA Phys. Nucl.
aller ,loc
si PROP=ISMRA
modifier PRP : Emprunt à domicile réservé aux étudiants de l'ISMRA
aller ,loc
modifier PRP :
aller ,loc

etiquette loc
si LOC=LA
modifier LCT : Salle du Libre-Accès - Cote : [DCG][COTE][FCG]
aller ,indice
si LOC=MAG
modifier LCT : Ouvrage conservé en Magasin - Numéro : [DCG][COTE][FCG]
aller ,indice
si LOC=A
modifier LCT : Salle Linnée - Cote : [DCG][COTE][FCG]
aller ,indice
si LOC=E
modifier LCT : Salle Pasteur - Cote : [DCG][COTE][FCG]
aller ,indice
modifier LCT :
aller ,indice


etiquette indice

imprimer [CLS]
imprimer [DCG][VM[NV]][FCG]
[TIRET][TIRET]
[DCG][ECTIT][FCG]
[ECAUT]
[ECEDREF]
[COLL]
[NT]
[ISB]
[NIV]

[BIBLIO]
[LCT]
[PRP]

Indice principal : [DCG][INDICE][FCG] - [NOMCLASS]

[TIRET][TIRET]
```

3 - La transformation des questions de l'utilisateur en questions acceptables par *Texto* est rendue possible par les opérations sur chaînes de caractères de *Logotel,* On peut utiliser les fonctions *avant X, après X, position X,* et la concaténation de chaînes. Le programme suivant permet d'éliminer les mots outils ou les signes de ponctuation de la question de l'utilisateur.

```
modifier TXT : [majusc [QUESUTIL]]

etiquette $1
modifier E : [position _ _ [TXT]]
si E=0
aller ,point
modifier TXT : [avant _ _ [TXT]]$[apres _ _ [TXT]]
aller ,$1

etiquette point
modifier E : [position . [TXT]]
si E=0
aller ,e1
modifier TXT : [avant . [TXT]]$[apres . [TXT]]
aller ,point

etiquette e1
modifier E : [position é [TXT]]
si E=0
aller ,e2
modifier TXT : [avant é [TXT]]E[apres é [TXT]]
aller ,e1

etiquette e2



modifier TXT : $[TXT]$

etiquette aux
modifier E : [position $AUX$ [TXT]]
si E=0
aller ,le
modifier TXT : [avant $AUX$ [TXT]]$[apres $AUX$ [TXT]]
aller ,aux

etiquette le
modifier E : [position $LE$ [TXT]]
si E=0
aller ,la
modifier TXT : [avant $LE$ [TXT]]$[apres $LE$ [TXT]]
aller ,le



etiquette tiret
modifier E : [position - [TXT]]
si E=0
aller ,2$
modifier TXT : [avant - [TXT]]$[apres - [TXT]]
aller ,tiret

etiquette 2$
modifier E : [position $$ [TXT]]
si E=0
aller ,$
modifier TXT : [avant $$ [TXT]]$[apres $$ [TXT]]
aller ,2$

etiquette $
modifier TXT : [gauche $ [TXT]]
modifier TXT : [droite $ [TXT]]
aller ,envoi
```

Le couple *Texto* + *Logotel* constitue ainsi un moyen puissant de constituer une banque de données et une interface d'utilisation. Toutefois, la mise au point des programmes est lourde. En particulier, l'absence d'éditeur de texte associé à *Texto* rend nécessaire de passer par l'éditeur de l'ordinateur et d'incorporer ensuite les programmes *Logotel* en mode *"ajout piloté". Logotel* ne propose aucune aide à l'écriture des programmes (indiquer la ligne erronée, indiquer l'absence d'un sous-programme appelé,...).

3.c - l'interface utilisateur

La conception de l'interface utilisateur a été régie par plusieurs principes, choisis parmi ceux qui ont été rappelés dans la première partie. En particulier, il semblait important de respecter les points suivants :

- Conserver une unité de manipulation des informations
- Utiliser le jeu complet de caractères du français
- Assurer la clarté des écrans délivrant l'information
- Offrir à l'utilisateur des moyens d'apprendre le fonctionnement

1 - Unité de manipulation des informations tout au long de l'application. Ce choix a été fortement inspiré par la démarche du vidéotex, notamment la définition des touches de fonctions de l'annuaire électronique. On a ainsi défini plusieurs opérations élémentaires, qui sont déclenchées par la même action sur le clavier, à tout moment du cheminement de l'utilisateur.

Ces opérations élémentaires sont les suivantes :

. *Revenir au menu de départ* (qui correspond à la séquence "* + SOMMAIRE" du vidéotex). Nous avons choisi de représenter cette opération par la lettre Z. L'utilisateur peut donc à tout moment arrêter son travail sur le système et recommencer une autre recherche. L'opération est aussi déclenchée par un BREAK (Control+C), ce que les étudiants en informatique ont tout de suite repéré.

. *Faire défiler une liste d'informations* : résultats d'une recherche documentaire, affichage du plan de classement, informations d'aide. Dans ce cadre, on définit deux opérations :

. *Défilement en avant,* correspondant à la fonction SUITE du vidéotex. Elle est représentée par le retour chariot sur un terminal. Il est

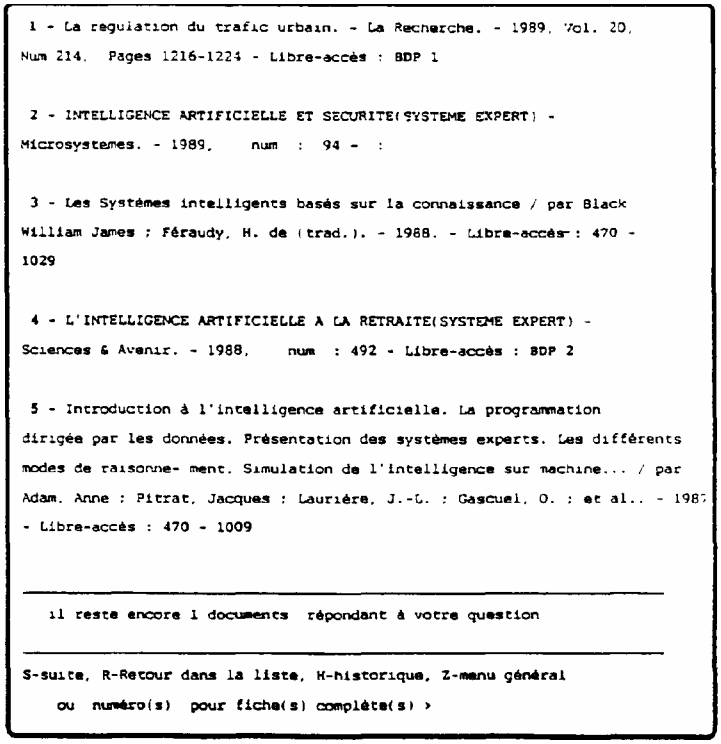
difficile d'indiquer cette opération pour des utilisateurs non avertis. Pendant un temps, nous utilisions l'indication classique des informaticiens : <RC>. Mais les utilisateurs écrivaient alors les deux lettres "RC" avec le clavier. Cette remarque montre que la réalisation d'une interface n'est pas seulement un problème conceptuel, mais doit s'attacher à des détails, parfois cocasses. Pour compenser cette limite, nous indiquons en toutes lettres <Retour Chariot>, ou S-Suite. Dans les deux cas, la manipulation <RC> ou S+<RC> conduit au même résultat, si bien que l'utilisateur habitué au système utilise spontanément le retour chariot pour *avancer* dans sa lecture.

. *Défilement en arrière,* pour revenir à l'information précédente, ce qui correspond à la touche RETOUR du vidéotex. Cette opération, relativement peu utilisée, est obtenue par la lettre R (R+<RC>, car le fonctionnement de *Logotel* nous contraint à faire suivre les commandes par le retour chariot ce qui correspond à une structure héritée mode ligne à ligne des terminaux TTY).

. *Quitter le catalogue* est obtenu par la commande explicite FIN. Cette méthode semble plus claire que d'indiquer une valeur numérique dans un menu (e.g. 0 - Quitter).

. *Les pages d'information* sur le système sont obtenues de même par la commande explicite AIDE.

Les opérations utiles sont rappelées sur chaque écran, en général dans les deux dernières lignes, les autres fonctions étant néanmoins accessibles.



2 - Utilisation du jeu complet de caractères du français, notamment de tous les signes diacritiques. Il est souvent inquiétant de voir les bibliothécaires sidérés, au sens premier, par l'informatique et perdre devant le dieu silicium leur sens commun, notamment dans la défense de la langue française. Le rôle des bibliothèques est de fournir une information, mais aussi de développer la culture du pays dans lequel elles existent. Dans le cas d'une bibliothèque universitaire, qui dépend donc du Ministère de l'Education Nationale, cette exigence est renforcée par le rôle éducatif de la bibliothèque. Ce n'est pas aux informaticiens de procéder à une réforme sauvage de l'orthographe, mais une opération politique et culturelle qui se traduit par des décisions au plus haut niveau de l'Etat. Si un catalogue informatisé est un instrument de meilleure diffusion des fonds des bibliothèques, il constitue en lui-même une forme de production d'un outil informationnel, et à ce titre doit s'organiser dans la langue du pays de diffusion. Notre langue comporte de nombreux signes diacritiques. C'est aussi le cas de presque toutes les langues européennes. Il faut en tenir compte dans l'élaboration de l'interface utilisateur.

La méthode consistant à remplacer toutes les informations par des lettres capitales, comme le catalogue de la bibliothèque universitaire de Bordeaux (GRACE), ou la banque de données *Téléthèses,* n'est qu'une mauvaise diversion. D'une part, le savoir-faire typographique intègre depuis des siècles les capitales accentuées pour conserver la richesse d'expression de la langue. D'autre part, la lecture d'un écran entièrement composé de caractères majuscules est très difficile. Le repérage des mots principaux et la division en zones de lecture n'est plus évidente.

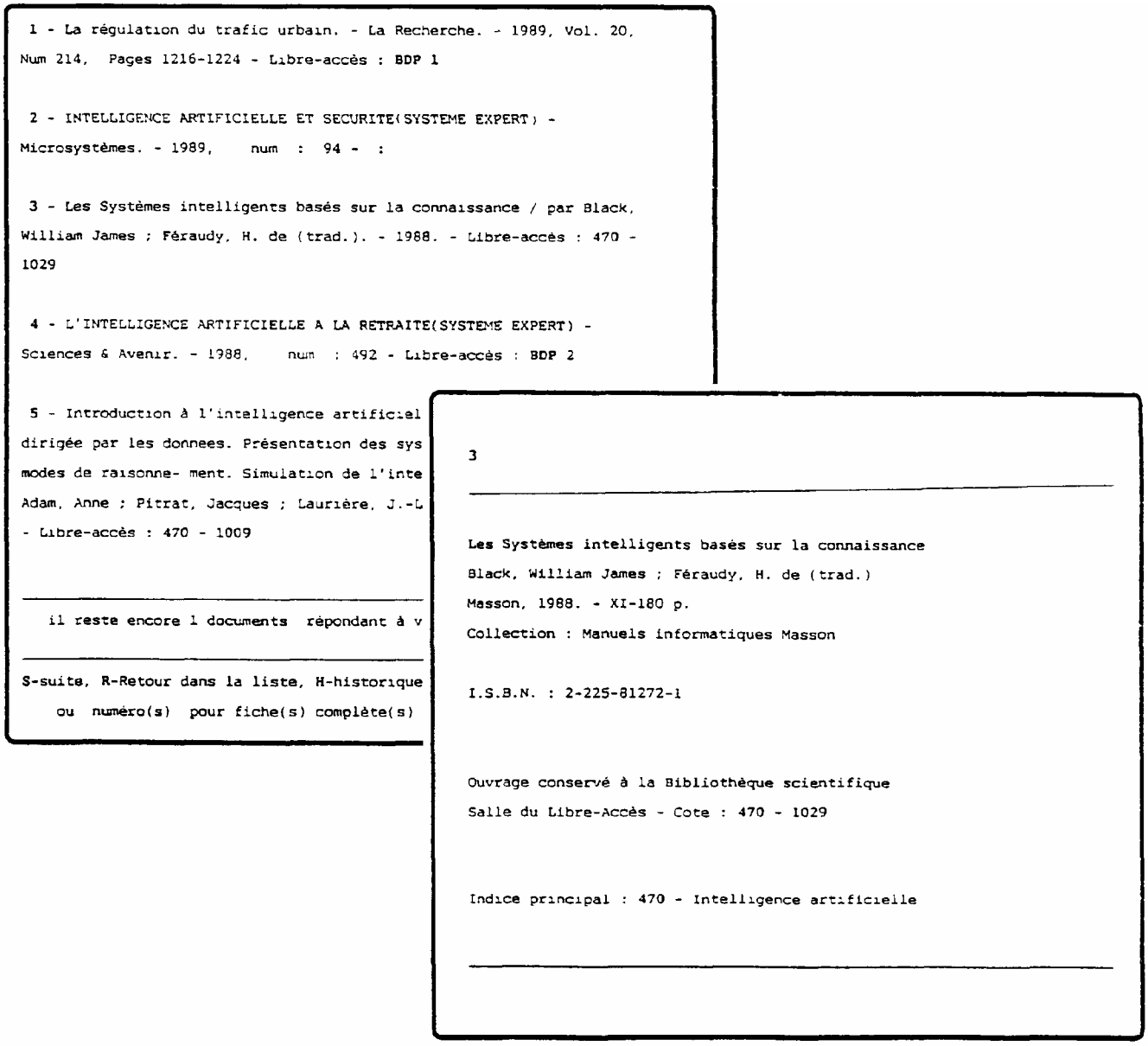
Notre volonté d'utiliser toute la richesse du jeu de caractères souffre toutefois d'une exception, dans le cas de l'utilisation du minitel 1B. En effet, la mémoire morte de cet appareil, qui se réduit au jeu ASCII défini par l'ISO, ne comporte pas les lettres surmontées d'un accent circonflexe ou d'un tréma. Pour les autres signes (à, ç, é, è, ù), *Texto* propose une roue de transcodage qui permet d'envoyer le caractère 7bits correspondant à chaque fois qu'un signe diacritique est lu par l'ordinateur serveur, dans son propre codage des caractères. Cette méthode des roues de transcodage permet d'utiliser le codage interne des caractères, propre à chaque type d'ordinateur, et de permettre des accès par des matériels hétérogènes.

3 - La conception des écrans vise avant tout à la clarté de la lecture. Celle-ci est obtenue en réduisant à l'essentiel les informations sur chaque document et en utilisant toutes les ressources typographiques des écrans *(jeu* de caractère en minuscules, surbrillance). L'utilisation de l'inverse vidéo a été évitée, car elle conduit à une lecture difficile. Loin de mettre en valeur, l'inverse vidéo tend à décomposer l'écran en zones, ce qui disperse la lecture.

On distingue deux formats de réponse :

. la liste des réponses en format court indique le titre, les auteurs et la cote de l'ouvrage. Elle peut suffire pour une recherche du document.

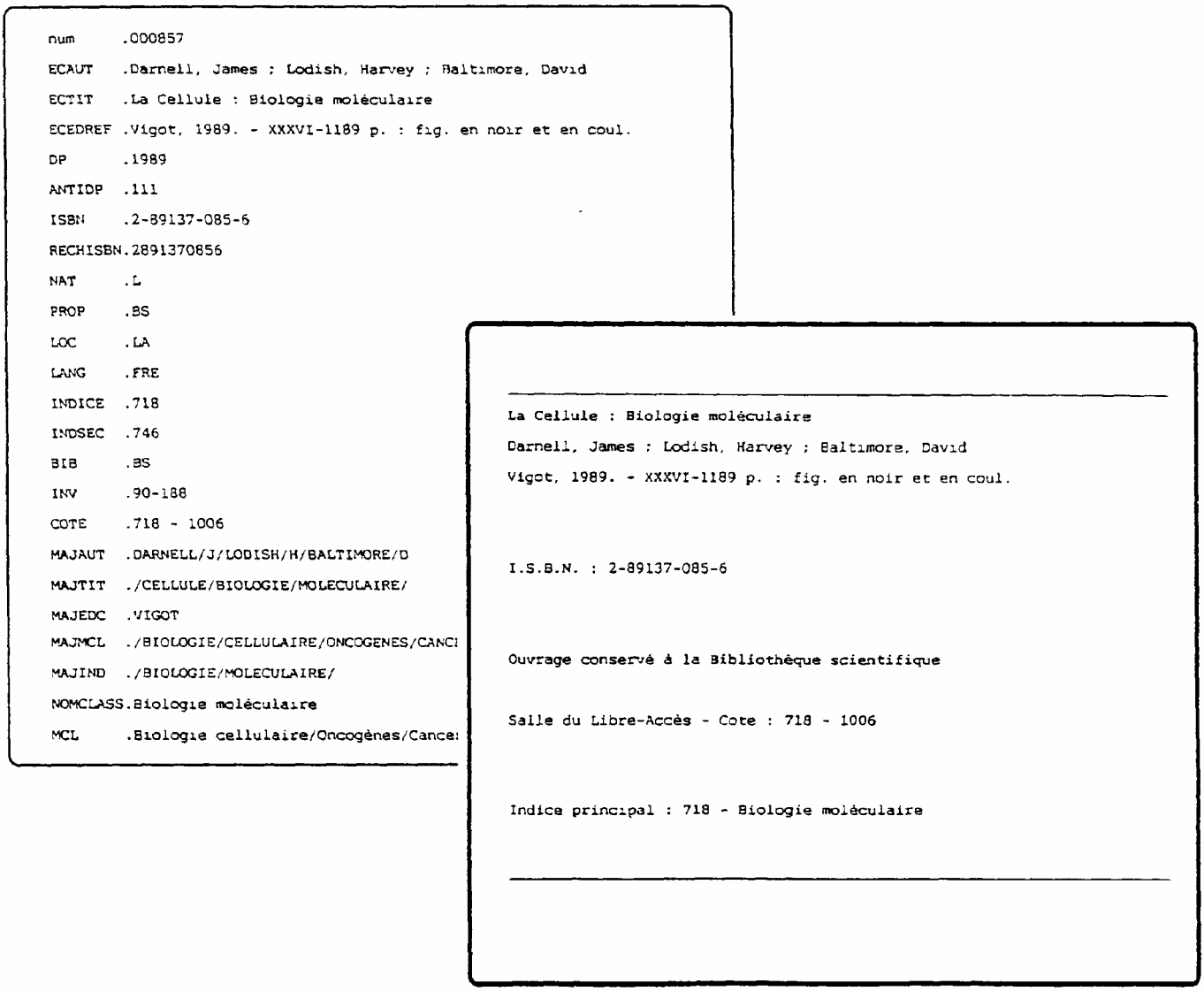
. la référence complète peut cependant être demandée par l'utilisateur pour obtenir plus de précisions.



La référence d'un ouvrage est complètement débarrassée de tout l'appareil documentaire permettant de le retrouver (mots-clés, indices secondaires,...). Au contraire, toutes les informations utiles qui apparaissent sous une forme codée dans le fichier documentaire (limitations sur le prêt, niveau d'étude, localisation...) sont transformées en expressions linguistiques complètes.

Le titre des documents apparaît en premier, mis en valeur par la surbrillance. Cela correspond à une bibliothèque scientifique, ou la recherche par sujet, en général exprimée par le titre, prime sur la recherche par auteur, L'indice principal de classement est indiqué, ainsi que sa signification en toutes lettres, ce qui permet un apprentissage du système : un livre intéressant un utilisateur renvoie ainsi à un code de recherche efficace.

Tout l'appareillage interne des données (nom des champs, codes internes...) est éliminé de la visualisation. L'exemple suivant montre la recomposition des données d'un ouvrage à l'écran.



4 - Guider l'utilisateur. Pour montrer à l'utilisateur comment le système traite sa question, on affiche les unitermes récupérés par l'analyse de sa requête, avec le nombre de documents correspondant. Cela permet ensuite de regarder l'un ou l'autre des ensembles correspondant à un terme ou à une combinaison. Cela permet aussi de montrer comment les questions sont traitées.

Cette explicitation du fonctionnement du système se double d'une approche progressive des difficultés de la recherche documentaire. Ainsi, les combinaisons booléennes, qu'il est indispensable de pouvoir réaliser dans un catalogue de bibliothèque ne sont proposées qu'à un deuxième niveau. Seul l'utilisateur averti, ou bien en ayant réellement besoin fera la démarche de construction d'une équation booléenne de recherche, à partir des ensembles obtenus.

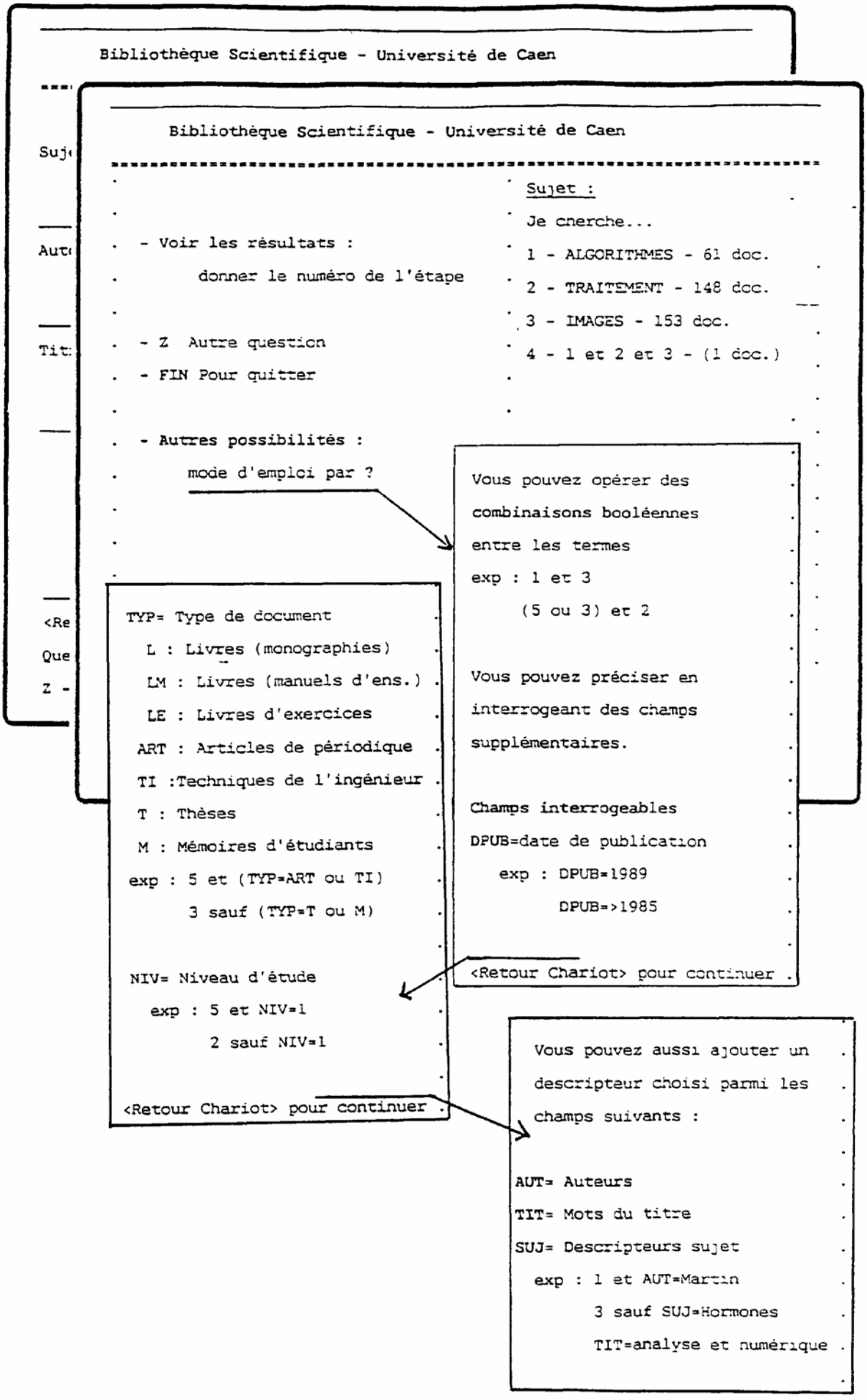
Le système distingue deux modes d'utilisation :

Le mode guidé reste plus limité. Le choix du niveau d'étude requis pour les documents va tendre à éliminer les articles ou les périodiques d'une recherche menée par un étudiant de premier cycle. Les combinaisons booléennes ne sont possibles qu'à partir des ensembles obtenus dans la première question.

Le mode en formulation libre sur champs typés repose sur la présentation d'une grille de recherche. Cet écran d'accueil va être le pivot de la recherche. Après expression des résultats, l'utilisateur peut préciser sa recherche en demandant un type spécifique de document, un niveau d'étude ou en ajoutant des critères de recherche (sujet, auteur, titre).

La partie droite de l'écran sert à renvoyer à l'utilisateur les termes de sa requête et le nombre de réponses correspondant. L'utilisateur peut aussi demander des explications sur les moyens de préciser sa recherche. Les explications apparaissent alors dans cette même partie droite, ce qui évite la perte de repère par "changement d'univers" entre l'appel d'aide et le message d'informations. Cette possibilité de préciser ses recherches lui est indiquée dès que le nombre de réponses dépasse 50, avec un "bip" sonore pour l'alerter.

La succession des écrans est présentée page suivante.



Il est apparu nécessaire de procéder en deux temps pour que l'utilisateur puisse indiquer ses besoins de précision spécifiques. La proposition, dès l'écran d'accueil, d'une grille de formulation trop précise (indiquant le type de document, l'année, le niveau d'étude...) conduit souvent à une surcharge d'informations fournies par l'utilisateur. Ainsi, l'utilisateur tend à préciser un sujet par des mots-clés alors qu'il a rempli les champs auteurs et titre, ce qui ne peut conduire qu'à un silence accru. Demander trop de précisions à l'utilisateur avant même qu'il ait vu ce que contenait le système le conduit à s'interroger plus que nécessaire devant l'écran pour remplir sa demande. De même, si la recherche porte sur le sujet des livres, on ne demande pas de précisions d'auteur ou de titre. Même s'il a la possibilité de ne rien indiquer, l'utilisateur tend toujours à remplir tous les champs qui lui sont proposés.

3.d - La saisie des documents

La saisie est réalisée sur micro-ordinateur. Le programme de saisie est écrit en *dBase.* Trois objectifs ont guidé sa réalisation :

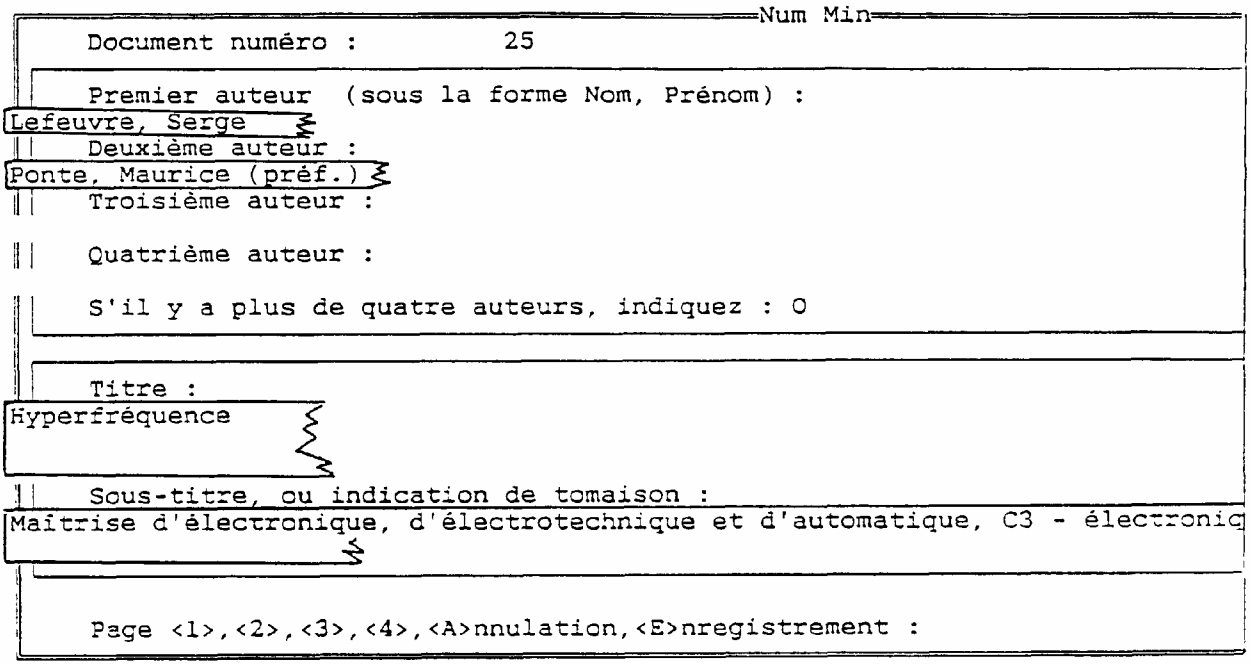
- permettre une saisie souple. La circulation entre les champs de saisie est entièrement libre. Les touches fléchées permettent de déplacer le curseur à l'endroit voulu, et de corriger les fautes qui auraient pu être faites. La saisie peut se faire en plusieurs temps. Les documents sont saisis et indexés, puis on corrige sur un listing et on approfondit l'indexation. On procède ensuite à la correction à l'écran.

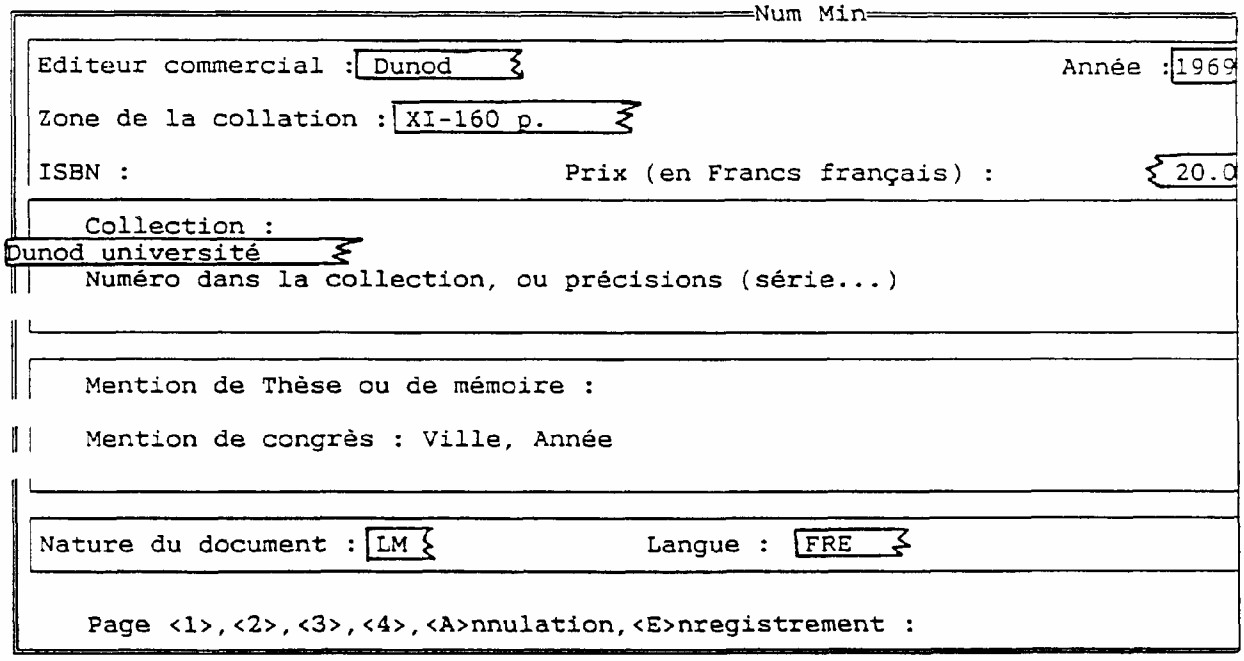
- permettre l'impression de listes de nouveautés, qui sont régulièrement envoyées dans les laboratoires.

- conserver la trace des cotes déjà attribuées aux livres. Cette exigence était obligatoire avec l'adoption des méthodes préconisées par la DBMIST (cotes de la forme : XXX - 1005, b). Si un document doit recevoir un numéro d'ordre dans un indice donné, il faut connaître en permanence le dernier numéro attribué dans cet indice. De même, la mention d'exemplaire étant précisée par une lettre, il faut pouvoir connaître la valeur maximale déjà attribuée, et la modifier si un nouvel exemplaire est reçu.

La saisie est divisée en 4 pages écran. A chaque fois qu'une page est remplie apparaît en bas de l'écran le choix permettant le changement de page, l'annulation ou l'enregistrement.

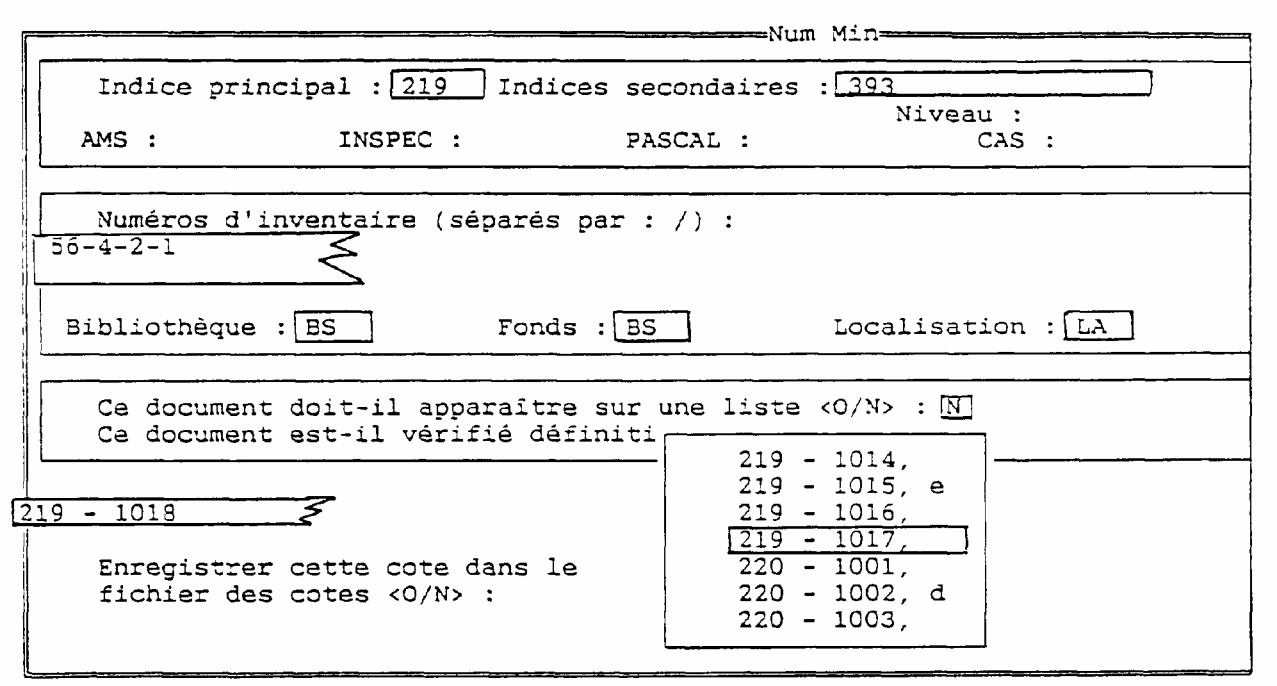
La première page regroupe les informations concernant le titre et l'auteur. La seconde page concerne l'édition et la collation, les mentions de collection ou de thèse.





La troisième page est consacrée aux informations locales. Dans le cours même de la saisie apparaît dans une fenêtre la liste des cotes déjà attribuées dans l'indice mentionné, ce qui permet de définir et d'enregistrer la nouvelle cote, sans devoir se reporter à un autre document (livre d'inventaire).

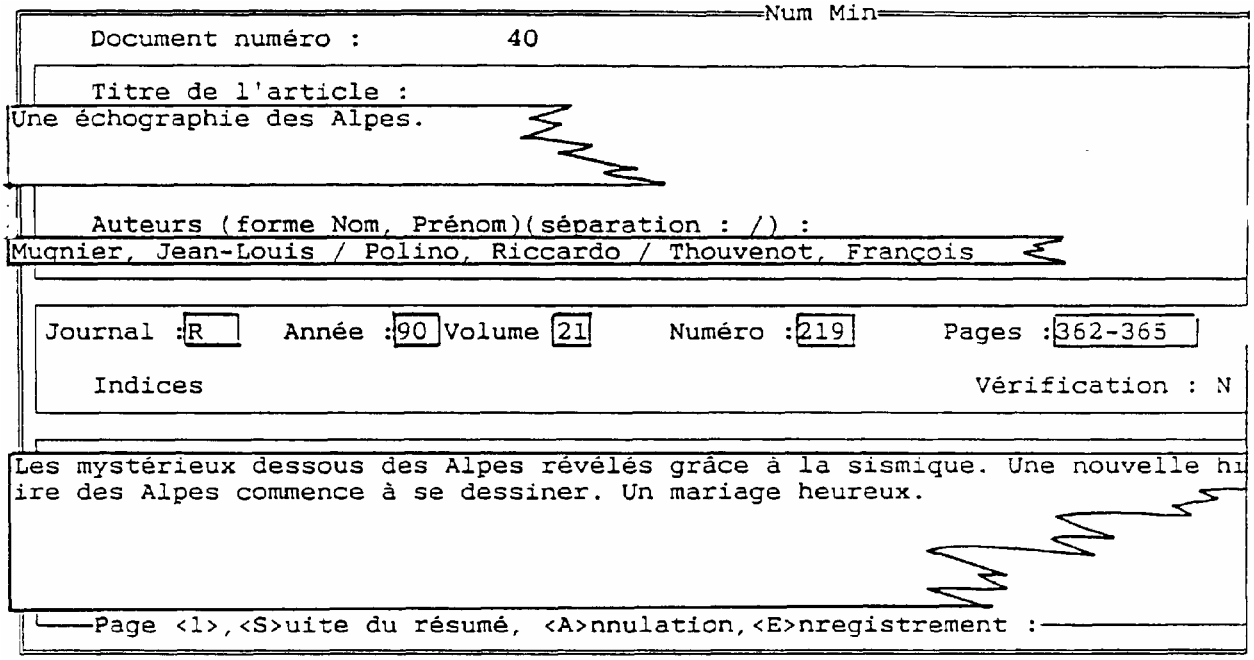
La quatrième page permet l'indexation. A l'origine apparaissait une liste des termes déjà employés, qui se positionnait en fonction des premières lettres saisies pour les mots-clés, l'objectif était d'utiliser le même vocabulaire pour tous les documents. Mais comme la liste n'était classée que par ordre alphabétique, sans liens sémantiques, que d'autre part l'indexation était en formulation libre, cette méthode avait plutôt tendance à ralentir la saisie, sans offrir d'avantage manifeste. L'idée a donc été abandonnée. C'est cet aspect historique qui induit le maintien de cette quatrième page. Lors de la prochaine correction de ce programme, la saisie des mots-clés apparaîtra dans la deuxième page.



Les documents enregistrés peuvent être imprimés, suivant une liste de travail (pour la correction) ou suivant une présentation plus complète pour la liste des nouveautés.

Un masque de saisie du même type a été adapté pour les articles de périodiques et le dépouillement des *Techniques de l'Ingénieur*. Il permet la saisie d'un résumé, et propose de conserver un certain nombre d'informations quand on passe d'un article à l'autre au sein du même périodique. Ce principe de faire réaliser par la machine tout ce qu'elle peut faire est souvent oublié, ce qui a pour conséquence d'alourdir le travail de saisie, et surtout de décourager les personnes qui l'accomplissent, en leur faisant faire des tâches répétitives.

Voici l'écran de saisie des articles :



Au moment du transfert des documents sur le *Vax* du Centre de Calcul, le programme de saisie produit un fichier texte au format ajout piloté. Ce fichier comporte en majuscules les champs d'indexation qui seront repris dans *Texto*. L'opération de préparation du fichier de transfert est relativement longue, mais elle n'est pas réalisée pour de nombreux documents. Le fichier texte est ensuite versé sur le *Vax* du Centre de Calcul par PC-Link, un logiciel de communication. Bien entendu, il arrive avec un mauvais codage des caractères accentués, et il a fallu développer un programme de transcodage.

Le fichier est ensuite versé sous *Texto*. Un programme d'accueil des nouveaux documents, écrit en *Logotel,* se charge, en traitement par lots durant la nuit, des opérations suivantes :

. traiter les caractères qui sont passés au travers du transcodage, comme le U-tréma majuscule (Ü) ou le C-cédille majuscule (Ç)... Ce traitement est indispensable. Il constitue la rançon de la volonté de conserver tous les signes diacritiques. On comprend en réalisant tous les problèmes liés à ces caractères pourquoi de nombreux concepteurs baissent les bras et abandonnent leur langue !

. insérer le séparateur d'articles (/) à la place des blancs dans les champs interrogeables, enlever les ponctuations et un certain nombre de termes du fichier (les articles, les prépositions...).

. fusionner le fichier ainsi traité avec le fichier principal.

. reclasser ce fichier en suivant l'ordre chronologique inverse

. reconstruire les index d'interrogation.

3.e - Conclusion

Ce catalogue informatique contient en octobre 1990 plus de 12 000 références. Sa mise en place a permis de procéder au passage des ouvrages en libre accès plus rapidement que tout autre type de traitement. Depuis le mois d'avril 1989 il est à disposition des étudiants sur la banque de prêt. Il est maintenant consulté régulièrement, et nous envisageons de proposer un deuxième terminal. Le fonds documentaire commence à être suffisamment conséquent pour rendre le service attendu par les étudiants. Il reste à intégrer certains livres de recherche et surtout à proposer le dépouillement des périodiques scientifiques pour accroître l'utilisation par les enseignants et les chercheurs. Cependant, cette information n'est vraiment déterminante que si nous parvenons à la produire rapidement lors de l'arrivée des périodiques. Si le projet du *Schéma directeur de l'information bibliographique* se met en place, on peut espérer que nous trouverons une source de notices qui puisse relever ce défi.

troisième partie

Quelques travaux de recherche

Les divers modèles développés dans la première partie constituent les fondations pour des projets de recherche et développement destinés à faciliter l'accès en ligne aux informations par des utilisateurs. La seconde partie a montré qu'un axe de travail partant des besoins des utilisateurs pouvait donner des résultats positifs y compris dans un environnement austère (indexation par unitermes, utilisation du modèle booléen strict, mode d'interaction ligne à ligne, langage de programmation fruste...). Cette troisième partie entend évaluer quelques perspectives unificatrices destinées à promouvoir l'utilisation efficace des banques de données.

Dans un premier temps on présentera trois axes de travail qui permettent d'avancer vers la résolution de nombreux problèmes d'accès aux informations. On montrera ensuite comment un système anté-serveur pourrait être mis en place, et les fonctionnalités nécessaires de ce type d'outil.

On peut distinguer trois axes de travail dans les recherches en cours :

- l'expertise, qui s'appuie sur une formalisation des opérations cognitives et pratiques mises en œuvre dans une opération de recherche documentaire. Cette formalisation peut être articulée dans un système multi experts.

- l'apprentissage concerne les capacités d'un système documentaire à évoluer en fonction de son expérience, c'est-à-dire à synthétiser les connaissances apportées par l'interaction avec l'utilisateur. On peut alors définir deux types d'apprentissage :

. l'apprentissage à court terme, dans lequel le système doit apprendre de l'utilisateur à mieux cerner la question en cours de traitement.

. l'apprentissage à long terme, qui permet au système de modifier ses connaissances sur le monde documentaire qu'il traite (connaissances sur la formulation des requêtes par les utilisateurs et connaissances cristallisées dans l'indexation des documents).

- la représentation, qui regroupe les travaux destinés à modifier les conditions d'insertion des documents dans le système documentaire. Cet axe de travail concerne plus particulièrement les concepteurs de systèmes documentaires (documentalistes, producteurs d'information). Nous présenterons dans ce cadre les travaux de l'équipe de recherche en sciences de l'information de l'Université de Caen concernant une méthode de représentation adaptée à des documents complexes, basée sur une approche topologique de l'espace sémantique.

L'évaluation des perspectives sur ces trois axes de travail doit nous permettre de tracer les grandes lignes des systèmes anté-serveurs qui assureront le passage des systèmes documentaires actuels, reposant sur le trépied *indexation par unitermes - modèle booléen de formulation des requêtes - langage de commande>* à des systèmes de nouvelle génération qui intégreront dans leur conception même les apports de la recherche en sciences de l'information, notamment les trois axes : expertise, apprentissage, représentation. Dans cette approche des anté-serveurs, il est nécessaire de préciser les critères d'évaluation de l'efficacité des systèmes proposés. Parce qu'ils travaillent sur les banques de données réelles de plusieurs millions de références, les anté-serveurs peuvent devenir des outils extraordinaires pour le test des modèles documentaires en grandeur réelle. Leur conception doit donc intégrer une architecture suffisamment ouverte pour permettre de comparer dans un même protocole expérimental plusieurs hypothèses.

I - Trois axes de travail significatifs

1 • L'expertise

Les systèmes experts constituent la face la plus visible des recherches en Intelligence Artificielle. Les fondements des systèmes experts sont parfois critiqués. On peut notamment remarquer que les experts humains n'utilisent pas un savoir-faire formalisé, basé sur l'application de règles, et que c'est justement

pour cela qu'ils sont des experts, alors que les débutants doivent plus scrupuleusement utiliser des règles de travail [AND87]. Cependant, le travail d'élaboration "d'experts" informatisés permet de mieux définir les limites du champ d'action d'un opérateur humain et l'enchaînement des opérations cognitives mise en œuvre dans une opération. En ce sens, les divers essais de construction de systèmes experts documentalistes et plus généralement des applications des recherches en intelligence artificielle à l'informatique documentaire permettent de mieux concevoir la division de l'opération de recherche documentaire en plusieurs blocs logiques, et d'opérer sur chaque bloc de compétences un apport de connaissances permettant de le réaliser automatiquement. Une synthèse de ces recherches est proposée par G. P. Zarri [ZAR88]. Nous étudierons ici deux "assistants intelligents" en recherche documentaire, les systèmes I^3R et EURISKO qui nous semblent offrir des perspectives de recherche intéressantes.

1.a - Présentation de I^3R

Le système I^3R *(Intelligent Intermediary for Information Retrieval)* de Bruce Croft [CRO87] constitue un modèle général d'articulation de plusieurs experts dans une architecture de *tableau noir (blackboard)*. Dans cette hypothèse, les diverses compétences sont divisées en plusieurs experts, et les échanges entre experts sont réalisés autour d'une structure de plan, gérée par un contrôleur *(scheduler)*. Le rôle du contrôleur est de déterminer l'activité la plus appropriée à chaque étape de la recherche, et de déclencher l'action de l'expert correspondant. Les données utiles et les résultats des opérations de chaque expert sont placés dans la mémoire active du tableau noir. Chaque inscription sur le tableau en modifie l'état, ce qui permet au contrôleur de lancer une autre opération. Les enchaînements d'actions sont déterminés par un plan, qui reprend les diverses étapes d'une recherche menée par un expert humain. Le plan est représenté par un arbre dont les nœuds sont les objectifs de l'étape (avec mention des experts efficaces dans la réalisation de cet objectif). Les transitions par défaut correspondent à une démarche qui obtient des résultats positifs à chaque étape. Si des problèmes sont rencontrés, le plan contient des transitions exceptionnelles à partir de chaque état de la recherche vers des objectifs complémentaires alternatifs et l'appel des experts correspondants.

I[3]Rdéfinit six experts :

. Le constructeur du modèle de l'utilisateur *(user model builder - UMB)* qui collecte des informations sur l'utilisateur, pour lui attribuer une place dans un des "stéréotypes" prévus. Il s'agit généralement d'informations sur des paramètres globaux (intérêt pour des recherches à forte couverture,...), mais aussi d'informations possédées par l'utilisateur sur le thème de sa recherche, qui peuvent servir à préciser le modèle de la requête.

. Le constructeur du modèle de la requête *(request model builder - RMB)* a pour tâche d'obtenir une formulation de la requête par l'utilisateur, et d'en déduire les descripteurs utilisés pour la recherche. Les questions sont posées en formulation libre, ou avec des opérateurs booléens, éventuellement avec des poids attachés aux concepts. Le constructeur du modèle de la requête se charge aussi de la reformulation en utilisant les jugements de pertinence de l'utilisateur pour extraire des informations complémentaires à partir des documents pertinents.

. Le gestionnaire de la base de connaissance *(domain knowledge expert - DKE)* propose à l'utilisateur le contenu de la base de connaissances concernant le domaine de la recherche en cours afin d'inclure de nouveaux termes dans le modèle de la requête. La base de connaissance est globalement structurée comme un thésaurus. Les concepts sont représentés par un schéma *(frame)* comportant trois types d'informations :

. le nom du concept

. des informations sur la manière de reconnaître un concept
dans le texte de la question. Les concepts sont définis à partir des termes présents dans la question, avec un certain degré de confiance par des règles du type :
Si radical de terme x *alors* Concept X *(Confiance :* c).

. des relations entre le concept considéré et d'autres concepts. Cette information permet de reformuler les questions. Les relations utilisées sont la synonymie, la généralisation (termes génériques, termes spécifiques d'un thésaurus), l'instanciation, la relation "partie_de", et une relation plus vague d'association entre concepts.

Les connaissances sont introduites graduellement dans la base de connaissances par les utilisateurs. Face à une question particulière, si le système possède des informations sur les concepts en œuvre, le gestionnaire de la base de connaissance les propose à l'utilisateur pour validation.

. Le contrôleur de la recherche *(search controller - SC)* assure l'exécution de la requête. I^3R fonctionne suivant le modèle probabiliste, et utilisant une base de données propre, permet aussi des recherches sur les agrégats de documents. Les documents retrouvés sont placés dans le constructeur du modèle de la requête, qui les fait évaluer par l'utilisateur pour une reformulation éventuelle de la requête.

. le générateur de parcours *(browsing expert - BE)* permet une navigation dans la base documentaire. Les connaissances acquises dans cette opération par le système permettent une reformulation de la question en réalisant une mise à jour du modèle de la requête. La navigation est essentiellement dirigée par l'utilisateur. Les liens entre documents sont de trois types :

. lien vers le document le plus proche (NN -*nearest neighbor)*
. liens vers les documents cités *(bib - bibliographie)*
. liens vers les documents citant le document examiné *(cite)*

Les parcours sont représentés graphiquement comme un réseau de nœuds et de liens. Le générateur de parcours fonctionne comme un système d'extension de l'activation dont le contrôle est entièrement dans les mains de l'utilisateur. Il ne peut suffire à lui seul pour effectuer une recherche productive, en raison des problèmes de désorganisation cognitive évoqués à propos des systèmes hypertextes. Son objectif est cependant de permettre d'acquérir de nouvelles informations sur les besoins de l'utilisateur pour modifier le modèle de requête.

. le journal *(explainer)* peut être appelé par l'utilisateur qui récupère alors la liste des règles utilisées par le système.

Cette architecture multi experts est utilisée dans un processus de requête à jugement de pertinence. L'utilisateur reste en permanence maître des actions du système, et en charge de nourrir la base de connaissances.

1.b - Présentation de EURISKO

Le système EURISKO, développé à l'université de Toulouse par Christine Barthes et P. Glize est destiné à l'accès en ligne aux serveurs. Il propose lui aussi une gestion basée sur la notion de "génération de plans", et de contrôle de raisonnement dans un monde évolutif [BAR89]. Ce système utilise les informations venant de la base de données, notamment le nombre de réponses à une question, pour évaluer chaque étape de la recherche et décider des actions à mener pour élargir ou au contraire préciser les questions.

Une enquête auprès de bibliothécaires (le projet EURISKO est soutenu par la DBMIST, et s'est déroulé en association avec la bibliothèque scientifique de l'université de Toulouse) permet de définir les moyens utilisés pour réaliser ces deux opérations fondamentales.

L'élargissement d'une requête est réalisé par :

- ajout de synonymes
- élargissement de la portée des troncatures
- substitution d'opérateurs (ADJ -> PHR -> même champ -> ET)
- élimination de restrictions de la recherche à certains champs
- élargissement aux termes en formulation libre
- substitution de termes trop spécifiques par des termes génériques
- élimination de termes trop spécifiques.

La précision est au contraire définie par :

- ajout de nouveaux concepts, éventuellement acquis lors de la visualisation
- limitation de la portée des troncatures
- substitution de termes génériques par des termes plus spécifiques
- limitation de la période de recherche aux dernières années
- restriction de la recherche aux champs "titre" et "descripteurs"
- suppression de certains synonymes
- restriction portant sur les types de documents ou les langues
- substitution d'opérateurs (OU -> ET -> PHR -> ADJ)

Le système EURISKO, dont l'objectif est d'interroger les bases de données réelles sur les serveurs commerciaux, est lui aussi basé sur l'articulation de plusieurs modules experts:

- le moteur d'inférence MIME
- la base de connaissances homme machine, qui travaille à la compréhension de la requête, exprimée en formulation libre (50 règles)
- la base de connaissances décrivant le processus de recherche d'information et l'accès aux banques de données (60 règles)
- un fichier décrivant les caractéristiques des banques de données et des serveurs, avec deux index par domaine et par serveur.
- un logiciel de communication et une carte modem.

La compréhension de la requête passe par la distribution des termes proposés par l'utilisateur dans un modèle de requête. Une requête est considérée comme la conjonction de concepts, les concepts étant une disjonction de termes ou d'expressions.

Cette démarche définit le profil de requête suivant :

<équation> = <concept> | ET <équation>
| ET <équation> <restriction>
<restriction> = SAUF <expression> I titre I descripteur I période...
<concept> =<expression> | OU <concept>
| OU <concept> <restriction>
<expression> = <mot> | (ET/SAUF) <mot>
| {ET/SAUF} <expression>
<mot> = <terme> | <syntagme> I code I abréviation
<syntagme> = <terme> | ADJ <syntagme>
<terme> = racine troncature

Une requête en formulation libre est répartie suivant ces objets afin de définir l'équation de recherche. Toutefois, l'utilisateur est sollicité pour confirmer

l'utilisation des opérateurs dans toutes les étapes de l'élaboration de l'équation de recherche.

I.c • Les leçons de l'expertise

L'architecture générale de I^3R nous montre qu'il est possible d'échanger des informations entre plusieurs modules autour d'un contrôleur de la recherche documentaire. Cette architecture semble la plus prometteuse si l'on veut réaliser des anté-serveurs intelligents. Elle permet de faire évoluer séparément, en fonction des recherches les divers experts. On peut cependant élever deux objections aux expériences en cours avec I^3R : la base de données associée est de faible envergure (la collection CACM de 1500 documents) mais surtout est gérée directement par le système, ce qui permet de proposer des modèles de recherche inaccessibles dans les systèmes en ligne (notamment la recherche par agrégats et le modèle de navigation).

Le système EURISKO nous apprend beaucoup sur les méthodes actuelles de travail des documentalistes qui pratiquent la recherche documentaire en ligne. Sa force est de se situer dans l'environnement réel des serveurs commerciaux. L'analyse qui a été menée avec les bibliothécaires permet de définir un cadre opératoire, car en phase avec les pratiques actuelles de la recherche documentaire. Il faut cependant souligner les limites : le bibliothécaire se comporte ainsi car les systèmes documentaires à sa disposition ne lui permettent pas de classer les documents par ordre de pertinence, et ne lui offrent aucun moyen d'évaluer le silence de sa requête (i.e. connaître les termes d'indexation utilisés dans les documents pertinents qui ne sont pas extraits par la requête). Les opérations décrites ne sont donc pas en elles-mêmes des moyens cohérents et théoriquement fondés de transformer une requête. Elles ne sont que des substituts, parfois efficaces, rendus obligatoires par l'état général des systèmes de recherche en ligne. En ce sens la démarche correspond à la vocation générale d'un système expert de remplacer un opérateur humain, sans intervenir sur l'environnement déterminant l'action de cet opérateur. La réalisation d'un anté-serveur intelligent doit s'appuyer sur cette analyse, mais aussi essayer de la dépasser en proposant un classement de pertinence des documents.

Il conviendrait alors de proposer une évaluation probabiliste de la

construction de la requête en fonction de son efficacité face à une banque de données réelles. Les opérateurs booléens proposés par l'analyse de la requête seraient validés par les résultats qu'ils permettent d'obtenir dans l'univers pragmatique défini par la banque de données.

2 • L'apprentissage

La notion d'apprentissage est fondamentale dans toutes les recherches informatiques modernes. Dans le domaine documentaire, cette préoccupation vient à la rencontre de deux critères qui justifient que l'on s'accorde à promouvoir des systèmes évolutifs :

- la définition de la question telle qu'elle est posée par l'utilisateur en début de recherche est susceptible d'évoluer en fonction de l'information qui lui est renvoyée par la banque de données. Dans le processus de recherche documentaire, l'utilisateur est confronté à deux problèmes : il cherche des choses qu'il ignore, et donc peut plus difficilement les formuler clairement, et il ne connaît pas le contenu de la banque de données, ni ses règles propres d'indexation. Cette position circonscrit le besoin d'un apprentissage immédiat. Cet apprentissage peut d'ailleurs être à double détente : le système doit "apprendre" ce qu'est le désir documentaire évolutif de l'utilisateur, et l'utilisateur doit "apprendre" ce que contient le système et les possibilités qu'il lui offre.

- la définition des documents en termes d'indexation est susceptible d'évoluer dans le temps. Il y a d'abord une évolution linguistique parallèle à l'évolution de la recherche, ou parallèle à l'évolution des points de vue, qui conduit à un affinement de termes qui à un moment donné peuvent être spécifiques, et devenir plus génériques. L'exemple de l'extension du champ couvert par des termes comme *"génétique", "intelligence artificielle", "banques de données"...* au cours des vingt dernières années est significatif. On trouverait de même des exemples dans les domaines de l'actualité (parler de *"choc pétrolier"* en 1973 ou maintenant n'a plus le même niveau de précision) ou des sciences sociales. Il y a ensuite un besoin de réduire le fossé entre l'indexation produite par analyse d'un document (analyse manuelle ou même automatique) et l'indexation qui serait

souhaitée pour ce document par les utilisateurs (i.e. les termes de la question de l'utilisateur qui faisaient souhaiter à cet utilisateur de retrouver ce document). Dans les deux cas, on peut définir un besoin d'apprentissage linguistique à long terme, qui permettrait au système de s'adapter au vocabulaire de l'utilisateur, au travers de son adaptation au vocabulaire d'un ensemble d'utilisation.

2.a - L'apprentissage à court terme : la reformulation

Si l'on considère le besoin documentaire de l'utilisateur comme étant un concept que le système doit apprendre [HALI89], on se rapproche plus près de la situation concrète de la recherche documentaire. La situation où la question originale représente exactement et univoquement le besoin de l'utilisateur est un cas particulier, vraisemblablement marginal.

Plusieurs arguments conduisent à ne considérer la question originale qu'au simple titre d'indicateur du besoin :

. il existe une frontière psychologique à exprimer brutalement un besoin. En fait, derrière un besoin documentaire se profile une interprétation prévisionnelle des résultats. Mais l'hypothèse de travail de l'utilisateur peut s'avérer erronée (par exemple croire qu'il n'y a pas ou peu de documents traitant sa question et se retrouver désemparé devant une masse trop importante même pour la survoler). Il est alors difficile de livrer entièrement à l'intermédiaire en information les implications de la recherche, surtout avant d'avoir évalué le comportement du système documentaire.

. un besoin documentaire s'exprime à la fois par des termes définissant le sujet de la recherche, mais aussi par la recherche d'un point de vue sur ce sujet, qui peut bien plus difficilement se résumer en quelques descripteurs. Par exemple, le point de vue *"applications en grandeur réelle dans des entreprises"* qui est à la source d'une recherche sur un sujet technique peut être contradictoire avec l'intérêt qui sera néanmoins porté à des revues de synthèse d'orientation plus fondamentale sur ce même sujet. Cette distinction entre le sujet et le point de vue est plus sensible encore dans le domaine des images. Une image peut être décrite par le sujet représenté (personnage, lieu, objets...), par la

morphologie (portrait, plan d'ensemble...), mais aussi par un aspect de connotation (allégorie de la tristesse, agitation, sérénité...). Ce dernier aspect est à la fois difficile à exprimer en descripteurs, mais plus encore n'apparaît pas forcément dans la conscience de l'utilisateur avant d'avoir vu des images et avant d'avoir lui-même constaté l'effet émotionnel de cette connotation.

. la recherche peut porter sur un sujet nouveau pour l'utilisateur, qui devra alors employer des termes qui ne sont pas forcément ceux de la spécialité, ni même le jargon professionnel propre au cercle restreint des auteurs publiant sur ce sujet. La recherche se déroulera alors "à partir" d'une position déjà connue, "vers" une description qui ne se formalisera qu'au cours de la recherche, éventuellement par une succession à intervalles répétés de recherches de plus en plus focalisées.

. l'utilisateur peut exprimer son besoin en termes très génériques, afin de se faire une opinion sur les capacités du système à répondre à sa demande. Cette attitude se retrouve devant l'intermédiaire humain, qui doit d'abord mettre en confiance l'utilisateur en l'assurant de sa capacité à traiter la question et à l'aider dans son besoin documentaire. Une démarche semblable est réalisée devant le système documentaire informatisé. C'est par exemple le cas d'étudiants demandant des livres *"de mathématiques"* au catalogue d'une bibliothèque scientifique.

. la formulation précise de la recherche peut être de l'ordre du "secret", correspondant à un besoin réel (notamment en milieu industriel) ou supposé (dire l'objet de sa recherche est déjà dévoiler les axes de travail). Cette obsession du secret peut s'avérer justifiée car il n'existe pas de code de déontologie des documentalistes ni des serveurs. Par exemple, des grandes entreprises ou des organismes militaires préfèrent acheter les fichiers documentaires aux producteurs et construire leur propre service de recherche documentaire *(CEDOCAR* en France) que de faire appel à des serveurs publies, éventuellement appartenant à des concurrents (le serveur *Dialog* fut longtemps propriété de Loockeed, ce qui devait singulièrement limiter la confiance de tous les industriels concurrents de l'aéronautique ou de l'armement).

Tous ces arguments plaident en faveur du fait suivant : le besoin documentaire ne s'exprime réellement qu'au cours du déroulement de la recherche. Le *"modèle de la requête"* doit donc être en permanence modifié pour s'adapter au besoin réel de l'utilisateur.

La pratique du jugement de pertinence *(relevance feedback)* est un des premiers moyens mis à la disposition du système pour apprendre de l'utilisateur le contenu exact de sa question. En jugeant les documents, l'utilisateur ne se constitue pas seulement son ensemble personnel de documents à extraire du système, il réoriente la recherche. L'utilisateur juge les documents, mais le système apprend à partir des descriptions documentaires. Cette distinction est encore plus présente dans le cas des images : l'utilisateur sélectionne un lot d'images parce qu'il les voit, sans connaître leur description, mais le système dispose alors d'un moyen de reformuler la question par analyse de ces descriptions.

La démarche traditionnelle d'utilisation du jugement de pertinence consiste à ajouter à la question les descripteurs qui appartiennent aux documents pertinents mais n'appartiennent pas aux documents rejetés. Un problème intervient cependant dans l'évaluation des raisons qui poussent à rejeter un document. Peut être est-ce tout simplement parce que le document est connu de l'utilisateur, peut être est-ce une mauvaise interprétation faite par l'utilisateur d'un document qui en fait aurait été pertinent, peut être est-ce lié au fait qu'un certain nombre de documents semblables ayant déjà été sélectionnés, l'utilisateur juge leur quantité suffisante... L'élimination de documents est plus aléatoire que la sélection, qui nous indique clairement les besoins de l'utilisateur.

Le jugement de pertinence est utilisé pleinement dans le système public *Dow Quest* [TRI89]. Ce système fonctionne sur une recherche sur fichiers signés en utilisant la puissance de calcul de la *Connection Machine.* La signature de la question est remplacée, après jugement de pertinence, par une nouvelle signature qui prend en compte l'ensemble des termes des documents (en occurrence des articles de presse et des dépêches d'agence). Il n'y a pas de pondération de l'importance des termes de recherche. Cela limite d'après Salton [SAL88b] les performances du système. Typiquement, tous les termes d'un document ne sont pas de bons moyens de modifier une question.

Pour que le système apprenne à partir du jugement de l'utilisateur, il faut pouvoir distinguer parmi les termes d'indexation des documents sélectionnés ceux qui permettent de réorienter au mieux la recherche. En ce sens, les termes présents uniquement dans les documents sélectionnés, ou en sens inverse ceux qui ne figurent que dans les documents rejetés sont significatifs. Les occurrences de termes ont aussi un pouvoir important. Enfin, l'évaluation de la Valeur de Discrimination d'un Terme (VDT) permet de sélectionner les termes les plus importants. L'objectif de cette analyse des descripteurs des documents jugés est de transformer une question de 7 à 10 termes en une question de 30 à 40 termes. Cette question plus riche permet alors, suivant le principe "question de qualité, réponse de qualité" *("quality in - quality out"* [CR087]), un meilleur calcul de pertinence formelle pour sélectionner et classer les documents les plus importants. Le jugement de pertinence permet d'obtenir des questions comportant un nombre suffisamment important de termes significatifs. Cela a pour effet de faire se confondre l'ensemble des documents pertinents pour l'utilisateur et l'ensemble des documents pertinents pour la question telle qu'elle est posée [CR089].

Une autre méthode d'exploitation du jugement de pertinence est proposée par le prototype RIVAGE développé à Nancy par Gilles Halin et Marion Créhange ([HALI89], [CRE89]). Le système gère un fonds d'images sur vidéodisque. Le jugement est porté par l'utilisateur sur un premier lot d'images obtenu par la question originale. Chaque image est décrite en utilisant un thésaurus, qui a été transformé pour constituer une taxinomie, c'est à dire intégrant un terme spécifique pour classer les documents qui ne sont pas prévus par les autres termes.

Le jugement de l'utilisateur va consister à transporter de l'information dans le réseau sémantique constitué par ce thésaurus. Chaque terme du thésaurus est affecté d'un *"niveau d'expression"* du besoin de l'utilisateur, qui varie dans [-1, +1]. A l'origine, tous les termes ont une expression de -1. On calcule un poids de pertinence pour chaque terme des descriptions en fonction du nombre d'images sélectionnées ayant ce terme et du nombre d'images rejetées ayant aussi éventuellement ce terme. Ce poids de pertinence est placé dans le thésaurus comme niveau d'expressivité, et propagé par une opération de moyenne sur tous les termes spécifiques. Le thésaurus porte alors la marque de

l'apprentissage réalisée par le système des besoins de l'utilisateur. Pour reformuler la question, le système passe en revue tous les termes du thésaurus et sélectionne ceux dont le poids d'expression dépasse un certain seuil.

L'apprentissage dans ce cas permet d'utiliser toute la puissance du thésaurus sans demander à l'utilisateur d'en connaître la structure, ni de devoir en descendre toutes les branches. La limite est cependant marquée par la dimension de ce thésaurus, dont le balayage doit être réalisé en entier. Enfin, le système doit s'appuyer sur un thésaurus déjà constitué, et pour l'instant mono hiérarchique.

Une dernière méthode permettant au système d'apprendre le besoin documentaire de l'utilisateur est d'offrir à ce dernier la possibilité de sélectionner des termes dans une liste de *"pistes"* [THO89], [VIC88]. Une piste est un terme ou une expression, appartenant à une liste pré définie, qui peut être déduit à partir des termes d'indexation présents dans le premier lot de documents sélectionnés (avant jugement de pertinence). Ne sont néanmoins considérés comme pistes et proposés à l'utilisateur que les termes qui ouvriraient une modification du nombre et du classement des documents que sélectionnerait la question reformulée. Ce dernier point est très important si l'on compare cette méthode avec les fonctions de tri statistique sur les descripteurs proposées par les serveurs commerciaux, par exemple la fonction ZOOM de *l'Agence Spatiale Européenne* ou la fonction MEMTRI de *Questel* +. Dans ce dernier cas, le système analyse les descripteurs (unitermes ou descripteurs composés) de tous les documents sélectionnés par une première question. On peut par exemple s'apercevoir que de nombreux documents sélectionnés par la recherche *"Beurre"* contiennent le terme *"Cacao",* et qu'il faut en demander l'élimination si l'on s'intéresse aux produits laitiers. Mais on peut aussi se retrouver avec un nombre important de termes inutiles, par exemple de termes qui ont un faible pouvoir discriminant dans l'ensemble de la banque de données, et dont la forte présence dans le lot des documents sélectionnés n'est pas significative (termes trop fréquents ou termes trop spécifiques).

Dans le système QUID que nous développons à l'université de Caen, et qui sera présenté dans le chapitre sur la représentation, les termes d'indexation sont d'un niveau "inférieur" aux termes proposés à l'utilisateur. On parle de

"micro caractéristiques" ou de *"caractéristiques élémentaires"*. Les calculs se font sur ces éléments et les pistes qui appartiennent à un vocabulaire limité sont ensuite déduites de la combinaison de ces éléments. Cette méthode permet l'aspect inverse de l'apprentissage : il s'agit alors de montrer à l'utilisateur des règles de fonctionnement du système afin de lui permettre de mieux maîtriser sa recherche. Une version simplifiée a été employée dans le catalogue de la bibliothèque scientifique de l'université de Caen, quand le système, au lieu de répondre un chiffre dénué de signification à des questions générales, comme *"mathématiques"* ou *"biologie",* propose le chapitre correspondant du plan de classement. Le langage documentaire n'est pas toujours le langage de l'utilisateur. Il reste cependant utile pour éviter les ambiguïtés ou les formulations (resp. indexations) incertaines. L'objectif est alors de permettre à l'utilisateur d'apprendre ce langage en fonction de ses besoins, tels qu'il les a exprimés dans sa première question. Cette méthode est plus efficace que de refuser la question *"sport"* (ou de proposer les 7 documents qui par hasard sont indexés à ce terme générique) parce que les ouvrages correspondant sont indexés à *"football", "tennis"* ou *"planche à voile"* comme dans le CD-Rom LISE.

David Blair [BLA90] considère que l'utilisateur placé face à un système documentaire se trouve en position d'apprendre un nouveau langage. S'appuyant sur les théories de Wittgenstein, il estime que nous n'apprenons pas un langage par des définitions ou par des explications, mais parce qu'on nous montre l'usage des expressions. Selon lui, la capacité d'un système documentaire à être "appris" par l'utilisateur dépend de sa capacité à :

. proposer des exemples judicieux d'utilisation des termes d'indexation. Ces termes sont souvent peu représentatifs en eux-mêmes, mais ne prennent valeur qu'en fonction des documents qu'ils indexent.

. proposer des exemples de recherches efficaces. On retrouve une des préoccupations des concepteurs d'hypertextes, notamment de Vannevar Bush, dont le *memex* devait conserver les traces des chemins empruntés par les autres utilisateurs. Il faudrait cependant que ces exemples soient choisis au plus près des préoccupations de la recherche en cours.

2.b • L'apprentissage à long terme

L'apprentissage à long terme est plus difficile à mettre en œuvre et peu de modèles peuvent s'attacher à ce travail. Un document est représenté par une liste de termes d'indexation. L'apprentissage à long terme peut donc suivre deux axes de travail :

. permettre au système d'apprendre la fonction des termes d'indexation dans la représentation et les relations qui s'installent et se modifient entre les mots du vocabulaire. Il s'agit d'obtenir des éléments de construction de thésaurus ou de réseaux sémantiques en fonction de l'expérience acquise par le système dans son domaine documentaire (expérience des indexeurs et expérience des utilisateurs). Cette connaissance permet alors d'agir au niveau de la fonction $\mathbf{f_q}$ de traduction de la question, en proposant des reformulations sémantiques à partir des termes d'une question d'utilisateur.

. permettre au système de modifier la description des documents pour la faire coïncider au mieux avec les besoins des utilisateurs. Cette connaissance existe par l'action qu'elle entretient sur la fonction $\mathbf{f_i}$ d'indexation des documents. C'est une connaissance qui intervient pour aider l'indexeur et qui agit sur la probabilité de satisfaction de l'utilisateur.

Plusieurs moyens sont envisagés pour élaborer ou affiner les compétences linguistiques du système. L'élaboration de la base de connaissances du système I^3R présentée ci-dessus est une forme directe d'apprentissage, qui consiste à demander à l'utilisateur des informations sur les termes qu'il emploie et à les insérer dans le réseau sémantique. On peut aussi envisager une méthode plus automatique.

1 - Construction interactive d'outils documentaires

Le système TEGEN développé à Munich par Güntzer *et al.* [GUN89] consiste par exemple à analyser les questions des utilisateurs pour obtenir des informations sur les relations entre les termes. Ce système fonctionne par l'application de règles d'acquisition basées sur l'observation de la syntaxe des

questions et des réactions de l'utilisateur aux propositions du système. L'acquisition de concepts et de relations est soumise à une approbation par l'utilisateur. Le système distingue les résultats intermédiaires du résultat final et se décompose en un processus d'acquisition et un processus de vérification, ces processus ne pouvant être réalisés par le même utilisateur.

Le thésaurus de TEGEN contient les relations : synonymes, termes associés, homonymes, déclinaisons morphologiques, termes génériques et spécifiques, liste négative (mots outils) et liste positive (vocabulaire contrôlé).

Les acquisitions réalisées automatiquement à partir de règles d'analyse des questions sont placées en résultat intermédiaire. Une règle d'acquisition est de la forme suivante :

Règle 14 : Si deux (ou plusieurs) termes x_i d'une équation de recherche X
sont combinés par l'opérateur OU
et
les termes x_i sont des descripteurs
et
le résultat de l'équation X est combiné par les opérateurs ET ou bien SAUF dans une autre équation de recherche avec d'autres termes y_i
alors
les termes x_i sont deux à deux considérés comme reliés par la relation "termes associés".

TEGEN contient 29 règles d'acquisition de ce type. Certaines conduisant à des résultats ambigus sont accompagnées de demandes explicites faites par le système à l'utilisateur.

Les résultats intermédiaires sont qualifiés par quatre attributs qui définissent la confiance que l'on peut porter aux règles acquises, et qui sont régulièrement mis à jour dans le processus d'apprentissage :

. *Poids d'apprentissage* qui indique le degré de confiance dans la relation

. *Statut d'apprentissage*
. *Nombre de confirmations* obtenues par le système pour un résultat intermédiaire
. *Nombre d'informations* d'un résultat intermédiaire (obtenu par analyse des réactions des utilisateurs).

L'ensemble de ce processus part de l'idée que les utilisateurs possèdent un savoir important dans les domaines qu'ils interrogent, et qu'ils peuvent le donner au système. On peut cependant penser qu'il convient de moduler cet apprentissage par des règles qui seraient acquises à partir des documents eux-mêmes. Le savoir des utilisateurs est orienté vers un but inconnu (recherche de documents), alors que le savoir synthétisé dans les documents est plus affirmé. Les relations sémantiques doivent aussi pouvoir être extraites des documents, soit par des opérations statistiques (occurrence de termes, co-citation de documents...), soit par des opérations linguistiques.

Dans ce dernier cadre, on peut souligner les approches de Gian Piero Zarri ([ZAR88], [ZAR90] et le système *RESEDA)* et de Lisa Rau ([RAU89] et le système *SCISOR)* dont l'objet est d'obtenir une représentation conceptuelle des documents de départ (une sorte de *"métadocument").* Ces deux systèmes ont pour objectif d'offrir directement à l'utilisateur une réponse à sa question telle qu'elle a été extraite sous une forme conceptuelle des documents ayant nourri le système. Dans les deux cas, les systèmes d'acquisition d'information sont dotés de modules "d'acquisition automatique des connaissances" destinés à faire face aux limites des réseaux sémantiques, qui ne peuvent actuellement traiter l'ensemble du savoir sur les entités lexicales.

2 - Modification interactive de l'indexation

Une autre démarche d'apprentissage à long terme concerne la capacité à modifier la description d'un document en fonction des utilisations qui en sont faites durant les différentes recherches consécutives qui ont jugé ce document pertinent. Si l'on admet, comme cela a été démontré dans la première partie, que la description documentaire est un art difficile, aux résultats aléatoires, on peut essayer d'améliorer la description dans un processus continu. Jean Tague, à la

suite d'une expérience sur 503 documents, 64 recherches et engageant 8 utilisateurs, estime que les résultats sont améliorés d'environ 25 % [TAG81].

L'utilisation des compétences extraites des questions des utilisateurs repose sur l'hypothèse qu'il existe des régularités dans la formulation des requêtes qui concluent à la pertinence de tel document. Ces régularités constituent les nouveaux termes d'indexation adaptés à ce document. Dans le même ordre d'idées, les termes qui ont été inefficaces dans une question (absents du lexique du système...) sont ajoutés aux documents retrouvés par l'utilisateur par une autre démarche, ce qui augmente le nombre de synonymes connus du système.

Ces principes ont été adoptés par Michael Gordon pour développer un algorithme génétique pour la redescription des documents [GOR88]. Dans son modèle, chaque document possède plusieurs descriptions, obtenues à partir de la description originelle et des modifications présentées par les utilisateurs (termes non utilisés présents dans les questions ayant jugées ce document pertinent). Lors de l'opération de recherche, on mesure la valeur de chacune des descriptions de ce document en fonction de la question ce qui permet de classer ces descriptions. Puis on remplace les descriptions les moins efficaces par une nouvelle description obtenue par concaténation de parties des descriptions les plus efficaces.

Cet algorithme est dit génétique, car comme dans un processus d'hérédité, il tend à promouvoir la partie la plus adaptée de la population (i.e. des descriptions du même document) en conservant des parties de descriptions des éléments les plus performants et en les combinant pour introduire de la variété dans les nouvelles générations de représentants. Ce système a été testé sur un très petit nombre de documents (18 documents) avec des résultats prometteurs. Il faudrait envisager son application sur des systèmes de plus grande importance. L'inconvénient de cette méthode est de prendre une place en mémoire importante, en raison de la conservation de plusieurs descriptions des documents. On peut cependant relativiser cet inconvénient car la mémoire est la partie de l'informatique qui coûte aujourd'hui le moins cher.

L'apprentissage à long terme est aussi présent dans certains modèles documentaires. Il est par exemple partie intégrante des modèles documentaires

connexionnistes Ainsi, le système AIR de Richard Belew, décrit plus haut, permet d'insérer des mots inconnus du système en tissant des relations avec les documents jugés pertinents par l'utilisateur qui a employé ces termes inconnus dans sa question. Le système permet aussi d'apprendre les liens entre termes et documents ou entre documents entre eux par modification à chaque recherche des poids de liaison. La somme de tous les poids entrant ou sortant d'un neurone est cependant constante, afin de s'assurer que les calculs restent bornés.

La possibilité de conserver des chemins dans l'information et de permettre à d'autres utilisateurs de les suivre est de même une application de l'apprentissage à long terme telle qu'elle est proposée par le modèle hypertexte.

3 • La représentation

L'apprentissage à long terme est aussi dépendant des modes de représentation des documents. Le langage est de loin la forme la plus souple et la plus précise pour exprimer les divers éléments d'une pensée. Cependant, le langage n'est pas le meilleur moyen de retrouver des documents, justement en raison de sa flexibilité, de l'invention permanente de termes et de structures. Les expériences sur la recherche *"en texte intégral"* [BLA85] prouvent que le taux de couverture reste faible, et que ce type de système ne peut pas proposer d'aide à la reformulation des questions par l'utilisateur tant les formulations linguistiques d'un même concept sont nombreuses et imprévisibles.

Il existe un vaste champ de recherche concernant la représentation des documents. Le modèle typique de la recherche documentaire est celui d'une liste de descripteurs associés à un document. Le système contient alors un représentant du document qui sert à la visualisation (référence bibliographique, texte, image,...) et un ensemble d'informations de recherche associées à ce documents (termes d'indexation, données structurées - auteur, date...). Certains termes ont éventuellement une présence conjointe dans les deux ensembles, comme les mots du titre ou du document lui-même pour une indexation en texte intégral, mais leur fonction peut néanmoins être totalement disjointe (par exemple le terme utilisé pour l'indexation est l'équivalent *en majuscules* du terme présent dans le document).

Le travail sur la représentation est un axe de travail fondamental en informatique documentaire. Il incorpore une large part de travail linguistique, notamment sur la capacité à extraire des descripteurs cohérents à partir des formes de surface d'un document, mais aussi pour exprimer les orientations particulières provoquées par l'articulation des mots dans les phrases, qui sont souvent mal exprimées par la simple juxtaposition des termes significatifs. Nous laisserons cependant de côté ce type de travail d'indexation automatique, qui a été évoqué dans la première partie, pour nous concentrer sur les modèles de représentation des termes d'indexation dans le système documentaire en présentant deux approches :

- une approche du type reconnaissance de forme qui permet une nouvelle
conception de la structure du représentant de document destiné à la recherche [LU90]

- une approche topologique basée sur la continuité des éléments de sens qui permet de définir des espaces sémantiques dans lesquels on peut traiter aussi bien les documents que les termes d'indexation ou les mots du vocabulaire de l'utilisateur ([THO89],[VIC89]).

3.a - Modifier la structure des représentants de documents

Dans la structure traditionnelle, un document est représenté par des termes d'indexation qui se réduisent à des mots isolés ou des expressions. Or les unités lexicales sont reliées entre-elles par des relations lexicales sémantiques, qui permettent de représenter d'une manière concise les diverses possibilités d'extension des questions d'utilisateur.

En se basant sur les recherches de Jean Tague [TAG81], Xin Lu [LU90] propose de s'appuyer sur quelques relations sémantiques qui peuvent s'exprimer en reliant directement les unités lexicales sans formulations syntaxiques (sans avoir besoin d'écrire des phrases explicatives).

Celles-ci sont regroupées en 5 types :

- Synonymie :
  - . Synonymie cognitive (fiddle : violon)
  - . Presque synonymie (citation : référence)
  - . Variation morphologique (man : men)

- Taxinomie :
  - . "taxonymy" (cheval : étalon)
  - . "co-taxonymy" (brebis : bélier)

- Partie Ensemble :
  - . "meronymy" (bras : main)
  - . Co-"meronymy" (palm : figure)
  - . Appartenance à un groupe (sénat : sénateur)
  - . Appartenance à une classe (prolétariat : travailleur)
  - . Appartenance à une collection (forêt : arbre)
  - . Entité et ses propriétés (habit : taille)
  - . Fait-en (pneu : caoutchouc)
  - . Vient-de (lait : vache)

- Hiérarchie directe (sans branchement)
  - . Chaîne (épaule : bras : coude : avant-bras)
  - . Hélice (printemps : été : automne : hiver)
  - . Rang (simple : double : triple)
  - . Gradation (petit : grand : énorme)
  - . Degré (médiocre : passable : excellent)

- Autres relations sémantiques :
  - . Entité et processus associés (bibliothèque : prêt de livres)
  - . Entité et entités associées (voiture : garage)
  - . Entité et personnes associées (hôpital : patients)
  - . Processus et entités associés (informatisation : ordinateur)
  - . Processus et processus associés (indexation : prise de décision)
  - . Processus et personnes associées (indexation : indexeur)
  - . Personne et processus associés (utilisateur : recherche)
  - . Personne et entités associées (bibliothécaire : livre)

Les relations hiérarchiques (taxinomie, tout et partie...) peuvent se diviser en relations de type *"chaîne"* et de type *"arbre"*. Un terme isolé est une relation hiérarchique réduite à un seul nœud. Les relations d'association n'ont pas de structure bien définie, mais peuvent être considérées comme un arbre particulier où les termes associés sont "fils" du concept considéré.

Dans cette hypothèse, un document n'est plus seulement représenté par des termes d'indexation, mais par trois types de représentants des concepts contenus dans le document :

. concepts de type point (termes isolés)
. concepts de type chaîne
. concepts de type arbre qui sont utilisés à partir du concept principal contenu dans le document, qui sert de racine à cet arbre.

La fonction de comparaison entre les documents et la question est alors conçue suivant le calcul d'une fonction de coût : le nombre minimal d'opérations nécessaires pour transformer le graphe de la question dans le graphe représentant le document.

Les opérations élémentaires nécessaires pour cette opération d'édition sont au nombre de trois :

. Changement d'une étiquette
. Elimination d'un sous graphe
. Insertion d'un sous graphe

Ce type de calcul est issu des recherches concernant la reconnaissance de forme et permet de mesurer la distance entre deux arbres. La distance est alors définie par algorithme, et non par une fonction comme dans le cas de la distance euclidienne. Cette proposition de représenter différemment les descripteurs d'un document en tenant compte des types de relations sémantiques semble ouvrir de larges perspectives, notamment parce qu'une méthode de calcul de pertinence permet d'exploiter la structure ainsi définie.

3.b - Une approche topologique

L'approche topologique correspond à la difficulté de représenter la sémantique de la langue par des éléments discrets (termes d'indexation, relations d'un réseau sémantique). La compréhension des expressions linguistiques se place dans un continuum de signification.

1 - représenter la polysémie

La polysémie du langage dépasse de loin la simple homonymie. Les homonymes (les trois sens du mot *baie,* ou les deux acceptions de *pêche)* sont en réalité, du point de vue de leur valeur sémantique des termes totalement différents. Il est alors possible dans un système informatisé de les considérer comme des entités différentes, de significations disjointes. Au contraire, la polysémie est une propriété plus complexe du langage, car c'est elle qui permet les variations de sens autour d'un tronc commun, et qui autorise la finesse et la précision du langage.

Par exemple [VTC88b], les diverses acceptions du mot *enfant* s'expriment dans un continuum de significations basé sur les deux sens principaux :

. humain jeune,
. descendance directe de première génération.

Le sens du terme *enfant* dans une situation considérée est alors désigné par le contexte. Par exemple dans l'opposition à l'adulte, la tranche d'âge concernée sera différente que dans l'opposition à bébé. Les utilisations métaphoriques *(enfants de la patrie, enfant de la rue...),* même si elles ont des sens particuliers s'appuient néanmoins sur la notion de descendance, exprimée dans le terme enfant. Une représentation discrète, accordant un sens spécifique à chaque expression perd de vue la ressemblance paradigmatique entre toutes les utilisations du mot enfant. Pour outrepasser ce problème, on peut concevoir qu'une expression linguistique occupe une région entière d'un *"espace de signification"* continu. Bernard Victorri [VIC88b] estime que le formalisme de la

géométrie différentielle est un modèle mathématique efficace pour représenter cette conception, et que l'implémentation informatique doit faire appel à des réseaux connexionnistes, qui peuvent se stabiliser autour de l'interprétation d'une expression la plus "vraisemblable" en fonction des indices linguistiques disponibles.

Les applications de cette conception topologique de représentation des mots du langage dans le domaine documentaire ont été testées autour de l'algorithme QUID développé dans notre équipe de recherche à l'université de Caen ([VIC89], [THO89]).

2 - QUID : la notion de référentiel sémantique

Dans le modèle QUID, on associe un *"espace sémantique"* à une banque de données. Cet espace sémantique est une variété différentielle, obtenu par la réunion de variétés différentielles élémentaires, chacune représentant une option sémantique utile pour l'analyse d'un document. Les espaces sémantiques élémentaires peuvent être dotés d'une structure continue (par exemple pour représenter le temps) ou discrète. La structure de l'espace sémantique dépend du point de vue que l'on adopte sur la banque de données.

Supposons que l'on ait une banque de données d'objets d'art, et que notre intérêt se limite à la date de création et au genre artistique. L'espace sémantique sera alors bidimensionnel, s'appuyant sur deux espaces élémentaires : l'un portant la flèche du temps (considéré comme continu, ou comme discret si l'on se contente de mentionner le siècle de création) et l'autre les genres artistiques. La précision de la représentation sur ce dernier axe est laissée à l'appréciation du concepteur : on peut concevoir une division en grands genres (peinture, sculpture,...) ou prévoir des raffinements (peinture sur bas relief...).

La donnée de l'espace sémantique associé à la banque de données permet de décrire, en les plaçant dans cet espace topologique, aussi bien les documents que les termes d'indexation. Un document représente une région dans l'espace sémantique. On peut même concevoir que cette région ne soit pas connexe, afin de représenter les divers aspects d'un document, et même éventuellement que la

"coloration" que le document porte sur une région soit pondérée en fonction de l'importance de cette région sémantique dans le document. Les termes d'indexation sont eux aussi rapportés à une région de l'espace sémantique, ce qui permet d'associer à chaque terme d'indexation plusieurs significations élémentaires, éventuellement "colorées", dans l'espace sémantique. Les termes d'indexation et les documents se recouvrent donc partiellement dans l'espace sémantique. Les termes du *"vocabulaire de l'utilisateur",* quand ils sont connus du système, sont eux aussi représentés dans cet espace sémantique. Le vocabulaire de l'utilisateur peut donc comporter les termes d'indexation, mais aussi des termes supplémentaires qui représentent une région connue par le système dans l'espace sémantique (par exemple des synonymes, des termes ambigus, des expressions trop générales pour conduire à une indexation efficace...).

Le mécanisme de recherche est alors le suivant :

- on associe aux termes de l'utilisateur une région dans l'espace sémantique, en utilisant les termes de la question qui sont connus du système (i.e. dont on connaît la décomposition sur l'espace sémantique)

- on calcule pour chaque document le *"coefficient de satisfaction"* du document pour la question, en évaluant le recouvrement entre la région représentée par le document et celle correspondant aux termes de la question. On définit pour cela une mesure, qui peut par exemple prendre la forme d'un coefficient de Dice ou de Jaccard.

- on propose à l'utilisateur pour jugement de pertinence la liste des documents classés suivant ce coefficient de satisfaction

- on propose la liste des termes d'indexation qui sont en recouvrement partiel avec les termes de l'utilisateur, ou avec les documents extraits, ou encore avec un représentant barycentrique des termes de l'utilisateur et des documents. Ces termes servent à proposer des pistes pour la reformulation par l'utilisateur.

La réponse à une question est donc composée d'une liste de documents et d'une liste de pistes (éventuellement plusieurs listes de piste si les espaces

sémantiques élémentaires sont totalement disjoints, par exemple une liste par sujet et une liste conjointe par aires géographiques). Les deux listes sont classées en fonction de la proximité des documents ou des termes d'indexation avec les termes de la question. La proximité est mesurée par le coefficient de satisfaction et mesure le recouvrement des régions correspondantes dans l'espace sémantique.

Ce processus a été implémenté sur un ordinateur SUN par Loïc Thomazo ([THO89]) pour gérer deux maquettes, l'une concernant un fichier de banques de données destiné à proposer à l'utilisateur des banques de données commerciales correspondant aux thèmes d'intérêt de sa recherche, l'autre concernant un guide touristique permettant de sélectionner un hôtel ou un restaurant. Dans les deux cas, l'espace sémantique est réduit à un vecteur binaire, chaque case du vecteur correspondant à une propriété des unités documentaires considérées. Les propriétés élémentaires *(micro-caractéristiques),* pour cette maquette, sont choisies dans une liste, ce qui est une forme limitée d'espace sémantique. Les termes de la liste ont cependant été articulés pour permettre d'exprimer le flou de certaines appartenance. On trouve par exemple *"biologie / zoologie", "biologie / environnement", "biologie / maritime",* mais aussi *"chimie / effets toxiques", "agro-alimentaire / aquaculture"* ou *"écologie / maritime",* ce qui donne une idée de la possibilité de disperser les représentations de certaines banques de données concernant la mer et les espèces marines, mais aussi de la capacité à ouvrir des pistes en décrivant dans le même espace sémantique des termes comme *"zoologie marine", "aquaculture", "pollution des mers",* ou même des termes généraux comme *"biologie".*

La structure de liste reste cependant une moyen limité de représenter les significations. Dans la nouvelle version en préparation du logiciel QUID, l'indexation sera réalisée à partir d'espaces sémantiques représentés en tableau, ce qui permet de s'appuyer sur la présentation graphique à l'écran de l'espace sémantique (ou du moins d'un espace sémantique élémentaire), l'indexation consistant à "cocher" des régions de cet espace.

Ce type de présentation s'accorde mieux avec les idées de thésaurus "à facettes" qui constituent une fondation théorique pour cette conception analogique de l'indexation. Il existe en effet une grande parentée entre ce modèle et les conceptions de Ranganathan ([MAN87],[COA88]). Ranganathan pensait que tous

les sujets pouvaient s'exprimer au travers de l'articulation de "facettes", notamment les cinq facettes *"Personnalité",* représentant l'entité du sujet traité, *"Matière",* représentant le matériau ou la méthode, *"Energie"* représentant les actions et les processus en œuvre dans le sujet traité, *"Espace"* et *"Temps"*. Dans la nouvelle version de QUID, l'indexation des banques de données par exemple sera réalisée à l'intérieur d'une grille, dont les lignes représenteront les sujets traités et les colonnes le point de vue porté sur ce sujet par la banque de données (les "facettes").

Prenons un exemple. Selon la couverture des banques de données, le sujet "Chimie" peut être traité suivant les points de vue

. "Aspect scientifique, recherche" *(Chemical Abstracts),*
. "Aspect propriété industrielle" *(Chemical Abstracts, Dérivent),*
. "Aspect Environnement" *(Chemical Hazards in Industry),*
. "Aspect économique, marché, produits" *(Chemical Business Newsbase),*
. "Aspect médical et toxicologique" *(Medline, Chemical Abstracts, Toxline)...*

La question d'un utilisateur *"Banques de données traitant de la* chimie " aura dans ce cadre une réponse constituée de deux listes :

- une liste de banques de données, la première étant celle qui couvrira le plus grand nombre d'aspects (la plus généraliste, vraisemblablement *Chemical Abstracts* dans notre exemple)

- une liste de pistes, qui indiquera les principaux aspects qui recoupent le sujet *"Chimie",* la première piste étant celle qui est la plus proche de la question de l'utilisateur. Cette liste de pistes doit permettre de passer du vocabulaire libre de l'utilisateur vers l'utilisation du vocabulaire contrôlé de la banque de données. Le jugement de pertinence s'effectue alors au travers des pistes, qui relancent la question.

L'espace sémantique de la banque de données ne se limite pas à un seul tableau à double entrée. On peut juxtaposer plusieurs tableaux, éventuellement hiérarchisés (le sujet *"chimie",* présenté au premier niveau, peut ouvrir un

tableau spécialisé avec les sous domaines *"chimie organique", "chimie minérale", "chimie analytique", "pétrole",* ...). Certains aspects peuvent être représentés sur une droite au lieu d'un plan (par exemple le type de documents traité par la banque de données : références, texte intégral, encyclopédies, données numériques....).

A la fin de l'opération d'indexation, les régions qui auront été pointées dans l'ensemble des espaces sémantiques élémentaires constitueront la représentation du document. Un vecteur sera alors construit par le système par réunion des vecteurs issus des espaces sémantiques élémentaires. Chaque case des tableaux élémentaires correspondant à une place définie dans le vecteur résultat, l'espace sémantique peut alors être considéré comme un *"référentiel"*.

Un certain nombre de mots et d'expressions seront aussi décrits dans ce référentiel, afin de constituer le *"vocabulaire utilisateur",* c'est à dire l'ensemble des termes reconnus dans les questions. Il s'agit donc d'une forme particulière du modèle vectoriel, dans lequel les éléments de base ne sont pas les termes d'indexation, mais des *"micro-caractéristiques",* définies au préalable, dans un espace entièrement maîtrisé par le concepteur.

Cette maîtrise de l'espace des significations élémentaires permet de nombreuses améliorations par rapport au modèle vectoriel traditionnel :

- les micros-caractéristiques, qui définissent l'espace vectoriel sont alors bien orthogonales, avec des significations qui ne se recoupent pas. Cela permet de compenser les problèmes d'indépendance des termes qui apparaissent dans tous les modèles vectoriels fondés sur les termes d'indexation.

- les mots du langage, leur inventivité, leur évolution...sont rapportés à un espace clos, ce qui permet de leur donner un sens non pas généraliste comme dans la communication linguistique, mais spécifique du but poursuivi par le système documentaire. Le terme *Pollution* sera décomposé en significations élémentaires ("micro-caractéristiques") non pas d'un point de vue général, mais en fonction de ce que l'on peut imaginer être l'objectif (ou les objectifs) d'un utilisateur qui pose cette question à la banque de données spécifique qui est indexée. Ainsi, *"Pollution"* ne sera pas décomposé dans le même espace selon

qu'il sert pour l'interrogation d'une banque de donnée sur les rivières, ou pour un fichier qui décrit lui même les banques de données.

- le vocabulaire contrôlé spécifique de la banque de données aura un sens précis vis à vis des documents, mais un sens évolutif vis à vis des termes des utilisateurs. Les termes du vocabulaire contrôlé ne sont plus les seuls termes acceptés pour les questions, mais au contraire les termes proposés pour un choix par l'utilisateur.

Un système QUID est alors un ensemble composite comportant :

- des documents constitués par leur forme visualisable (texte ou image). Ce document est géré en dehors du système de recherche documentaire. A l'intérieur de ce système, il correspond à une adresse (une clé unique). En ce sens, un système QUID peut être interface avec tout type de système de stockage (vidéodisque, système d'archivage...)

- la description de chaque document dans l'espace sémantique défini par le concepteur de la banque de données. Cette description est rapportée à un vecteur défini dans le référentiel constitué par la réunion de tous les espaces sémantiques élémentaires (c'est une propriété des variétés différentielles de créer une nouvelle variété différentielle par l'opération de réunion). Ce vecteur peut être binaire (une micro-caractéristique est présente dans le document ou non) ou pondéré (expression du niveau auquel est traité l'élément sémantique dans le document).

- un certain nombre de mots interrogeables, qui peuvent décrire spécifiquement le document mais qui ne sont pas décomposables avec une précision jugée suffisante dans le référentiel. Dans l'exemple de la description des banques de données, un fichier spécialisé sur l'acupuncture sera décrit dans l'espace sémantique référentiel (i.e. retrouvé par la requête *Médecine*). Cet espace ayant été prévu pour toutes les banques de données, la spécificité du terme acupuncture risque de disparaître. Ce terme est donc traité à part dans un index, (i.e. le document sera retrouvé par *acupuncture,* mais les pistes seront proposées à partir de la description du fichier dans l'espace sémantique, ce qui permettra de proposer *médecine).*

- une liste de *termes d'indexation* qui correspondent à un vocabulaire contrôlé. Ces termes seront utilisés pour proposer des pistes de reformulation. Le calcul de proximité de ces termes avec les mots de la question, ou avec les documents jugés pertinents par l'utilisateur sera possible car ces termes d'indexation sont définis eux aussi au sein de l'espace sémantique constitué par le référentiel.

- une liste de termes du *vocabulaire utilisateur* qui sera constituée par la description d'un certain nombre de termes dans le référentiel. Ces termes ne font pas partie du vocabulaire contrôlé, et donc ne sont pas proposés comme piste. Ces termes, une fois qu'ils sont "connus" par le systèmes sont eux aussi décrits dans le référentiel. Ce vocabulaire est modifiable à volonté, car il n'influe pas sur les capacités de réponse du système. Il sert au contraire à assurer une meilleure couverture des recherches. Ainsi, un terme inconnu posé dans une question pourra être automatiquement rajouté au vocabulaire utilisateur, en le décrivant dans le référentiel, son poids affecté à chaque micro-caractéristique étant la moyenne de tous les poids affectés aux documents et aux pistes jugés pertinents lors de la recherche.

On calcule le coefficient de satisfaction à partir du recouvrement des termes et des documents dans le référentiel.

On définit dans le référentiel adopté :

. un document $D_i := < x_{ij} >$

. un terme d'indexation $T_i := < y_{ij} >$

. un terme du vocabulaire utilisateur $V_i := < z_{ij} >$ (les termes d'indexations T_i constituent un sous-ensemble du vocabulaire utilisateur).

Dans ces expressions, x_{ij}, y_{ij}, z_{ij}, représentent la valeur de la $j^{ème}$ micro-caractéristique respectivement dans le document, dans le terme d'indexation et dans le vocabulaire utilisateur. *j* varie de 1 à k, k étant le nombre total de micro-caractéristiques.

Une question d'utilisateur est obtenue en faisant la réunion de tous les termes connus du système présents dans la formulation de la question. Si la

question est formée des *n* mots M_i, la fonction de traduction qui permet de définir la requête acceptable Q s'exprime :

Q = réunion (M_i inter V_i) = <r_j>, r_j; représentant la valeur de la j^{ieme} micro-caractéristique.

Cette opération est rendue possible par l'existence d'un fichier inverse qui pour chaque terme utilisateur donne la valeur du vecteur correspondant dans le référentiel. On a alors : $r_j = \max(z_{ij})$ pour les termes connus du système.

Le coefficient de satisfaction d'un document D_i pour la question Q peut être calculé par un coefficient de Dice. Celui-ci correspond au rapport du nombre de micro-caractéristiques en commun sur le nombre total de micro-caractéristiques concernées par le document et par la question si l'on ne prend pas en compte les micro-caractéristiques de poids nul :

$$CS(D_i, Q) = 2 \frac{\sum_{j=1}^{k} r_j x_{ij}}{\sum_{j=1}^{k} r_j + \sum_{j=1}^{k} x_{ij}}$$

On calcule de même le coefficient de recouvrement des termes d'indexation par rapport à la question (équivalent à une "projection" sur le vocabulaire contrôlé) :

$$Pr(T_i, Q) = 2 \frac{\sum_{j=1}^{k} r_j y_{ij}}{\sum_{j=1}^{k} r_j + \sum_{j=1}^{k} y_{ij}}$$

La "réponse" à la question est alors formée de deux listes :

- les documents D_i classés dans l'ordre de $CS(D_i, Q)$, si ce coefficient dépasse un certain seuil, laissé au choix de l'utilisateur (i.e. définissant le taux de couverture de la recherche).
- les termes d'indexation T_i, classés dans l'ordre de $Pr(T_i Q)$, si ce coefficient dépasse un certain seuil, pour éviter de disperser l'utilisateur avec des termes éloignés de sa première requête.

La sélection de pistes dans cette liste provoque une modification de la question de l'utilisateur, qui peut devenir :

. soit l'intersection entre les termes d'indexation et la question originale : $r_j = \min(r_j, y_{ij})$

. soit le remplacement par la réunion des termes d'indexation choisis : $r_j = \max(y_{ij})$

. soit une combinaison linéaire entre la question originale et les termes choisis : $r_j = \min(1, r_j + a.y_{ij})$, a étant choisi dans [0,1].

Il faudra tester le système en grandeur réelle pour juger de la meilleure formule.

3 - Exemple

Les grilles ci-dessous représentent un exemple d'indexation en mode QUID. Il s'agit d'un exemple outrageusement simplifié, pour que peu de micro-caractéristiques entrent en jeu. Dès que le nombre de micro-caractéristiques devient important, il faut disposer de plusieurs grilles enchâssées. Ce référentiel simplifié est destiné à recevoir la description de banques de données. On voit alors que des banques de données qui pourraient sembler proches avec des termes d'indexation comme *Chemical Abstracts* et *Chemical Hazard in Industry* ne correspondent pas à des régions semblables de l'espace sémantique :

	Chimie	Biologie	Agriculture	Toxicologie	Médecine
Aspect scientifique	1	0,3	0,2	0,4	
Aspect propriété indus.	1		0,1		
Aspect économique					
Aspect médical	0,1				
Aspect juridique					
Aspect environnement	0,2		0,1		

CHEMICAL ABSTRACTS

	Chimie	Biologie	Agriculture	Toxicologie	Médecine
Aspect scientifique	0,3		0,1	0,7	
Aspect propriété indus.					
Aspect économique	0,5	0,3		0,5	
Aspect médical	0,8			0,7	
Aspect juridique	0,4				
Aspect environnement	1	0,8	0,3	1	

CHEMICAL HAZARDS IN INDUSTRY

Les termes d'indexation *produits chimiques et pollution* sont eux aussi décrits dans ce référentiel

	Chimie	Biologie	Agriculture	Toxicologie	Médecine
Aspect scientifique	1		0,8	1	
Aspect propriété indus.	1	0,3	0,3		0,3
Aspect economique	1	0,4	0,2	0,5	
Aspect medical	1	0,2		0,5	
Aspect juridique	1			0,5	
Aspect environnement	1	0,5	0,4	0,8	0,7

PRODUITS CHIMIQUES

	Chimie	Biologie	Agriculture	Toxicologie	Médecine
Aspect scientifique	1	1	0,5	1	0,5
Aspect propriété indus.	0,2		0,1		
Aspect economique	0,5	0,5	0,3	1	
Aspect medical	0,7	0,7	0,5	1	
Aspect juridique	1	1		1	
Aspect environnement	1	1	1	1	

POLLUTION

La question *"pollution des rivières par les produits chimiques"* aura dans ce cas plus de chance de privilégier *Chemical Hazard in Industry* que *Chemical Abstracts,* mais cette dernière banque de données ayant un coefficient de satisfaction non nul pourra être proposée à l'utilisateur.

4 - Vers des banques de compétences

Le processus décrit ci-dessus semble particulièrement bien adapté pour organiser des "banques de compétences". Une compétence est un élément difficile à définir par un ensemble de termes, qui peut se représenter de façon éclatée, avec des points forts qui ne sont pas obligatoirement connexes.

Par exemple, les *"compétences"* d'un employé peuvent se situer dans des domaines très divers : dans sa tâche quotidienne, mais aussi dans le sport, dans l'activité associative ou syndicale, dans les connaissances annexes qui sont les siennes : dactylographie ou informatique même si ce n'est pas son emploi principal...

Les *"compétences"* d'une banque de données représentent les sujets couverts, les aspects sous lesquels ils sont traités, le type de documents contenus, la régularité de la mise à jour, l'aire géographique couverte, éventuellement des éléments plus "subjectifs" comme la qualité de l'indexation, la disposition d'un thésaurus en ligne... Ces "compétences" ne peuvent pas s'exprimer en quelques mots-clés, mais plus encore, chaque mot-clé définit sans nuance le domaine couvert, alors qu'il peut être marginal dans la banque de données.

Les *"compétences"* d'une petite annonce de vente d'une automobile ou d'un logement se jouent aussi sur de nombreux domaines. Par exemple, une *super 5 ,* une *AX* et une *205* ont en commun la compétence d'être des "petites voitures". Il y a dans ce domaine plus de relations entre ces voitures qu'entre une *205* et une *205 GTI Turbo.* La description des annonces doit alors permettre de proposer une 205 à la fin de la liste des *super 5* et *R5,* même si la question était *"Je cherche une R5 ayant moins de 100 000 kilomètres".*

Nous sommes actuellement en train de ré-écrire le logiciel QUID en fonction des réflexions présentées ci-dessus. La première application de ce modèle sera appliquée à la description en grandeur réelle des banques de données diffusées par tous les serveurs commerciaux, dans le cadre de la réalisation d'un anté-serveur intelligent. Pour résoudre un besoin documentaire, il faut sélectionner une banque de données. Il est indispensable qu'un anté-serveur puisse proposer plusieurs options à un utilisateur. La phase de sélection de la (ou des) banques de données adéquates sera réalisée avant la formulation de la question par QUID.

II - Les anté - serveurs

Les anté-serveurs et leur place dans le processus de diffusion des banques de données ont été présentés dans la première partie. Les recherches concernant les futurs modes d'accès à l'information ont ensuite été développées. L'objet de ce chapitre est de concevoir les fonctionnalités des anté-serveurs qui peuvent être développés aujourd'hui.

Dans ce cadre, l'anté-serveur est conçu comme le lien tissé entre les systèmes booléens traditionnels fonctionnant avec des langages de commande et les systèmes documentaires intelligents, gérant différemment les données (systèmes connexionnistes, hypertextes, utilisant les techniques d'agrégation de documents...) et permettant des interactions différentes avec l'utilisateur. Cette présentation part de l'expérience acquise dans la conception et l'expérimentation d'un anté-serveur intelligent développé par la société *Triel,* en liaison avec l'université de Caen. Il s'agit d'une synthèse des idées portées au sein de l'équipe de recherche en sciences de l'information de l'université de Caen.

Avant de décrire les fonctionnalités d'un anté-serveur, il importe de repartir des besoins des utilisateurs. Il existe une importante marge de manœuvre qui mène des systèmes booléens à langage de commande vers les systèmes entièrement automatiques. Les anté-serveurs devront remplir ce gouffre pas à pas. De nombreuses améliorations peuvent être apportées aux systèmes booléens et aux serveurs professionnels, qui devraient servir aussi bien l'utilisateur que l'intermédiaire en information.

1 - Typologie des besoins de l'utilisateur

Dans un article récent [BAT90], Marcia Bates s'interroge sur la part des actes de recherche documentaire qui peut être prise en charge par le système lui-même. Son travail permet de définir deux types d'approche du problème :

- le niveau de compétence du système.

Cinq niveaux sont avancés :

. *niveau 0 :* toutes les activités de recherche sont dans les mains de l'utilisateur

. *niveau 1* : le système indique les diverses activités possibles, même celles qui ne peuvent être exécutées directement (ce niveau correspond aux systèmes d'aide en ligne : fonctions "help" ou "guide")

. *niveau 2 :* le système exécute certains enchaînements d'activités suivant les commandes reçues par l'utilisateur.

. *niveau 3 :* le système contrôle le processus de recherche et suggère des activités de recherche, soit quand l'utilisateur demande des suggestions, soit quand le système identifie un besoin. A ce niveau apparaissent les premières applications utilisant les méthodes de l'intelligence artificielle.

. *niveau 4 :* Le système exécute automatiquement les diverses activités requises par la recherche, et informe l'utilisateur, ou lui fournit directement les résultats.

- une typologie des activités mises en œuvre dans la recherche.

Quatre types d'activité sont repérés :

. l'action *(move)* représente l'acte élémentaire (sélectionner un terme de recherche, utiliser un opérateur booléen...). Les autres activités peuvent se traduire par des enchaînements d'actions.

. la tactique *(tactic)* est un ensemble d'actions qui repose sur le besoin de réagir face au système, sans considérations stratégiques. Les tactiques s'appliquent dans plusieurs cadres :

. *tactiques de formulation :* rendre la recherche plus spécifique, plus générale, réduire le nombre de termes de recherche...

. *tactiques portant sur les termes :* généraliser ou spécifier les termes, notamment en utilisant un thésaurus, utiliser plusieurs écritures, essayer des variantes...

. *tactiques portant sur l'interaction :* reprendre l'historique d'une recherche, élargir ou rétrécir l'objectif de la recherche...

. les stratagèmes *(stratagem)* désignent les opérations qui partent de points désignés comme productifs et utilisent cette connaissance pour élargir la recherche (les méthodes à jugement de pertinence par exemple). Quelques exemples de stratagèmes sont le survol du sommaire d'un périodique repéré par le grand nombre de bonnes réponses publiées dans ce journal ; le suivi du réseau des citations ; le suivi des publications d'un auteur significatif dans le domaine choisi...

. les stratégies *(strategy)* désignent les plans établis et corrigés pour conduire l'ensemble de la recherche.

Cette approche permet de construire un tableau à double entrée, et de pointer les apports à chacun des types d'activités par des systèmes de différents niveaux. Cela conduit à une situation évolutive.

Par exemple, un stratagème comme le survol de sommaires de certains journaux productifs dans un domaine est un stratagème largement répandu dans les systèmes manuels (i.e. par les usagers des bibliothèques). Il peut être mis en œuvre, avec de grandes difficultés, par un utilisateur bien formé dans les systèmes commerciaux actuels, mais on appréciera sa réalisation par des systèmes automatiques. L'emploi de ce stratagème peut être suggéré, même si la réalisation de l'équation de recherche correspondante est laissée aux soins de l'utilisateur (niveau 1), réalisée automatiquement sur demande (niveau 2), proposée si les résultats de la recherche y invitent (niveau 3) ou intégrée directement aux méthodes de recherches comme moyen de classer les documents selon l'ordre de pertinence (niveau 4). Ce stratagème peut facilement être proposé (si par analyse automatique des titres le système souligne la pertinence accentuée d'un titre) ou réalisé sur demande (l'analyse statistique n'est menée qu'en fonction de la demande, et la recherche puis l'organisation des sommaires sont prises en charge par le système). Il est beaucoup plus difficile de l'intégrer dans l'ensemble des calculs probabilistes qui permettent de classer les documents.

Cette nature évolutive des aides apportées à l'utilisateur doit guider la conception des anté-serveurs. L'anté-serveur est alors vu comme un instrument pouvant remplir des tâches diverses dans le cours d'une recherche, et pas

seulement comme un instrument tout entier subordonné à un but et un concept stratégique unique. Un anté-serveur peut grandement améliorer l'interaction avec le système documentaire, tout en restant sous le contrôle de l'utilisateur, qui peut exprimer le besoin (ou suivre une suggestion) d'une activité qui s'éloigne un peu de la stricte orthodoxie stratégique qui mène de la question à la liste classée de documents. Ces remarques pourraient se généraliser à d'autres activités utiles pour la recherche documentaire (généralisation de termes, élargissement ou spécialisation de la recherche, variantes orthographiques...) qui peuvent être incorporées aux services que peut rendre un anté-serveur sans devoir être immédiatement et totalement englouties par le système, qui les appliquerait sans en référer à l'utilisateur.

2 - Les fonctionnalités d'un anté-serveur

Pour les besoins de l'exposé, on doit considérer l'anté-serveur comme une boîte noire, placée entre le système documentaire et le terminal de l'utilisateur. Les problèmes informatiques soulevés par l'implémentation des diverses fonctionnalités ne seront pas abordés dans ce travail, qui se concentre sur la définition des tâches que devrait remplir un anté-serveur aujourd'hui. C'est le travail du concepteur informatique du système de se baser sur une architecture modulaire permettant d'incorporer ces tâches une par une dans l'ensemble des fonctions de l'anté-serveur. La programmation objet et le développement de modules échangeant des informations par messages semblent être une voie particulièrement efficace.

2.a. - échanges avec le terminal de l'utilisateur

Les fonctionnalités d'un anté-serveur ne peuvent pas être soumises à tel ou tel type de terminal. L'anté-serveur se concentre sur les problèmes d'ergonomie cognitive dans le domaine documentaire (organisation de l'interaction avec l'utilisateur, traitement de la question, classement des réponses....). Les activités de ces anté-serveurs peuvent être transmises à divers types de terminaux, depuis des terminaux minitel jusqu'à des stations de travail graphiques. Il en résultera une plus ou moins grande facilité de travail pour l'utilisateur (lecture des

informations, place disponible sur l'écran, disposition simultanée ou cachée des listes d'actions possibles...) mais cela correspondra aux mêmes fonctionnalités pour l'anté-serveur. Typiquement, si l'anté-serveur est capable de proposer une liste de pistes à un utilisateur pour relancer (préciser ou élargir) sa recherche, elle pourra être cochée par une souris dans une fenêtre spécifique sur station de travail *SUN* ou un *Macintosh,* alors que le possesseur d'un minitel devra appeler la liste par la touche de fonction "GUIDE" et indiquer les numéros sélectionnés avec la touche "ENVOI". Cependant, du point de vue des fonctionnalités documentaires de l'anté-serveur, les résultats seront similaires. Il en va de même des actions permettant le jugement de pertinence, de la présentation de "boutons" pour lancer diverses actions, ou de la nécessité de choisir dans une liste de commandes....

D'un point de vue cognitif, l'écran de l'utilisateur doit posséder cinq *"zones"*. Ces zones seront diverses fenêtres sur une station de travail, ou plusieurs parties de l'écran d'un terminal TTY. Chaque partie correspond à un *type* d'activité, qui peut être mis en avant (choisi par l'utilisateur, ou proposé par le système). Cette division en zones permet d'éviter les effets de perte de repères liés à la succession d'écrans. Quand une zone est privilégiée, les autres zones restes actives et accessibles, soit en changeant de fenêtre de travail, soit par basculement d'une zone à l'autre sur un écran plus petit.

Les zones utiles nous apparaissent être les suivantes :

. La zone d'échange.

Cette zone est utilisée tantôt pour formuler les questions, tantôt pour lire les documents qui sont présentés à l'utilisateur pour qu'il en juge la pertinence. C'est la zone principale de l'écran pendant la recherche documentaire.

La saisie des questions se fait sous deux formes : une saisie en formulation libre sur champs typés (sujet, auteur, titre, date... ou d'autres champs spécifiques adaptés à la banque de données considérée) et une saisie complémentaire suivant la demande de l'anté-serveur (langues connues, formulation en mode texte de la question, type de documents souhaités...). La saisie complémentaire a deux fonctions : obtenir de nouvelles formulations de la même question pour permettre

des calculs plus puissants et permettre d'occuper l'utilisateur pendant la première phase d'interrogation des serveurs (connexion, transfert de requêtes, combinaisons booléennes, calculs de pertinence formelle).

Cette même zone est utilisée pour présenter les documents extraits, dans un format court, adapté à un jugement par l'utilisateur. Comme il a été souligné tout au long de cette thèse, le jugement de l'utilisateur est un moteur essentiel de l'interaction avec le système. Même sans ré utiliser pleinement les informations issues de ce jugement, cette méthode permet de ne conserver dans la "réserve" de l'utilisateur que les documents qui l'intéressent.

. la zone de réserve.

Cette zone permet à l'utilisateur de constituer son "panier" de documents sélectionnés. En format réduit, par exemple rapportée à un coin d'écran, elle est juste constituée d'un "compteur", qui indique le nombre de documents sélectionnés par l'utilisateur. Il s'agit alors de documents pertinents, c'est-à-dire placés dans la réserve après jugement par l'utilisateur.

En format agrandi (i.e. prenant la majeure partie de l'écran sur un terminal, ou devenant la fenêtre active sur un autre système), la zone de réserve permet deux types d'opérations :

. *la lecture des documents sélectionnés.* Les documents apparaissent alors en entier. Dans le cas d'un texte intégral de document, les phrases comportant les termes des questions de l'utilisateur peuvent être présentées en surbrillance, éventuellement sélectionnées pour un survol plus rapide.

. *les opérations de post-traitement sur les documents sélectionnés.* Si l'on admet, comme on l'a souligné dans la première partie de cette thèse, que la recherche documentaire n'est qu'un élément d'une opération de prise de décision, on doit considérer le besoin de l'utilisateur de réaliser divers traitements sur les documents sélectionnés. Ces traitements doivent être choisis par l'utilisateur, éventuellement sur liste de propositions du système. Les traitements peuvent évoluer, et les capacités de l'anté-serveur à offrir des services

à valeur ajoutée sur les lots de documents sélectionnés pourront augmenter graduellement.

On peut repérer les travaux utiles suivants :

- impression et envoi des documents,
- tri des documents suivant un critère choisi par l'utilisateur,
- envoi des documents sur un ou plusieurs réseaux de messagerie, éventuellement avec un en-tête écrit par l'utilisateur,
- impression d'étiquettes pour publipostage, triées suivant les codes postaux,
- transformation des documents dans un format spécifique pour une récupération par l'utilisateur sur son système local. Cette opération est fondamentale si l'anté-serveur doit permettre la récupération de notices par les bibliothèques (réseau bibliographique),
- impression des informations selon des formats spécifiques (tableaux, tris, impression avec des balises typographiques....)

. la zone de l'historique.

Cette zone permet d'indiquer à l'utilisateur les opérations que réalise, ou qu'a déjà réalisées, l'anté-serveur. En format réduit, cette zone fait défiler les termes des questions et les occurrences telles qu'elles sont obtenues par interrogation des serveurs. Certaines opérations ne seront cependant pas montrées à l'utilisateur, notamment les diverses conjectures établies à partir des éléments statistiques, qui peuvent parfois être sémantiquement incompréhensibles. Il est important que le travail réalisé par l'anté-serveur soit présenté à l'utilisateur, notamment pendant la première phase, qui précède la lecture des premiers documents. Cette phase correspond à une période d'attente inactive. Elle peut apparaître très longue à l'utilisateur. Pour autant, cette

présentation de l'activité de l'anté-serveur n'occupe pas forcément l'ensemble de l'écran (à moins que l'utilisateur ne le souhaite).

En format agrandi, cette zone sert à montrer l'historique de la recherche. Si l'anté-serveur, comme cela semble nécessaire, interroge plusieurs banques de données en même temps, l'historique est répété pour chaque fichier. Avec le développement des capacités de l'anté-serveur, on peut imaginer une présentation graphique de cet historique, notamment en utilisant des diagrammes de Venn, afin de préciser pour l'utilisateur l'interprétation des connecteurs booléens.

C'est aussi au niveau de la zone de l'historique que l'utilisateur peut choisir diverses tactiques, ou stratégies : suppression ou ajout de termes de recherche, élargissement de tel concept, reprise à partir de telle question...

. la zone des pistes.

Cette zone est destinée à recevoir les informations calculées par le système permettant la reformulation des questions. Il existe en effet plusieurs utilisations possibles du jugement de pertinence de l'utilisateur. En particulier, celui-ci peut permettre d'obtenir une liste classée des termes d'indexation (ou des mots du titre) les plus fréquemment utilisés dans les documents pertinents, afin de sélectionner des pistes de reformulation (un équivalent plus sophistiqué des fonctions ZOOM, MEMTRI ou GET des serveurs). On peut agir de même, en fonction des banques de données considérées avec les codes de classement, ou les codes spécifiques (codification des entreprises, des concepts biologiques....). Enfin, si la banque de données dispose d'un thésaurus en ligne, les propositions qui peuvent être extraites de ce thésaurus seront présentées à l'utilisateur dans cette zone.

On pourra aussi proposer à l'utilisateur certaines opérations de butinage qui sont possibles même dans le cadre booléen : sommaire de périodiques, réseau d'auteurs et de co-auteurs...

En format réduit, la zone des pistes peut simplement indiquer par un signe que des possibilités de reformulation ont été repérées par le système, ce qui n'est pas vrai au tout début de la recherche.

. la zone des commandes.

Plus traditionnelle, cette zone rappelle les commandes disponibles à tel ou tel moment de la recherche. Elle permet notamment de sélectionner la zone active dans les systèmes qui ne disposent pas d'une souris.

2.b - échanges avec les serveurs

L'anté-serveur agit comme un unificateur des diverses sources d'information électronique. Certaines banques de données ne sont distribuées que par certains serveurs, voire sous le seul format vidéotex. Pourtant, l'utilisateur doit retrouver toutes les informations sous une forme comparable. L'anté-serveur doit donc être un transcodeur de protocoles : transcodage des caractères et des commandes.

Les anté-serveurs actuels *(Easynet, Geomail...)* sont capables d'interroger plusieurs banques de données en même temps pour un même utilisateur. Cette opération doit être étendue à des banques de données sous plusieurs formats (vidéotex, serveurs sous divers langages de commandes...).

Une opération de niveau 1 dans la classification de Marcia Bates consiste à réaliser la traduction des commandes d'un jeu à l'autre, pour permettre à la personne formée sur un serveur d'interroger, tout en restant seul maître de l'interaction, les divers autres serveurs. En ce sens, les commandes des serveurs ont un représentant interne, qui peut s'appuyer sur la norme minimale *"Common Command Language"* ([LEC89], [WILC88]), et des tables de transcodage. Cette opération ne peut simplement porter sur les "verbes" des commandes, mais doit aussi s'appliquer à la syntaxe des commandes, notamment de visualisation, et aux qualifications en préfixe ou en suffixe.

Mais ces opérations ne prennent vraiment leur sens que si le système est capable de connaître les diverses banques de données avec lesquelles il doit échanger. Les noms de champs, les méthodes d'écriture des données dans les champs,.... varient d'une banque de données à l'autre. Un anté-serveur doit posséder des connaissances de ce type. C'est un avantage appréciable, même si les

banques de données sont disponibles sur le même serveur. Par exemple, les systèmes de recherche sur des groupements de banques de données *(Questcluster, Dialog OneSearch...)* posent la même question sur tous les fichiers sélectionnés, ce qui conduit parfois à des résultats incompréhensibles. Cette connaissance des méthodes propres à chaque banque de données est d'autant plus importante que les recherches sur les codes et les diverses informations normalisées (langage contrôlé, numéros de registre, indexation numérique...) sont souvent les plus performantes. Par exemple, une recherche de brevets qui n'utiliserait pas la Classification Internationale des Brevets aurait peu de chances d'obtenir un taux de couverture intéressant. Or chaque banque de données écrit différemment le même code C.I.B. !

Les connaissances sur les banques de données sont utilisées pour transformer (filtrer) les informations qui sont élaborées dans un langage pivot propre à l'anté-serveur. On peut ainsi distinguer plusieurs types de champs génériques, qui seront ensuite traduits dans les divers champs des banques de données. Un champ générique pouvant d'ailleurs correspondre à plusieurs champs d'une banque de données particulière.

Une liste des champs génériques pourrait s'appuyer sur l'énumération suivante :

. *titre.* L'équivalent dans les banques de données annuaires ou répertoires est le nom des sociétés concernées par la recherche.

. *auteur.* Pour les banques de données de brevets, les champs correspondants sont, dans l'ordre, le champ inventeur, puis le champ déposant.

. *sujet.* On retrouve l'équivalent du lexique de base *(basic index)* des serveurs commerciaux. Le champ sujet se hiérarchise en fonction de la précision, depuis le champ correspondant au langage contrôlé jusqu'aux mots du résumé. Il suffira de connaître les noms dans l'ordre hiérarchique de précision pour s'adapter à chaque fichier.

. *date.* La précision dans la date dépend du type de banque de données (le jour pour les dépêches d'agences de presse, l'année pour les bibliographies).

. *langue.*

. *type de document.*

. *code de classement.* Même si les codes de classement ne sont pas, en général, posés dans la question de l'utilisateur, notamment parce qu'il n'existe pas d'instruments de navigation permettant de les utiliser pour formuler la question, les informations codifiées peuvent être recueillies et traitées au niveau du jugement de pertinence. On peut traiter de manière équivalente les codes de produits, les codes de concept, les codes de secteur industriels...

. *source bibliographique.* Connaître l'organisation de ce champ dans les diverses banques de données permet de récupérer l'information nécessaire pour proposer la lecture des sommaires des revues les plus importantes.

Cette liste ne permet pas de tirer partie de toutes les richesses d'indexation de certaines banques de données. En revanche, elle est suffisamment souple pour s'adapter à un grand nombre de cas. Chaque type de champ pose des problèmes particuliers, correspondant à des stratégies de recherche différentes. Elaborer les stratégies pour ces champs principaux permet d'acquérir un savoir-faire qui pourra plus aisément s'intégrer sur d'autres types de champs.

3 - Les modules de l'anté - serveur

Les modules de l'anté-serveurs doivent remplir les diverses opérations permettant de transformer la question d'un utilisateur en une requête acceptable par les serveurs professionnels. L'architecture informatique de l'anté-serveur doit permettre d'intercaler des modules évolutifs, afin de mettre en place pas à pas les méthodes documentaires évoluées : reformulation, jugement de pertinence, classement des documents en fonction de leur pertinence formelle...

En effet, une amélioration considérable pourrait être apportée aux systèmes commerciaux si les documents étaient classés avant d'être présentés à l'utilisateur ([COOP88], RADE88]). D'une part le jugement de pertinence pourrait

s'appliquer sur les premiers documents de la liste des réponses, d'autre part, les effets de fatigue seraient bien moins sensibles. En suivant la liste des documents, l'utilisateur fixerait lui-même son taux de couverture, en s'arrêtant lorsqu'il s'estimerait satisfait.

Il est impossible de traiter ici des problèmes rencontrés par chacun des modules. De nombreux éléments pourraient être tirés des expériences articulant plusieurs systèmes experts (cf. le chapitre sur "l'expertise"). Nous nous centrerons sur trois points :

. la traduction de la question de l'usager
. les méthodes permettant de classer les documents
. les méthodes permettant de proposer des listes de pistes pour la reformulation.

3.a - les questions de l'utilisateur

Dans un premier temps, l'utilisateur écrit sa question à l'intérieur d'un masque de saisie. Comme cela a souvent été souligné, on ne peut prendre pour une formulation exacte du besoin documentaire la question telle qu'elle est écrite la première fois. Il est donc intéressant, pendant que le système traite cette première formulation de demander des précisions sur le besoin documentaire, en proposant une saisie en texte libre. Les redondances avec les questions de la première grille de saisie seront des indices d'importance, et les différences seront signes de précisions.

Les utilisateurs doivent pouvoir poser leurs questions en formulation libre. Cependant, la formulation sur un champ typé garde toujours un aspect contraint, qui incite l'utilisateur à se limiter à quelques termes. Les compétences linguistiques de l'anté-serveur peuvent évoluer et permettre une meilleure interprétation des questions d'utilisateur. Un niveau élémentaire peut consister à éliminer les mots outils, suivant un dictionnaire, et à repérer les divers termes de la question, afin d'établir l'équation de recherche à partir de calculs probabilistes.

Gérard Salton propose une série d'heuristiques allant dans ce sens pour remplacer les divers connecteurs en fonction des résultats des recherches

([SAL82], [SAL88a]). Ces méthodes pourraient être appliquées par un expert particulier, équivalent du *"constructeur du modèle de la requête"* du système I^3R. La méthode proposée par Gérard Salton est entièrement basée sur une démarche statistique, sans analyse linguistique de la requête.

La démarche est la suivante :

1 - Proposer à l'utilisateur de formuler librement sa requête et d'indiquer le nombre *m* de documents qu'il souhaite retrouver.

2 - Décomposer la requête en unitermes, après élimination des mots-outils.

3 - Evaluer pour chaque terme T_i le nombre n_i de documents correspondants dans la banque de données, et le poids associé. Le poids est défini en fonction de la fréquence par

$$w_i = 1 - \frac{n_i}{N}$$

où N est le nombre total de documents de la banque de données.

4 - Evaluer la fréquence probable et le poids correspondant de toutes les conjonctions de paires de termes (A et B). La fréquence n_{ij} de la paire T_i *et* T_j est approximativement

$$n_{ij} = \frac{n_i \cdot n_j}{N}$$

et le poids correspondant $w_{ij} = 1 - \frac{n_{ij}}{N}$

On peut aussi se limiter aux paires ayant une corrélation significative.

5 - Evaluer de même les fréquences probables et les poids correspondants des conjonctions de triplets (T_i *et* T_j *et* T_k). On évalue :

$$n_{ijk} = \frac{n_i \cdot n_j \cdot n_k}{N^2} \qquad w_{ijk} = 1 - \frac{n_{ijk}}{N}$$

On peut limiter cette opération aux triplets fortement corrélés

6 - Evaluer le nombre de documents *d* qui seraient retrouvés par une disjonction de tous les termes (T_1 *ou* T_2 *ou... ou* T_n). On calcule $d = \Sigma n_i$.

7 - Si $d>m$, remplacer les termes de poids les plus faibles par la paire formée de la conjonction de termes qui ne figurent pas dans les unitermes restant.

On a $d = \Sigma n_i + \Sigma n_{jk}$.

On procède de même tant que $d>m$ ou jusqu'à ce que la requête soit entièrement composée de paires ou avec des termes uniques dont la fréquence est inférieure à la fréquence minimale des paires qui ne comprennent pas ce terme.

8 - Si la question devient entièrement composée de paires, répéter la même opération avec des triplets de termes qui ne sont pas compris dans les paires restantes après élimination des paires de plus faible poids.

On a alors $d = \Sigma n_i + \Sigma n_{jk} + \Sigma n_{lmp}$. On arrête le processus quand $d<m$.

Cette démarche est entièrement basée sur les occurrences des termes dans la banque de données. Elle ne prend pas en compte le sens des mots. Testée sur une banque de données de 1033 documents, elle semble donner meilleure satisfaction qu'une recherche formulée manuellement [SAL88a]. Il serait intéressant d'appliquer ce processus sur une banque de données de taille réelle. Comme nous le soulignerons plus loin, ce devrait être un des rôles des anté-serveurs de permettre ce type d'expérience.

Cette méthode n'est pas exclusive d'une autre approche plus sémantique ou syntaxique, qui remplacerait l'élément uniterme du processus par une expression déterminée par des procédés linguistiques. Cela semble d'ailleurs important dès que l'on doit traiter des requêtes ayant un grand nombre de termes. Par effet combinatoire, le nombre de paires et de triplets devient alors très important. Il importe donc que ce type de démarche soit intégré dans un processus contrôlé. Comme dans les systèmes de jeux informatiques, une telle méthode de calcul probabiliste devient un appui dans une prise de décision, même s'il semble aléatoire de s'en remettre uniquement à la "force brute" du calcul.

Nous menons actuellement, avec le concours de linguistes, une étude sur les formulations des utilisateurs devant un système documentaire. Pour cela, nous avons écrit un "simulateur", qui permet de poser la question sur champs

typés, puis sur un champ de texte. L'étudiante linguiste associée est ensuite présente quand l'utilisateur pose sa question au bibliothécaire, afin d'étudier les différences de formulation, mais aussi d'essayer de caractériser les phrases et les modèles de requête.

Réfléchir sur mon propre travail m'a permis d'éclairer une partie des méthodes de recherche que je mettais en œuvre depuis longtemps sans avoir pu les synthétiser. Dans la pratique, une question d'utilisateur est décomposée par l'intermédiaire suivant trois types de descripteurs :

. les *"objets"* qui représentent les formulations les plus spécifiques : nom de produit, d'espèces vivantes,...

. les *"méthodes"* qui indiquent les processus permettant de transformer les "objets" ou l'environnement qui permet de relier les "objets" dans une même question.

. les *"concepts"* qui sont les représentants du point de vue suivant lequel est posée la question.

On s'aperçoit ensuite qu'une question doit être formulée par intersection des *"objets"*. Le nombre de réponses obtenu peut à lui seul être plus faible que le nombre souhaité par l'utilisateur. On ne précise par les *"méthodes"* que si cela s'avère nécessaire. Enfin, les *"concepts"* ne sont utiles que pour tailler à la hache dans un ensemble de résultats trop important. Les "concepts" sont en général mal définis par des termes précis, et de plus ne sont pas souvent décrits par l'indexeur. Ils se rapprochent plus des notions pragmatiques.

3.b - Les méthodes pour classer les documents

Même imparfait, un classement par ordre approché de pertinence est un moyen de satisfaire au mieux l'utilisateur, même avec des banques de données correspondant au modèle booléen. Par exemple, un document qui possède tous les termes d'une disjonction est certainement plus important que celui qui n'en possède qu'un seul. De même, les documents qui contiennent les termes de plus faible fréquence peuvent être les meilleurs candidats à la satisfaction de l'utilisateur.

Bruce Croft et Pasquale Savino [CR088] proposent la démarche suivante pour classer les documents, en admettant que le nombre de documents qui sera balayé par l'utilisateur est m, compris entre 20 et 50 (éventuellement fixé par l'utilisateur) :

1 - Calculer le poids de chaque terme T_j de la requête, en appliquant l'inverse de sa fréquence. On peut par exemple utiliser, comme le suggère [ROB76], la formule :

$$w_j = \log \frac{N}{n}$$

où N représente le nombre total de documents contenus dans la banque de données et n_j le nombre de documents contenant le terme T_j.

2 - Utiliser cette liste suivant l'ordre décroissant des poids pour connaître les documents correspondants. Pour le premier terme T_1, donner le poids w_1 à tous les documents dont les adresses sont contenues dans le fichier inverse à l'entrée T_1. Placer la référence et son poids dans une table de sortie.

3 - Ajouter le poids du terme examiné T_j au score obtenu par les documents déjà présents dans la table de sortie.

4 - Quand le nombre de documents présents dans la table est supérieur au nombre *m* souhaité, le poids du document placé en fin de liste est le pire qui peut être obtenu. Les documents suivants ne sont donc plus ajoutés dans la table si le poids total des termes restant est plus faible que ce poids minimum.

5 - Quand tous les termes ont été traités, les documents restant dans la table sont triés par ordre décroissant.

On trouve plusieurs suggestions de traitements de ce type, qui s'appuient sur les faibles indications dont dispose le système (i.e. la simple mention de la fréquence des termes, ou plus exactement du nombre de documents contenant un terme donné). On regardera par exemple [MOR82] et [RADE88] pour des variations sur ce type de calculs.

3.c - proposer des pistes pour la reformulation

Cet axe de travail est un moyen efficace de laisser l'utilisateur modifier lui-même sa question pour l'adapter à ses besoins documentaires. La capacité à sélectionner un terme dans une liste est bien plus grande que la capacité à imaginer les termes représentant un besoin documentaire.

Les serveurs proposent déjà certains moyens de lister des termes. On peut ainsi utiliser les fonctions de tri des serveurs professionnels (ZOOM, GET, MEMTRI,...). Ces fonctions agissent cependant sur les documents extraits par la recherche, mais pas sur les documents pertinents. Les listes obtenues, comme les listes de noms d'auteurs prises dans le lexique des auteurs (afin de repérer les divers prénoms, et éventuellement les diverses écritures du nom d'un même auteur) ou les listes de mots-clés obtenus par les fonctions de tri constituent cependant une bonne approche. L'anté-serveur doit cependant être capable de repérer les éléments d'un document électronique (référence, texte,...) et d'opérer lui-même des tris statistiques. Cela permet d'effectuer des calculs même si les serveurs ne proposent pas de fonctions statistiques *(Dialog, serveurs vidéotex...}* et surtout de réaliser ces calculs uniquement sur les documents pertinents, afin d'obtenir des termes permettant de reformuler globalement la question.

Les opérations de tri les plus importantes sont dans l'ordre :

- tri sur les mots-clés et les mots du titre
- tri sur les codes de classement. Parce que l'utilisateur ne connaît pas les méthodes d'indexation des banques de données, il est d'autant plus important d'obtenir des informations à partir des documents sélectionnés.

Les plans de classement ou les instruments permettant de sélectionner les codes ne sont pas en général disponibles en ligne (exception : les codes SIC -*Standard Industry Classification* - et C.I.B. -*Classification Internationale des Brevets* - qui correspondent à des banques de données particulières). La reformulation employant ces méthodes doit donc passer par une analyse d'un premier lot de documents. On appréciera dans ce cadre la définition associée en clair des codes, comme le pratique par exemple *BIOSIS.*

- tri sur les titres de périodiques, notamment si le survol de sommaires est mis en place.

4 - Quelques remarques sur les anté - serveurs

Il semble utile de clore ce chapitre sur les anté-serveurs par deux types de remarques, portant sur les conditions de travail lors de la recherche documentaire, et sur les possibilité d'utiliser les anté-serveurs comme instruments pour tester des algorithmes et méthodes proposés en science de l'information.

4.a - les conditions du travail de recherche

A l'heure actuelle, la recherche documentaire est trop souvent organisée comme une course contre la montre. Le mode de facturation des serveurs professionnels est largement la cause de cette pratique. Organiser et diffuser l'information coûte très cher. Il ne s'agit pas de remettre ceci en cause, ou du moins pas dans le fond, même si on doit s'interroger sur les aides qui peuvent être apportées à certaines catégories d'utilisateurs afin de favoriser la démocratie d'accès aux ressources informationnelles, et dans ce cadre sur le besoin d'élargir les diverses aides de l'Etat (réductions dans les bibliothèques universitaires) ou des organismes spécialisés (ANVAR, Chambres de Commerce...). Plus précisément, il s'agit de concevoir un mode d'accès qui permette à l'utilisateur de se sentir libre de mener sa recherche, d'étudier en liaison avec le système les documents sélectionnés, et de payer au prorata de sa satisfaction.

C'est ce raisonnement qui a poussé le serveur de *l'Agence Spatiale Européenne* à mettre en place une facturation qui élimine le facteur "temps passé au terminal", en s'appuyant sur le couplage d'un forfait et d'un prix plus élevé porté sur les documents conservés. Cet exemple est en passe d'être suivi par le serveur *DIMDI*. En attendant qu'il se généralise, il doit être le mode de facturation de référence pour les anté-serveurs. Il ne sert à rien de créer des fonctions sophistiquées pour classer les documents, proposer des pistes ou aider à la reformulation si l'utilisateur ne peut pleinement en profiter pour PENSER au travail documentaire qu'il est en train de réaliser. Le système documentaire ne

doit pas être considéré seulement comme un réservoir d'informations, mais aussi comme une aide informatisée à l'ensemble des prises de décisions nécessaires à la recherche documentaire.

Cette préoccupation doit être présente dès le début de la conception de l'anté-serveur, notamment pour organiser les espaces de travail en mémoire adaptés à cette importance des opérations de lecture et de réflexion, en proportion plus longues que les calculs automatiques. L'anté-serveur doit aussi organiser l'accès aux serveurs afin de minimiser les coûts liés à la durée de la connexion.

4.b - un nouvel instrument d'évaluation

On a souvent souligné la difficulté de l'évaluation des systèmes documentaires. La majeure partie des études menées pour comparer diverses propositions le sont sur des banques de données de faible taille. Or étant donnée les caractéristiques du langage, notamment la polysémie des termes et les multiples possibilités de création de syntagmes nominaux, il est impossible de déduire des conclusions valables sur des grandes banques de données à partir d'exemples sur des banques de faible taille.

En offrant un accès à des banques réelles, constituées de plusieurs millions de références, en minimisant les coûts de connexion aux serveurs, les anté-serveurs peuvent devenir le moyen de tester plusieurs hypothèses en obtenant des évaluations relatives. Les taux de couverture et de précision sont des valeurs absolues, déterminées en fonction des questions. Leur calcul se heurte à la difficulté à connaître et même à évaluer le nombre de documents pertinents qui ne sont pas extraits (le "silence" documentaire). En revanche, il est plus aisé d'effectuer des comparaisons entre plusieurs méthodes et d'obtenir un classement d'efficacité. Du moins s'agit-il d'une hypothèse qui mérite d'être suivie.

Dans ce cadre, la comparaison se ferait pour une même question, dans la même formulation et sur la même banque de données. Les modules de classement des documents, ou les calculs statistiques seraient cependant différents. Même si

certaines méthodes entraînent des temps de calcul plus élevés, ce facteur pourrait être pris en compte dans l'évaluation.

Il faut en effet dépasser le fonctionnement déconnecté entre la recherche pointue sur de petites banques de données et la commercialisation de banques de données riches de plusieurs millions de références sans aucune évaluation des taux de satisfaction des utilisateurs. Les anté-serveurs ont une place à prendre dans ce processus.

quatrième partie

Conclusion

et

bibliographie

Cette thèse a permis d'aborder les problèmes de la recherche documentaire, à la fois du point de vue théorique (la modélisation), du point de vue pratique (la réalisation d'un catalogue de bibliothèque) et du point de vue des perspectives opérationnelles dans l'environnement réel (les axes de recherche et la réalisation d'anté-serveurs). On doit cependant constater une différence frappante entre les points de vue développés et la pratique actuelle de la documentation.

L'enjeu de la continuation de ce travail est de combler ce fossé. Un point de vue a traversé en permanence tout ce document : la recherche documentaire est une activité complexe, profondément humaine, marquée par l'incertain, l'aléatoire, la difficulté à exprimer le besoin documentaire, et pourtant la nécessité d'obtenir satisfaction de ce besoin.

La recherche documentaire, à l'époque de la circulation immédiate des informations, à l'époque où les instruments d'aide au travail intellectuel sont en train de changer rapidement et profondément, est un excellent modèle des activités dans lesquelles l'ordinateur sert à augmenter les capacités de travail intellectuel de l'opérateur humain. Les découvertes de la recherche documentaire sont appelées à se diffuser rapidement dans des domaines connexes, comme la bureautique ou les instruments d'aide à la décision, puis dans l'ensemble des instruments de production et de diffusion du savoir : réseaux de messageries, instruments de rédaction collective de documents, réseaux de diffusion de l'information.

La relation entre l'utilisateur et l'ordinateur dans le domaine documentaire devrait posséder une couleur particulière : l'ordinateur n'est plus seulement un instrument de mémoire (stockage massif des information, rapidité de recherche), ni même un instrument d'enregistrement et de préparation de documents (traitement de texte, édition électronique) mais devient un allié dans le travail intellectuel. De ce point de vue, les capacités de lecture sont déterminantes. On ne peut penser correctement que si l'on peut lire correctement, avec le temps qui nous est nécessaire, avec la précision qui nous semble souhaitable, avec la capacité d'annoter, de comparer, de juxtaposer, de classer... Le livre a porté longtemps ces demandes. La découverte de la forme *codex,* l'organisation des premières bibliothèques ont modifié profondément les modes de pensée des

hommes du Moyen-âge. Ouvrir plusieurs livres en même temps est un facteur d'émancipation intellectuelle fantastique. L'autonomie du lecteur s'en trouve grandement favorisée. L'imprimerie, en multipliant la diffusion de l'écrit, puis l'invention du "livre de poche" par Alde Manuce, en favorisant l'appropriation individuelle des écrits, ont parachevé ce travail de l'écrit. Les mots écrits deviennent les compagnons quotidiens de la pensée libre.

Or si nous analysons la situation des systèmes documentaires informatisés, nous retrouvons une situation qui n'est pas sans rappeler celle qui précéda la diffusion du livre. La recherche documentaire informatisée est encore vécue comme une activité séquentielle, avec une succession chronologique d'événements. Elle se rapproche plus de la lecture des manuscrits en rouleaux : on doit lire le texte (i.e. effectuer la recherche documentaire) suivant un ordre pré établi. On ne peut comparer, annoter, s'arrêter pour penser, revenir en arrière, s'appuyer sur des documents pour reprendre différemment la recherche...

Notre objectif est donc ambitieux : proposer un nouveau mode d'utilisation des fonds documentaires informatisés, qui soit adapté aux besoins des utilisateurs et non pas aux contraintes de la gestion documentaire. Pour autant, il ne s'agit pas d'une révolution copernicienne : les instruments nécessaires existent déjà. Ils sont appliqués dans de nombreux autres domaines de l'informatique, notamment, et ce n'est certainement pas innocent, dans le travail de production du texte. Les traitements de texte utilisent de plus en plus le multi fenêtrage (comparaison, annotation, déplacement), la présentation d'écrans lisibles (choix des caractères, typographie), la capacité à mélanger plusieurs types d'informations (texte, graphiques, images...).

Cet objectif doit s'appuyer sur plusieurs axes :

- un travail de persuasion envers le secteur de la diffusion et de l'organisation des fonds documentaires. Montrer que les systèmes documentaires informatisés peuvent proposer d'autres approches que les catalogues ou les bibliographies imprimées. L'informatique nous donne d'autres moyens, nous devons nous en servir pour d'autres buts. En ce sens, le recherche en science de l'information ne doit pas s'enfermer dans la répétition de modèles, mais se

confronter à la dimension réelle des fonds documentaires, et à la pratique réelle des indexeurs et des utilisateurs.

- un travail d'élaboration de modèles adaptés à certains besoins documentaires spécifiques. Le modèle QUID ne cherche pas à résoudre d'emblée les problèmes attachés aux banque de données textuelles, mais au contraire se définit comme un moyen de traiter des cas particuliers, que nous nommons les *"banques de compétences"*. Dans ce cadre, il offre un moyen de penser la description documentaire suivant d'autres voies, totalement différentes des pratiques habituelles, mais cependant corrélées avec des notions avancées par des praticiens de la documentation. En particulier, l'héritage de Ranganathan nous permet d'envisager des descriptions documentaires d'un mode différent.

- un travail de liaison entre les pratiques actuelles et les modèles souhaitables. Il sera difficile de transformer rapidement les instruments d'organisation des fonds électroniques, et de diffusion des informations sur le réseau. Les serveurs professionnels n'ont pas suivi les recherches des sciences de l'information depuis plusieurs dizaines d'années. Les banques de données, même si elles sont souvent insatisfaisantes, sont des produits industriels qui sont destinés à durer encore longtemps. Il faut donc opérer en dessus de ces instruments pour établir la liaison entre la situation actuelle et les modes de lecture spécifiques adaptés à la recherche documentaire. C'est le rôle des anté-serveurs. Le pouvoir de ce type d'instrument doit s'appuyer non pas sur les méthodes des serveurs en les recouvrant d'un vernis "convivial" (ce qui est la tâche des frontaux grand publics des serveurs), mais au contraire de partir des propositions faites à l'utilisateur pour modifier sa pratique documentaire et de transcrire ce besoin en fonction des méthodes des serveurs. Bien entendu, il y aura de nombreuses étapes entre la situation actuelle et la généralisation de ce type d'instruments. Mais des expériences simples et s'appuyant sur les modèles déjà diffusés, comme celle qui a été menée à la bibliothèque scientifique de l'université de Caen peuvent montrer que ce n'est pas un travail irréaliste. Il suffit souvent de peu de choses pour transformer profondément la capacité d'un utilisateur à aborder les systèmes documentaires.

Ces trois axes sont au confluent des recherches fondamentales, (l'élaboration de modèles et les développements informatiques) et des préoccupations pratiques, issues de l'expérience professionnelle. Il s'agit là d'un

positionnement habituel des sciences de l'information, qu'il faut défendre et qui doit nous encourager à œuvrer pour le développement et la reconnaissance de ce secteur de recherche.

Au long de cette thèse, nous avons étudié divers modèles et plusieurs expériences de systèmes documentaires. Au cours de ce travail, nous avons vu émerger de la réflexion trois nouveaux regards sur la recherche en sciences de l'information qui peuvent servir de repères pour reprendre la recherche :

- un modèle plus général de système documentaire
- un modèle plus dynamique de l'espace documentaire
- une conception élargie de l'aide intelligente aux utilisateurs.

Le modèle de système documentaire que nous avons utilisé doit aujourd'hui être élargi pour prendre en compte l'aspect dynamique de la recherche documentaire. Dans ce cadre, le concept de reformulation, tel qu'il a été défini dans le texte de cette thèse, occupe une place centrale. On doit considérer qu'un système documentaire se juge principalement au travers des instruments de reformulation qu'il offre à l'utilisateur. Cette place centrale de la reformulation se conçoit mieux encore si l'on inclus dans le modèle du système documentaire les intrants et les extrants qui sont réellement en jeu :

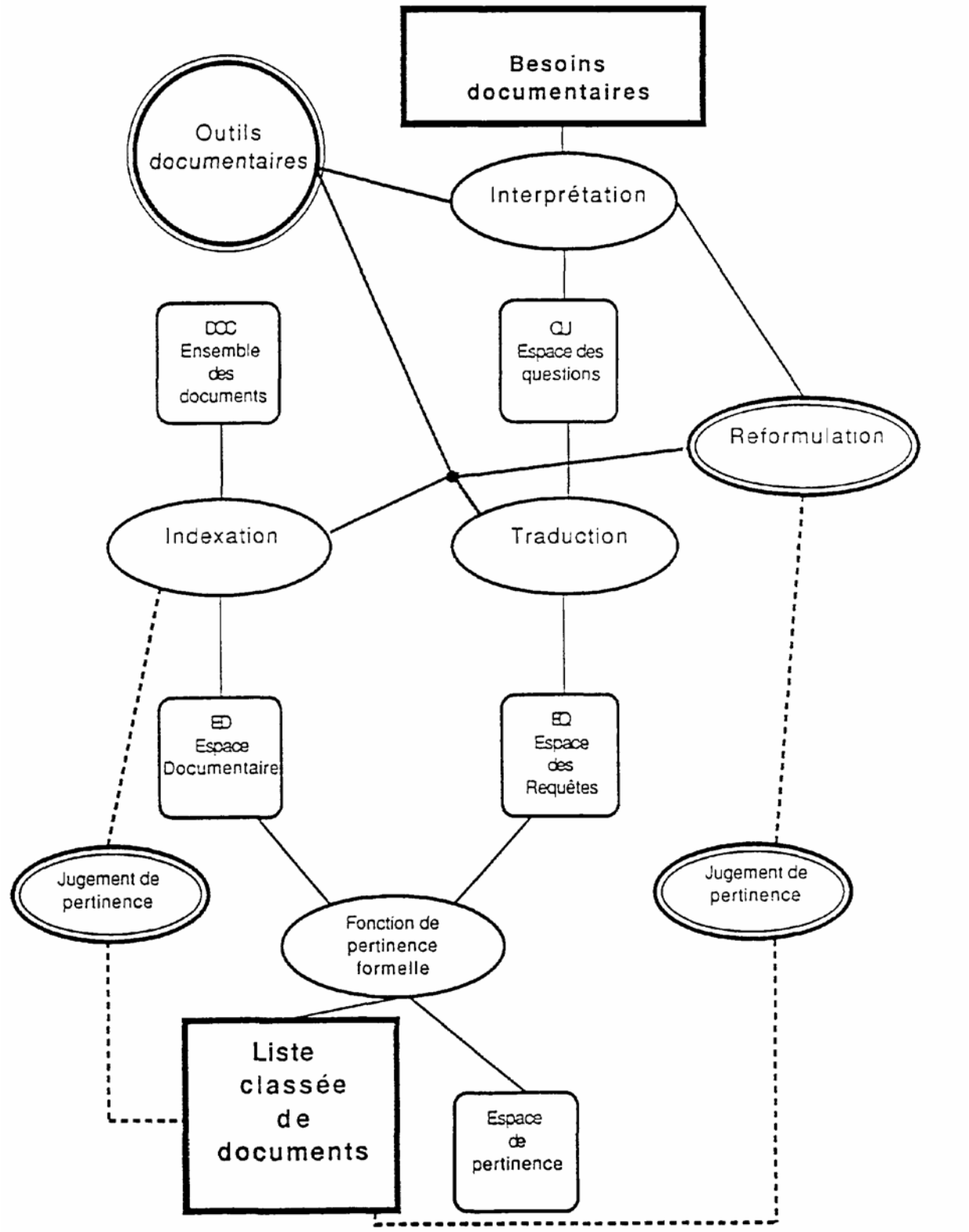
- *le besoin documentaire,* en ce qu'il diffère souvent de la "question de l'utilisateur", sans que celui-ci en soit directement conscient (phénomène d'ancrage, découvertes intervenant dans le cours même de la recherche...)

- *la liste classée de documents,* qui constitue le résultat final de l'opération de recherche, et qui doit donc servir de "boussole" au cours de la recherche (jugement de pertinence, proposition de pistes de reformulation...).

De même, les outils documentaires (thésaurus, classifications, réseaux sémantiques...) doivent être intégrés au modèle du système documentaire, en ce qu'ils sont des instruments pour l'indexation des documents, la traduction des questions et la reformulation. Dans ce sens, les outils documentaires ne sont plus conçus dans l'optique traditionnelle d'aide à l'indexeur et à l'intermédiaire, mais comme des instruments dynamiques entièrement associés au système documentaire considéré. Leur insertion dans le système doit être transparente à l'utilisateur. Il semble aussi nécessaire que ces instruments soient en

permanence enrichis à partir des connaissances des utilisateurs et des connaissances que le système peut déduire des pratiques auxquelles il donne lieu (nouveaux termes, synonymes, liens entre termes ou entre documents...)

Ce modèle général de système documentaire pourrait se schématiser dans la représentation suivante :

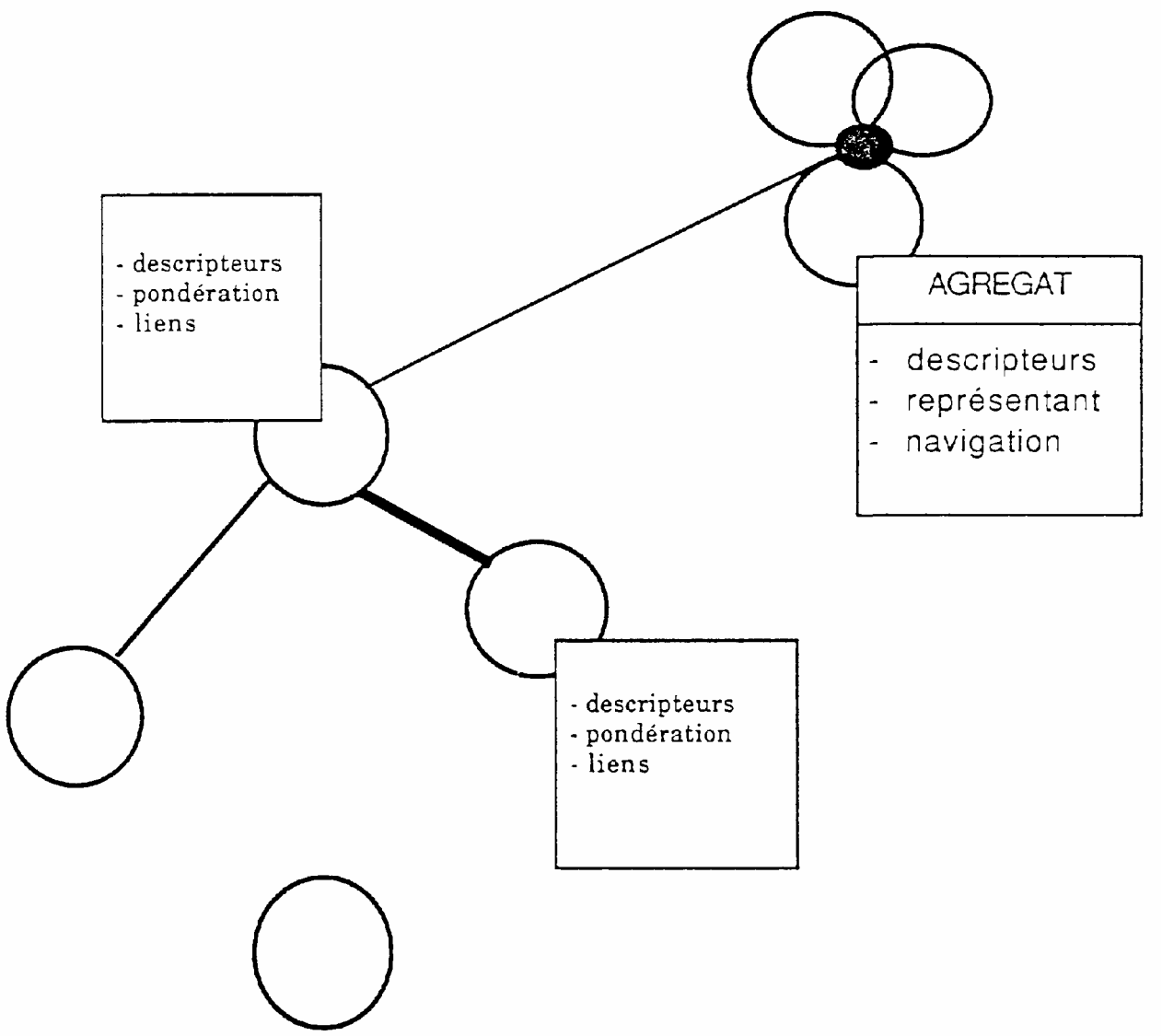


Dans les systèmes actuels, l'espace documentaire est rapporté à une liste de termes d'indexation, chaque document ayant des liens avec certains éléments de cette liste de termes. Les descripteurs associés au document sont produits par une analyse du document suivant les composantes que constituent les termes d'indexation connus par le système. Or la réalité des documents est différente : les documents sont directement en relation entre eux *(intertextualité* d'un point de vue conceptuel et *paratexte* du point de vue éditorial), et c'est seulement au second degré qu'ils sont en relation avec des termes d'indexation. Ce qui pénalise les systèmes actuels, c'est de s'appuyer sur un espace documentaire abstrait défini par des mots pour représenter un espace documentaire réel défini par des documents. Dès lors, ces systèmes deviennent soumis aux fluctuations des représentations (inconsistance de l'indexation, invention linguistique permanente, notamment au niveau des expressions composées, polysémie...). Les limites des modèles vectoriels et probabilistes sont certainement à rechercher dans cette confusion originelle.

On pourrait donc concevoir un espace documentaire plus dynamique en associant à chaque document une liste ouverte de descripteurs pondérés et des liens avec d'autres documents. La "force de ce lien" détermine la transmission des descripteurs d'un document à l'autre. Des documents très rapprochés constituent alors des agrégats, qui au niveau conceptuel sont eux aussi considérés comme des objets spécifiques auxquels sont associés une liste pondérée de descripteurs (ceux-ci deviennent alors valables pour tous les documents de l'agrégat en raison de la grande force des liens) et des liens avec d'autres agrégats ou d'autres documents. L'indexation devient alors une opération de rapprochement de documents, qui se traduit de façon dynamique par une modification de la liste des descripteurs associés (ou du moins des pondérations). Ainsi, au lieu de définir un référentiel global constitué par l'ensemble des termes connus du système, on s'appuie sur des relations locales (ce qui se traduit par une moindre sensibilité à la polysémie et aux ambiguïtés). La recherche documentaire dans un tel modèle s'appuie pour sa part sur deux méthodes : la navigation au sein des agrégats (apport du modèle hypertexte), et la capacité du système à fournir une réponse au niveau de précision correspondant à la requête de l'utilisateur. Si un terme est trop général, il sera alors associé à un agrégat, mais pas à l'ensemble des documents de cet agrégat. Cela se traduira par la présentation des agrégats de niveau inférieur (équivalent des "pistes de reformulation"). En sens inverse, un terme très particulier n'obérera pas la

possibilité de retrouver des documents proches (par navigation, mais aussi parce qu'au moment de l'insertion d'un document dans le système, ses termes d'indexation auront été "transmis" suivant les liens aux documents proches. Enfin, un système conçu sur ce modèle peut apprendre à partir des recherches menées dans le système : si deux documents sont jugés pertinents simultanément un certain nombre de fois, leur "force de liaison" se trouve augmentée, avec des conséquences sur l'indexation de chaque document par transfert des termes (modification de leur pondération) et par constitution éventuelle d'agrégats (passage des termes d'indexation du niveau du document au niveau supérieur).

Une représentation de ce modèle dynamique de l'espace documentaire pourrait suivre le schéma ci-dessous. Dans ce modèle, l'indexation comme la recherche peuvent être considérés comme équivalent à un déplacement d'un document dans le tissu, des documents et des agrégats constituant l'espace documentaire.



Enfin, en ce qui concerne l'aide intelligente qui peut être apportée aujourd'hui à l'utilisateur des banques de données, il faut se placer d'un point de vue plus général, pour couvrir ce qui est la situation réelle de l'utilisateur face à un système. On peut résumer ce besoin d'aide en quatre actions :

- *exprimer* le besoin documentaire face au système, ce qui requiert que l'utilisateur puisse apprendre le fonctionnement du système et sa zone de compétence.

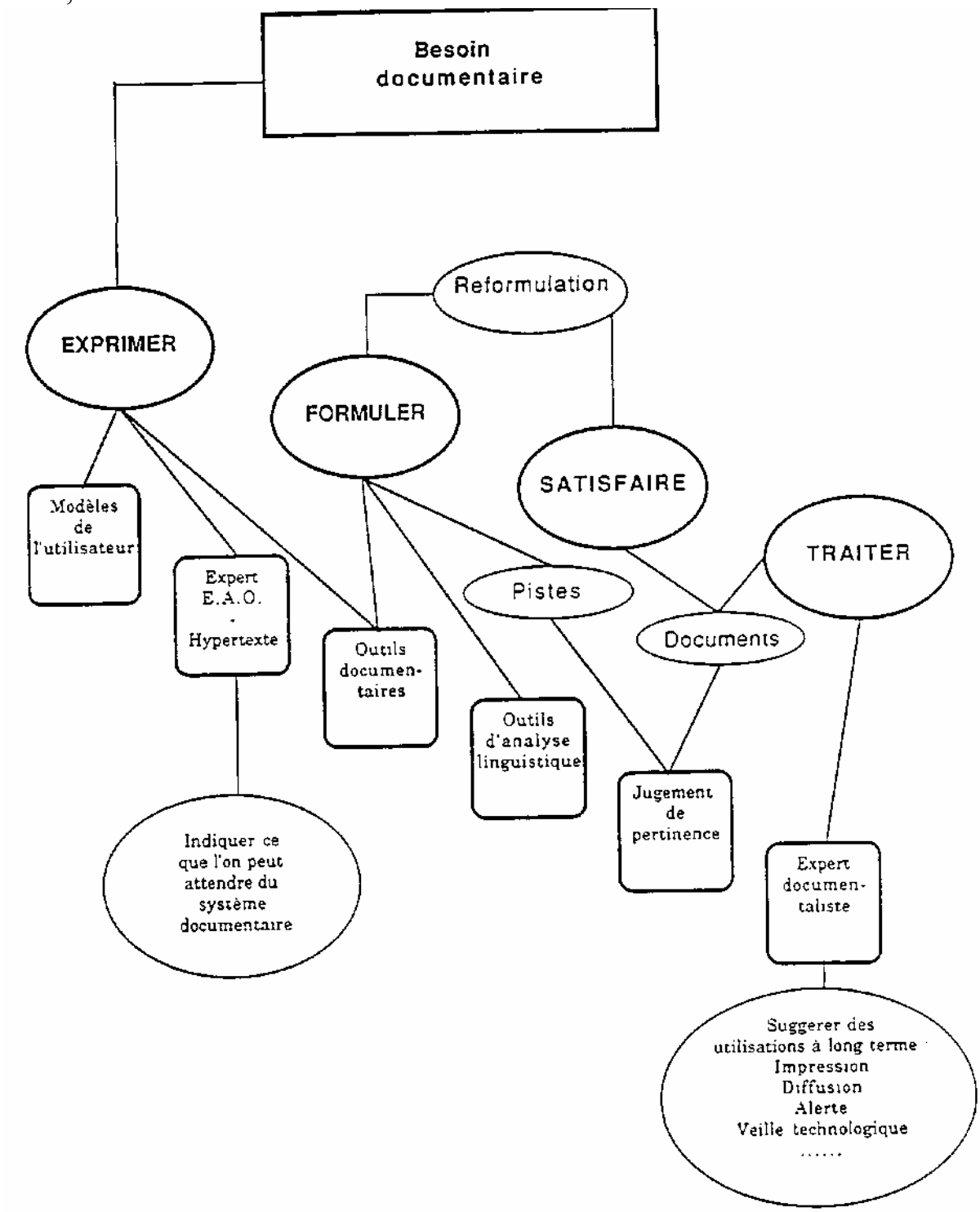
- *formuler* ce besoin documentaire. C'est la phase d'interprétation de la première requête et la phase dynamique de reformulation.

- *satisfaire* le besoin en proposant une liste classée de document et une liste de pistes de reformulation et en permettant le jugement de pertinence sur ces deux listes.

- *traiter* les documents recueillis par la recherche (impression, télédéchargement, conversion de formats...), mais aussi traiter la suite à donner à cette requête, en suggérant des utilisations à long terme (profils d'alerte, veille technologique, méthode d'obtention des documents...).

Aujourd'hui, le travail destiné à fournir une aide intelligente est souvent limité aux deux opérations de formulation et de satisfaction. Dans ce cadre, on insiste sur la reformulation et sur le jugement de pertinence. Il ne faut cependant pas oublier les deux autres actions, qui sont pour l'utilisateur les réelles portes d'entrée et de sortie du système documentaire. Cette conception élargie de l'aide intelligente à l'utilisateur se traduit par la mise au point d'un certain nombre d'outils (modèles de l'utilisateur, outils linguistiques, outils documentaires...) et par la capacité du système à suivre les évolutions de l'utilisateur dans le système et à offrir des moyens de repérer les comportements inadéquats pour proposer des voies de résolution. Or souvent, les limites proviennent de ce que l'utilisateur ne conçoit pas clairement les objectifs du système et éprouve des difficultés à définir pour lui-même les objectifs de sa recherche. C'est aussi à cette difficulté conceptuelle qu'il convient de s'attaquer, ou du moins dans un premier dont il convient de tenir compte dans la définition des fonctionnalités d'un système d'intermédiaire intelligent (anté-serveur, catalogue de bibliothèque accessible en ligne, Disque Optique Compact...).

Le schéma suivant représente cette conception élargie de l'aide à l'utilisateur,



Ces trois orientations constituent un programme de recherche qui s'appuie sur un travail théorique (renforcer la modélisation des systèmes documentaires), un travail prospectif (proposer un modèle plus dynamique de l'espace documentaire et le mettre en œuvre avec les moyens modernes de l'informatique comme la programmation objet et les interfaces graphiques) et un travail appliqué (définir l'aide intelligente dont a besoin l'utilisateur).

Bibliographie

[AFN84] AFNOR (Association Française de Normalisation). - *Documentation. Catalogage des monographies. Description Bibliographique minimale.* Fascicule Z 44-072 : 23 p. ; septembre 1984.

[AFN85] AFNOR (Association Française de Normalisation). - *Documentation. Liste d'autorité de matières. Structure et règles d'emploi.* Fascicule Z 47-200 : 14 p. ; mars 1985.

[AFN86] AFNOR (Association Française de Normalisation). - *Documentation. Indexation analytique par matière.* Fascicule Z 44-070 : 15 p. ; août 1986.

[AKS88] Akscyn, Robert M. ; McCracken, Donald M. ; Yoder, Elise A. - KMS : a distributed hypermedia system for managing knowledge in organizations. *Communications of the ACM.* 31(7) : 820-835 ; 1988.

[AND87] Andler, Daniel. - Progrès en situation d'incertitude. *Le Débat.* 47 : 5-25 ; 1987.

[ANDE73] Anderla, Georges. - *L'information en 1985 : une étude prévisionnelle des besoins et des ressources.* OCDE ; 1973.

[BAR89] Barthes, Christine. - Spécification du contrôle du raisonnement en mode évolutif : Application à la recherche documentaire en ligne. *Thèse. Université de Toulouse.* 1989.

[BAT86] Bâtes, Marcia J. - Subject access in online catalogs : a design model. *Journal of the American Society for Information Science.* 37(6) : 357-376 ; 1986.

[BAT90] Bâtes, Marcia J. - Where should the person stop and the information search interface start ? *Information Processing & Management.* 26(5) : 575-591 ; 1990.

[BEG88] Begeman, Michael L. ; Conklin, Jeff. - The right tool for the job : even systems design process fall within the realm of hypertext. *Byte.* Octobre 1988 : 255-266 ; 1988.

[BEL89] Belew, Richard K. - Adaptive information retrieval : using a connectionist representation to retrieve and learn about documents. *ACM-SIGIR Forum.* Special Issue : 11-20 ; 1989.

[BELK87] Belkin, N.J. ; Croft, W. Bruce. - Retrieval techniques. In : Williams, Martha E. (Ed) *Annual Review of Information Science and technology.* 22 : 109-145 ; 1987.

[BEN89] Benharrat, Alia. - Les stratégies de l'offre dans l'industrie des banques de données. *Documentaliste.* 26(2) : 233-237 ; 1989.

[BIR89] Biru, Tesfaye ; El-Hamdouchi, Abdelmoula ; Rees, Rodney S. ; Willet, Peter. - Inclusion of relevance information in the term discrimination model. *Journal of Documentation.* 45(2) : 85-109 ; 1989.

[BLA84] Blair, David G. - The data-document distinction in information retreival. *Communication of thé ACM* . 27(4) : 369-374 ; 1984.

[BLA85] Blair, David C. ; Maron, M.E. - An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM.* 28(3) : 289-299 ; 1985.

[BLA88] Blair, David C. - An extended document retrieval model. *Information Processing & Management.* 24(3) : 349-371 ; 1988.

[BLA90a] Blair, David C. - *Language and representation in information retrieval.* Amsterdam : Elsevier Science Publishers. 335 p. ; 1990.

[BLA90b] Blair, David C. ; Maron, M.E. - Full-text information retrieval : further analysis and clarification. *Information Processing & Management.* 26(3) : 437-447 ; 1990.

[BLAN89] Blanc-Montmayeur, Martine. - OPAC ou à la trinité : l'indispensable langage naturel. *Bulletin des Bibliothèques de France.* 34(1) : 60-62 ; 1989.

[BOI89] Boisard, Geneviève. - Le coût du catalogage. *Bulletin des Bibliothèques de France.* 34(4) : 330-339 ; 1989.

[BORG86] Borgman, Christine L. - Why are online catalogs hard to use ? Lessons learned from information-retrieval studies. *Journal of the American Society for Information Science.* 37(6) : 387-400 ; 1986.

[BOS83] Bossuat, Marie-Louise. - L'UAP (Universal Availability of Publications) : L'Accès Universel aux Publications. *Bulletin de l'Association des Bibliothécaires de France.* 119 : 7-8 ; 1983.

[BOV87] Bovey, J. D. ; Brown, P. J. - Interactive document display and its use in information retrieval. *Journal of Documentation.* 43(2) : 125-137 ; 1987.

[BR078] Broadbent, D.E. ; Broadbent, M. H. P. - The allocation of descriptor terms by individuals in a simulated retrieval system. *Ergonomics.* 21(5) : 343-354 ; 1978.

[BROW86] Brown, P.J. - Interactive documentation. *Software-Practice & Experience.* 16(3) : 291-299 ; 1986.

[BROW88] Brown, P.J. - Linking and searching within hypertext. *Electronic Publishing.* 1(1) : 45-53 ; 1988.

[BUS45] Bush, Vannevar. - As we may think. *Atlantic Monthly.* 176(1) : 101-108 ; 1945.

[CAI89a] Cailloux, J. M. - *Séminaire sur l'utilisation de l'OSI pour les bibliothèques.* Luxembourg : Commission des Communautés Européennes, 1989. 126 p.

[CAI89b] Cailloux, J. M. ; Casimir, C. - *OSI model for library applications : a tutorial.* Luxembourg : Commission des Comunautés Européennes, 1989. 135 p.

[CAN87] Can, Fazli ; Ozkarahan, Esen A. - Computation of term/document discrimination values by use of the Cover Coefficient concept. *Journal of the American Society for Information Science.* 38(3) : 171-183 ; 1987.

[CAN89] Can, Fazli ; Ozkarahan, Esen A. - Dynamic cluster maintenance. *Information Processing & Management.* 25(3) : 275-291 ; 1989.

[CANT86] Canter, D. ; Powell, J. ; Wishart, J. ; Roderick, C. - User navigation in complex database systems. *Behaviour and information technology,* 5(3) : 249-257 ; 1986.

[CAR88] Carroll, David M. ; Pogue, Christine A. ; Willett, Peter. - Bibliographic pattern matching using the ICL Distributed Array Processor. *Journal of the American Society for Information Science.* 39(6) : 390-399 ; 1988.

[CAT86] Cater, Steven C. - The topological information retrieval system and the topological paradigm : a unification of the major models of information. *Thesis. Ph. D. The Louisiana State University and Agricultural and Mechanical Col. (U.S.A.)* . 1986.

[CHA89] Charton, Ghislaine ; Dalbin, Sylvie ; Monteil, Marie-Gaëlle ; Verillon, Monique. - Indexation manuelle et indexation automatique : Dépasser les oppositions. *Documentaliste.* 26(4-5) : 181-187 ; 1989.

[CHAU86] Chaumier, Jacques. - *Systèmes d'information : marché et technologies.* Paris : Entreprise Moderne d'Edition. 117 p. ; 1986.

[COA88] Coates, E.J. - Raganathan's thought and its significance for the mechanisation of information storage and retrieval. *Herald of Library Science.* 27(1-2) : 3-14 ; 1988.

[CON87] Conklin, Jeff. - Hypertext : an introduction and survey. *Computer.* 20(9) : 17-41 ; 1987.

[CON89] Conklin, Jeff ; Begeman, Michael L. - gIBIS : a tool for all reasons. *Journal of the American Society for Information Science.* 40(3) : 192-199 ; 1989.

[C0089] Cooper, Lorraine K.D. ; Tharp, Alan L. - Inverted signature trees and text searching on CD-Roms. *Information Processing & Management.* 25(2) : 161-169 ; 1989.

[COOP68] Cooper, William S. - Expected Search Lengths : a single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *Journal of the American Society for Information Science.* 19 : 30-41 ; 1968.

[COOP88] Cooper, William S. - Getting beyond Boole. *Information Processing & Management.* 24(3) : 243-248 ; 1988.

[COU76] Courrier, Y. - Analyse et langage documentaire. *Documentaliste.* 13(5-6) : 178-189 ; 1976.

[COUR89] Courtial, Jean-Pierre ; Law, John. - A co-word study of artificial intelligence. *Social Studies of Science.* 19(2) : 301-311 ; 1989.

[CRE89] Créhange, M. ; Foucaut, 0. ; Halin, G. ; Mouaddib, N. ; Foucaut, J.F. - Semantics of the user interface for image retrieval : possibility theory and learning techniques. *Information Processing & Management.* 25(6): 615-627; 1989.

[CRI88] Crigean, Janey K. ; Manson, Gordon A. ; Willett, Peter ; Wilson, Georges A. - Efficiency of text scanning in bibliographic databases using microprocessor-based multiprocessor networks. *Journal of Information Science.* 14(6) : 335-345 ; 1988.

[CR086] Croft, W. Bruce. - Boolean queries and term dependencies in probabilistic retrieval models. *Journal of the American Society for Information Science.* 37(2) : 71-77 ; 1986.

[CR087] Croft, W.Bruce. ; Thompson, R.H. - I3R : a new approach to the design of document retrieval systems. *Journal of the American Society for Information Science.* 38(6) : 389-404 ; 1987.

[CR088] Croft, W. Bruce ; Savino, P. - Implemanting ranking strategies using text signatures. *ACM Transactions on Office Information Systems.* 6(1) : 42-62 ; 1988.

[CR089] Croft, W.B. ; Lucia, T. J. ; Cringean, J. ; Willet, P. - Retrieving documents by plausible inference : an experimental study. *Information Processing & Management.* 25(6) : 599-614 ; 1989.

[CROU90] Crouch, C. J. - An approach to the automatic construction of global thesaurii. *Information Processing & Management.* 26(5) : 629-640 ; 1990.

[DAC90a] Dachelet, Roland. - Etat de l'art de la recherche en informatique documentaire : la représentation des documents et l'accès à

l'information. In : *Le document électronique, Cours INRIA, 11-15 juin 1990.* Rocquencourt : INRIA, 1990.

[DAC90b] Dachelet, Roland. - Hypertexte et hypermedia : Documents, Informations, Connaissances. In : *Le document électronique, Cours INRIA, 11-15 juin 1990.* Rocquencourt : INRIA, 1990.

[DAN86] Daniels, P. J. - Cognitive models in information retrieval : an evaluative review. *Journal of Documentation.* 42(4) : 272-304 ; 1986.

[DBM88] DBMIST (Direction des Bibliothèques, des Musées et de l'Information Scientifique et Technique. Ministère de l'Education Nationale). - *Recommandations concernant le traitement des documents acquis et leur mise à la disposition des lecteurs. Bibliothèques Universitaires, Services de Documentation. Circulaire 88-1/9 du 22 juillet 1988.* 14 p. ; 1988. (le texte de cette circulaire est repris dans [SANS88]).

[DEB89] Debili, Fathi ; Fluhr, Christian ; Radasoa, Pierre. - About reformulation in full-text 1RS. *Information Processing & Management.* 25(6) : 647-657 ; 1989.

[DES85] Deschâtelets, Gilles. - L'homo médiaticus vs l'interface masquée : un combat à finir. *Documentation et bibliothèques.* 31(2) : 55-66 ; 1985.

[DEW89] Deweze, André. - *Informatique documentaire.* Sème Ed. Paris : Masson, 1989.

[DOL88] Doland, V.M. - The hermeneutics of hypertext. *12th International Online Information Meeting. Proceedings. Tome 1. Londres. Décembre 1988.* Learned Information. 75-82 ; 1988.

[DOU88] Dou, Henri ; Hassanaly, Parina ; Quoniam, Luc ; Pullino, J. ; Le Coadic, Yves. - La dispersion des sources documentaires. Mesure et évolution. Problèmes de décision. *Revue Française de Bibliométrie.* 11-29 ; 1988.

[DUR87] Durand C. - Vers le neuro-ordinateur. *MicroSystèmes.* Octobre 1987 : 85-95 ; 1987.

[DUS89] Dussert-Carbone, Isabelle. - Comparaison entre les normes françaises et les règles anglo-américaines de catalogage. *Bulletin des Bibliothèques de France.* 34(4) : 352-361 ; 1989.

[EAS89] Eastman, Caroline M. - Handling incrementally specified boolean queries : A comparison of inverted and signature file organizations. *Information Processing & Management.* 25(1) : 27-38 ; 1989.

[ERL87] ERLI. - Accès naturel à une base de données textueles : Indexation automatique et interrogation simplifiée. *Note interne.* 1987.

[FAL85] Faloutsos, Christos. - Access methods for text. *Computing Surveys*. 17(1) : 49-74 ; 1985.

[FIDE88] Fiderio, Janet. - A grand vision. *Byte.* 237-244 ; October 1988.

[FOG87] Fogelman-Soulié, Françoise. - Méthodes connexionistes pour l'apprentissage, in : Pastre, Dominique (Ed.) : *Intelligence artificielle. Actes des journées nationales PRC-GRECO.* Toulouse 1988.

[FOS89] Foss, Carolyn L. - Tools for reading and browsing hypertext. *Information Processing & Management.* 25(4) : 407-418 ; 1989.

[FUH89] Fuhr, Norbert. - Models for retrieval with probabilistic indexing. *Information Processing & Management.* 25(1) : 55-72 ; 1989.

[FUR83] Furnas, G.W. ; Landauer, T.K. ; Gomez, L.M. ; Dumais, S. T. - Human factors and behavorial science : statistical semantic : analysis of the potential performance of key word information system. *The Bell technical journal.* 62(6) ; 1983.

[FUR86] Furnas, G.W. - Generalized fish eye views. *ACM-SIGCHI Bulletin.* Avril 1986 : 16-23 ; 1986.

[FUR87] Furnas, G. W. ; Landauer, T. K. ; Gomez, L. M. ; Dumais, S. T. - The vocabulary problem in human-system communication. *Communications of the ACM.* 30(11) : 964-971 ; 1987.

[GAR86a] Garfield, Eugene. - ISI's master list of title words provide a special perspective on science and scholarly activity. Part 1 : The lexicography of the Unique Word Dictionnary. *Current Contents.* 27 : 3-8 ; July 1986.

[GAR86b] Garfield, Eugene. - ISI's master list of title words provide a special perspective on science and scholarly activity. Part 2 : Comparative etymology of neologisms and research fronts. *Current Contents.* 28 : 3-10; July 1986.

[GAR90] Garfield, Eugene. - KeyWords Plus : ISI's breakthrough retrieval methods. Part *1. Current Contents.* 32 . 6 août 1990 : 5-9 ; 1990.

[GHE90] Gherissi, Houssen ; Langlet, Franck. - Réalisation d'une interface vidéotex pour la consultation du catalogue de la bibliothèque de l'Université de Caen. *Mémoire de licence. Université de Caen.* 19 p. ; 1990.

[GOR88] Gordon, Michael. - Probabilistic and genetic algorithms for document retrieval. *Communications of the ACM.* 31(10) : 1208-1218 ; 1988.

[GRI86] Griffiths, Alan ; Luckhurst, H. Claire ; Willett, Peter. - Using interdocument similarity information in document retreival systems. *Journal of the American Society for Information Science.* 37(1) : 3-11 ; 1986.

[GUN89] Gûntzer, U. ; Jùttner, G. ; Seegmùller, G. ; Sarre, F. - Automatic thésaurus construction by machine learning from retrieval sessions. *Information Processing & Management,* 25(3) : 265-273 ; 1989.

[HAL88] Halasz, Frank G. - Reflections on Notecards : seven issues for the next généraltion of hypermedia systems. *Communications of the ACM.* 31(7) : 836-852 ; 1988.

[HALI89] Halin, Gilles. - Apprentissage pour la recherche interactive et progressive d'images : processus EXPRIM et prototype RIVAGE. *Thèse. Université de Nancy 1.* 337 p. ; 1989.

[HAR90] Harman, Donna ; Candela, Gerald. - A very fast prototype retrieval system using statistical ranking. *ACM-SIGIR Forum.* 23(3-4) : 100-110 ; 1990.

[HIN86] Hinton, G. E. ; McClelland, J. L. ; Rumelhart, D. E. - Distributed representations, p. 77-109. In : Rumelhart, David E. ; McClelland, James L. (Eds.) *Parallel Distributed Processing : Exploration in the microstructure of cognition - Vol. 1 : Foundations.* Cambridge (Ma) : MIT press, 1986.

[HOP82] Hopfield, J.J. - Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academv of Science.* 79 : 2554-2558 ; 1982.

[HUD85] Hudrisier, Henri. - Le mariage des boîtes en carton et du vidéodisque : les mutations d'une banque d'image. *Bulletin du Centre des Hautes Etudes Internationales en Informatique Documentaire.* 20 : 45-55 ; 1985.

[HUD86] Hudrisier, Henri. - L'imageur documentaire S.E.P. in *L'imaginaire numérique.* Paris : Hermès, p. 136-155 ; 1986.

[INT75] INTERMARC (M) : Format bibliographique d'échange pour les monographies : Manuel. Paris : *Bibliothèque Nationale.* 1975.

[IRI89] Irish, Peggy M. ; Trigg, Randall H. - Suporting collaboration in hypermedia : issues and expériences. *Journal of the American Society for Information Science.* 40(3) : 192-199 ; 1989.

[IWR90a] Online chemistry titans clash. *Information World Review.* 49 : 1-2 et 4 ; 1990.

[IWR90b] Focus : Towards real-time. *Information Worl Review.* 49 : 19-23 ; 1990.

[JAK88] Jakobiak, François. - *Maîtriser l'information critique.* Paris : Ed. d'Organisation. 225 p. ; 1988.

[KNI90] Knight, Kevin. - Connectionist ideas and algorithms. *Communication of the American Computing Machinery.* 33 (11) : 59-74 ; 1990.

[KOC63] Kochen, Manfred. - On natural information systems : pragmatic aspects of information retrieval. *Meth. of Inf. in Med.* 2(4) : 143-147 ; 1963.

[KOV86] Koved, Larry ; Shneiderman, Ben. - Embedded menus : selecting items in context. *Communications of the ACM.* 29(4) : 312-318 ; 1986.

[KRA85] Kraft, Donald H. - Advances in information retrieval : where is that /**&@0 record ? *Advances in computers.* 24 : 277-318 ; 1985.

[KW085] Kwok, K. L. - A probabilistic theory of indexing and similarity measure based on cited and citing documents. *Journal of the American Society for Information Science.* 36(5) : 342-351 ; 1985.

[KW089] Kwok, K. L. - A neural network for probabilistic information retrieval. *ACM-SIGIR Forum.* Special Issue : 21-30 ; 1989.

[LAI82] Laine, Sylvie. - Extraction et sélection de descripteurs complexes dans un ensemble de textes pour leur indexation automatique. *Thèse. Université de Lyon I.* 1982.

[LAM85] Lambert, Jill. - *Scientific and Technical journals.* - C. Bingley. 191 p. ; 1985.

[LEC86a] Le Crosnier, Hervé. - *La micro-informatique : un nouveau secteur de la bibliothèque.* Ed. du Cercle de la Librairie. 175 p. ; 1986.

[LEC86b] Le Crosnier, Hervé. - Bientôt des livres sur disques compacts. *Livres-Hebdo.* 37 : 96-100 ; 1986.

[LEC87a] Le Crosnier, Hervé. - Librarians and user education for online bibliographie retrieval. *IATUL Quaterly.* 1(2) : 102-106 ; 1987.

[LECSTb] Le Crosnier, Hervé. - Faut-il croire au CD-Rom ? *Livres-Hebdo.* 32.35 : 74-75 ; 1987.

[LEC88a] Le Crosnier, Hervé. - *L'édition électronique : Publication assistée par ordinateur, Information en ligne, Médias optiques.* Paris : Ed. du Cercle de la Librairie. 286 p. ; 1988.

[LEC88b] Le Crosnier, Hervé. - A média neuf, produits nouveaux. *Livres de France.* 96 : 101-104 ; 1988.

[LEC88c] Le Crosnier, Hervé. - CD-Rom : Le décollage du marché. *Livres-Hebdo.* 39 : 136-140 ; 1988.

[LEC88d] Le Crosnier, Hervé. - *Pour une banque de données répartie des collections antropologiques régionales.* Caen : Section Fédérée de Basse-Normandie de l'Association des Conservateurs des Collections Publiques de France. 70 p. ; 1988.

[LEC89] Le Crosnier, Hervé. - *Les banques de données : Etat de l'art et perspectives de développement.* Paris : Triel. 400 p. ; 1989.

[LEL89] Le Loarer, Pierre. - Opacité et transparence des catalogues informatisés pour l'usager. *Bulletin des Bibliothèques de France.* 34(1) : 64-77 ; 1989.

[LELU86] Lelu, Alain ; Rosenblatt, D. - Représentation et parcours d'un espace documentaire. Analyse des données, réseaux neuronaux et banques d'images. *Les Cahiers de l'Analyse des Données.* 11(4) : 453-470 ; 1986.

[LELU89] Lelu, Alain ; Georgel, Albert. - Un modèle neuronal d'apprentissage de données documentaires, in : *Les systèmes d'informations élaborées. Congrès. Société Française de Bibliométrie Appliquée. Ile-Rousse,* p. 41-55 ; 1989.

[LEM89a] Le Marée, Joëlle. - Les OPACS sont-ils opaques : la consultation des catalogues informatisés à la BPI du Centre Pompidou. *Bulletin des Bibliothèques de France.* 34(1) : 78-85 ; 1989.

[LEM89b] Le Marée, Joëlle. - *Dialogue ou labyrinthe ? La consultation des catalogues informatisés par les usagers.* Paris : Bibliothèque Publique d'Information, 1989.

[LEV89] Lévine, Pierre ; Pomerol, Jean-Charles. - *Systèmes interactifs d'aide à la décicion et systèmes experts.* Paris : Hermès, 1989.

[LOS88a] Losee, Robert M. - Parameter estimation for probabilistic document-retrieval models. *Journal of the American Society for Information Science.* 39(1) : 8-16 ; 1988.

[LOS88b] Losee, Robert M. ; Bookstein, Abraham. - Integrating boolean queries in conjonctive normal form with probabilistic retrieval models. *Information Processing & Management.* 24(3) : 315-321 ; 1988.

[LU90] Lu, Xin. - Document retrieval : a structural approach. *Information Processing & Management.* 26(2) : 209-218 ; 1990.

[LUC88] Lucarella, Dario. - A document retrieval system based on nearest neighbour searching. *Journal of Information Science.* 14 : 25-33 ; 1988.

[LUS86] Lussato, Bruno ; Messadié, Gerald. - *Bouillon de culture.* Paris : Laffont. 260 p. ; 1986.

[MACR89] MacRoberts, Michael H. ; MacRoberts, Barbara R. - Problems of citation analysis : a critical review. *Journal of the American Society for Information Science.* 40(5) : 342-349 ; 1989.

[MAN87] Maniez, Jacques. - *Les langages documentaires et classificatoires : conception, construction et utilisation dans les systèmes documentaires.* Paris : Editions d'Organisation, 1987.

[MAR88] Martin, Henri- Jean. - *Histoire et pouvoirs de l'écrit.* Paris : Librairie Académique Perrin. 517 p. ; 1988.

[MAR060] Maron, M. E. ; Kuhns, J. - On relevance, probabilistic indexing and information retrieval. *Journal of the American Computing Machinery. 1 :* 216-244 ; 1960.

[MAR088] Maron, M. E. - Probabilistic design principles for conventional and full-text retrieval systems. *Information Processing & Management.* 24(3) : 249-255 ; 1988.

[MCA89] McAleese, Ray. - Navigation and Browsing, p. 6-44. In : McAleese, Ray (Ed.). - *Hypertext : theory into practice.* Blackwell Scientific Publications, 1989.

[MCC86] McClelland, J. L. ; Rumelhart, D. E. ; Hinton, G. E. - The appeal of PDF. p. 25-31. In : Rumelhart, David E. ; McClelland, James L. (Eds.) *Parallel Distributed Processing : Exploration in the microstructure of cognition - Vol. 1 : Foundations.* Cambridge (Ma) : MIT press, 1986. (cet article est partiellement traduit en français dans *Le Débat* 47 : 45-64 ; 1987.)

[MEA88] Meadow, Charles T. - Back to the future : making and interpreting the database industry timeline. *Databases.* 14-31 ; October 1988.

[MEAU88] Meaudre, Ethel. - Les banques de données juridiques françaises. *Mémoire : DESS Information et documentation : Institut d'Etudes Politiques de Paris.* 1988.

[MEL89] Melot, Michel. - La bibliothèque : un centre d'information. *ARBIDO-R.* 4(4) : 94-98 ; 1989.

[MET88] Metzner, Patrick ; Leyrat, Jacques ; Le Crosnier, Hervé. - Les banques de données bibliographiques : un exemple de création. *Informations Chimie.* 294 : 307-309 ; 1988.

[MIC88] Michelet, Bertrand. - L'analyse des associations. *Thèse. Université de Paris VII.* 1988.

[MIE89] Miège, Bernard. - *La société conquise par la communication.* Grenoble : Presses de l'Université de Grenoble. 228 p. ; 1989.

[MIN89] Ministère de la culture, de la communication, des grands travaux et du bicentenaire. France. - Schéma directeur de l'information bibliographique. Rapport final, rédigé par Annie Gourdier et François Pellegri. juillet 1989. (de larges extraits de ce rapport sont publiés dans *Bulletin des Bibliothèques de France* 34(4) : 288-311 ; 1989.)

[MINS88] Minsky, Marvin Lee ; Papert, Seymour A. - *Perceptrons : an introduction to computational geometry.* Cambridge : MIT-Press, 1988 (new éd.). 292 p.

[MIQ89] Miquel, André. - *Les bibliothèques universitaires. Rapport au ministre d'Etat ministre de l'Education Nationale de la Jeunesse et des Sprots.* Paris : La documentation française. 79 p. ; 1989.

[MIT89] Mitev, Nathalie ; Hildreth, Charles H. - Les catalogues interactifs en Grande-Bretagne et aux Etats-Unis : systèmes et interfaces. *Bulletin des Bibliothèques de France.* 34(1) : 22-47 ; 1989.

[MOR82] Morissey, J. - An intelligent terminal for implementing relevance feedback on large operationnal retrieval systems. *Lecture Notes in Computer Science.* 146 : 38-50 ; 1982.

[NEL88] Nelson, Ted. - Unifying tomorrow's hypermedia. *12th International Online Information Meeting. Proceedings. Tome 1. Londres. Décembre 1988.* Learned Information. 1-8 ; 1988.

[NYC89] Nyce, James M. ; Kahn, Paul. - Innovation, Pragmaticism, and technological continuity : Vannegar Bush's Memex. *Journal of the American Society for Information Science.* 40(3) : 214-220 ; 1989.

[OGD87] Ogden, William C. ; Sorkenes, Ann. - What do user say to their natural language interface ? In : Bullinger, H-J. and Shackel, B. *(Eds.) Human-Computer Interaction - INTERACT'87.* Elsevier, 1987.

[PAL89] Palierne, Catherine. - Logiciels hypertexte : simple d'emploi, mais pour quoi faire ? *Micro-systèmes.* Mai 1989 : 183-186 ; 1989.

[PAQ89] Paquel, Norbert. - Mémoire Technique - Schéma d'élaboration d'un catalogue collectif national des fonds documentaires. *Document Interne.* Paris : Cabinet Norbert Paquel. 1989.

[PEN87] Pennel, Patrice ; Lupovici, Catherine ; Denis, Anne-Marie. - Le plan catalogue. *Bulletin des Bibliothèques de France.* 32(2) : 118-132 ; 1987.

[PENI90] Pénicaut, Nicole. - Formation professionnelle chez Renault : Dlétrisme, la méthode Renault. *Libération.* 22 mai 1990.

[PER88] Personnaz, Léon ; Dreyfus, Gérard ; Guyon, Isabelle. - Les machines neuronales. *La Recherche.* 19 (204) : 1362-1371 ; 1988

[PERR83] Perry, S. A. ; Willett, Peter. - A review of the use of inverted files for best match searching in information retrieval systems. *Journal of Information Science.* 6 : 59-66 ; 1983.

[POG87] Pogue, Christine A. ; Willet, Peter. - Use of text signatures for document retrieval in a highly parallel environment. *Parallel Computing.* 4 : 259-268 ; 1987.

[POL88] Pollit, Arthur Steven. - MenUSE for Medicine : End-user browsing and searching of MEDLINE via the MeSH thesaurus. *International Forum on Information and Documentation.* 13(4) : 11-17 ; 1988.

[POR76] Porcher, Louis. *Vers la dictature des médias ?* Paris : Hatier. 80 p. ; 1976.

[PUJ89] Pujo, Pascal. - Développement d'une interface conviviale pour l'interrogation en langage naturel d'une base de données avec utilisation des concepts et des moyens de l'intelligence artificielle. *Thèse. Université Paris XI* ; 1989.

[RAD88] Radasoa, Hary Pierre. - Méthode d'amélioration de la pertinence des réponses dans un système de bases de données textuelles. *Thèse. Université de Paris Sud (Orsay).* 1988.

[RADE88] Radecki, Tadeusz. - Probabilistic methods for ranking output documents in conventionnal boolean retrieval systems. *Information Processing & Management.* 24(3) : 281-302 ; 1988.

[RAGS 6] Raghavan, Vijay V. ; Wong, S.K.M. - A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science.* 37(5) : 279-287 ; 1986.

[RAS87] Rasmussen, Edie M. ; Willett, Peter. - Non-Hierarchic document clustering using the ICL Distributed Array Processor. *Processing of the tenth international ACM-SIGIR Conference on research and development in Information Retrieval.* New Orleans : Association of Computing Machinery. 132-139 ; 1987.

[RAS88] Rasmussen, Edie M. ; Downs, Geoffrey M. ; Willet, Peter. - Automatic classification of chemical structure databases using a highly parallel array processor. *Journal of Computational Chemistry.* 9(4) : 378-386 ; 1988.

[RAS89] Rasmussen, Edie M. ; Willet, Peter. - Efficiency of hierarchic agglomerative clustering using the ICL distributed array processor. *Journal of Documentation.* 45(1) : 1-24 ; 1989.

[RAU89] Rau, Lisa F. ; Jacobs, Paul S. ; Zernik, Uri. - Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing & Management.* 25(4) : 419-428 ; 1989.

[REI85] Reiter, Martha B. - Can you teach me to do my own searching ? Or tailoring online training to the needs of end-user. *J. Chem. Inf. Comput. ScL* 25 : 419-422 ; 1985.

[RIC86] Richter, Noë. - *La pratique de l'indexation.* Le Mans : Bibliothèque de l'Université du Maine, 1986.

[RIC87] Richter, Noë. - *Grammaire de l'indexation décimale.* Le Mans, Bibliothèque de l'Université du Maine, 1987.

[RIC88] Richter, Noë. - *Grammaire de l'indexation alphabétique.* Le Mans, Bibliothèque de l'Université du Maine, 1988.

[RIC90] Richter, Noë. - *Les langages documentaires encyclopédiques : guide pratique d'indexation à l'usage des documentalistes de l'enseignement et des candidats aux examens et concours des bibliothèques et de la documentation.* Marigné (Sarthe) : La queue du Chat, 1990.

[ROB76] Robertson, S. E. ; Sparck-Jones, Karen. - Relevance weighting of search terms. *Journal of the American Society for Information Science.* 27(3) : 129-146 ; 1976.

[ROB82] Robertson, S. E. ; Maron, M. E. ; Cooper, W. S. - Probability of relevance : a unification of two competing models for document retrieval. *Information Technology : research and development.* 1(1) : 1-21 ; 1982.

[ROBE84] Roberts, Norman. - The pre-history of the information retrieval thesaurus. *Journal of Documentation.* 40(4) : 271-285 ; 1984.

[ROY87] Roy, Richard. - *Classer & indexer : introduction à l'indexation documentaire.* Le Mans : Bibliothèque de l'Université du Maine, 1987.

[SAL76] Salton, Gérard. - SMART. In : Belzer, Jack ; Holzman, Albert ; Kent, Allen (Eds.) *Encyclopedia of computer science and technology.* New- York : Marcel Dekker, 1976.

[SAL81] Salton, G. - A blueprint for automatic indexing. *ACM-SIGIR Forum.* 16(2) : 22-38 ; 1981.

[SAL82] Salton, G. - A blueprint for automatic boolean query processing. *ACM-SIGIR Forum.* 17(2) : 6-25 ; 1982.

[SAL83] Salton, Gerard ; McGill, Michael J. - *Introduction to modern information retrieval.* McGraw-Hill. 448 p. ; 1983.

[SAL85] Salton, G. ; Fox, E.A. ; Voorhees, E. - Advanced feedback methods in information retrieval. *Journal of the American Society for Information Science.* 36(3) : 200-210 ; 1985.

[SAL86] Salton, Gerard. - Another look at automatic text-retrieval systems. *Communications of the ACM.* 29(7) : 648-656 ; 1986.

[SAL88a] Salton, G. - A simple blueprint for automatic boolean query processing. *Information Processing & Management.* 24(3) : 269-280 ; 1988.

[SAL88b] Salton, Gerard ; Buckley, Chris. - Parallel text search methods. *Communications of the ACM.* 31(2) : 202-215 ; 1988.

[SAL90] Salton, Gerard ; Buckley, Chris ; Smith, Maria. - On the application of syntactic methodologies in automatic text analysis. *Information Processing & Management.* 26(1) : 73-92 ; 1990.

[SAN89] Sandison, Alexander. - Thinking about citation analysis. *Journal of Documentation.* 45(1) : 59-64 ; 1989.

[SANS88] Sansen, Jean-Raoul. - L'accès aux documents dans les bibliothèques universitaires. *Bulletin des Bibliothèques de France.* 33(6) 456-466.

[SCA89] Scacchi, Walt. - On the power of domain specific hypertext environments. *Journal of the American Society for Information Science.* 40(3) : 183-191 ; 1989.

[SCH86] Schwering, Julie. - Market opportunities for CD-Rom. *Actes du premier colloque français sur le CD-Rom et ses applications. Versailles, juin 1986.* GFFIL. 155-163 ; 1986.

[SEN89] Sen, B.K. ; Karanjai, A. ; Munshi, U.M. - A method for determining the impact factor of a non-SCI journal. *Journal of Documentation.* 45(2) : 139-141 ; 1989.

[SENA90] Senach, Bernard. - Evaluation ergonomique des interfaces homme-machine : une revue de la littérature. *Rapport de recherche INRIA.* n° 1180 ; mars 1990.

[SHN89] Shneiderman, Ben ; Brethauer, Dorothy ; Plaisant, Catherine ; Potter, Richard. - Evaluating three museum installations of a hypertext system. *Journal of the American Society for Information Science.* 40(3) : 172-182 ; 1989.

[SIN89] SINORG. - BRS, progiciel de recherche documentaire en texte intégral. *Document de présentation.* 1989.

[SOK76] Sokal, Robert R. - Cluster analysis. In : Belzer, Jack ; Holzman, Albert ; Kent, Allen (Eds.) *Encyclopedia of computer science and technology - Tome 5.* New- York : Marcel Dekker, 1976.

[STA86b] Stanfill, Craig ; Kahle, Brewster. - Parallel free-text search on the Connection Machine system. *Communications of the ACM.* 29(12) : 1229-1239 ; 1986.

[ST087] Stone, H. S. - Parallel querying of large databases : a case study. *Computer,* 20(10 octobre 1987) : 11-21 ; 1987.

[SWA60] Swanson, Don R. - Searching natural language text by computer. *Science.* 132(3434) : 1099-1104 ; 1960.

[SWA89a] Swanson, Don R. - Online search for logically-related noninteractive medical literatures : A systematic trial-and-error strategy. *Journal of the American Society for Information Science.* 40(5) : 356-358 ; 1989.

[SWA89b] Swanson, Don R. - A second example of mutually isolated medical literatures related by implicit, unnoticed connections. *Journal of the American Society for Information Science.* 40(6) : 432-435 ; 1989.

[SWA90] Swanson, Don R. - Integrative mechanisms in the growth of knowledge : a legacy of Manfred Kochen. *Information Processing & Management.* 26(1) : 9-16 ; 1990.

[TAG81] Tague, Jean M. - User-responsive subject control in bibliographic retrieval systems. *Information Processing & Management.* 17 : 149-159 ; 1981.

[TES89] Test sur les bases bibliographiques. Rapport du Bureau Marcel Van Dijk, rédigé par Michèle Lenart et Michel Labastide. *Bulletin des Bibliothèques de France.* 34(4) : 312-325 ; 1989.

[TS086] Tsouria-Belaïd, Lahouaria. - Contribution méthodologique à l'interrogation d'un système documentaire "grand public". *Thèse. Université de Toulouse (Sciences).* 1986.

[TH089] Thomazo, Loïc. - Recherche documentaire non-booléenne, (description d'un algorithme). QUID : une application. *Mémoire de DEA. Université de Caen.* 1989.

[THOM90] Thompson, Paul. - A combinaison of expert opinion approach to probabilistic information retrieval. Part 1. The conceptual model. *Information Processing & Management.* 26(3) : 371-382 ; 1990.

[TRI89] Trivette, Donald B. - What's different about DowQuest ? Ease, content & test retrieval. *DowLine.* second quarter. 1989.

[VEI85] Veilex, Florence. - Approche expérimentale des processus humains de compréhension en vue d'une indexation automatique des résumés scientifiques : application à un corpus de géologie. *Thèse de troisième cycle* . Université de grenoble 2, 1985.

[VER89] Verity. - *Topic Product Overview.* Document publicitaire. Verity Inc. (1530 Plymouth. Mountain View. California).

[VIC88a] Victorri, Bernard ; Liscia, Bruno. L'interface Homme-Machine sur Minitel : un défi pour l'intelligence artificielle. In *Colloque Systèmes experts et télématique. Paris. 28 et 29 janvier 1988.*

[VIC88b] Victorri, Bernard. - Sémantique et variétés différentielles. *Les cahiers du LIVC.* 11p.; 1988.

[VIC89] Victorri, Bernard ; Le Crosnier, Hervé ; Boyreau, Gilbert ; Thomazo, Loïc. - Recherche documentaire non-booléenne et algorithmes massivement parallèles. *Journées PARUSI. Paris. 18-19 mai 1989.* 1989.

[VIR88] Virbel, Jacques. - Linguistique quantitative. In *Encyclopaedia Universalis* ; 1988.

[VOO86] Voorhees, Ellen M. - Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing & Management.* 22(6) : 465-476 ; 1986.

[VRI79] Van Rijsbergen, C. J. - *Information Retrieval.* London : Butterworth, 1975.

[WAD88] Wade, Stephen ; Willet, Peter ; Robinson, Bruce ; Vickery, Brian ; Vickery, Alina. - A comparison of knowledge-based and statistically-based techniques for reference retrieval. *Online Review.* 12(2) : 91-108 ; 1988.

[WAL87] Waltz, David L. - Applications of the Connection Machine. *Computer.* 20(1) : 85-97 ; 1987.

[WATT68] Watt, W.C. - Habitability. *American Documentation.* 338-351 ; julv 1968.

[\VHI89] White, Howard D. ; McCain, Katherine M. - Bibliometrics. *Annual Review of Information Science and Technology (ARIST).* 24 : 119-186 ; 1989.

[WIL84] Willett, Peter. - A note on the use of nearest neighbors for implementating single linkage document classifications. *Journal of the American Society for Information Science.* 35(3) : 149-152 ; 1984.

[WIL88] Willett, Peter. - Software and hardware techniques for string searching in serial document databases. *World Patent Information.* 10(2) : 120-129 ; 1988.

[WIL89] Willett, Peter. - Textual and chemical information processing using parallel computer hardware. *Journal of Information Science.* 15 : 223-236 ; 1989.

[WILC88] Wilcox, R. O. ; Quinn, M. E. ; Jensen, I. N. - The Telebase implementation of Common Command Language. *12th International Online Information Meeting. Proceedings. Tome 1. Londres. Décembre 1988.* Learned Information. 507-513 ; 1988.

[WOO89] Woodin, S. J. - Environmental effects of air pollution in Britain. *Journal of Applied Ecology.* 26 : 749-761 ; 1989.

[YAK87] Yakubovitz, Z. - Associative retrieval system. *Contemporary Topics in Information Transfer.* 4 : 209-214 ; 1987.

[ZAD84] Coping with the imprecision of the real world : an interview with Lotfi A. Zadeh. *Communications of the ACM.* 27(4) : 304-311 ; 1984.

[ZAR88] Zarri, Gian Piero. - Etat de l'art : les nouvelles tendances de l'informatique documentaire. *Bulletin du C.I.D.* 32 : 11-40.

[ZAE90] Zarri, Gian Piero. - A knowledge representation language for large knowledge bases and "intelligent information retrieval systems". *Information Processing & Management.* 26(3) : 349-370 ; 1990. "

[ZIN90] Zink, Steven D. - Planning for the perils of CD-Rom. *Library Journal.* 51-55 ; February 1, 1990.

[ZL077] Zloof, M. M. - Query By Example, a database language. *IBM systems journal.* 16(4) : 324-343 ; 1977.