Numéro d'ordre: 34699 1999

THESE

Présentée

Devant **l'UNIVERSITE CLAUDE BERNARD – LYON 1**

Pour l'obtention du

Diplôme de Doctorat de Science de l'Information et de la Communication

Présentée et soutenue publiquement le

17 décembre 1999

Par

Madjid IHADJADENE

LA RECHERCHE ET LA NAVIGATION DANS UN SYSTEME DE RECHERCHE D'INFORMATION GRAND PUBLIC

Le cas des hypercatalogues sur l'Internet

JURY :

Richard Bouché	Professeur en sciences de l'information et de la communication, Enssib	*Directeur de thèse*
Jean-Paul Metzger	Professeur en sciences de l'information et de la communication, Université Lyon3	*Rapporteur*
Jacques Perriault	Professeur en sciences de l'information et de la communication, Université Paris 10	*Rapporteur*
Michael Buckland	Professeur school of information management and systems, University of California, Berkeley	
Hubert Fondin	Professeur en sciences de l'information et de la communication, Université Bordeaux III	
Steven Pollitt	Directeur du Cedar, school of computing and mathematics, University of Huddersfield	

Table des matières

Liste des figures

Liste des tableaux

Introduction

L'introduction de l'Internet a complètement bouleversé le monde de l'informatique documentaire. Elle a aussi favorisé la multiplicité des ressources documentaires grand public. Dans les bibliothèques, elle a eu beaucoup de conséquences dont deux nous semblent importantes pour la recherche d'informations :

- Doter les catalogues en lignes (OPACs) d'une interface graphique conviviale et hypertextuelle à travers le Web : les WWW-OPACs, constituant ainsi un hypercatalogue.
- La possibilité d'un accès à distance à ces WWW-OPACs .

L'expansion des systèmes de recherche de l'information (SRI) grand public (catalogues en lignes, bibliothèques numériques, moteurs de recherches) entraîne non seulement une multiplication et une diversification des usagers mais aussi une hétérogénéité croissante des documents (monographies, rapport, multimédia, pages Web,…). Cette double évolution n'a pas modifié l'objectif fondamental de la recherche d'information: le repérage de l'information pertinente avec le maximum de précision.

Plutôt que de parler d'usagers « grand public », nous parlerons dans cette thèse de système d'information grand public. Aujourd'hui, le minitel, l'Internet ou les catalogues de bibliothèques peuvent être employés par des usagers sans aucune médiation, ni aucune formation aux techniques documentaires.

Si l'usage des SRI spécialisés par les documentalistes ou des experts d'un domaine, a été étudié de longue date, en revanche il n'en était rien pour l'emploi d'un catalogue en ligne par des usagers en situation réelle, en France. Dans la conclusion de l'enquête PARINFO, les auteurs (Hassoun &Roger, 1994) du rapport ont indiqué "*que dans le cas des catalogues en lignes, il est primordial de pouvoir recueillir des données sur leur usage réel*". Trois ans seulement après ce projet national, nous effectuons la première étude du genre en France afin de mieux comprendre les stratégies de recherche et de navigation des usagers et d'étudier le problème de la surabondance de l'information pour pouvoir proposer un système de recherche enrichi de nouvelles fonctionnalités. Nous avons pris en compte trois facteurs aussi bien dans les différentes études d'usages que nous avons effectué que dans la conception du nouveau système :

- Le problème du vocabulaire dans la recherche d'information
- La navigation dans les SRI
- La surabondance d'informations

Le problème du vocabulaire dans la recherche d'information

La recherche documentaire est une tâche assez complexe qui nécessite la mise en relation d'un besoin d'information imprécis et le contenu d'un catalogue. Pour mener sa recherche, l'usager doit en être en mesure de maîtriser trois types de savoir-faire (Borgman, 1996): le premier est procédural et relève de la maîtrise instrumentale du système comme celles du clavier, des commandes et des touches de fonction. Le second est d'ordre sémantique : comment définir et formuler ce qu'on l'on cherche d'une façon compréhensible par le catalogue ? comment utiliser les opérateurs booléens ? comment employer les différentes options du système ? Le dernier est d'ordre conceptuel. Il concerne la capacité de l'usager à définir son besoin d'information, à élaborer une stratégie de recherche pour affiner et/ou élargir les résultats obtenus. Actuellement, ce sont ces deux derniers qui posent le plus de problèmes.

Les usagers d'un catalogue, comme ceux qui utilisent les moteurs de recherches, ne sont pas que des professionnels de la documentation. Interroger un catalogue est une tâche particulièrement difficile car l'usager ne possède pas toujours les termes nécessaires qui lui permettent d'exprimer ses besoins d'informations. La recherche documentaire est un processus interactif qui se présente comme une suite de formulations et de reformulations de requêtes jusqu'à la satisfaction des besoins d'information. La requête initiale permet rarement d'aboutir à un résultat qui réponde entièrement à l'attente de l'usager. Celui-ci peut en effet avoir des difficultés à exprimer clairement les concepts sur lesquels porte sa recherche. L'efficacité d'une recherche documentaire nécessite souvent la connaissance de la façon dont la base documentaire a été indexée. La possibilité de visualiser une sélection de documents et de termes et donc de compléter sa recherche, ouvre de nouvelles perspectives pour les usagers d'un système de recherche grand public.

La navigation dans les SRI

Toutes les études effectuées sur les catalogues en ligne traditionnels mais aussi les SRI spécialisés aboutissent à la nécessité et l'urgence d'inclure des options de navigation.

Plusieurs chercheurs ont présenté la navigation non linéaire comme la solution la plus performante pour les usagers novices. Parmi toutes les solutions qui sont proposées pour améliorer la recherche dans les catalogues, l'hypertexte est celui qui a été reçu avec le plus d'enthousiasme et aussi celui qui a été le plus objet de critiques. Pour les premiers, la navigation dans l'espace des concepts et dans la base des documents rend le catalogue plus accessible. La navigation hypertexte permet de découvrir l'information au fil des opérations sémantiques et permet à l'usager d'utiliser l'information présente à l'écran pour lancer d'autres recherches. Les seconds contestent fortement cette approche. Selon eux, il est difficile de l'inclure dans les catalogues opérationnels. De plus, l'usager va se perdre dans l'hyperespace « l'hyperspace ». La généralisation du Web a bouleversé complètement cette dichotomie. Les catalogues sont maintenant sur l'Internet. Le premier objectif de cette thèse est donc d'étudier ce changement afin de savoir dans quelle mesure il fait évoluer les usages et les stratégies de recherche. Il s'agit aussi d'observer le comportement des usagers dans différents contextes pour voir en quoi la navigation facilite la recherche d'information ?

La surcharge d'information

L'augmentation constante du volume d'information dans le monde, la mise en réseau et l'interconnexion des système de recherche d'information rend de plus en plus aigu le problème de la surcharge d'information. Il existe un décalage entre les demandes formulées et les possibilités du catalogue. Les usagers expriment leurs besoins à un niveau plus générique que celui des vedettes matières de la base bibliographique. Par conséquent, leurs interrogations aboutissent souvent à des surcharges d'information. Ce problème fut longtemps négligé dans les études sur les catalogues en ligne. Si l'on excepte la thèse de Larson (1986), il y a très peu de travaux sur ce domaine de recherche.

L'intérêt du prototype CATHIE (CATalog Hypertextuel Interactif et Enrichi) que nous avons développé est qu'il propose à l'usager un ensemble d'outils et de solutions pour mieux préciser ses demandes et filtrer les réponses obtenues. CATHIE associe la richesse des vocabulaires contrôlés (RAMEAU et la classification décimale de Dewey), les possibilités de visualisation et de navigation de l'hypertexte et la puissance du modèle probabiliste.

Plan de la thèse

Ce mémoire de thèse est organisé en trois parties. Dans la première partie (chapitres 1, 2 et 3) nous présentons d'abord les deux paradigmes qui dominent actuellement la recherche en informatique documentaire puis nous abordons en détail la description des hypercatalogues.

La deuxième partie est particulièrement consacrée à l'évaluation du comportement des usagers. Nous nous attacherons dans le chapitre quatre à étudier le corpus recueilli à la bibliothèque publique d'information (BPI) pour mieux comprendre les interactions verbales entres les usagers et les bibliothécaires. Cette partie portera aussi bien sur la structure du dialogue, que sur les stratégies utilisées par les bibliothécaires pour parvenir à de meilleurs résultats de recherche. Bien entendu une attention toute particulière sera accordée au processus de définition et de sélection des termes.

Les trois chapitres suivants précisent la partie empirique de la thèse. Nous analyserons les erreurs effectuées, le taux d'échec et de surcharge pour chaque point d'accès ainsi que les tactiques mises en œuvre par les usagers pour contourner ces deux problèmes. Le huitième chapitre décrit le processus de navigation des usagers. Pour cela, nous utiliserons l'analyse des traces informatiques (les fichiers logs) comme principale méthode de recueil de données. L'analyse des comportements des utilisateurs qui accèdent à distance fera l'objet du neuvième chapitre.

Dans la dernière partie, nous terminerons l'exposé de ce travail par quelques précisions sur la réalisation informatique, tant au niveau de l'architecture générale de CATHIE qu'à celui du fonctionnement spécifique. En conclusion, nous verrons les limites et les extensions possibles de cette étude.

Les problèmes liés à la recherche et la navigation dans un SRI sont interdisciplinaires par nature. Notre travail a touché particulièrement à trois spécialités : la bibliothéconomie, l'interaction homme-machine et l'informatique documentaire.

Partie I : état de l'art

Chapitre Un : Quel paradigme pour les systèmes de recherche d'information ?

L'évaluation des systèmes de recherche d'information (SRI) est un thème récurrent et relativement ancien puisqu'il date des années soixante. Nous situerons cette évaluation selon les deux cadres de pensée (paradigmes) qui dominent actuellement la recherche en informatique documentaire:

1. Le paradigme système
2. Le paradigme usager ou qualitatif

Le premier, connu sous le nom de paradigme physique, est le " the matching paradigm". Il constitue depuis plus de trente ans, le paradigme dominant en informatique documentaire. D'une façon générale, ses partisans considèrent que ce sont les fonctions de traitement de l'information, notamment celles de l'appariement entre les requêtes et les descriptions des documents, qui constituent le cœur du système. A l'inverse de ce courant de recherche, les tenants du paradigme usager considèrent que l'attention doit être davantage portée sur les besoins de l'usager et son environnement. Pour Dervin (1986), les usagers sont actifs et construisent leurs connaissances lors de leurs recherches. Elle considère qu' il faut examiner le système d'information uniquement d'un point de vue de l'usager. Il s'agit désormais d'étudier et d'évaluer comment les usagers définissent et reconnaissent leurs besoins d'information dans différentes situations. Comment les formalisent-ils? Comment présentent-ils leurs besoins au système ? Comment utilisent-ils les fonctions du système pour satisfaire à leurs besoins d'information ?

Meunier (1997) présente une autre typologie. Il considère les actions cognitives d'accès au contenu comme un système de traitement d'information (STI) . Il présente cinq modèles : cybernétique, documentaire, linguistique, intelligence artificielle, emérgentiste.

1.1. Le paradigme système : « the matching paradigm ».

L'idée sous-jacente à ces systèmes, est que le degré d'appariement entre les termes de la base et ceux de la requête de l'usager permet d'indiquer la pertinence des documents retrouvés. L'objectif essentiel consiste à améliorer les performances de recherche selon deux mesures: le bruit et la précision. Les chercheurs ont constamment amélioré les procédés d'appariement, au travers de divers modèles.

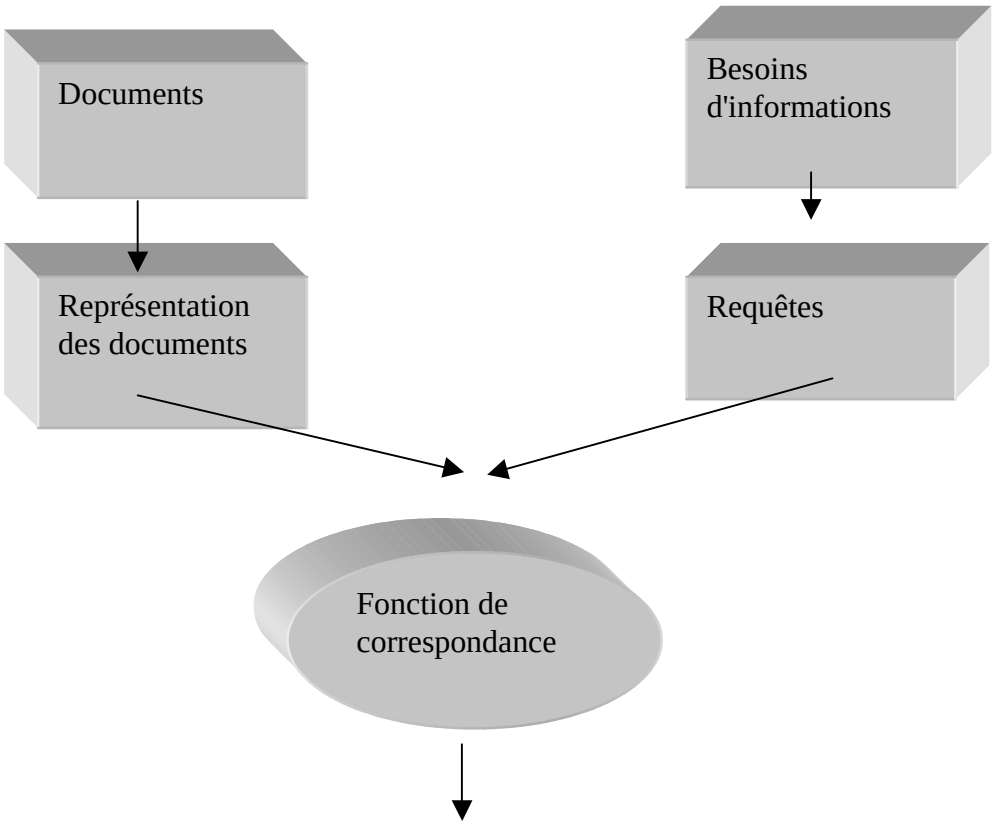


Figure 1 : le modèle d'un SRI classique

1.1.1. Description de quelques modèles de RI

Voici d'une façon condensée la description des principaux modèles qui dominent encore les SRI opérationnels.

1.1.1.1. Le modèle booléen ou ensembliste

Dans ce modèle , un document (D) est représenté par un ensemble de termes (ti). Une requête est une expression logique composée de termes assemblés par les opérateurs logiques ET, OU et SAUF . L'approche booléenne se définit par:

a) une indexation des documents par un ou plusieurs descripteurs non pondérés.

b) une formulation de requête qui se base sur les trois opérateurs booléens:

b1) la conjonction (connecteur ET), exige que les termes soient présents simultanément dans la description d'un document.

b2) la disjonction (connecteur OU) exige qu' au moins un des termes soit présent dans la description des documents retrouvés.

b3) la négation (connecteur SAUF) permet d'éliminer les documents possédant un terme particulier

c) les documents sont retrouvés suivant la présence ou l'absence des termes utilisés dans l'équation de recherche.

L'approche booléenne utilise le mode d'appariement exact qui consiste à ne restituer que les documents répondant exactement à la requête. L'évaluation de la fonction de correspondance R entre un document D et les différentes formes de requêtes, est donnée par les équations suivantes :

- $R(D, t_1 \text{ ET } t_2) = R(D, t_1) . R(D, t_2)$
- $R(D, t_1 \text{ OU } t_2) = R(D, t_1) + R(D, t_2) - (R(D, t_1) . R(D, t_2))$
- $R(D, t_1 \text{ SAUF } t_2) = R(D, t_1) . (1 - R(D, t_2))$
- $R(D, t) = \begin{cases} 1 \text{ si t appartient au document D} \\ 0 \text{ sinon} \end{cases}$

Les limites de l'approche booléenne sont bien connues:

- Le succès d'une recherche booléenne dépend essentiellement du degré de maîtrise de la logique booléenne. Le sens des opérateurs booléens (ET, OU, SAUF) est diffèrent du sens qu'ils ont généralement dans la langue quotidienne. De plus, le poids relatif des concepts au niveau de la requête ou à l'intérieur du document n'est pas pris en compte.
- Les documents ne sont pas classés et leur nombre est difficile à contrôler.

- Les documents importants mais qui ne sont pas indexés par les termes de la requête, ne sont pas sélectionnées

Pour remédier aux inconvénients du modèle booléen, Salton (1983) a mis au point une technique dite "booléenne étendue". Le principe de cette approche est d'attribuer aux termes de l'équation booléenne des poids et de considérer les opérateurs de l'équation comme des distances entre les requêtes et les documents. La plupart des systèmes de recherche opérationnels sont actuellement basés sur l'approche booléenne traditionnelle.

1.1.1.2. Le modèle vectoriel

Dans ce modèle les termes d'indexation forment une base de l'espace vectoriel dans lequel chaque document est un vecteur représenté par une combinaison linéaire de ces termes:

$D= a_{i1}t_1 + a_{i2}t_2 + ...+ a_{in}t_n$

Les coefficients a_{in} sont les poids du terme Tj pour le document D .

Une requête Q est aussi formalisée par un vecteur de termes de recherche pondérés:

$Q=b_{i1}t_1 + b_{i2}t_2 + ...+ b_{in}t_n$

b étant le poids du terme t dans la requête.

Durant le processus d'évaluation de la pertinence d'un document en regard de la question Q, le système sélectionne les documents D dont la similarité avec la requête est la plus grande. Pour calculer cette similarité, on peut utiliser diverses méthodes (Salton,1983) comme celle du cosinus, de Jaccard, de Dice ou le produit d'Inner. La fonction cosinus consiste à mesurer le cosinus de l'angle des deux vecteurs Q ET D. Les documents sont alors retrouvés en fonction de leur proximité avec la question . Les réponses sont classées et restituées en fonction de cette mesure. Les travaux de (Larson,1992) sur l'utilisation de ces méthodes montre que la mesure de similarité qui correspond au Cosinus donne les meilleurs résultats en terme de précision et de rappel.

Cette possibilité de jugement de pertinence est utilisée pour reformuler automatiquement les questions en prenant en compte les documents pertinents et non pertinents.. Le prototype SMART (ftp://ftp.cs.cornell.edu/pub/smart) est un exemple du modèle vectoriel. L'affirmation de l'indépendance des termes d'indexation rend ce modèle trop réducteur.

1.1.1.3. Le modèle probabiliste

Ce modèle est basé sur le fondement de la théorie de la décision où le but est de calculer la probabilité que le document soit pertinent par rapport à une demande formulée par un usager (Van Rijsbergen, 1979). Etant donné une requête Q, elle découpe la banque de données en deux ensembles: les documents pertinents et ceux qui ne le sont pas. La probabilité que la réponse soit pertinente étant donné D, est exprimée sous une forme conditionnelle :

P (pertinent / D) = P1

D'après le théorème de Bayes:

$$P1 = \frac{Ppertinent(D) \times P(pertinent)}{P(pertinent) \times Pp(D) + P(n\ per) \times Pn.p(D)}$$

Pp (D) et Pn.p (D) représentent respectivement la probabilité pour D d'être un document pertinent ou non pertinent.

P(pertinent) et P(n per) sont respectivement la probabilité a priori de trouver des documents pertinents et non pertinents.

Pour un corpus donné, le calcul de P1 est réduit à celle de Pp (D) et Pn.p (D). Un moyen de les estimer consiste à établir les probabilités de pertinence et de non-pertinence d'un document en fonction des probabilités des termes contenus. On estime cette probabilité selon la pondération des mots de la question et des termes d'indexation.

Le poids des termes est souvent déterminé en fonction :

- De la fréquence relative du terme T dans le document
- De la fréquence relative du terme T dans la collection
- De la taille du document

Nous verrons dans le chapitre neuf (cf 10.3) , qu'il existe une grande diversité dans les procédés de calcul de la pertinence d'un document par rapport à une requête. Nous utiliserons l'une de ces méthodes pour introduire un classement des notices bibliographiques dans notre prototype. Les études de (Larson,1992) ont montré que ces méthodes donnent de meilleurs résultats que les méthodes vectorielles.

1.1.1.4. Les modèles linguistiques

Cette approche fait appel aux techniques de l'intelligence artificielle pour analyser et comprendre les concepts contenus dans les documents. Le but est de rendre l'accès à l'information plus aisé, en permettant à l'utilisateur de formuler sa requête en langage naturel, évitant ainsi à l'usager novice de connaître et de maîtriser l'usage des opérateurs booléens. L'analyse linguistique peut être décomposée en cinq niveaux :

- Un niveau morphologique pour identifier les mots en tant que chaîne de caractères.
- Un niveau lexical qui effectue le reconnaissance des formes diverses des mots (conjugués , dérivés).
- Un niveau syntaxique qui propose un compte rendu de l'agencement des mots dans une phrase et indique les structures des phrases.
- Un niveau sémantique qui identifie le sens du mot selon son emplacement dans la phrase. Une analyse sémantique utilise généralement une base de connaissance comprenant des relations sémantiques entre les termes.
- Un niveau pragmatique qui identifie le contexte des phrases.

L'évaluation de la requête a pour objectif de mettre en correspondance les termes fournis par l'usager et ceux du système (terme d'indexation). Cette opération est réalisée via un appariement syntaxique et sémantique. Une fois cette transformation réalisée (le passage du langage naturel au langage de représentation du système), on aboutit à une expression booléenne de termes d'indexation permettant la sélection des documents. Ces derniers sont ordonnés selon la valeur de la mesure de correspondance avec la requête.

Remarque

Les chercheurs ont constamment amélioré les procédés d'appariement, en proposant divers modèles pour l'appariement et le traitement des requêtes. On peut citer :

- L'approche connexionniste (Boughanem,1992)
- Les systèmes experts (Alberico,1990)
- L'approche logique et sémantique (Puget, 1993)

- Le raisonnement à base de cas (Smail,1997)
- Les algorithmes génétiques (Korfhage,1998)
- L'hypertexte (Wolfram,1996)

1.1.2. Evaluation des SRI

Depuis le début des années 60 jusqu'aux récentes évaluations de TREC[1] (Text Retrieval Experiment Conference) , plusieurs études sont menées pour évaluer les SRI. De nombreuses variables sont employées pour mesurer leurs performances. Depuis, les travaux de Cleverdon (Cleverdon,1966), les partisans de l'approche système ont toujours privilégié les mesures de performance que sont le rappel et la précision.

- **Le rappel**

Le rappel mesure la capacité du système à trouver, pour une requête, tous les documents pertinents : il s'agit de la proportion de documents pertinents retrouvés par rapport à l'ensemble de documents pertinents existant dans le système.

- **La précision**

La précision mesure la capacité du système à trouver, pour une requête, uniquement les documents pertinents. C'est le pourcentage des réponses correctes parmi les documents retrouvés.

1.1.3. Critiques et limites de l'évaluation classique des SRI.

Divers chercheurs ont montré les limites de l'évaluation de ces systèmes de recherche et ont rejeté une partie de ces travaux. La nature de ces critiques varie selon les auteurs et selon le but poursuivi. Pour les uns, l'objectif de ces critiques est d'apporter une amélioration aux systèmes existants et proposent divers facteurs souvent négligé dans les évaluations , d'autres remettent en cause en totalité l'approche des SRI (Dervin,1986).

On peut regrouper ces limites en trois catégories

[1] http://www.nist.org/trec

1.1.3.1. L’absence de l’usager dans le processus d’évaluation

L'une des premières et sans doute la plus importante critique est liée à l'absence des usagers dans le processus d'évaluation des systèmes de recherches classiques. Ces derniers utilisent des critères autres que ceux du rappel et de la précision lorsqu'ils initient ou terminent une session de recherche. De plus, ces évaluations ne tiennent pas compte du contexte dans lequel se fait la recherche puisqu’elles ne sont pas effectuées en situation d’utilisation réelle. Par conséquent, ni les besoins d’informations des usagers, ni leur système de pertinence ne sont effectivement pris en compte.

1.1.3.1.1. Le besoin d'information

L'analyse des besoins d'information n'est pas faite ou sous estimée puisqu'on considère, à tort, que les usagers arrivent avec un besoin spécifique et stable. On fait l'hypothèse que l'usager sait ce qu'il veut, et qu'il est à même de préciser les mécanismes de fourniture d'information (Le Coadic, 998).

Les partisans de l'approche usager considèrent que le besoin d’information n’est pas toujours défini et fixe, mais au contraire évolutif. Pendant la recherche, survient parfois l’apparition de nouveaux besoins d’information au gré des différentes informations rencontrées (Le Coadic,1998), (Kulthau,1993), (Dervin,1983). Ces derniers considèrent que :

- Le besoin d'information est difficile à évaluer et à satisfaire. Un obstacle se dresse entre le besoin d'information et l'usager à la recherche d'information. Une infime portion des besoins d'information s'exprime naturellement : ce sont les besoins qui se traduisent par une demande formelle. Mais le plus souvent l'usager n'arrive pas à exprimer ce besoin. Le dialogue avec un documentaliste lui permettra de s'exprimer et de préciser son besoin d'information.
- L’objectif principal d’un système d’information est d’ aider l’usager à formuler et à diagnostiquer son besoin d’information. Pour le professionnel, il s’agit de comprendre ce qui conduit un usager à rechercher de l’information.
- Dans de nombreux cas, l'usager n'est pas conscient de son besoin d'information. Dervin (1986) constate "*que les usagers ne se servent d'un système d'information que lorsqu'ils ont reconnu qu'ils ne savent pas quelque chose, qu'il y a une lacune (brèche) dans leurs connaissances*".
- Il existe une variété de besoins d'information.

- Le besoin d'information évolue dans le temps.

1.1.3.1.2. Le problème de la pertinence

De nombreux travaux théoriques et expérimentaux ont pour but de définir les caractéristiques de la pertinence en recherche d'information (Saracevic, 1996), (Schamber, 1994). Les recherches sur la définition de la pertinence sont anciennes, puisqu'elles datent des années 50. Le Coadic (1993) présente et synthétise une chronologie de ces recherches du début des années 1950 jusqu'à l'année 1992. Comme, il le constate, on peut regrouper ces études en deux groupes. Le premier groupe renvoie à la notion de pertinence objective (point de vue du système) et le second groupe rassemble les notions de pertinence subjective (point de vue de l'usager).

Pour les tenants du paradigme système, un "bon" système est celui qui trouve tous les documents pertinents avec un minimum de références non pertinentes. Plusieurs auteurs dont (Saracevic,1996) et (Shamber,1994) ont montré que les critères de performance (rappel/précision) sont insuffisants pour déterminer la pertinence d'une recherche. Dans l'approche traditionnelle on présume que les jugements de pertinence :

- sont des indicateurs d'efficience des SRI
- sont stables et ne varient pas dans le temps
- peuvent être réduits à un jugement binaire : un document est soit pertinent, soit non pertinent. Or, il est difficile de fixer un seuil de pertinence à partir duquel un document est pertinent.
- sont attribués indépendamment les uns des autres. Les jugements de pertinence doivent donc être nuancés. La pertinence d'un document est relative à l'ensemble des documents consultés

Les usagers utilisent des critères autres que ceux du rappel et de la précision qui restent étrangers à leurs préoccupations, lorsqu'ils initient ou terminent une session de recherches. Ainsi, pour arrêter une recherche d'information, les variables telles que le temps, le coût , le nombre de documents, la langue des documents jouent souvent un rôle déterminant (Harter,1997), (Wang,1994).

Les usagers effectuant la même requête et se trouvant confrontés aux mêmes ensembles de documents, peuvent avoir des avis différents sur la pertinence des documents. Leur jugement tient compte en effet de plusieurs facteurs liés au contexte de recherche (leur niveau de

connaissances, leur expérience du domaine, etc.) et à la nature de leur besoin en information. Une grande diversité de facteurs entre en jeu lorsqu'un individu évalue la pertinence d'un document. Pour Schamber (1994), la pertinence est subjective , cognitive et multidimensionelle . L'auteur a identifié quatre vingt facteurs qui agissent ou influent sur la pertinence d'un document.

Il existe plusieurs facteurs qui interviennent dans l'évaluation de la pertinence des réponses. Seuls une infirme partie peut être intégré dans le SRI. La fonction d'appariement ne prend pas en compte la complexité de la pertinence utilisateur. Dans le paradigme physique, les méthodologies d'évaluation des SRI reflète ce décalage qui existe entre les connaissances sur la complexité de la pertinence utilisateur et la modélisation réduite pauvre de cette pertinence dans les systèmes. En France, Denos (1997) a reprit une partie des études américaines sur la pertinence et propose dans sa thèse d'étendre la fonction de correspondance. Elle propose de fournir des moyens interactifs qui permettent à l'usager de contrôler le sens que prennent les critères formulés par le SRI et de visualiser les diverses combinaisons de ces critères. Parmi ces critères, certains font référence à des caractéristiques des documents qui sont rarement indexées comme par exemple la solidité scientifique d'un document textuel ou de la qualité esthétique d'un document iconographique

On ne peut pas parler d'un seuil de pertinence d'un document mais d'un système de pertinence de l'usager. Mucchielli (1993) notait à juste titre que le système de pertinence d'un individu *"est un état psychologique de prédisposition mettant en cause le cognitif, l'affectif, le perceptif et le comportemental. Il est en fonction de l'ensemble des problèmes spécifiques qui préoccupent l'individu, des projets qu'il a et qui forment son orientation de vie au moment où on le considère"*. Il y a donc une relation forte entre le projet d'un usager, ses buts et une perception sélective des facteurs qui peuvent agir sur son système de pertinence. Dans sa thèse, Wang (1995) illustre bien cette relation dans la sélection des documents.

1.1.3.2. Absence de l'interaction

Ces études ne prennent pas en compte le caractère dynamique et interactif d'une recherche d'information. Un modèle d'évaluation qui néglige l'interaction est irréaliste et inadéquat pour les SRI d'aujourd'hui. On peut considérer que depuis la généralisation de l'architecture client/serveur, les interfaces des SRI sont graphiques. Depuis 1994, ces interfaces sont de

plus en plus hypertextuelles. Il a fallu attendre la quatrième conférence de TREC pour aborder d'une façon limitée cet aspect interactif . Pour Balpe (1996) : « *Il semble difficile d'extrapoler de telles mesures (précision et rappel) à des systèmes navigationnels, où de nombreuses stratégies d'exploration sont possibles, et où en situation de travail réelle l'exploration en cours peut influer sur la nature et la formulation du problème initial* ».

On peut toutefois signaler les efforts entrepris actuellement dans le projet européen MIRA[2] (Multimedia Information Retrieval Applications) pour introduire d'autres méthodes d'évaluation issues de l'ergonomie cognitif ou du domaine des interfaces homme machine.

1.1.3.3. Les méthodes utilisées et les résultas obtenus

Un autre critique récurrente pour ces systèmes est liée aux conditions d'expérimentation de ces systèmes puisqu'il y a une simplification de la réalité. Les collections-test sont très limitées. Le calcul du rappel présuppose que l'on puisse donner le nombre de documents pertinents présents dans une collection, pour une requête donnée. Pour des environnements comme le Web, il est impossible d'estimer ce nombre. De plus, ce critère repose sur l'hypothèse que le calcul du nombre de documents pertinents peut se faire d'une manière objective. Or, plusieurs études ont montré le caractère subjectif de la pertinence

Comme ces évaluations sont effectuées sur de petites bases, elles ne permettent pas de transférer les résultats obtenus dans un contexte plus large (Blair,1984), (Harter,1997). C'est pourquoi ces résultats n'ont pas eu de grand impact sur le développement des SRI opérationnels. Pour Harter (1997) : « *the power of research, especially experimental research, is in its ability to explain, predict, and even control the phenomenon under investigation. Yet experimental research in IR using a Cranfield-based approach fails to possess even the weakest of these. It is probably for this reason that the findings of studies based on Cranfield intruments have played a relatively minor role in the development of commercial and other operational IR systems"*. Il est évident que, les problèmes posés par le catalogue de l'enssib (moins de 30.000 notices) sont différents de ceux de Melvyl (7 millions de notices).

[2] *http://www.dcs.gla.ac.uk/mira*

1.2 Les paradigmes usagers

Ces paradigmes considèrent l'usager, son comportement, son intention comme un élément central (en complément de l'importance de l'appariement) du processus de recherche d'information. Alors que les pionniers de l'approche système sont en majorité des ingénieurs et scientifiques comme Salton, Clervedon etc., les tenants de la seconde approche témoignent de l'arrivée de chercheurs issus d'autres disciplines telles que la psychologie, les sciences éducatives, les sciences sociales, la communication ou les sciences cognitives. Dans une étude basée sur la méthode de "co-citation " des 120 chercheurs les plus cités en science de l'information sur une période allant de 1972 à 1995, White (1998) montre bien cette distinction qui existe entre les partisans des deux approches.

Saracevic (1997a) explique que cette divergence est d'ailleurs liée en grande partie à la définition et au statut de " l'information" dans ces disciplines.

Pour lui, les chercheurs en science de l'information doivent considérer l'information comme un message traité d'un point de vue cognitif par un usager dans un contexte donné :

" *information involves not only messages (first sense) that are cognitively processed (second sense), but also a context, a situation; task, problem-at-hand, the social horizon, human motivations, intentions. For information science in general, and IR in particular, we have to use the broadest interpretation of information, because users and use are involved and they function within a context. That's what the field and activity is all about. That is why we need to consider IR in the broader context of information science*".

A l'inverse du paradigme système, il n'existe pas un seul paradigme usager mais plusieurs. Il est d'ailleurs difficile de situer les travaux de recherches entre ces différentes approches à l'intérieur du paradigme usager. Hert (1995) en a établi une typologie que l'on peut résumer en deux catégories générales :

- les approches cognitives.
- Les approches dynamiques.

D'un point de vue méthodologique, ces approches s'inspirent souvent des méthodes ethnographiques et socio-cognitives.

Les tenants de ces approches cherchent à expliquer la façon dont les usagers organisent leur pensées et leur activité. Ils soutiennent que les caractéristiques de l'usager représentent des

mesures primaires à étudier et que celles-ci restent constantes durant l'interaction avec le système. Chen (1992) considère que l'usager traite l'information à partir des représentations qu'il a du monde. Ces représentations peuvent être erronées. Il conçoit l'activité de l'usager comme une activité de résolution de problème. Ce courant de recherche est aussi représenté par les travaux Allen, Egan, Hewins, Shaw. D'autres auteurs (Shaw, Kerr, Fischoff…etc.) soutiennent que les performances d'un système sont aussi influencées par l'interface homme machine. Pour ces auteurs, un changement d'interface affecte le comportement de l'usager. Il s'agit d'étudier l'effet de certaines variables telles que la couleur, l'utilisation des icônes, le clignotement en complément du calcul, le taux d'erreurs, le temps de réponse, etc. Enfin, on peut aussi inclure dans cette catégorie, les travaux sur les différences individuelles. L'hypothèse sous-entendue est que les écarts dans les performances d'un SRI sont attribués aux différences individuelles des usagers (expériences, objectifs, etc.). Ce courant de recherche est représenté notamment par les travaux de Borgman.

Reprochant aux autres paradigmes d'avoir négligé la dimension sociale et culturelle de l'usager, les tenants de l'approche dynamique ou holistique considèrent que l'usager a une représentation située, déterminée en partie par sa formation, sa pratique et son histoire.

La notion d'action situé établit un cadre de recherche pour concevoir les activités cognitives dans le monde réel (Leplat, 1977). Conein (1994) présente deux interprétations de la notion d'action située; la première met l'accent sur la compréhension de l'action et la communication sociale; la seconde met l'accent sur la perception et l'organisation spatiale (environnements): "*définir une action comme située signifie généralement que l'on doit concevoir l'organisation de l'action comme un système émergent in situ de la dynamique des interactions. Mais cette dynamique peut en effet résulter de deux processus: soit de la compréhension que chaque participant a des actions de l'autre, soit de la perception des indices provenant directement de l'environnement immédiat*". Il existe plusieurs courants de l'action située allant de la théorie de l'activité (Vygotsky), la sociologie cognitive, l'éthnométhodologie jusqu'à à la cognition distribuée .

Selon Theuron (1994) " *l'activité est construite par les acteurs en situation, compte tenu de leur état et de leur culture, en fonction de leur activité antérieure dans cette même situation ou dans d'autres, y compris hors travail. D'autre part, ce qui pertinent dans cette situation, qu'il s'agisse d'éléments de l'environnements physique, de l'environnent sociale ou des taches, n'est pas prédéterminée, mais résulte de l'activité elle même*".

Beaudouin-Lafon (2000) a introduit le terme d'informatique situé (situated computing) pour décrire une approche de conception des systèmes interactifs qui prenne en compte le contexte et les situations d'usage. La notion de situation et du contexte est pour Beaudouin-Lafon, un des enjeux majeurs des interfaces post-WIMP.

Reprochant aux tenants de l'approche cognitive d'avoir négligé la dimension sociale et culturelle dans leurs analyse, un ensemble de chercheurs dont Fidel, Hert et Kuhlthau tentent de suivre une approche holistique. Il s'agit de considérer la cognition humaine comme multidimensionelle, située socialement et construite. Concernant l'évaluation[3], Hert (1995) et (Harter & Hert, 1997) recensent plusieurs approches et méthodes d'évaluation allant de l'analyse des erreurs, d'études d'utilisabilité jusqu'aux études longitudinales.

Fidel (1985), Hert (1995) et Kuhlthau (1993) suggèrent d'évaluer l'ensemble du processus de recherche des usagers et les stratégies qu'ils mettent en œuvre. Ils recommandent d'effectuer des études longitudinales et de considérer le contexte réel de l'activité des usagers dans sa complexité. Pour Hert (1995), la consultation d'un catalogue en ligne est déterminée par une série d'actions situées. Par action située, elle sous entend que le comportement de l'usager et ses actions ne sont pas complètement déterminés. Elle a établit un modèle de recherche d'information et a catégorisé l'ensemble des facteurs qui influent ce processus que nous avons repris dans le tableau ci-dessous.

> « *An OPAC interaction is a situated series of actions on the part of the user. By situatedness is meant that as a user moves through an interaction, his or her actions are not completely predetermined, instead elements of the situation are utilized to determine action. It is possible to understand the interaction in terms of a set of elements associated with the user, the project or problem in which the user was engaged, and the system response*."

[3] l'édition de janvier 1996 de JASIS est consacré à l'évaluation des SRI.

Elements Associated with the Respondent	Description	Elements Associated with the Respondent	Description
Attitudes concerning the role of the system in the information-seeking process	How respondents viewed what the system could do for them as well as how they would use the system	Mood or emotional state	Respondent's state of mind or predominant emotion
Attitudes concerning the role of the interaction in the information-seeking process	Respondent's perceptions of or beliefs about the role of the interaction in the larger process in which he or she was engaged.	Subject area of the project	The subject area as expressed by the respondent (without indication of particular facets)
Knowledge of facets of the topic area	Respondent's expressions about particular facets of the subject of the project	Specific requirements of the project	Requirements of the project which are imposed by sources external to respondent (e.g., the instructor)
Knowledge of system contents	Knowledge (or lack of knowledge) of what was in the system in terms of subject areas covered and types of materials	Point in the process	Point in the intellectual or logistical process of the project
Knowledge of system functionality	Knowledge (or lack of knowledge) of how to use particular commands as well as a general understanding of how the system operated	Whether the project was a group project or not	Respondent indication that he or she was working with a group on the project
Knowledge of other available resources	Knowledge that other resources outside of SUMMIT might be useful in project or problem	Nature of retrieved sets	Features of the sets retrieved with a particular search command
Knowledge of the specific requirements of the project	Knowledge (or lack of knowledge) about how to proceed in a project or what a project entailed	Status messages from the system	Messages which provide feedback when the respondent inputs an invalid command or gets no hits
Experience with SUMMIT or similar systems	Indication that experience with SUMMIT or another system judged similar by the respondent was influencing action	Investment in interaction or information-seeking process	The time available for searching as well as how important the person considered the project
Respondent expectations	Respondent expectatations about themselves or the system and its contents		

Tableau 1: éléments de situation identifiés par Hert

Kuhlthau (1993) propose un modèle qui décrit le processus de recherche selon six étapes (initiation, sélection d'un thème, identification, exploration, formulation, collection, présentation). Elle vérifie ce modèle d'un point de vue empirique en effectuant des études longitudinales durant plus de quatre ans**.** Elle incorpore dans ce modèle trois dimensions qui sont communes aux six étapes : la dimension affective, cognitive et physique. Concernant la dimension émotionnelle, elle montre que les usagers ont des attentes différentes mais surtout que leur comportement, leur perception de la difficulté de recherche et leur état évolue durant le processus de recherche (Tableau 2). A chaque stade, l'usage évolue de l'incertitude à la satisfaction ou à l'insatisfaction. Ses travaux ont montré qu'il est important qu'une relation de confiance s'établisse entre l'usager et le bibliothécaire pour mener à bien la recherche d'information. Elle propose de différencier l'aide que le système et/ou le bibliothécaire apportent selon les étapes de la recherche.

	affective	Actions
initiation	Incertitude, inquiétude, appréhension	Navigation dans les étagères, discussion avec d'autres collègues
sélection	Optimisme et anxiété	Recherche préliminaire dans la bibliothèque, consultation de médiateurs;
exploration	Confusion, augmentation du doute	Location de documents; lecture des documents trouvés; prise de notes, bibliographie
formulation	Clarté, confiance en soi	Lecture de notes
collection	confiance	Recherche de documents spécifiques, prendre des notes détaillées pour les mettre en bibliographie;
présentation	Satisfaction et/ou désappointement (déception)	Confirmer les documents pertinents trouvés et spécifier la bibliographie finale

Tableau 2: évolution de l'état et des actions des usagers durant le processus de recherche.

Plusieurs modèles ont été proposés tenant compte des avancées théoriques et expérimentales de l'approche orientée usager. On peut citer :

- Le modèle de Belkin (1995) : « the episode model »
- Le modèle cognitif de Ingwersen (1996)
- le modèle de Saracevic (1997a) « The stratified model » (figure2)

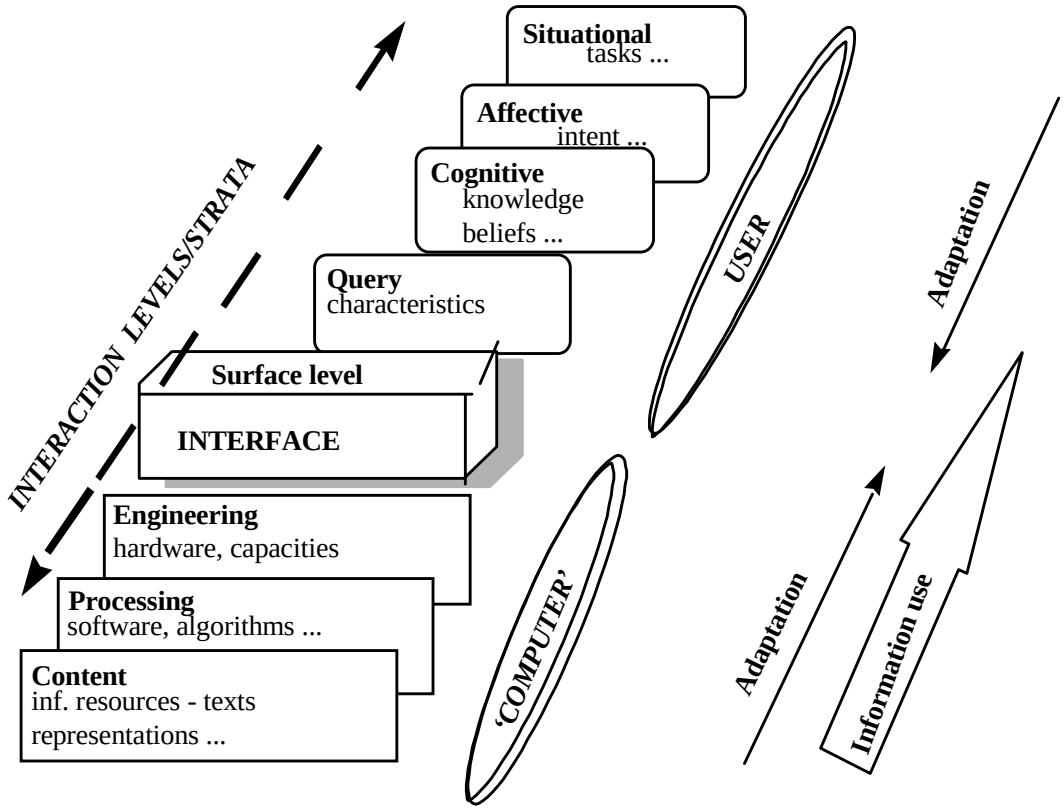


Figure 1. Elements in the stratified model of IR interaction

Figure 2 : le modèle de Saracevic

Le modèle de Saracevic a le mérite de proposer une vision globale du processus de recherche et d'interaction. Il faut considérer le contexte réel de l'activité des usagers dans sa dynamique

et sa complexité sans faire l'impasse sur l'impact des opérations de traitements de l'information (indexation, classification) et des modes d'appariements utilisés. Ce modèle permet de situer les travaux de recherches selon différentes couches :

- l'étude du contexte comme les tâches à effectuer, la précision des besoins d'information.
- la dimension psycho-sociale et socio-identitaire de la recherche qui comprend par exemple l'incertitude, les motivations et l'intention de l'usager.
- les aspects cognitif où l'usager interagit avec les représentations du documents.
- L'outil (ici l'ordinateur en tant que machine), ses capacités, sa puissance, …etc.
- Le traitement de l'information et les algorithmes d'appariements mis en œuvre.
- les propriétés de la requête de l'usager
- Enfin, les documents et leur représentations.

Comprendre le processus général de repérage de l'information, nécessite donc de prendre en compte la totalité de ces facteurs et les relations entre les différentes strates du modèle.

Dans notre thèse, on s'intéresserait particulièrement à trois strates : l'analyse des requêtes, l'évaluation de l'interaction et l'amélioration des algorithmes sans négliger les facteurs liés au contexte pour expliquer la démarche des usagers.

Conclusion

Depuis une dizaine d'années, un puissant courant de recherche s'est constitué autour des travaux de Dervin pour appliquer une approche orientée usagers aux SRI. Comme dans toute étude sur les usages, c'est *« l'homme qui est ici au cœur de l'investigation et non l'appareil. Ce parti pris suppose que l'on tienne le plus grand compte des contextes psychologiques, sociologiques, culturels, économiques, si l'on veut comprendre comment s'établit et se propage l'usage d'un appareil*" (Perriault,1989). Dans la majorité des études orientées usagers, on remarque qu'au delà de recommandations générales, rares sont ceux et celles qui ont conçu de tels systèmes orientés usagers ou à défaut développé des méthodes de conception de tels systèmes. Les partisans de l'approche « orientée usagers » se rendent compte de la difficulté d'introduire leurs résultats et observations dans la conception de nouveaux SRI. Comme le souligne Saracevic (1997b) "*we failed to bridge the gap between human and system side in designs of IR. This is still the major job for the next generation of information scientists and a major agenda item for the field as a whole*". Il ne s'agit plus de privilégier une approche par rapport à une autre (système vs usager) mais de les adopter ensemble. Ceci implique une réflexion sur les méthodes d'analyse mais aussi une confrontation des résultats des deux approches notamment en ce qui concerne l'évaluation des systèmes d'information. Dans son livre sur les usages Perriault (1989) montre l'intérêt et l'urgence de combiner les enquêtes statistiques avec des études de microanalyse pour *« construire des schémas d'enquête et des indications qui cernent de plus près, en les quantifiant, les usages, dans leur réalité et dans leur diversité »*. Suivant la tradition des études d'usages des SRI opérationnels, nous allons utiliser l'analyse transactionnelle (analyse des traces informatiques) comme principal moyen de collecte de données que nous compléterons par deux autres méthodes : les questionnaires et l'analyse des verbalisations des usagers. Les résultats de ces études empiriques constituent le point de départ pour la conception d'un nouveau catalogue de bibliothèque.

Chapitre Deux : les catalogues en lignes : état de l'art

2.1. Définition et objectifs d'un catalogue

2.1.1 Définitions

Un catalogue est un ensemble de notices catalographiques des documents d'un fonds documentaire rédigées selon des principes normalisés et classées afin de faciliter les recherches des utilisateurs . Un catalogue interactif ou OPAC (Online Public Access Catalog) est un système d'information bibliographique permettant l'interrogation directement par les utilisateurs finals. Ils ont été mis en œuvre dès le début des années 60 dans un certain nombre de bibliothèques , notamment dans les pays anglo-saxons

La première fonction du catalogue fut d'abord l'inventaire, puis est venue celle de décrire matériellement les documents de la bibliothèque. Il n'est pas le seul moyen offert aux usagers pour accéder à ce qu'ils cherchent. Un accès spatial est offert dans les collection en libre accès par le classement et par sa signalisation

Provensal (1997) constate que: " *la théorie des catalogues a permis de passer de la simple liste d'inventaire classée par ordre d'entrée (apparue dès les bibliothèques de l'Antiquité) aux répertoires classés par titre, par auteur puis dans le milieu du XVIIeme siècle au classement dictionnaire qui mêle ces différents accès. Au XIXeme siècle les fiches cartonnées rendent possible les classements et les accès multiples ».* D'un point de vue conceptuel , un catalogue interactif est un système de recherche d'informations bibliographique qui a les caractéristiques suivantes:

- Une base de données de notices bibliographiques décrivant les documents répertoriés dans le fonds des bibliothèques.
- Une interface utilisateur qui gère le dialogue entre l'usager et le système
- Une fonction d'indexation et d'interrogation de la base de données.
- Un ensemble de référentiels, langages documentaires, qui servent à décrire les différents champs d'une façon normalisée.

Dans le passé, il existait une différence significative entre les SRI bibliographiques et les catalogues en ligne (tableau 3).

	Catalogues en ligne	**SRI**
Population concernée	- public hétéroclite et non formé	- public spécialisé et formé aux techniques documentaires
médiation	- utilisation de l'OPAC sans intermédiaire	- interrogation des SRI avec l'aide d'un intermédiaire (bibliothécaire, documentaliste, archiviste)
Base de données	- le fonds documentaire est encyclopédique - contient des monographies et d'articles - les enregistrements d'une notice bibliographique sont souvent dépourvus de résumés ou de tables de matières - utilisation d'un format, MARC en général	- le contenu de la base est spécialisé - composée d'articles et de texte intégral - les notices documentaires contiennent des résumés.
Vocabulaires contrôlés	- emploi d'une liste d'autorité et de classifications encyclopédiques (classification décimale de Dewey, classification décimale universelle …etc)	- emploi d'un thésaurus
Fonction d'appariement	- système booléen	- système booléen mais aussi probabiliste et vectoriel
Coût	- consultation libre et gratuite	- consultation souvent payante

Tableau 3 : comparaison entre un catalogue et un SRI.

Actuellement, on peut affirmer que, hormis le contexte d'usage des catalogues (base encyclopédique d'ouvrages, usagers occasionnels), on ne constate pas de différence entre les deux systèmes. Les résultats des travaux sur les SRI peuvent être donc adaptés pour un perfectionnement de l'accès matière aux catalogues en ligne.

2.1.2. Les objectifs d'un catalogue

Charles Ami Cutter (1904) est sans doute le premier à énoncer les objectifs d'un catalogue :

1 : Rechercher : un catalogue doit permettre de retrouver un livre à partir d'un seul des ses éléments connus (auteur, le titre ou le sujet).

2 : Regrouper : il doit pouvoir montrer ce que possède la bibliothèque sur un certain auteur ou sur un certain type de littérature (dans un domaine donné)

3 : Assister : il doit aider et guider l'usager dans le choix d'un ouvrage d'après son édition, par ses caractéristiques physiques ; son caractère (littérature ou documentaire).

" 1 : to enable a person to find a book of which either :

the auteur
the person } *is known*
the subject

2 : to show what the library has :

by a given author
on a given subject
in a given kind of literature

3 : to assist in the choice of a book

as to its edition
as to its character (literary or topical)"

2.2. Les OPACs : le point de vue technique

Décrire un catalogue en ligne d'une façon technique, c'est d'abord examiner le noyau autour duquel s'organise le catalogue : le langage documentaire RAMEAU et le format MARC. Nous nous intéresserons dans cette thèse à RAMEAU.

2.2.1. La liste d’autorité RAMEAU (Répertoire d'Autorité-Matière Encyclopédique et Alphabétique Unifié)

C'est à partir de 1987 que la DBMIST et la bibliothèque nationale se sont associées pour gérer et diffuser en France un Répertoire des Autorités Matières Encyclopédiques et Alphabétiques Unifiés : RAMEAU. Auparavant, certaines bibliothèques utilisaient des listes de termes internes , d'autres n'offraient pas d'accès matière. RAMEAU est une liste d'autorité matière, cela signifie qu’elle est un référentiel servant à la description normalisée du champ SUJET des notices bibliographiques. RAMEAU a pour origine la liste des vedettes matières de l'université de Laval à Québec, qui est une traduction et une adaptation de la liste de vedettes matières de la Bibliothèque du Congrès de Washington: LCSH (Library Congress Subject Heading). Cette liste est gérée conjointement par la bibliothèque nationale de France et l'ABES. RAMEAU permet une indexation dite" précoordonnée" , ceci signifie que le sujet du document est décrit de façon synthétique au moyen de phrases assemblées par une syntaxe. Cette phrase est appelée une vedette matière construite (VMC) . Les éléments qui la composent sont les différentes entrées de la liste d'autorité. Selon leur place dans la vedette d'indexation, ces entrées ont le statut de tête de vedette ou de subdivisions (sujet, géographique, forme , chronologie). Il existe deux types de subdivisions (affranchies et non affranchies). Par subdivision affranchie, on entend subdivision applicable à une ou plusieurs catégories de vedettes. L'emploi de ces subdivisions étant libre, elles ne figurent pas obligatoirement dans la liste d'autorité sous toutes les vedettes sous lesquelles on peut les employer. Par subdivisions non affranchies, on entend des subdivisions spécifiques à la tête de vedette et inséparables de celle-ci. Elles sont toujours entrées sous la vedette qui les régit. Cette subdivision peut être propre à une seule vedette ou commune à un petit nombre de vedettes. Les vedettes matières construites comportant de telles subdivisions sont obligatoirement entrées dans RAMEAU.

RAMEAU comprend en plus du vocabulaire , les indications et règles qui permettent de construire des vedettes-matières. Il ne contient pas toutes les combinaisons que l’on peut rencontrer dans un tel fichier. Afin d'assurer l’homogénéité et la cohérence de l’indexation, il est impératif de respecter les règles de construction des vedettes matières. Dans une bibliothèque tout en permettant une description d’ouvrages issus d’un fonds encyclopédique, RAMEAU permet une indexation très précise. C’est l’un des avantages d’une liste précoordonnée.

2.2.1. 1. Structure et syntaxe générale

La syntaxe, selon laquelle les entrées sont assemblées dans RAMEAU , est en effet porteuse de sens : elle permet de préciser quels rôles les différents éléments de description que sont les entrées jouent les uns par rapport aux autres. Une vedette matière comporte toujours une *tête de vedette* qui exprime l'essentiel du sujet. Elle peut être un nom commun, un nom de personne physique ou morale (collectivité), un nom de lieu, un titre uniforme...etc. Elle peut aussi contenir après le premier élément retenu:

- *Des éléments rejets* :Hodgkib, Maladie de.

RAMEAU emploie la précision pour lever la polysémie d'un terme.

- *Des qualificatifs* : droit social (droit européen)
- *Des sous- vedettes* : médias**droit européen

La tête de vedette peut être suivie d'une ou plusieurs subdivisions: on parle alors de vedette matière construite (VMC) . Le choix de ces subdivisions obéissent à des règles de construction très précises. Pour décrire le contenu du document , le catalogueur est amené à construire une ou plusieurs vedettes matières. Il cherchera à établir les VMC les plus précises, suivant ainsi un des principes de l'indexation matière : la règle de spécificité. Cutter (1904) recommandait " *enter a work under its subject-heading, not under the heading of a class which includes that subject*". Le concept de spécificité est équivoque et ne peut être précisé qu'en fonction de facteurs tels que la nature de la collection et le public visé.

La syntaxe de RAMEAU est trop complexe. En conséquence, les VMC sont très longues et très éloignées du langage naturel. Il faut toutefois signaler et saluer le grand effort qui se fait actuellement par l'équipe de la BNF qui privilégie de plus en plus les expressions plutôt que les chaînes construites se rapprochant ainsi du langage naturel. Par exemple la vedette matière *Notaires ** Archives* devient *Archives notariales.*

L'ordre des subdivisions est généralement le suivant:

> *Tête de vedette **subdivision sujet **subdivision géographique **subdivision chronologique sujet** subdivision de forme.*

Un livre relatif à un congrès sur l'histoire du droit bancaire en Suisse au 19ème siècle donnera la vedette matière construite (VMC) suivante:

Banques ** Droit ** Suisse ** histoire 19th ** Congrès ;

Les subdivisions de sujet, données après la tête de vedette ou après une autre subdivision de sujet, servent à préciser un aspect particulier du sujet. Un sujet traité selon des points de vue différents nécessite la rédaction de vedettes matières distinctes, avec une même tête de vedette et une subdivision sujet différente. La subdivision géographique exprime le point de vue géographique. Utilisées sous les vedettes sujets ou géographiques, les subdivisions chronologiques permettent d'exprimer la période envisagée dans le document à indexer. Une VMC ne comportera pas plus d'une subdivision chronologique. On en distingue deux sortes :

- Les subdivisions chronologiques spécifiques, propres à un lieu ou à un sujet
- Les subdivisions chronologiques communes, qui peuvent être employées là ou il n'y a pas de découpage chronologique prévu dans RAMEAU.

Parfois, ces subdivisions chronologiques ne servent qu'à séparer et partager une documentation trop abondante. C'était l'un des objectifs de leur introduction dans la liste LCSH.

Enfin, les subdivisions de forme permettent d'expliciter la présentation matérielle d'un document. Elles précisent la forme particulière des documents se rapportant à un sujet. Elle n'est pas obligatoire. Une VMC ne comporte qu'une seule subdivision de forme. Les principales subdivisions de forme sont: Biographies, Bibliographies, Dictionnaires; Examens; Manuels d'enseignements; Statistiques, Congrès, Guide touristique; Index; Journées d'étude, Logiciel; Traité, Iconographie...etc.

2.2.1. 2. Les renvois dans RAMEAU

La liste d'autorité RAMEAU admet trois catégories de relations permettant ainsi de relier les différents termes de la liste. On retrouve les relations hiérarchiques (générique et spécifique), la relation d'équivalence et celle d'association.

Les renvois d'équivalences:

RAMEAU réduit l'usage du vocabulaire en proscrivant l'emploi des synonymes ou quasi-synonymes. Seuls les termes retenus peuvent servir à l'indexation. Les termes exclus font l'objet de renvois d'équivalences, sous la mention "VOIR".

Exemple :

OPACs **VOIR** Catalogues en ligne

Catalogues informatisés **VOIR** Catalogues en ligne

Dans une notice , tout terme précédé de **EP** (employé pour) est synonyme ou quasi-synonyme de la vedette d'autorité. Il faut cependant remarquer que les renvois d'exclusion (formes rejetées) ne portent que sur les autorités et non sur les vedettes matières construites.

Exemple:

Catalogues en ligne

- EP Catalogues automatisés
- EP Catalogues informatisés
- EP Catalogues interrogeables en ligne
- EP CIEL (catalogues)
- EP Online public access catalogs
- EP OPAC

Les relations d'associations

Dans une notice, tout terme précédé de **TA** est associé à la vedette d'autorité, réciproquement, on trouvera le lien symétrique dans la notice correspondante.

Exemple:

Indexation (documentation)

TA Indexation (documentation)

TA Synonymes et antonymes

TA Vedettes-matière

Les relations hiérarchiques

Elles sont de deux types:

- **Les relations génériques**: dans une notice, tout terme précédé de TG est générique de la vedette d'autorité
- **Les relations spécifiques**: dans une notice, tout terme précédé de TS est spécifique de la vedette d'autorité

Ce sont deux liens inverses.

Exemple :

Logement

. TS Autoconstruction

. TS Discrimination dans le logement

. TS Fonctionnaires ** Logement

. TS Habitations ** Possession

. TS Logement rural

. TS Ménages

. TS Minorités ** Logement

- TG Etablissements humains
- TG Problèmes sociaux

L'introduction des relations (TG, TA, TS) dans les listes d'autorité est récente. Ce n'est qu'à partir de 1986 que la bibliothèque du congrès a décidé de remplacer les renvois « voir » et « voir aussi » par les relations de synonymie, d'association et de hiérarchie.

Le tableau suivant présente les caractéristiques et les différences principales entre une liste d'autorité et les thésaurus (Hudon, 1993).

Listes de vedettes matière	Thésaurus
Vocabulaire contrôlé pour l'indexation et la recherche de documents dans les collections documentaires	Vocabulaire contrôle pour l'indexation et la recherche d'information dans les documents.
Outil de type pragmatique dont l'organisation interne reflète celle d'une collection[4] documentaire *(literacy warra*nt)	Outil de type philosophique dont l'organisation reflète celle d'un champs d'études ou d'une discipline
Peut décrire l'ensemble des connaissances, en introduisant une structure hiérarchique lâche. (*encyclopédiques*)	Doit décrire un domaine spécialisé de la connaissance, puisque la structure hiérarchique doit être solide et rigoureuse
Unité de base (vedette matière) représente un sujet, c'est à dire une combinaison de concepts.	Unité de base (descripteur) représente un concept unique
Unité de base (vedette matière) est pleinement signifiante	Unité de base (descripteur) n'est signifiante que dans les rapports (*relations*) qu'elle entretient avec d'autres descripteurs
Créé et structuré pour des bibliothécaires et peu utilisé par l'usager	Créé et structuré pour une grande variété d'utilisateurs d'information .

Tableau 4 : comparaison entre une liste d'autorité et un thésaurus

2.2.2. Rameau et les notices bibliographiques

Dans cette partie, nous analysons les travaux de (Markey,1994) relatifs aux possibilités d'un appariement entre les vedettes matières telles qu'elles existent dans la liste d'autorité LCSH et les vedettes matières construites dans une grande bibliothèque (celle de l'Université de Michigan). Comme RAMEAU, La liste LCSH fournit des règles pour construire des vedettes matières mais ne liste pas l'ensemble des termes déjà crées dans le catalogue, de telle sorte

[4] Celle de la bibliothèque du congrès pour le LCSH et celle de la BNF pour RAMEAU

que souvent , le catalogue peut contenir des vedettes qui ne sont pas répertoriées dans la liste d'autorité. Markey (1994) montre que la majorité des vedettes de la liste LCSH n'admettent pas de subdivisions (70.4%) alors que dans une base bibliographique c'est l'inverse (presque 86.3%). Elle montre qu'il y a tout au plus 45 % d'appariement entre les sujets existants dans l'index d'un catalogue de grande bibliothèque et la bande des LCSH.

	% LCSH-mr (n=128367)	% notices bibliographiques (n=34279)
Tête de vedette (a)	70.4%	13.7%
a + une subdivision	25.9%	36.3%
a+2 subdivisions	3.5%	37.7%
a+3 subdivisions	0.2%	11%
a+4 subdivisions	0.0 %	1.2%
a+5 subdivisions		0.1%
a+6 subdivisions		0.0%
Total	100 %	100%

Tableau 5 : appariement entre la liste LCSH et une base documentaire d'après (Markey,1994)

Nous avons vérifié ces données par rapport à RAMEAU, et dans un contexte différent (une bibliothèque spécialisée : celle de l'enssib). Nous avons analysé un échantillon de 900 notices bibliographiques du catalogue de l'ENSSIB. Comme le montre le tableau suivant, plus de 60 % des vedettes matières admettent des subdivisions (Ihadjadene, 1998c).

type de subdivisions	nombre	pourcentage
A	490	37%
a+1 (ax, ay, az)	627	47.4%
a+2 (axy, axx, axz...)	184	13.9%
a+3 et plus	22	01.7%
Total	1323	100%
nombre moyen de vedette par notice	1.47	
nombre moyen de mots par vedette matière	3.1	

Tableau 6 : catégories des subdivisions dans le catalogue Loris de l'enssib.

a: tête de vedette, x: subdivisions sujet ou de forme, y: subdivisions géographiques, z: subdivisions chronologiques.

En faisant un appariement « exact » : c'est à dire faire correspondre le contenu de la vedette matière construite et une vedette matière de la liste d'autorité , Markey (1994) aboutit à un taux de 6.9 %. Dans une expérience similaire mais au contexte différent, Liddy (1993) aboutit à presque 10 % d'appariement exact. Les raisons qui expliquent l'échec d'appariement sont nombreuses et variées :

- certaines vedettes n'existent pas dans les versions papier de la liste LCSH (musique,).
- la pratique des subdivisions affranchies.
- des erreurs dans la construction des vedettes matières (soit un champ MARC qui est incorrect, soit à cause d'un oubli d'un code du champ $a, $x, $y,$z ou d'un oubli d'une parenthèse pour les vedettes avec qualificatifs.

En conséquence, si on met en ligne la totalité de la liste LCSH (ou de RAMEAU) , les possibilités d'un appariement exact sont assez faibles : entre 7 et 10 %. Cela veut dire aussi qu'au total, seul 10 % des vedettes matières construites peuvent utiliser la richesse sémantique de la liste (relations, notes, indice de classification...etc.) et des possibilités de navigation dans la liste d'autorité.

2.3. Les différentes générations de catalogues en lignes

Hildreth (1993) distingue trois générations d'OPACs qui correspondent à différents principes d'automatisation de l'accès au catalogue.

2.3. 1. Les catalogues de première génération

La première génération de catalogues consultables en ligne fut avant tout un produit dérivé de l'informatisation dans le début des années 1970 des fonctions de catalogages et de prêt. Ces produits coexistaient pendant longtemps avec les catalogues sur fiches. Ils furent donc des versions simplifiées des catalogues traditionnels sur papier, dont ils adoptèrent la structure rigide et les modes d'accès. Cette première génération se caractérise par des outils d'accès très rudimentaires. Ils simulent la recherche d'information dans les catalogues traditionnels (microfiches). Ils n'offraient que la possibilité d'un appariement exact par phrases d'où leurs efficacité pour retrouver des données exactes (pour chercher des ouvrages déjà connus), ils permettaient de retrouver chaque champ bibliographique tel qu'il apparaît dans l'enregistrement en effectuant un appariement exact entre la phrase qui est entrée par l'usager et ce qui est contenu dans les enregistrement. . Par contre, l'usager a des difficultés à trouver une réponse s'il ne connaît pas à l'avance l'ordre exact de la phrase (nom de l'auteur, le titre exact,...etc.). Pour chercher un document, il faut en connaître le titre en entier et saisir les mots de la requête dans l'ordre. Pour trouver l'ouvrage de Ingwersen[5] , l'usager doit écrire cette équation " TIT = Information Retrieval Interaction". De plus, ces catalogues requièrent des règles précises de succession des écrans et d'entrées des données. L'accès aux différents types de recherches passe par le respect d'une syntaxe d'entrée des données. L'usager peut passer d'un mode de recherche à un autre, en tapant le code fonction correspondant (TIT, AUT/TIT...etc.), mais s'il saisit par exemple l'ensemble de l'intitulé du code correspondant (TITRE), le système ne le reconnaît pas et la recherche n'aboutit pas.

Ces catalogues présentent donc les défauts suivants :

- l'emploi exclusif du langage spécialisé des bibliothécaires

[5] P. Ingwersen: "Information Retrieval Interaction ». London: Taylor Graham, 1992.

- absence de messages d'aide pour les usagers
- format unique d'affichage des notices bibliographiques
- illisibilité des écrans
- impossibilité d'améliorer la recherche à partir des premiers résultats obtenus
- le manque de points d'accès comme la recherche sujet
- une interface peu ergonomique: ils n'offraient pas de messages d'aides, ils utilisaient un langage spécialisé (celui des bibliothécaires) dans le dialogue avec l'usager.
- les possibilités de navigation sont limitées sinon inexistantes.
- pas de recherche par mots clés

Les premières études effectuées sur ce type de catalogue montrent que la consultation n'est pas plus performante que celle d'un catalogue manuel. Certains auteurs estiment même qu'ils étaient moins efficaces que les catalogues manuels. Il était plus facile de se servir manuellement des fiches que de manipuler un clavier ou de lire sur un écran d'un catalogue de première génération. Malgré ces faiblesses, les catalogues de première génération apportaient un confort à la vitesse d'accès dans de grandes masses de données. Ils offraient déjà des informations sur la disponibilité des ouvrages. Pour Hudon (1998) ce fut un des facteurs clé de la rapide popularité du catalogue en ligne. Devant les difficultés d'un usager à formuler clairement une requête faute de connaissance de la syntaxe, voire des options de recherche proposées par le catalogue de première génération, il est apparu nécessaire de permettre la recherche par un formulaire ou par un menu. Les améliorations apportées aux catalogues en ligne ont permis de passer des OPACs de première génération, dans lesquels la recherche se fait par phrases (accès précoordonné), simulant la recherche dans les catalogues manuels, aux OPACs de seconde génération, influencés par les SRI où l'accès se fait par mots (accès postcoordonné).

2.3. 2. Les catalogues de deuxième génération

Les catalogues de deuxième génération ont intégré les fonctionnalités des SRI comme la recherche booléenne (recherche postcordonnée), la multiplication des critères de recherches, l'utilisation de la troncature et des opérateurs de proximité, les possibilités de limite et de tri des recherches, l'accès par un vocabulaire contrôlé , l'affichage des vedettes matières et des

renvois "voir" et "voir aussi". Actuellement, la majorité des catalogues opérationnels en Europe et en Amérique du nord, sont de deuxième génération : la structure des notices, leur contenu et les principales clés d'accès sont ceux de la première génération alors que les options de recherche et d'affichage sont issus des SRI. Avec la recherche postcoordonée (par mots-clés plutôt que par vedettes-matières exactes ou par titre complet) l'usager n'a plus besoin de connaître des phrases exactes. En plus de l'introduction des possibilités de recherche booléennes, il y a eu des améliorations au niveau de l'interface homme-machine:

- les notices bibliographiques pouvaient désormais être visualisées sous différents formats de présentation, soit en abrégé ou complet.
- une solution au problème de surcharge d'information consiste à introduire des facteurs options de limitations, soit par date de publication, par langue, par type de bibliothèque. Ces facteurs sont rarement utilisés.
- des améliorations ont également été portées au niveau de l'interface et du dialogue comme l'introduction des interfaces graphiques, la possibilité d'interrogation par commande et par menu, la disponibilité de plusieurs formats d'affichages, l'affichage de l'historique de recherche et une amélioration du contenu des aides en lignes et des messages d'erreurs.

L'ensemble des éditeurs de logiciels documentaires offre ce type de système comme GEAC[6], DRA[7], EVER[8] ou VTLS[9].

2.3. 3. Usages des catalogues de deuxième génération

Les chercheurs en science de l'information et en bibliothéconomie , conscients des limites et des contraintes des catalogues, et pour mieux comprendre les tactiques et stratégies mises en œuvre par les usagers lors de la consultation des catalogues en ligne, ont mené des recherches sur le comportement des usagers. Ces études anglo-saxonnes dans la majorité des cas, ne concernent en général que les catalogues de deuxième génération. Les chercheurs se basent généralement sur l'analyse des traces informatiques comme moyen de recueil de données.

[6] Http://www.geac.com

[7] http://www.dra.com

[8] http://www.ever.fr

[9] http://www.vtls.com

Ces travaux qui ont fait l'objet de nombreuses publications, sont à l'origine des réflexions sur les catalogues de troisième génération.

2.3. 3. 1. Les besoins et préférences des usagers

Il est difficile , sinon hasardeux, de classer les besoins des usagers d'une bibliothèque puisque la nature et le contexte de ces besoins sont mal définis : c'est pourquoi nous parlerons plutôt de préférences des usagers. Bates (1990) trouve que les usagers souhaitent d'abord, contrôler le processus de recherche. Ercegovac (1989) constate que 45% des usagers désirent l'affichage des relations sémantiques de la liste LCSH et 42% d'entre eux l'incorporation des tables de matières ou des index d'un livre dans le catalogue. Elle note qu'ils souhaitent que les options de contrôle de l'affichage des résultats (possibilité de classer les résultats par date, par auteur, par titre...) soient mieux adaptées selon le contexte de recherche.

2.3. 3. 2. Choix des ressources documentaires

L'une des premières difficultés qu'appréhendent les usagers, est le problème d'identification des bases documentaires à interroger. Ils n'ont pas une vision claire sur leurs contenus, ni sur leurs structures. Les usagers se heurtent aussi au choix des ressources documentaires (monographies, périodiques, ...etc.). Le nombre croissant des catalogues sur l'Internet et leurs l'interconnexion par le biais de la norme Z39.50 va sans doute accroître ces difficultés.

2.3. 3. 3. Fonctionnalités utilisées et problèmes rencontrés

Nous avons réalisé une synthèse des travaux effectués pour connaître quels sont les points d'accès utilisés dans un catalogue en ligne, que nous avons regroupé dans l'annexe1. Bien qu'il soit hasardeux d'établir une comparaison entre ces études , on peut cependant souligner la prépondérance de l'accès sujet dans les catalogues en ligne de deuxième génération. On note que près de deux tiers des sessions de recherches concernent seulement un seul mode de recherche (sujet, auteur, titre, ..etc.). Cependant, Larson (1991) constate que l'accès sujet a régressé au dépends d'une recherche par mots clé. Il a effectué une analyse transactionnelle sur le catalogue MELVYL . Il conclut que la recherche sujet a diminué de 2.15% par an sur une période de six ans alors que la recherche par mots du titre a augmenté de 2.15%. Pour lui, plus la base s'agrandit, plus les causes des échecs et de surcharge d'information augmentent.

La majorité des catalogues ont introduit soit la troncature, les opérateurs booléens, des critères de limitation comme la date, la langue des documents ou le choix de la collection à interroger. Malheureusement, les usagers emploient rarement ces fonctionnalités avancées.

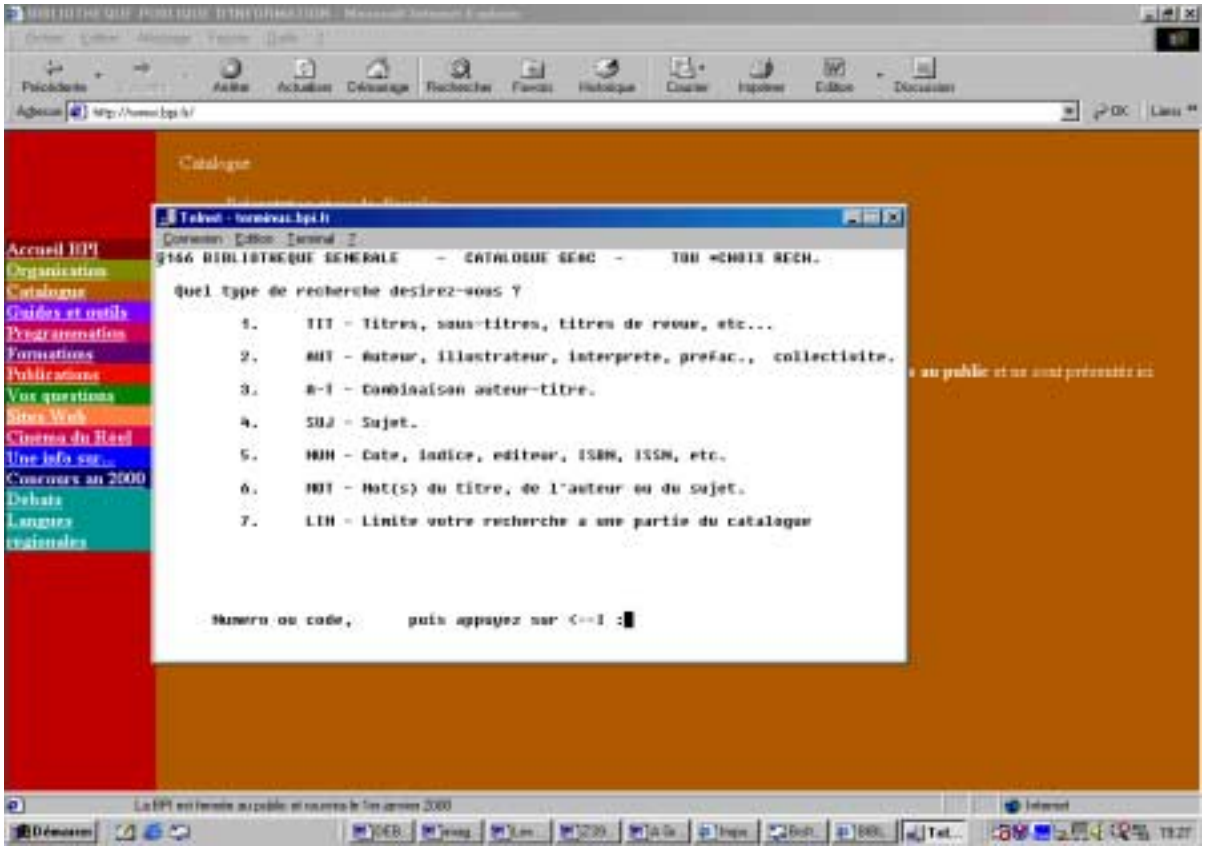


Figure 3 : interface d'accès dans un catalogue de deuxième génération (le cas de la bpi)

2.3. 3. 4. Problèmes dans la formulation des requêtes.

Les difficultés de formulation des requêtes présentent trois facettes différentes :

Le choix des termes

Les usagers ne sont pas familiers avec les langages documentaires utilisés (LCSH ou les classifications). Lorsqu'il est disponible en version papier, les usagers ne consultent pas les listes d'autorité. Ils préfèrent plutôt utiliser les mots qui leurs viennent à la tête : « *In fact, most subject queries used whatever (terms) popped into the searcher's mind*" » (Markey, 1983). Cette dernière constate que les usagers ne consultent pas la liste d'autorité (LCSH) lorsqu'elle est disponible en ligne [10]. Deux raisons peuvent expliquer ce

[10] La majorité des bibliothèques ne mettent pas à la disposition des usagers cette liste d'autorité..

rejet. La plus évidente est que les usagers ignorent la notion de vedette matière et l'utilité d'une liste d'autorité. De plus, comme le dévoilent Butkovich et al. (1989), les usagers n'assimilent pas bien la structure et l'organisation de la liste (LCSH). Le choix des termes ne se fait pas selon un plan de recherche, ni selon une approche méthodique ; au contraire, il résulte d'une démarche d'essai-erreur. Le chois des termes est rendu encore plus difficile du fait que les usagers n'ont souvent qu'une vague idée de ce qu'ils cherchent et qu'ils n'ont aucune connaissance sur la façon dont sont indexés les ouvrages d'un catalogue.

L'écriture de requête booléenne :

Les usagers ont encore des difficultés à maîtriser l'usage des opérateurs booléens (ET, OU, SAUF) et des parenthèses qui leur sont associées. Ensor (1992) constate qu'ils ne sont pas encore bien familiarisés avec les opérateurs booléens. Elle trouve que la maîtrise de l'outil informatique est un facteur important dans la facilité d'usage d'un catalogue en ligne. Peters (1989) observe que dans 73.5% des cas, l'usage de l'opérateur booléen ET entre les différents points d'accès (une recherche multicritère) aboutit à un échec (zéro réponse) alors que dans 84.8% l'usage de ET dans une recherche sujet entre deux vedettes matières aboutit aussi à zéro réponse. L'utilisation des opérateurs booléens et des parenthèses, ne sont pas intuitifs pour un usager occasionnel.

Compréhension de la structure du catalogue (interface de dialogue)

L'usager se perd dans la succession des écrans du catalogue, il ne distingue pas les différentes commandes et les touches à activer pour effectuer sa recherche. Ceci est particulièrement le cas pour les catalogues dont l'interface est en mode caractère. La saisie des requêtes est rendue encore plus difficile du fait de l'incertitude de l'usager sur le format des données. Ceci est particulièrement vrai pour la saisir des noms d'auteurs, les dates, titres et le sujet.

2.3. 3. 5. Le problème des échecs

Par échec, nous entendons le fait qu'une requête n'aboutisse pas, c'est à dire qu'il y a "zéro réponse". Depuis plusieurs années, les travaux Markey (1984) ont montré qu'entre 23% et 45 % des recherches aboutissent à des échecs. Ce problème n'a pas changé de façon significative avec l'introduction des catalogues de deuxième génération. Markey (1989) trouve que 34%

des recherches par mots de sujet n'aboutissent pas à l'affichage de notices et que 53% des recherches par mots du sujet à une notice pertinente. Nous avons regroupé les résultats de ces études dans l'annexe1.

En utilisant des ordinateurs, les usagers font inévitablement des erreurs de type mécanique (erreurs d'utilisation du clavier, des touches de commandes) mais aussi des erreurs syntaxiques et typographiques (conjonctions, ellipses ou ponctuation incorrecte). Borgman (1986) montre par exemple que 13 % des commandes utilisées contiennent des erreurs typographiques ou logiques alors que Seaman (1992) trouvent que 40% des erreurs des usagers proviennent de l'écriture d'un nom d'auteur ou titre erroné et que 9% des erreurs sont dûes à ce que le catalogue ne supporte pas une recherche des abréviations. Ces deux types d'erreurs sont à l'origine des échecs de prés d'un quart des recherches, mais la principale cause réside dans la difficulté de correspondre le langage des usagers et celui des langages documentaires utilisés lors de l'indexation (Hunter, 1991), (Peters, 1989) et (Seaman, 1992).

2.3. 3. 6. Difficulté à développer une stratégie de recherche.

Plusieurs travaux ont montré que dans les catalogues de deuxième génération, les ressources sont sous-utilisées. Les usagers se confinent souvent à des recherches simples. Ils n'emploient pas les possibilités fournies par le système et modifient rarement sa stratégie de recherche. Ils emploient un terme général qui aboutit le plus souvent à l'affichage de plusieurs réponses et ils emploient rarement des termes synonymes pour améliorer la précision des recherches.

Non seulement les usagers ont des problèmes pour formuler leurs requêtes, mais leurs tentatives de transformation des requêtes n'aboutissent pas. Tolle (1985) montre que 15% des usagers abandonnent leurs recherches après affichage d'un message d'erreur, à la suite d'une formulation inadaptée au système. Les usagers acceptent souvent cette situation et n'essaient pas d'examiner la cause des échecs. Shenouda (1990) trouve que la majorité des modifications effectuées par les usagers est soit d'ajouter, soit supprimer un terme. D'une façon générale, ne pas prendre en compte les étapes précédentes (et les résultats des recherches précédentes) est une caractéristique des usagers d'un catalogue (Seaman, 1992) , (Hildreth, 1993). L'évolution de la formulation de la demande au cours de la recherche est donc très faible. Ercegovac (1989) montre que 46% des usagers ont des difficultés à élargir leurs recherches pour augmenter le nombre de réponses affichées..

Chen (1991) constate par exemple que les deux tactiques les plus courantes (plus de 80% des cas), consistent à feuilleter les écrans ou à utiliser la stratégie d'essai et du retour au début. Ils n'emploient pas toutes les possibilités du système. Pour lui, cette difficulté à développer des stratégies de recherches constitue d'ailleurs la principale différence entre un usager "grand public" et un intermédiaire. Nous faisons l'hypothèse dans cette thèse que plus l'interactivité offerte par le catalogue est importante, plus les possibilités de modification des stratégies de recherche seront employées.

2.3. 3. 7. Des problèmes pour manipuler les résultats de recherche.

La surabondance d'informations

Les modes d'accès et d'indexation qui fonctionnent dans le catalogue de « taille moyenne » sont inadéquats pour les catalogues des grandes bibliothèques. Un problème lié à ce facteur taille est celui de l'affichage d'un grand nombre d'informations : le problème de surcharge d'information. Deux cas de figures peuvent se présenter :

- la surcharge d'information dans l'affichage de l'index alphabétique des sujets.
- la surcharge d'information dans l'affichage des références bibliographiques.

Une requête sujet sur le droit administratif dans le catalogue de la bpi (bibliothèque publique d'information) nécessite l'affichage de plusieurs écrans.

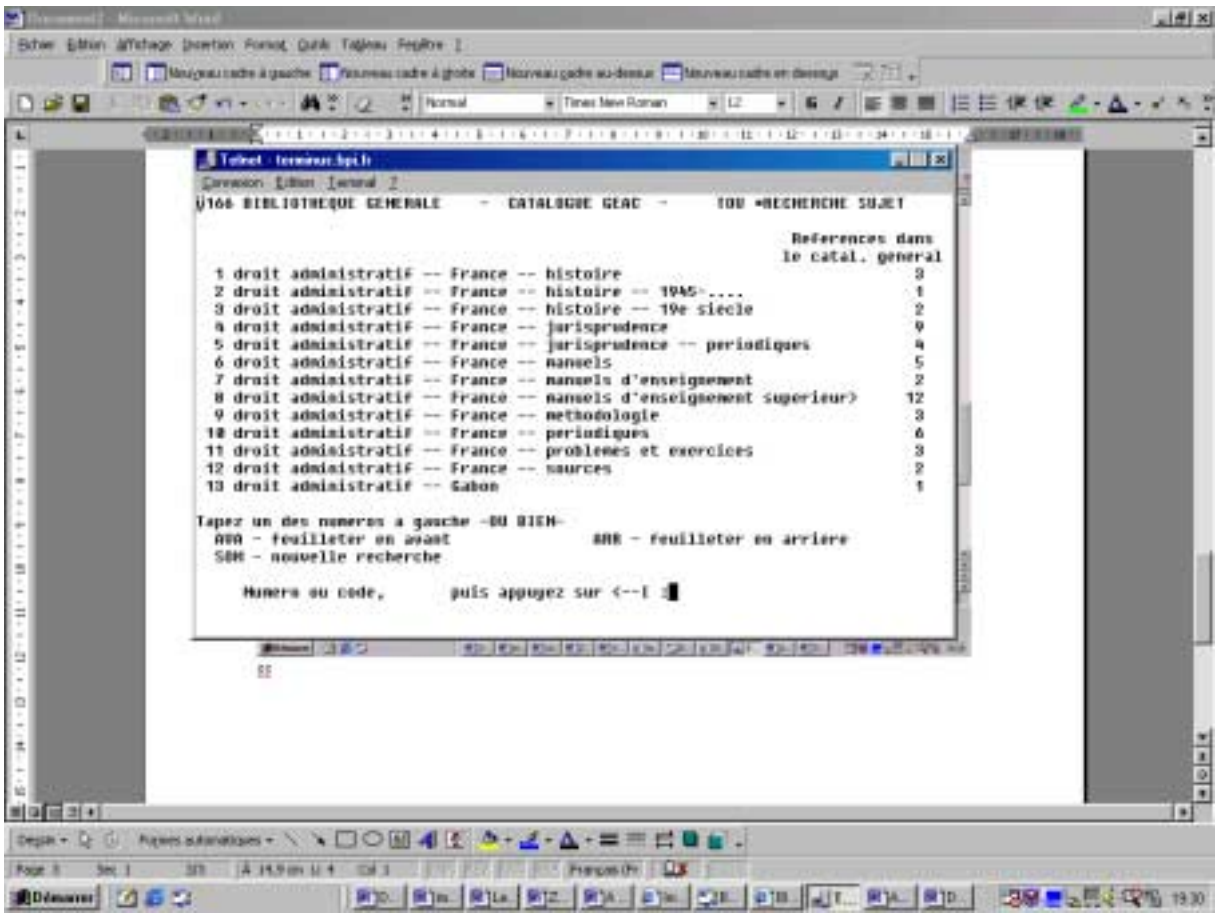


Figure 4 : recherche sur droit administratif (bpi)

Des études passées ont montré déjà que lorsque les usagers trouvent un grand nombre de réponses (entre 22% - 53% de fois) , 26% d'entre eux ont des difficultés à réduire leurs résultats (Matthews,1982) , (Kern-Simerenko, 1983). Wiberly et Dougherty (1988) soulignent que cette surcharge d'information peut produire un effet de saturation. Ceci rebute les usagers à utiliser les catalogues en ligne. Bates a observé dès 1984 que l'affichage de 30 réponses peut être considérée comme une surcharge d'information (Bates, 1984). Elle soutient d'ailleurs que ce nombre (30 réponses) peut être considérée comme le seuil à ne pas dépasser, même dans une recherche sur les moteurs de recherche (Bates-1996). Dans un grand catalogue comme celui de l'université de Californie (MELVYL), le nombre moyen d'affichage est 181 notices (Wiberly & Dougherty, 1988). Larson (1986) trouve pour sa part que ce nombre est 77.5, mais les usagers n'affichent que 9.1 notices. En 1992 , les usagers de MELVYL ont toujours ce problème de surcharge d'information : ils obtiennent en moyenne plus de 100 notices mais n'affichent que 15 notices par session (Berger, 1992). Il faut noter que les options et les aides proposées par la majorité des catalogues classiques pour réduire

ce problème,, sont inadéquats pour un usager grand public. Souvent, l'usager ne peut pas déterminer à l'avance quelles sont les caractéristiques (langue , date d'édition, domaine,…etc.) des réponses qu'il va obtenir.

La pertinence des références obtenues

Parfois, il s'avère difficile pour l'usager à déterminer si les réponses obtenues sont satisfaisantes ou non ? Ceci est dû principalement à la pauvreté informationnelle du contenu des notices MARC. Cette situation explique pourquoi les usagers consultent en général les premières références affichées (Hancock,1994**).** Nous avons montré ailleurs que l'enrichissement des notices par les tables de matières faciliterait la tâche de sélection d'un document (Ihadjadene, 1998c). Une telle amélioration fournit à l'usager des éléments plus complets sur le contenu des références et facilite ainsi son jugement et son choix.

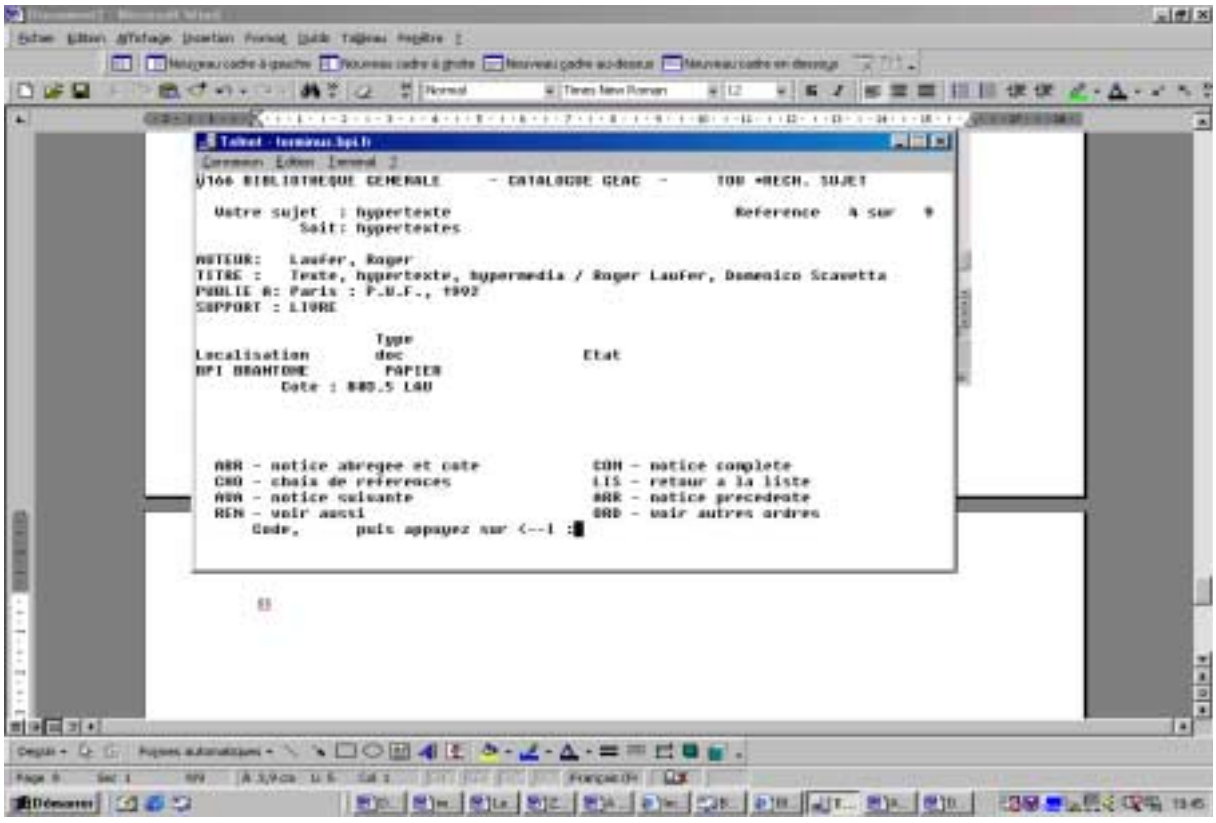


Figure 5 : affichage d'une réponse dans un catalogue de deuxième génération (le cas de la bpi)

2.3. 3. 8. Ergonomie des interfaces des catalogues de première et de deuxième génération

L'ergonomie des interfaces des catalogues en ligne fut négligée dans les études sur les catalogues de première ou de deuxième générations.

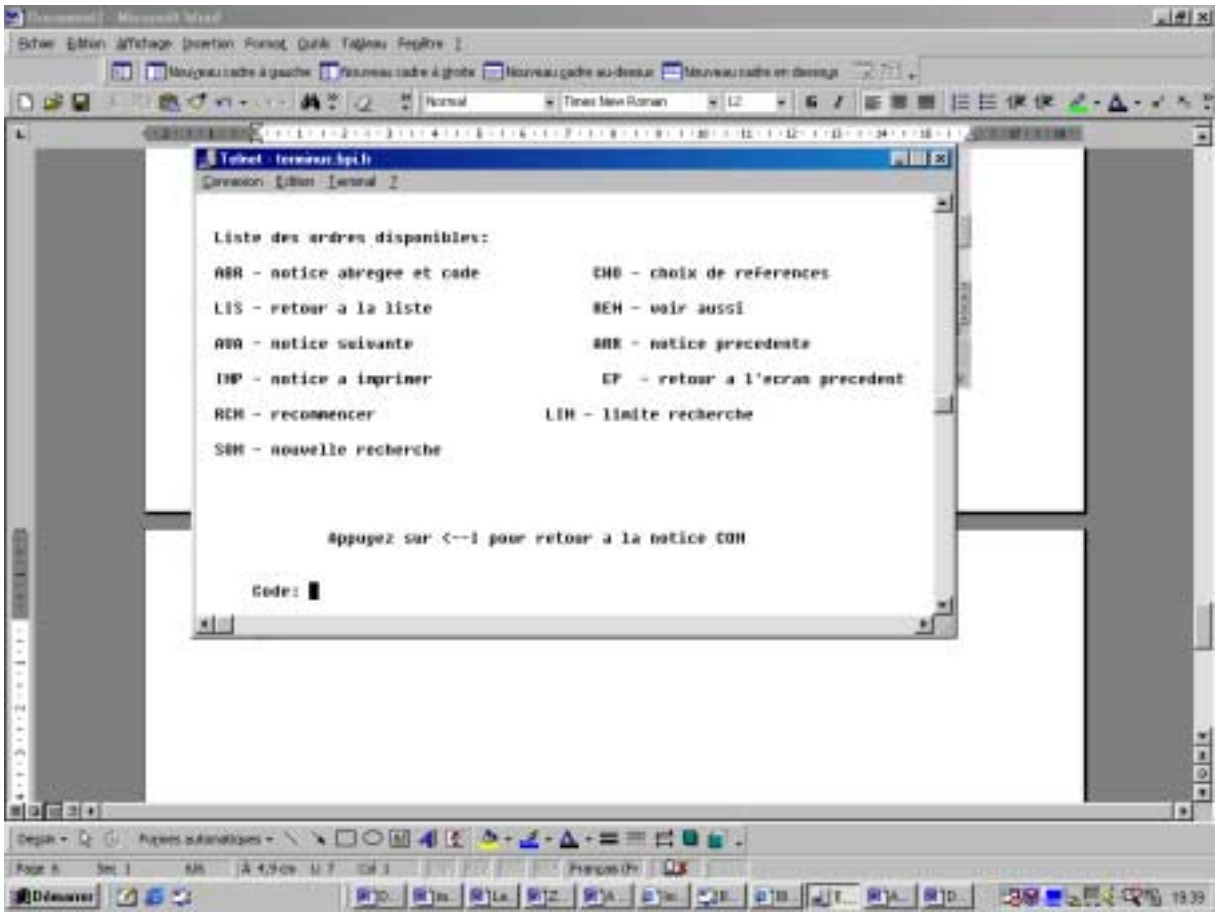


Figure 6 : commandes d'accès dans un catalogue de deuxième génération
(le cas de la bpi)

Les rares travaux que nous avons consulté (Hildreth, 1989) , (Mathewes,1982) montrent qu'en majorité les catalogues présentent les défauts suivants

- Les écrans sont souvent illisibles , à cause de la multiplication des fonctionnalités offertes à l'usager : mauvaise présentation de l'information , la terminologie employée est inintelligible : l'emploi généralisé des abréviations et de codes sont incompréhensibles pour le lecteur novice.
- Le processus de recherche est discontinu et il est impossible pour les usagers de conserver l'historique de leurs recherches ou de sauvegarder leurs résultats

(Le Marec, 1989). La succession des écrans dans les catalogues de deuxième génération est un obstacle au bon déroulement d'une recherche. L'usager a des difficultés à circuler à travers les différentes fonctions du système. Il se perd souvent dans la succession des écrans.

- Les écrans d'aides et les touches de fonction sont difficilement accessibles.
- Le langage de commande est difficile pour les usagers novices et l'usage du clavier est une source potentielle d'erreurs.

2.3. 4. Les catalogues de troisième génération

Divers auteurs (Hildreth, 1993) , (Markey, 1994) , (Frost, 1994) ont présenté les caractéristiques d'un catalogue qu'ils désignent sous le nom générique de catalogue de troisième génération. Ces outils, en plus des fonctionnalités classiques des catalogues de deuxième génération, possèdent de nouvelles caractéristiques (Ihadjadene, 1998a) :

Nouvelles fonctions d'accès et de navigation:

- interrogation en langage naturel
- technique d'aide à la recherche en ligne et à la traduction de la requête
- techniques de recherche non-booléennes
- affichage des documents dans l'ordre de pertinence (pour faciliter la sélection des documents, de nouveaux systèmes basés sur des techniques statistiques proposent l'affichage des documents dans l'ordre de pertinence)
- reformulation interactive
- recherche multilingue
- navigation hypertextuelle
- intégration des classifications , des listes d'autorités et des mots-clés

Enrichissement du contenu de la base bibliographique:

- enrichissement du contenu des notices MARC (résumé, table de matières, augmenter le nombre de vedettes matières...etc.)
- signalement des différents types de documents (sommaire de revues, articles, dossiers, document numérique...etc.)

Mise en réseau avec d'autres catalogues :

un accès à d'autres bases de données et à des catalogues en lignes distants. La généralisation de l'architecture client/serveur basée sur la norme Z39.50 va sans doute accroître cette tendance.

Filtrage coopératif : offrir des outils de personnalisation de la recherche, permettre des recherches coopératives entre les différents usagers.

Ces caractéristiques sont mises en évidence aussi bien dans de récents prototypes[11] que dans quelques systèmes commerciaux.

2.3. 4. 1 Les arbres de décisions

Comme la majorité des OPACs existants incorpore différentes stratégies d'interrogation par sujet (sujet alphabétique, mots sujet, utilisation des opérateurs booléens, affichage des renvois, etc.), il est nécessaire de développer un mécanisme permettant d'orienter le type de recherche selon les caractéristiques des requêtes des utilisateurs : c'est la base des arbres de décisions. L'hypothèse sous-jacente est que chacune des stratégies permette de retrouver des notices bibliographiques différentes. Les arbres de décision sont implantés dans le prototype ASTUTE (Markey,1996) et dans un catalogue opérationnel comme COPAC[12]. Cherry (1994) a testé l'efficacité d'une de ces stratégies en effectuant des recherches par mots du titre et par mots du sujet chaque fois qu'un accès sujet (feuilletage alphabétique ou par phrases) échoue . Elle a trouvé que le rappel augmente d'une façon significative.

Lorsque les usagers saisissent seulement un terme, la probabilité de correspondre avec les termes du référentiel sont grandes. Par contre dans les autres cas, cette probabilité diminue d'une façon très significative (Markey,1994). Il est donc important au système, de suggérer des termes selon les caractéristiques des requêtes de l'usager. Le système ASTUTE (Markey, 1996) permet de choisir l'une de ces approches:

- l'affichage alphabétique des vedettes matières
- la recherche exacte (recherche par phrase)
- la recherche par mots clé

(Markey, 1996) a évalué l'efficacité de ASTUTE. L'intérêt des travaux de Markey se situe à deux niveaux. Le premier concerne la méthodologie utilisée (analyse du fichier log et questionnaires), le second est relatif aux résultats qualitatifs obtenus. Lors de son expérimentation (528 sessions de recherche), elle montre que les performances du système, en terme de précision, sont légèrement améliorés. Les nouveautés apportées par ASTUTE sont rarement utilisées, surtout la possibilité d'élargir la recherche par le biais des

---

[11] Gerry McKirnan a recensé une partie des projets de catalogues de troisième génération dans son site web CYBERSTACKS:"http://www.public.iastate.edu/~CYBERSTACKS/onion.html

[12] http://www.copac.uk/

termes généraux (TG) ou d'afficher les subdivisions de forme. Elle signale que les usagers ne font pas de recherche exhaustive et qu'ils rencontrent des difficultés pour naviguer dans le catalogue. Il y a aussi un effet de saturation: avant d'aboutir à l'affichage d'une réponse, plusieurs étapes intermédiaires se succèdent ce qui rebute l'usager. Markey conclut qu'il sera nécessaire d'ajouter des fonctions à ASTUTE lui permettant de suggérer à l'usager de nouveaux termes pour poursuivre sa recherche. La principale limite de cette étude est liée à la nature de la base de données utilisée lors de l'expérimentation. Les problèmes de surcharges d'information n'ont pas été pris en compte. La deuxième limite concerne les catégories conceptuelles[13] qu'elle a définit manuellement pour faciliter la navigation dans la liste des vedettes matières construites. Ceci exige de revoir l'ensemble des pratiques d'indexation, ce qui se révèle coûteux.

2.3. 4. 2. OPACs et système experts

Les systèmes experts n'ont pas été appliquées à de larges bases documentaires multidisciplinaires. Dans le cas des catalogues, seuls (Khoo,1998) et (Chen,1992) ont entrepris d'exploiter cette technique. Nonobstant, l'usage des systèmes expertes pour les catalogues permet de questionner sur l'opportunité et les choix effectués au niveau de leurs outils de représentation et de leurs mécanismes d'inférence. On peut se poser la question sur la nature de cette "expertise" et les moyens de son acquisition. Ainsi dans le système E-referencer[14], Khoo (1998) a développé deux stratégies de recherches, la première pour une meilleure formulation de la requête, la deuxième pour la reformulation. Dans la première, le système E-referencer permet de remplacer les mots vides d'une requête par l'opérateur booléen ET, d'ajouter la troncature à droite pour chaque terme et d'utiliser un opérateur de proximité entre les termes (ceux qui ne sont pas déjà liés par l'opérateur ET). Par exemple, la requête " expert systems in library reference service" sera traduite par le système en " find expert?system?AND librar?refer?servic?". Ces expertises ne sont pas le fruit d'un examen du processus de recherche des spécialistes. Vickery et Brooks (1986) , qui ont développé l'un des premiers systèmes experts de recherche d'information (PLEXUS) demeurent sceptiques vis à vis de l'approche système expert qui présente plusieurs

[13] En plus des quatre subdivisions de la LCSH, elle a définit manuellement une dizaine d'autres types de subdivisons de sujet.

[14] Il est accessible à l'URL suivante : http://islab.sas.ntu.sg:8000/E-Referencer

faiblesses. Pour eux, il est inutile de consacrer beaucoup de temps à modéliser le contenu de banques de données, sans égard aux caractéristiques cognitives de l'utilisateur et à ses capacités d'expression. Nous verrons au chapitre quatre que ces derniers emploient diverses tactiques en fonction de nombreux facteurs (le besoin de l'usager, la nature de la base documentaire, les réponses du système...etc.). Nous pensons aussi que la nature de cette expertise est très variée ; elle concerne aussi bien la connaissance du vocabulaire contrôlé, un ou plusieurs domaine de recherche (histoire, informatique,...etc.), les techniques d'interrogation, les sources d'informations mais aussi le fonds documentaire. Les expérimentations de Khoo (1998) ne montrent pas d'ailleurs une amélioration de la recherche.

2.3. 4. 3. Accès en langage naturel

L'accès en langage naturel aux bases documentaires spécialisées a fait l'objet de plusieurs recherches. Au début des années 80, plusieurs auteurs voyaient dans le traitement en langage naturel, la solution "idéale" pour en finir avec les problèmes d'accès aux catalogues en ligne. Malheureusement, cette solution n'a jamais été validée, ni testée dans des catalogues opérationnels. Si les techniques de traitement linguistique peuvent être efficaces dans un domaine circonscrit, le problème devient plus compliqué dans le cas d'un système encyclopédique. On peut cependant améliorer la consultation et l'accès par le biais de certaines techniques qui font appel à une analyse sémantique ou morpho-syntaxique pour pouvoir rectifier automatiquement les fautes de saisie ou d'orthographe (Le Loaer, 1993). Les OPACs qui offrent ce type de correction sont rares. Il semble aussi intéressant d'intégrer ces analyseurs morpho-syntaxiques pour permettre d'élargir les recherches aux formes dérivées (pluriel vs singulier, féminin vs masculin, adjectifs...etc.).

2.3. 4. 4. Enrichissement du contenu des notices bibliographiques

L'un des facteurs qui influe d'une façon importante sur la pertinence de l'accès sujet est l'insuffisance du nombre de vedettes matières par notice. Alors que dans les bases de données bibliographiques, le nombre des descripteurs par notice se situe entre 10 et 20; dans les OPACs il est inférieur à deux. Sur un échantillon de 900 notices bibliographiques du catalogue de l'ENSSIB, nous avons trouvé 1,47 de vedettes par notices. Cette insuffisance de l'indexation est particulièrement évidente pour une part des ouvrages d'une bibliothèque: les œuvres de collaboration. Il s'agit d'ouvrage qui contiennent deux ou plusieurs parties écrites

par plusieurs auteurs. Dans ces livres, beaucoup de parties ne sont pas pleinement représentées à travers les points d'accès des catalogues. Pourtant, souvent elles sont succinctement décrites dans les tables des matières. Identifier ces parties devrait fournir un point d'accès supplémentaire à l'information. Divers éléments tels que la langue de publication, la date de publication, la politique d'acquisition suivie, influent sur le nombre d'œuvres de collaboration. En tenant compte de ces critères, on peut penser que le nombre moyen d'œuvres de collaboration dans une collection varie de 12 à 20 % (Ihadjadene, 1998c).

Études	**Echantillon**	**Pourcentage**
Hoffmann	4098	21.3 %
Weintraub	375	12 %
Poulsen	887	24 %
Poulsen	698	18 %
Ihadjadene	420	26%

Tableau 7: pourcentage d'œuvres de collaboration

Depuis une vingtaine d'années, des bibliothécaires et des spécialistes en recherche de l'information ont effectué des études sur l'opportunité d'enrichir les notices bibliographiques par les tables de matières. Du projet novateur SAP (Subject Access Project) effectué en 1977 , à l'étude de PALINET 1996 , une cinquantaine de travaux ont été publiés sur ce sujet

Les parties d'un document ont chacune plus ou moins de valeur pour la recherche d'informations. Afin d'évaluer la pertinence des tables de matière pour la recherche d'informations, la majorité des études d'évaluation que nous avons consultées utilise des critères de performance tels que le rappel et la précision.

En 1982 Settel et Cochrane , cité dans (Ihadjadene, 1998c) , conduisent une étude pour déterminer si en ajoutant des mots et des phrases extraits des tables des matières, cela améliore l'accès au sujet et le taux de rappel pour l'utilisateur. Deux types d'enregistrements sont comparés.

- enregistrements non enrichis
- enregistrements enrichis à l'aide de mots ou phrases extraits des tables des matières

Le deuxième type d'enregistrement double le taux de rappel sur le domaine des sciences sociales et le triple en sciences humaines par rapport au premier type d'enregistrement. Ces auteurs en concluent donc que l'addition de termes extraits des tables des matières augmente significativement le taux de rappel, et évite ainsi le silence. Si les tests effectués par Dillon et l'équipe d'ADFA , cités dans (Ihadjadene, 1998c), confirment cet accroissement du taux de rappel, ils notent cependant une légère diminution du taux de précision:

	ESP (ADFA)	Dillon
Rappel		
base normale	44 %	17 %
base enrichie	75 %	26 %
Précision		
base normale	88,6 %	71 %
base enrichie	70,16 %	59 %

Tableau 8 : efficacité de la recherche d'information dans un catalogue enrichi

La deuxième raison qui justifie l'enrichissement des notices par les tables de matières est de faciliter le choix des documents à l'écran. En effet, les utilisateurs rencontrent des difficultés pour sélectionner un document. Les notices bibliographiques contiennent très peu d'informations sur le contenu du document (auteur, sujet, titre, éditeur, etc.). L'enrichissement des notices bibliographiques par les tables de matières peut aider l'utilisateur à mieux sélectionner les documents. Les tables de matières donnent à la fois une vue d'ensemble et permettent d'identifier les parties du document. L'utilisateur ayant la possibilité de visualiser la table de matières d'un ouvrage, pourrait mieux juger à l'écran la pertinence de ces références sans aller aux rayons. Lorsqu'on sait que presque le quart des collections des bibliothèques universitaires en France ne sont pas en libre accès, on mesure l'importance de cet enrichissement. Cette difficulté de sélectionner les documents grandit lors d'un accès à distance (par Telnet ou par le Web). Il existe bien sur d'autres possibilités d'enrichissement, notamment la possibilité d'inclure les articles ou des ressources électroniques dans le catalogue ainsi que la possibilité d'ajouter (images de la page de couverture et les sommaires). Si les possibilités d'enrichissement ont été bien admises par les bibliothécaires, il n'en demeure pas moins très peu d'expériences d'envergure dans ce domaine. En France, une récente initiative de l'association des bibliothécaires français (ABF)[15] a permis de regrouper les bibliothécaires, les fournisseurs d'informations bibliographiques (qui souhaitent proposer des ressources supplémentaires, notamment images de couverture, tables des matières, 4^{e} de

[15] L'auteur est membre de l'observatoire de l'abf (http://www.abf.fr)

couverture, présentation de l'auteur) et des fournisseurs de logiciels intégrés de bibliothèque pour étudier la possibilité d'enrichir les catalogues de bibliothèques.

2.3. 4. 5. Un nouveau modèle d'indexation: les graphes conceptuels

Le modèle des graphes conceptuels de SOWA est un modèle de représentation de connaissances de type « réseaux sémantiques » permettant une représentation sous forme graphique. D'une façon générale, un graphe conceptuel est défini comme un graphe qui a deux sortes de nœuds:

- Les nœuds concepts qui représentent des entités, des états, des attributs, des événements
- Les nœuds relations conceptuelles qui symbolisent les liens existant entre deux concepts.

Ainsi, les indexations des documents et les requêtes des usagers seront des graphes. L'utilisation de tels graphes permet de représenter non seulement les concepts qui apparaissent dans le document mais aussi les relations sémantiques qui existent entre ces concepts. Ce choix des graphes conceptuels est justifié par son pouvoir d'expression c'est-à-dire ses capacités à exprimer de la sémantique profonde et des connaissances déductives. En France, il existe quelques SRI spécialisés basés sur ce modèle, notamment le système Infodiab (Puget,1993). Nous pensons que l'intérêt des travaux de Genests (1999) est de généraliser cette approche à une base encyclopédique et à une liste d'autorité comme RAMEAU. A l'heure actuelle, ils n'ont pas encore développé de SRI basé sur ce modèle. Il reste à savoir comment implémenter ce modèle pour les catalogues en ligne opérationnels et quels sont les changement à apporter à ces catalogues dans le processus d'indexation.

2.3. 4. 6. Un nouveau modèle d'appariement: OKAPI

Les travaux sur OKAPI sont nombreux et variés. Les résultats des études effectuées sur ce prototype ont beaucoup influencé les travaux et le développement des catalogues en ligne et les systèmes de recherche en texte intégrale. Depuis 1983, les chercheurs de la City University, notamment (Walker, 1989) et (Hancock, 1994) travaillent sur un OPAC de troisième génération capable de classer les références dans un ordre de pertinence en s'appuyant sur le modèle probabiliste. Il existe d'autres prototypes basés sur ce modèle comme CHESHIRE (Larson, 1996) . Fox (1993) a développé un catalogue, MARIAN, qui

s’appuie sur le modèle vectoriel. L'intérêt de OKAPI ne se limite pas à la possibilité de classer les résultats de recherches. L'équipe qui a développé OKAPI a testé un ensemble de fonctionnalités, qui restent absentes dans la majorité des catalogues opérationnels, notamment les différentes possibilités du feedback et de reformulation (figure7). L'utilisateur peut relancer la recherche en choisissant les références qui correspondent à son besoin. OKAPI reformule la question en utilisant les descripteurs et les mots du titre apparaissant dans les notices sélectionnées. Hancock (1992), qui a expérimenté cette fonctionnalité et montre que prés d'un tiers d'usagers sur les 858 sessions étudiées, ont utilisé le bouclage de pertinence. La principale raison invoquée par les usagers pour le non usage du relevance feedback, est leur satisfaction devant les résultats obtenus. Une partie d'entre eux indique cependant qu'il ne saisissent pas comment effectuer cette reformulation. Lorsqu’un usager décide de sélectionner la commande "more"[16], aucun moyen ne permet de savoir s'il veut élargir ou affiner sa recherche. Récemment, (Hancock , 1997) a développé des versions interactives de cette option pour faciliter le contrôle du processus de reformulation .Les résultats de son sont décevants, seuls 11% des usagers ont opté pour la reformulation interactive. La majorité des usagers ont eu des difficulté pour ajouter ou rejeter des termes. Le second axe de recherche des auteurs du système OKAPI, porte sur l'usage des thesaurus pour formuler et élargir les requêtes des usagers (Jones, 1995). Ils ont testé l'intérêt d'inclure un thesaurus graphique (INSPEC) pour faciliter le choix des termes. Ils constatent que les usagers ne sélectionnent qu'une faible partie (10%) des termes affichés par le thesaurus sans réelle préférence pour l'une des trois relations sémantiques de INSPEC (TG, TS, TA).

[16] c'est une commande qui permet au système de retrouver des documents similaires.

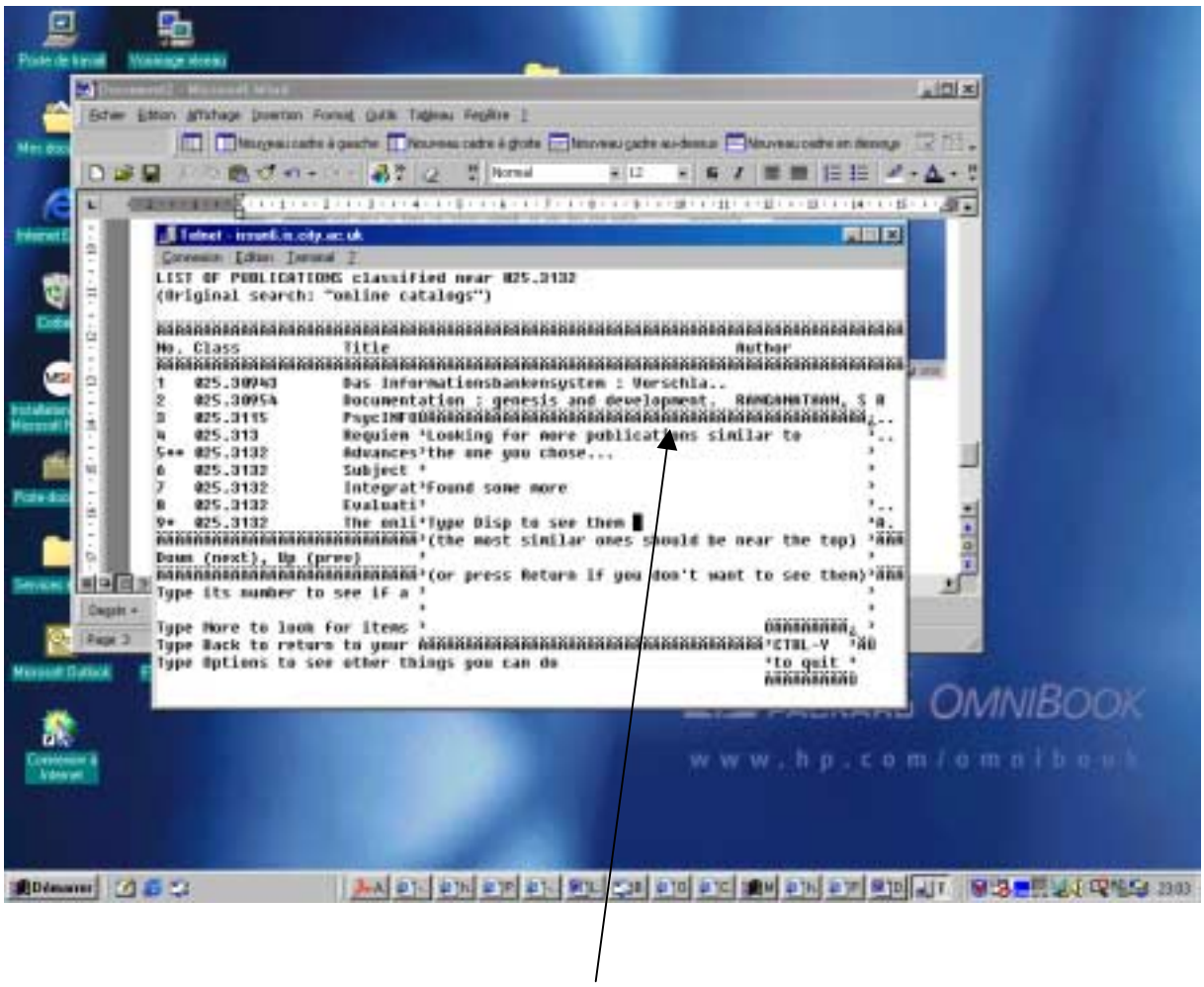


Figure 7 : le feedback dans OKAPI

2.3. 4. 7. Le filtrage collaboratif[17]

Une des solutions aux problème de la recherche d'information, consiste à partager entre usagers l'évaluation du contenu des documents : c'est le principe de la recherche d'information collaborative. Sur l'Internet, cette solution est devenu faisable . Des systèmes tels que GroupLens et Firefly recommandent des choix de livres ou d'articles fondées sur les appréciations de personnes partageant des intérêts similaire. Les étiquettes du métadonnées PICS (platform for internet content selection) facilitent ce filtrage collaboratif de

[17] Cette section est le fruit de la réflexion avec nos collègues Twidale et Nicols lors de la rédaction du projet Portulan , puis de DEBORA (projet européen sous la coordination du Pr Bouché) , ainsi que d'un mémoire bibliographique que nous avons dirigé sur ce thème.

l'information. Le site (http://www.sims.berkeley.edu/resources/collab/) recense plus d'une vingtaine de ces système de filtrage collaboratif. Ces travaux se situent dans un cadre de réflexion théorique globale sur la « cognition situé et distribuée ».

Longtemps, l'utilisation du catalogue d'une bibliothèque est considérée comme une activité solitaire, et il est fait peu de cas dans les études sur les catalogues, des aspects sociaux de système d'information. Plusieurs types d'interactions collectives entre utilisateurs ont été observés:

- Un groupe d'étudiants travaillant ensemble autour d'un catalogue, discutant de leurs idées et planifiant les actions à réaliser
- Des étudiants travaillant sur des terminaux adjacents, discutant de ce que chacun fait, comparant leurs résultats. Ils se penchent souvent sur le terminal du voisin et parfois se groupant autour de l'un d'entre eux.
- Des usagers travaillant individuellement sur des terminaux adjacents, se tournant quelques fois vers le voisin pour demander l'aide. C'est peut-être la forme de collaboration la plus répandue dans les bibliothèques.
- Des personnes travaillant individuellement sur des terminaux non adjacents se mettant à l'écoute des autres.

Le prototype SOPAC (Sugamoto,1995) est un système qui gère une forme limitée de ce genre d'interaction. Il permet aux usagers et aux bibliothécaires de communiquer entre eux en temps réel.

Kantor & Koenig (1994) ont proposé un système ANLI[18] permettent deux modes d'accès. Le premier est traditionnel, c'est l'accès par requêtes booléennes, le second permet aux usagers de naviguer dans un réseau de liens hypertextes constitués par les commentaires des usagers. Après une requête, l'usager aura accès aux commentaires d'autres usagers proposant telle ou telle référence intéressante traitant du thème recherché. L'usager peut bien sûr choisir le type de commentaires auxquels il veut accéder, par exemple ceux des bibliothécaires ou d'enseignants. L'une des propositions les plus intéressantes dans ce domaine est sans doute celle de PORTULAN. L'idée principale de PORTULAN émanait du professeur Bouché qui constatait que la recherche d'information met en œuvre plusieurs savoirs: le savoir construit dans le document, le savoir mis en œuvre par les bibliothécaires à travers une description

[18] Cette proposition ne fut pas l'objet d'un développement informatique. ANLI est juste une proposition de recherche comme Portulan.

sémantique du document, et enfin le savoir construit par l'utilisateur qui peut être communiqué à d'autres usagers. Le but de PORTULAN était de développer des outils graphiques qui mettent en évidence ces différents types de savoir. L'échec de PORTULAN (mais aussi de ANLI) était dû, à notre avis, à l'inadéquation entre ces objectifs et les moyens informatiques de l'époque: l'introduction de langages de programmation indépendamment des plates-formes comme JAVA permet maintenant de mettre en œuvre aisément ce genre d'application. Le seul prototype qui admet le filtrage collaboratif et la possibilité de partage de commentaires est ARIADNE (Twidale,1998) (figure 8).

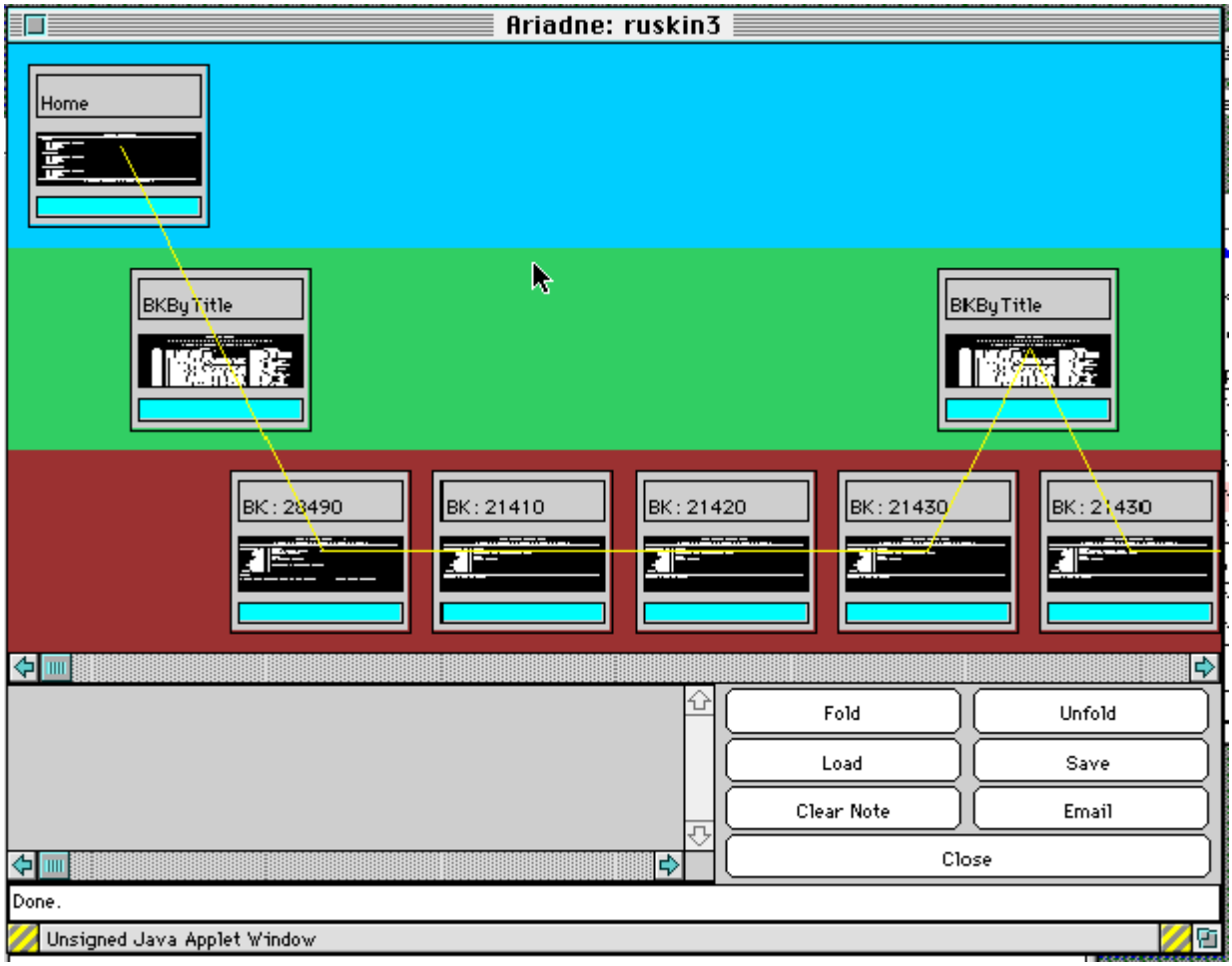


Figure 8 : le prototype ARIADNE

Développé avec java, le système enregistre les requêtes et leurs résultats et fournit alors une visualisation du processus de recherche qui peut être partagé par d'autres usagers. Ce processus peut être sauvegardé et réaffiché à l'écran lors d'une autre session. Ce domaine de recherche est encore à ses débuts. Beaucoup de questions restent sans réponses, notamment

celles relatives à l'usage de cette technologie, au contrôle ou à la modération des commentaires émis par les usagers et à son intégration dans un catalogue en ligne opérationnel.

2.3. 5. Les limites des études d'usages sur les catalogues en ligne

Les études d'usage effectuées sur les catalogues de deuxième génération présentent plusieurs limites. La majorité de ces travaux ne s'intéressent qu'à l'accès sujet et aux problèmes qu'il pose. Or les diverses parties d'un catalogue sont indissociables : il n'est pas pertinent de considérer l'accès par sujet indépendamment de l'interrogation par les autres critères. L'interrogation sujet peut se transformer en une interrogation par titre ou auteur. Nous verrons par la suite que c'est une tactique parfois efficace lorsqu'il y a un échec. Concernant la surabondance de l'information, encore très peu d'études s'intéressent à ce problème, selon les différents points d'accès, autres que l'accès sujet. Hormis les récents travaux sur OKAPI, l'ensemble des études effectuées sur les catalogues opérationnels présentent une interface en mode caractère qui ne facilite pas le dialogue de l'usager. Pour mettre en œuvre un certain nombre de stratégies de recherche (reformulation, navigation entre les documents) la maîtrise de l'outil informatique est primordiale. Notre hypothèse de travail est que la disponibilité de catalogue avec des interfaces hypertextes pourra faciliter la mise en œuvre d'un certain nombre de stratégies de recherche. L'affichage très rare des notices complètes rend difficile l'utilisation de la reformulation . Or ce problème découle aussi de la non maîtrise de l'outil informatique. L'analyse des difficultés et des erreurs est aussi discutable. Aucune analyse sur les tactiques mises en œuvres pour corriger leurs erreurs n'a été effectuée. Nous verrons dans les chapitres suivants les différentes tactiques mises en place dans les WWW-OPACs. Majoritairement, ces travaux sont également limités à cause des fonctions réduites des catalogues étudiée notamment l'appariement exacte. Or, les catalogues actuels offrent en général un accès par mots clé. Enfin, a notre connaissance, aucune étude empirique n'a été menée sur la navigation hypertextuelle dans les catalogues en ligne.

2.4. Les catalogues en France

Cette étude a pour but de recueillir des données empiriques sur les OPACs des bibliothèques universitaires françaises en général, et plus particulièrement, de faire le point sur l'accès sujet .

2.4.1. Situation des bibliothèques universitaires

Après avoir effectué un état de l'art concernant l'accès par sujet dans les OPACs, nous avons adapté et enrichi un questionnaire (voir annexe 2) déjà testé en Grande Bretagne (Crawford, 1993). Nous avons envoyé un questionnaire aux directeurs des bibliothèques universitaires et des services communs de documentation (SCD) entre la période de Décembre 1996 et Mars 1997. Ce questionnaire est composé de six sections (catalogues manuels, OPACs et fonds documentaire, stratégies d'indexation, stratégies de recherche, affichage de l'index et des renvois, aides disponibles). Sur les 93 bibliothèques universitaires[19] (BU) que nous avons recensées, nous avons reçu 84 réponses (47 réponses des SCD, 27 des sections et 10 sont inexploitables). Dans notre échantillon, 65 BU ont un OPAC. En plus de l'envoi du questionnaire nous avons eu accès aux catalogues des BU qui existent sur l'Internet. Le travail de vérification et de correction des données s'est réalisé en téléphonant aux responsables des bibliothèques universitaires. Selon les points du questionnaire, nous avons décidé de regrouper les réponses des SCD avec celles des sections ou de les traiter séparément. Les comptages ne sont donc pas aussi homogènes que nous le voulions, mais la valeur des informations reçues nous permet de donner une vue générale sur l'accès par sujet dans les BU

[19] Le terme bibliothèque universitaire est donc utilisé pour désigner aussi bien un SCD qu'une section d'un SCD.

2.4.2. Analyse du questionnaire

2.4.2. 1. Les Catalogues manuels:

Sur les 74 BU, nous avons trouvé que quarante sept (47) BU conservent encore un accès à travers des supports traditionnels (fiches ou microfiches). Actuellement, l'accès manuel ne concerne qu'une partie du fonds documentaire. Ceci impose aux utilisateurs d' effectuer une double recherche (manuelle et automatisée) pour trouver l'information voulue , ce qui est parfois contraignant.

catalogue papier	9
OPAC	27
les deux	38
total	74

Tableau 9 : typologie des catalogues en France

L'accès manuel se fait d'une façon complémentaire par un catalogue dictionnaire et un catalogue systématique . Les deux modes d'accès les plus fréquents sont l'accès par auteur (47 BU) et l'accès dictionnaire matière (46 BU).

catalogue dictionnaire auteur	47
catalogue dictionnaire matière	46
catalogue systématique	13
catalogue dictionnaire titre	6
autres catalogues	9

Tableau 10 : modes d'accès

Neuf BU possèdent en plus, soit un catalogue topographique, soit un fichier de microfiches OCLC ou enfin un catalogue de thèses et mémoires classés par année.

Dans les catalogues manuels, les bibliothécaires emploient deux stratégies d'indexation principales , celle basée sur RAMEAU et une indexation libre par une liste de descripteurs maison (22 établissements)

RAMEAU	24
MESH	7
Autres	22

Tableau 11 : langages documentaires utilisés dans les catalogues manuels

2.4.2. 2. Les catalogues en ligne (OPACs)

Soixante cinq (65) ont fait le choix d'un système intégré de gestion de bibliothèques qui admet un module OPAC. Sept BU n'ont pas un module OPAC séparé mais un CD-ROM (CD AURUC). Les systèmes les plus utilisés sont ceux de Dynix et de GEAC.

On observe que l'introduction des OPACs dans les BU est récente. On peut relever que plus des 2/3 des BU ont introduit un OPAC ou changé de système depuis 1990.

Qualité du fonds documentaire:

De l'indexation et la conservation classique des monographies et des thèses , le fonds des BU tend à être de plus en plus multimédia.

Les collections des BU sont donc d'une grande diversité comme l’indique le tableau suivant:

monographies	65
thèses et mémoires	55
conférences	42
périodiques	42
images animées (films, vidéo...)	26
documents sonores	23
microfiches	21
logiciels	7
documents électroniques	3
autres	8

Tableau 12 : qualité du fonds documentaire

La catégorie « autres » inclut essentiellement les cartes géologiques (7 établissements).
Si L'OPAC signale la diversité de la collection de la bibliothèque, il ne permet pas cependant d'effectuer des recherches sur des parties de documents. Or ce sont ces parties qui intéressent l'usager.
On a observé que la taille des fonds documentaires varie considérablement selon les BU. L'ancienneté, mais aussi l'implémentation géographique explique ces disparités.

Les notices informatisées se présentent sous différents formats. Nous remarquons une avancée des BU en matière de normalisation puisqu'on utilise en majorité des formats compatibles MARC.

Unimarc	28
LC marc	24
Marc SIBIL	03
Autres	07
non réponses	03

Tableau 13 : format des notices

On s'est intéressé au pourcentage de fonds documentaire qui est informatisé. Sur les 47 SCD, nous observons qu'une partie importante de ces fonds ne sont pas consultables sur un support informatisé. Il s'agit surtout des BU qui ont un volume important.

<25 %	10
25 -50%	9
50- 75%	4
> 75 %	15
non réponses	7

Tableau 14 : pourcentage du fonds informatisé

Malgré l'effort considérable qui a été fait dans le domaine de la rétroconversion, notamment le programme de 1991 (Renoult,1994) , les BU sont loin de disposer d'un catalogue entièrement informatisé.

Catalogage et indexation sujet

La majorité des bibliothèques pratique les deux types de catalogage (courant et dérivé). Le catalogage en local tend de plus en plus à se réduire et la mise en service du futur système universitaire va sans doute accroître cette tendance.

Le catalogage dérivé à partir des trois réservoirs (BN-OPALE, OCLC, SIBIL) est pratiqué sur une large échelle. On observe une augmentation des BU qui utilisent le réservoir BN-Opale

Les langages documentaires les plus répandus sont RAMEAU et la classification décimale de Dewey. Nous pouvons constater l'usage fort répandu de la classification Dewey qui a profité de la déréglementation de 1988 (Renoult,1994). Six bibliothèques utilisent plus d'une classification, ce qui pose souvent des problèmes d'incompatibilité liés à l'existence de référentiels hétérogènes.

Rameau	65
Classification de Dewey	19
Classification décimale universelle	4
Classification de la bibliothèque du Congrès	4
National library of médecine	2
Classification de Cunnigham	2
Classification maison	2

Tableau 15 : langages documentaires utilisés dans les OPACs

L'introduction rapide et généralisée de la liste RAMEAU est bien reçue par les bibliothécaires car elle permet d'harmoniser les pratiques d'indexation . Pour la majorité d'entre eux, c'est la garantie que les lecteurs n'auront pas à s'adapter en passant d'une bibliothèque universitaire à une autre. Toutefois, RAMEAU n'est pas exempt de certains problèmes dont:

- Nécessité d'un accès plus naturel: la plupart des bibliothécaires trouve l'accès sujet à travers RAMEAU un peu difficile pour les lecteurs non avertis qui ont besoin d'une interface, permettant de passer facilement du langage naturel à un langage assez structuré du type RAMEAU.
- Insuffisance des renvois: les bibliothécaires déplorent l'absence des renvois ou leur insuffisance en nombre car « l'utilisation des renvois permet de pallier au manque du langage naturel »
- Le problème de l'hétérogénéité: l'emploi de plusieurs langages documentaires (par exemple RAMEAU, MESH) , crée des conflits.
- Fonds spécialisé: lorsqu'une partie du fonds est spécialisé, «RAMEAU est insuffisant car les vedettes sont très larges, ce qui nous conduit à affiner l'indexation par des termes libres ».
- Nombre de vedettes: souvent les bibliothécaires regrettent l'insuffisance du nombre de vedettes par notice.

Les clés d'accès

L'informatisation des catalogues a permis d'offrir plus de clés d'accès aux utilisateurs. Cette augmentation donne désormais à l'usager des possibilités inconnues avec les formes traditionnelles des catalogues. 68% des usagers utilisent des clés d'accès inexistantes dans les catalogues manuels (Peters, 1993).

Pour la majorité des bibliothécaires, la disponibilité d'une recherche par mots clé (mots du titre, mots de sujet et mots dans toute la notice) permet de contourner la difficulté d'une recherche sujet classique.

auteur (nom de personne)	65
titre	61
mots de titre	60
auteur (non de collectivité)	58
mots de sujet	57
sujet (feuilletage alphabétique)	55
indice de classification,	35
auteur/titre	31
autres	32

Tableau 16 : clés d'accès dans les OPACs

La catégorie « autres » comprend éditeur, collection, isbn, n°d'inventaire, n°de la notice OCLC, lieu d'édition. Les cinq BU qui se sont dotées de l'OPAC AB6, peuvent offrir à leurs utilisateurs d'effectuer des recherches en texte intégral.

Recherches booléennes, recherches avancées.

Quarante huit BU offrent un accès booléen à leurs catalogues. Comme nous le montre le tableau suivant, cela varie selon les points d'accès et selon les opérateurs booléens proposés :

points d'accès /opérateurs booléens	ET	OU	SAUF	ET implicite
auteur	31	33	25	26
sujet	31	32	24	30
titre	30	32	25	31
mots clés	31	29	25	25
Indice de classification	14	16	14	10
Autres points d' accès	8	7	5	6

Tableau 17 : recherche booléennes dans les OPACs

Nous avons vu , que pour pallier au problème de la formulation des requêtes, diverses solutions issues de l'intelligence artificielle (système experts, logique floue, recherche probabiliste, langage naturel...etc.) sont proposées . Rares sont les OPACs qui incorporent ces techniques avancées; en effet un seul des catalogues de notre échantillon est doté d'une recherche phonétique et seule la troncature est vraiment présente dans les OPACs.

points d'accès /opérateurs booléens	Troncature	opérateurs d' adjacence
auteur	49	7
sujet	47	9
titre	47	8
mots clés	34	5
indice de classification	19	3
Autres points d'accès	6	2

Tableau 18 : recherche avancée dans les OPACs

La navigation dans les catalogues en France

Un autre mode de recherche qui a du succès chez les usagers est le butinage de l'index. L'incorporation des liens (synonymie, hiérarchie, association) existants dans cette liste d'autorité a été souvent présentée par les bibliothécaires comme l'une des solutions pour résoudre le problème de l'accès sujet dans les catalogues. Cinquante cinq BU permettent un affichage de l'index mais la majorité des catalogues actuels n'exploitent pas les structures de renvois de la liste RAMEAU.

Pour les bibliothécaires, les raisons qui expliquent l'absence des renvois sont:

- L'incapacité du système informatique à gérer les renvois.
- Le manque de personnels.
- Les différents modes de recherche existants.
- rendent l'affichage de renvois redondants.
- Des raisons budgétaires.

Modes de diffusion.

Le travail en réseau est une pratique importante et courante de la communauté universitaire. Un certains nombre d'utilisateurs des BU (notamment les chercheurs et étudiants de 3° cycle) ne peuvent se contenter d'un accès local. L'accès distant à d'autres ressources est devenu primordial pour eux. Ils peuvent désormais interroger des milliers de catalogues accessibles sur l' Internet à partir de leurs postes de travail. La diffusion du protocole de recherche et de repérage Z39.50 va sans doute accentuer cette tendance. Environ la moitié des BU offre un accès distant à leurs ressources documentaires par le biais de WWW.

Formation des usagers

La plupart des bibliothécaires estime que la formation des utilisateurs et la disponibilité des bibliothécaires sont des moyens de contourner la difficulté d ‘accès à un OPAC car les aides en ligne ne sont pas suffisantes. Voici les différentes aides dont dispose l’usager dans les BU :

- disponibilité des bibliothécaire (95%)
- aide en ligne sur écran (78%)
- aide mémoire et brochure (53%)
- session de formation (49%)
- panneau d'affichage (43%)
- didacticiels (3%)

Beaucoup de bibliothécaires estiment que le contenu de la formation doit lui, aussi évoluer. Il ne doit plus s'articuler seulement sur les aspects bibliothéconomiques, mais prendre de plus en plus en compte des aspects techniques de téléchargement, d'impression, de formatage et d'accès à distance.

Conclusion

Dès le début des années 90, un effort considérable et soutenu a été réalisé pour l'informatisation des BU. Les résultats de cette étude montrent que cet effort est maintenu depuis sept ans, aussi bien au niveau de la rétroconversion que de celui de la normalisation. D'après les réponses que nous avons reçues, la majorité des catalogues de notre échantillon sont de deuxième génération. La disponibilité sur le marché de systèmes commerciaux basés sur le modèle probabiliste, la diffusion de la norme Z39.50, l'enrichissement des notices et l'introduction du WWW sont quatre facteurs qui vont certainement améliorer la recherche dans les catalogues en ligne. Nous verrons dans ce qui suit un de ces développement : la navigation dans les catalogues : les hypercatalogues.

Chapitre trois : Les hypercatalogues

L'un des développements les plus importants de ces dernières années est l'intégration des techniques hypertextuelles dans les catalogues interactifs. Afin d'améliorer la recherche et la navigation dans les catalogues, nous avons suggéré d'améliorer la conception de ces systèmes selon deux niveaux (couches) (figure.9).

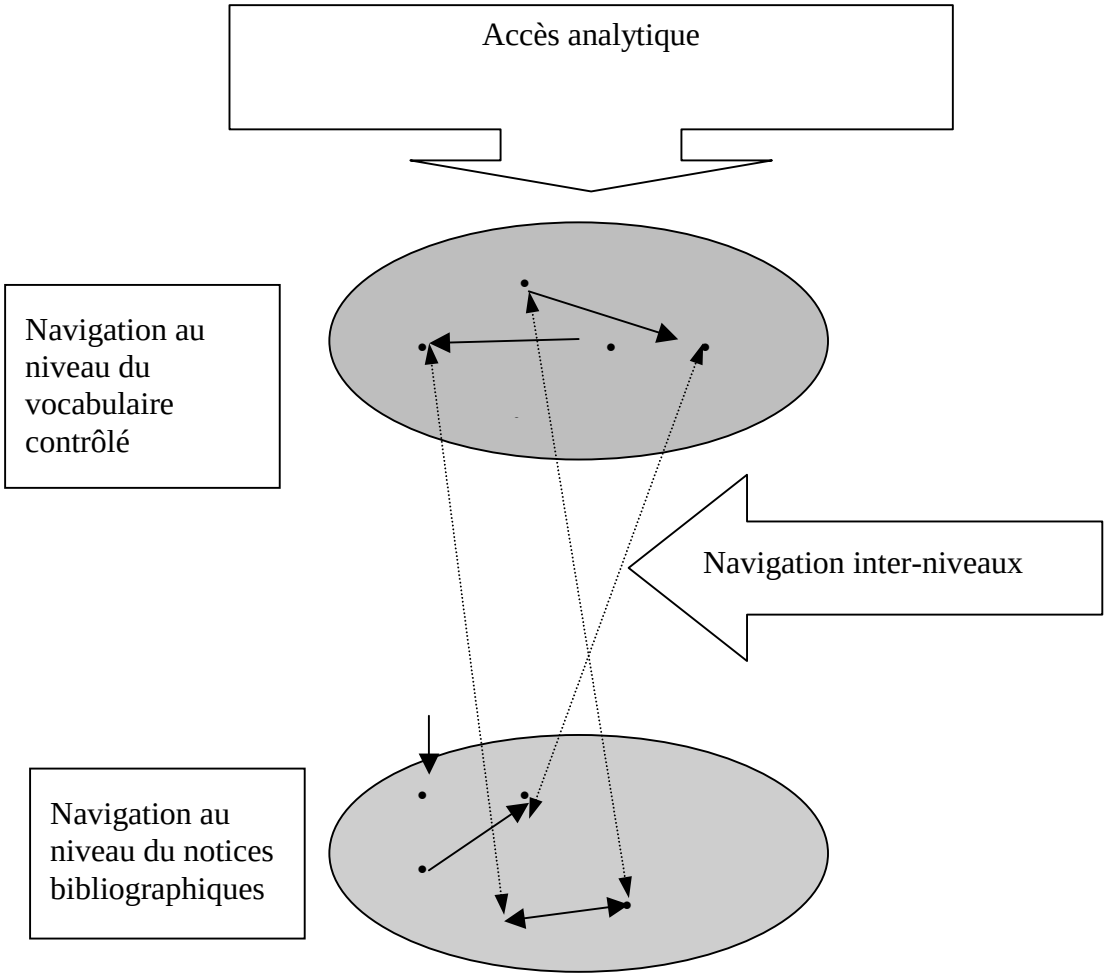


Figure 9 : schéma d'un hypercatalogue

Le premier niveau concerne la navigation dans les langages documentaires permettant à l'usager de se situer dans l'espace conceptuelle, le second permet à l'usager d'explorer l'espace des documents.

Les objectifs de ce chapitre sont :

- définir et préciser la notion d'hypercatalogue
- analyser les possibilités de représentations hypertextuelles des langages documentaires
- étudier la possibilité d'une représentation hypertextuelle de la liste d'autorité RAMEAU.
- effectuer une typologie des relations existantes dans les hypercatalogues.
- examiner les différents types d'hypercatalogues et leurs limites.

3.1 Hypertextes et langages documentaires

L'hypertexte est approprié dans des situations où un grand volume d'informations est partagé en plusieurs fragments dans lequel chaque fragment est en relation avec un autre, et dont l'utilisateur n'a besoin que d'un petit volume d'informations à un moment donné. Les vocabulaires contrôlés (thesaurus, classifications, listes d'autorité) conviennent bien à ce type d'organisation d'information. Il existe deux façons d'exploiter l'hypertexte et les vocabulaires contrôlés .

- Utiliser des vocabulaires contrôlés comme un outil d'aide à la navigation dans les hypertextes.
- Concevoir et développer des interfaces hypertextes pour permettre une navigation dans les vocabulaires contrôlés au niveau des systèmes de recherche d'informations.

Nous nous intéressons dans cette partie à ce dernier point et spécifiquement aux catalogues en ligne. L'hypothèse sous entendue dans la représentation d'un vocabulaire contrôlé sous forme hypertextuelle est de faciliter aux utilisateurs la connaissance les descripteurs ou des termes sujets utilisés dans une base de données. La stratégie de recherche s'en trouve simplifiée, car cela permet à l'utilisateur de désigner les termes dont il a besoin plutôt que de les chercher mentalement. La plupart des travaux sur les vocabulaires contrôlés ont porté particulièrement sur leurs méthodes d'organisation et de présentation mais n'ont pas pris en compte leurs usages en situation réelle d'interrogation. On sait peu de choses sur la manière dont sont effectivement utilisés les langages documentaires par les usagers.

3.1.1. Classifications et hypertexte

Certains auteurs ont cherché à rendre clair à l'usager l'organisation hiérarchique des connaissances en faisant des classifications (CDU, DDC, LCC) un outil de navigation. Divers catalogues en ligne offrent un accès sujet par le biais des indices de classification. Ce type d'accès pose deux inconvénients majeurs:

- Ces accès sont différents d'un catalogue à un autre.
- L'usager doit connaître l'indice exact qui représente son centre d'intérêt.

Dujol (1986) a montré aussi le décalage qui existe entre les instruments mis à la disposition des lecteurs d'une bibliothèque, les cotes et leurs usages réelles. Pour la majorité des usagers, les cotes sont un code dont ils ne comprennent pas le sens. Ils ne font pas la relation qui existent entre l'indice et l'intitulé. Elle constate que *" ce n'est pas une logique de la classification qui les guide, mais une logique de plan cartographique*".

Cette séparation du mode de symbolisation (notation) du langage d'interrogation est pour (DE Grolier, 1989) un des plus importants problèmes à résoudre si l'on veut aboutir à des systèmes plus satisfaisant que les systèmes actuelles. Elle est réalisé dans quelques catalogues en ligne qui permettent un accès à travers les libelles et non plus par le biais des indices de classification. Nonobstant, nous pensons cette séparation n'est pas suffisante. Pour que les classification jouent leurs rôles d'outils d'accès et de navigation à des fonds documentaires hétérogènes, il est faut que les systèmes qui les supportent respectent un ensemble de facteurs (Iyer, 1995)

- Permettre un accès à travers les libelles (mots clés).
- Améliorer l'indexation des libelles des classes et sous classes à travers une indexation en chaîne "chain indexing".
- Permettre un accès en multilingue: les efforts de traduction de la Dewey en plusieurs langues (anglais, français, arabe, hébreu, ...) constitue un effort important dans ce support multilingue.
- Intégrer la recherche analytique et la navigation . Nous pensons avec Nielson[20] que l'une des raison qui expliquent le succès de Yahoo (http://www.yahoo.com) est la facilité de

[20] Why Yahoo is good (http://www.usneit.com/jakob).

navigation hiérarchique et l'intégration des deux modes de consultations « recherche et navigation ». La classification est complétée par une recherche par mots clés qui permet de trouver directement l'indice correspondant à une notion, sans à voir à parcourir toute l'arborescence.

L'usage des classifications permet, d'une part d'assister les utilisateurs pour construire une requête, l'élargir ou la spécifier ; d'autre part il permet à l'utilisateur de "placer" les termes de sa recherche dans un voisinage sémantique. C'est pour résoudre ces deux problèmes, que divers projets ont émergé dès 1986 ayant pour objectif principal de tester l'efficacité de la mise en ligne des schémas de classification comme outil de recherche et de navigation sujet. Avec le développement récent d'un format pour les classifications (USMARC DATA CLASSIFICATION), il est possible d'améliorer l'accès et la navigation des systèmes de classification en ligne et d'utiliser pleinement les classifications comme outil d'accès sujet dans les OPACs.

3.1.1. 1. Hypertexte et classification Décimale de Dewey (DDC)

Par sa structure même, la classification Dewey est sujette à une représentation hypertextuelle. Les dix classes et leurs différentes sous-classes sont les nœuds et les différents liens entre une classe et une sous classe sont des liens hiérarchiques bidirectionnels. L'usager peut ainsi naviguer d'une classe à une sous classe et vice versa. Le projet Dewey Decimal Catalog initié à OCLC par Markey (1990) permet ce type de navigation hiérarchique. De plus, il incorpore les index ; le système permet ainsi différents stratégies de recherche dont l'une est une recherche par mot clé. Markey a testé l'efficacité de ce système et a démontré qu'il améliore le rappel et qu'on trouve par ce biais des notices introuvables par d'autres types d'accès. Après DDC , d'autres prototypes permettant une navigation dans les tables ont été construits. Les deux plus importants sont : DORS (Dewey Online Retrieval System) de Svenious (1991) et SLC (System Library Catalog) de Borgman (1995). L'élément le plus novateur de DORS est sans doute la disponibilité d'un "chain index" généré automatiquement à partir des différentes tables de la DDC. Alors que l'index relatif permet un feuilletage à l'intérieur d'une hiérarchie, le "chain index", en affichant un terme selon diverses perspectives, permet un feuilletage entre les différentes hiérarchies (classes). Développé sous hypercard, le projet SLC de Borgman présente une interface graphique autorisant une navigation hiérarchique dans les tables. Les dix classes de la DDC et les différentes sous-classes sont présentées sous forme de

rayons et d'étagères d'une bibliothèque. Pour Borgman, naviguer dans la classification revient à se " déplacer" entre les rayons d'une bibliothèque. L'usager s'y déplace en cliquant sur le rayon qu'il veut explorer. Le rayon s'affiche alors avec ses étagères qui correspondent aux sous-classes de la Dewey. En cliquant sur telle étagère, la sous-classe se détaille en ses composants de plus bas niveau, jusqu'à ce que l'on parvienne aux ouvrages. SLC incorpore des aides à la navigation comme le retour arrière, ou la fonction historique. Par contre il n'offre aucun accès par index, ni une recherche classique (booléenne). L'évaluation de l'utilisation du système par des enfants, en comparaison avec des OPACs classiques, n'indiquent que très peu de différences entre SLC et les systèmes classiques. Pollit (1997) présente une autre approche pour faciliter la navigation à travers la classification Dewey. Il déconstruit chaque indice en trois sujets (facettes), donnant ainsi à l'usager la possibilité de voir plusieurs « vues » de la base bibliographique. On peut enfin citer les travaux de (Allen,1994) avec son prototype HOPAC.

3.1.1. 2. Hypertexte et classification Décimale Universel (CDU)

Dans le cadre du projet européen HYPERLIB[21], une interface hypertextuelle à la CDU a été réalisée . Hyperlib permet un accès, soit en spécifiant un mot clé et le système affiche alors les classes contenant ce mot, soit en butinant d'une classe générale en sous-classes selon les hiérarchies existantes. Deux prototypes ont été développés, l'un permettant l'accès avec une interface VT100 , l'autre un accès de type WWW.

3.1.1. 3. Hypertexte et Classification du Congrès

Cette classification, contrairement à la classification de Dewey, a fait l'objet de peu de représentations hypertextuelles. Hildreth (1993) a développé un prototype qui offre la possibilité de lier des vedettes matières de la liste d'autorité de la bibliothèque du Congrès avec les indices de la classification du Congrès. Après une requête, le système permet d'afficher pour chaque notice bibliographique, en plus des vedettes matières, des termes extraits des classes. Ce prototype n'offre aucune présentation générale des différentes classes et sous-classes sous forme hypertexte

[21] http:/ / lib.ua.ac.be

Remarque

L'émergence du Web ainsi que le succès des moteurs de recherches "thématiques" tels que Yahoo ont stimulé l'usage de la Dewey ou de la CDU comme mode d'accès. Différents sites proposent ce type d'accès à leurs collections par l'intermédiaire des classifications, offrant ainsi une alternative à l'accès par mot-clef. C'est dans ce contexte qu'ont émergé des projets pour permettre une navigation à travers les classifications dans les catalogues de bibliothèque. On peut citer notamment le catalogue de l'Ircam[22] (figure 10) . Ces travaux n'ont malheureusement pas pris en compte les avancées théoriques, ni les résultats empiriques des premières études, notamment celles de Markey (1990) ou de Sveniuous (1990). Dans tous les cas, l'usager doit avoir une idée du thème dans lequel peut être classée l'information recherchée. Il ne faut pas non plus s'attendre à trouver instantanément la réponse à sa question. Aucun des catalogues des bibliothèques universitaires en France n'offre à l'heure actuelle cette possibilité. Nous pensons que la mise en ligne des classifications peut constituer une stratégie de recherche complémentaire à l'accès par les vedettes matières. Pour une meilleure représentation des classifications dans un environnement informatisé, on se doit de respecter un ensemble de critères de qualité comme ceux préconisés par Iyer (1995). Si les classifications en ligne peuvent jouer un rôle dans l'accès et la navigation , elles peuvent aussi jouer un rôle dans le classement et le filtrage des réponses. Nous verrons dans le chapitre neuf l'application d'un tel accès pour notre prototype CATHIE

[22] http://www.médiathèque.ircam.fr/catalogue

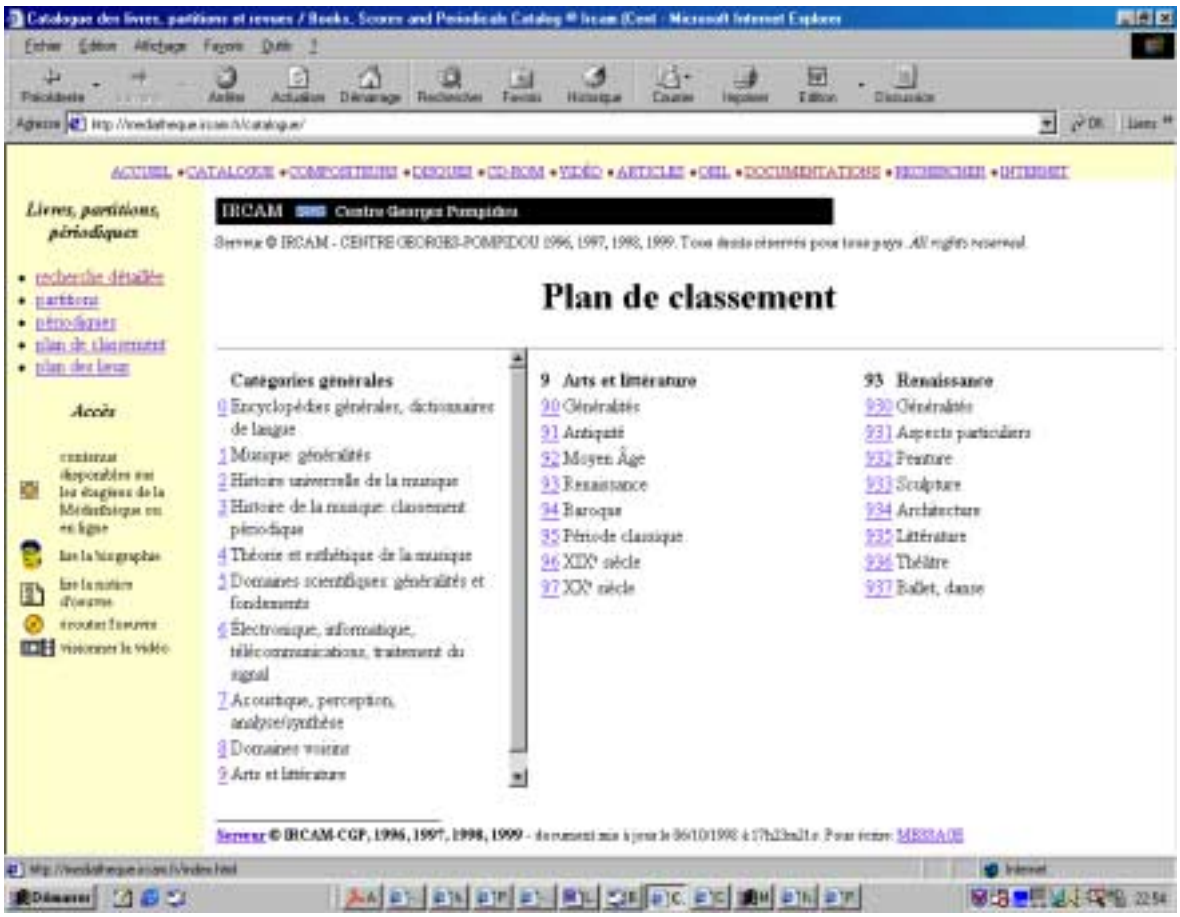


Figure 10 : plan de classement du catalogue de l'ircam

3.1.2. Thesaurus et Hypertexte

Les thésaurus sont rarement présents dans des OPACs. Seules quelques bibliothèques spécialisent utilisent des thésaurus pour l'indexation de leurs fonds documentaires.
Le contenu d'un thésaurus peut être représenté en trois modes principaux : la première présentation est alphabétique ; la seconde est systématique (organisation en domaines ou disciplines et organisation par facettes), enfin la dernière présentation est graphique. Les deux derniers types sont accompagnés d'un index alphabétique
Pour permettre une forme de navigation dans les thesaurus en ligne, Jones (1995) suggère trois types d'implémentations informatiques :

- thésaurus manipulables par menus déroulant,
- thésaurus sous forme de cartes graphiques,
- thésaurus sous forme d'hypertextes,

L'une des premières représentations d'un thésaurus sous forme hypertexte a été mise en place à Aberdeen (Mcleese,1989). L'usager a une vue panoramique du domaine qui l'intéresse et peut se focaliser sur un détail comme avec un téléobjectif.

Pollard (1993) a développé une version hypertexte du thésaurus de la base ERIC qui est constituée de plus de 10.000 termes (descripteurs et non descripteurs) et de plus de 66.000 relations. L'accès initial au thesaurus se fait par un index de mots-clefs (terme descripteur ou non). Chaque fois qu'il y a appariement entre un terme de la requête et un terme du thesaurus, le système affiche celui-ci avec toutes ses relations (terme général, terme spécifique, terme associé). Chaque fois qu'un utilisateur clique sur un terme, une vue du thesaurus est affichée avec toutes les termes associés au terme sélectionné. Ce système présente deux moyens d'aides à la navigation , le retour arrière, et la recherche par index de mots-clefs. Johnson et Cochrane (1995) a développé une interface hypertextuelle au thésaurus INSPEC. C'est une représentation sous forme d'arborescence hiérarchique. L'avantage principal de cette représentation hypertextuelle vient de la présentation de deux sections lors de l'affichage (figure11) . La partie gauche de l'interface présente à l 'utilisateur un index KWIC (keyword in context) ainsi que l'ensemble des termes associés (TA) affichés sous forme d'un nuage flottant autour du mot de la requête. La partie droite du système présente la liste des termes

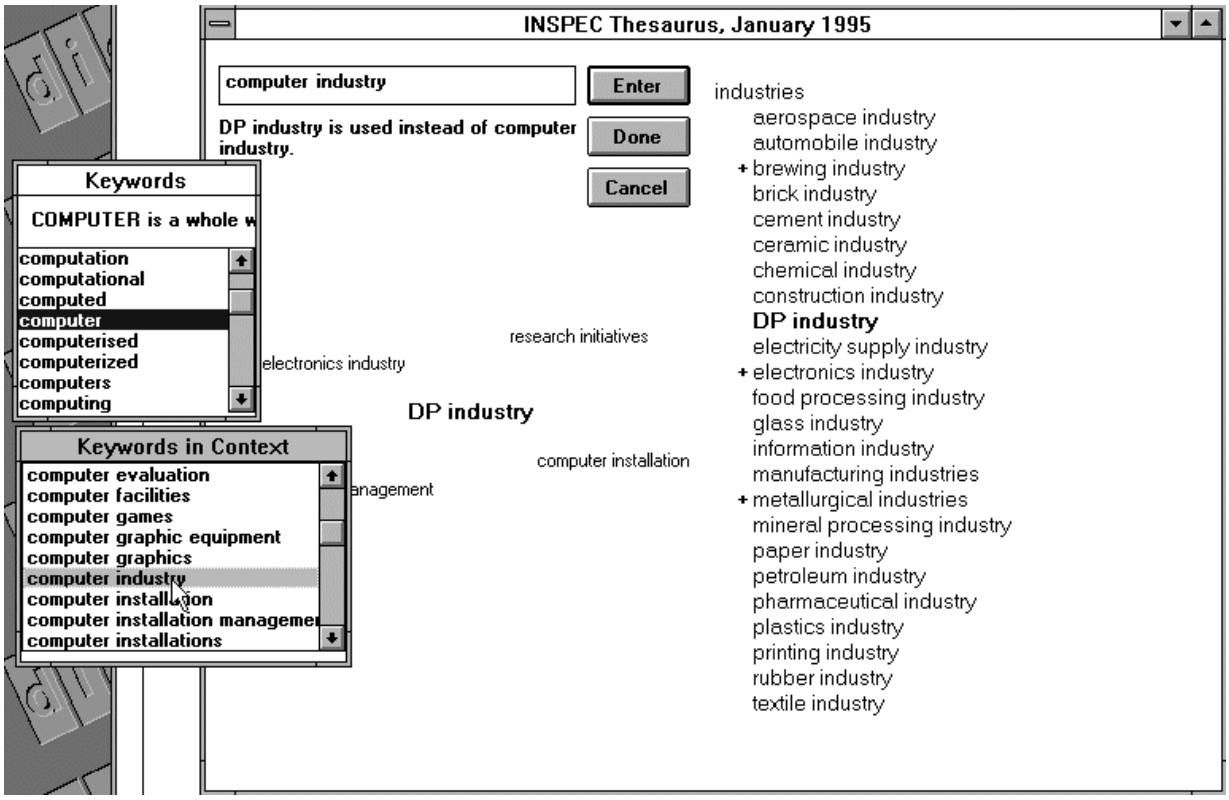


Figure 11 : représentation du thésaurus Inspec sous forme d'hypertexte.

reliés hiérarchiquement avec le terme choisi. Johnson utilise deux signes (+,-) pour indiquer ces relations hiérarchiques. L'absence de ces deux signes signifie l'absence d 'un terme spécifique. Pour atténuer le problème de désorientation, Johnson a introduit un système de signets permettant à l 'utilisateur de placer les termes du thesaurus qui lui semblent importants pour une utilisation future . Les premières expérimentations effectuées par Johnson démontrent que les usagers préfèrent grandement visualiser le nuage de termes reliés autour de l'entrée demandée, plutôt que de regarder la disposition hiérarchique plus traditionnelle. D'autres éditeurs offrent la possibilité de représentation hiérarchique. Par exemple, pour interroger la base de données MIDLINE , l'éditeur OVID[23] a implémenté une interface hypertextuelle sous forme d'une arborescence au thésaurus MESH (medical subject heading)

[23] http://www.ovid.com

Comme pour les classifications, Internet a renforcé cette représentation hypertextuelle. Il existe quatre types modes de consultation des thesaurus sur l'Internet[24] :

- Les thésaurus qui offrent uniquement la possibilité de consulter une liste alphabétique
- Les thésaurus qui offrent uniquement la possibilité de consulter une liste thématique
- Les thésaurus qui offrent la possibilité de faire uniquement la recherche
- Les thesaurus qui offrent ces trois possibilités (alphabétique, thématique et recherche).

La majorité des logiciels documentaires opérationnels offrent l'une de ces possibilités.

Les travaux sur l'usage des thésaurus aboutissent à des résultats qui sont parfois contradictoires. Pour Chen (1992) , par exemple, la consultation du thésaurus est l'une des stratégies de recherche des professionnels qu'il faut implémenter sur les systèmes destinés au grand public alors que Spink (1997) montre les thésaurus ne sont pas souvent utilisés (19%). Elle montre que l'utilisation du thésaurus n'aboutit pas d'ailleurs à de meilleurs résultats de recherche.

McAlese et Duncan (1988) présentent deux thésaurus graphiques à des enfants d'école élémentaire. L'un est un thésaurus papier et comporte de nombreux types de relations définies à partir de propositions des enfants. Le second est informatisé et fait appel aux relations classiques TG, TS, TA. Ces réseaux sémantiques sont présentés graphiquement sur un support papier. Les auteurs observent que leur utilisation conduit à corriger certaines erreurs d'interrogation souvent signalées chez les usagers (pas d'emploi de synonymes, termes trop génériques). Ils notent cependant que les relations autres que les relations hiérarchiques sont peu utilisées et que même ces dernières ne sont pas employées aussi complètement qu'elles pourraient l'être . Dans son étude sur l'usage du thésaurus Inspec, Jones (1995) a montré que les usagers n'ont pas de préférence pour l'une des trois relations (TG, TS, TA). Seul 10% a réellement employé les termes du thésaurus.

3.1.3. Hypertexte et listes de vedettes matières

Dans un catalogue, la liste d'autorité n'apparaît pas clairement. Les utilisateurs qui cherchent une vedette matière ne peuvent interroger cette liste que s'ils savent qu'elle existe. L'un des

[24] http://www.darmstadt.gmd.de/~lutes/thesalpha.html

vœux récurrents des bibliothécaires est de pouvoir disposer en ligne de la liste d'autorité et de ses renvois de synonymie et de hiérarchie (Ihadjadene,1998a). Une représentation graphique et hypertextuelle d'une liste d'autorité peut alors aider l'usager à rentrer dans le vocabulaire d'accès mais surtout lui permet de d'élargir ou de spécifier ses besoins d'informations. En effet, diverses données expérimentales montrent que les requêtes des utilisateurs sont communément soit très larges, soit très spécifiques (Markey,1994).
On peut ainsi exploiter l'hypertexte à deux niveaux :

- Pour afficher graphiquement les vedettes matières et les différents renvois.
- Pour élargir le vocabulaire d'accès à l'ensemble des termes non retenus. Ceci peut s'opérer en élaborant un réseau de liens hypertextuels reliant des termes des titres ou des tables de matières à des descripteurs. Il s'agit en fait d'élargir les renvois de synonymie déjà existants.

Le prototype ILSA (Improving library Subject Acces) de Micco (1991) intègre l'hypertexte comme une interface frontale aux interrogation en langage naturel: Les mots clés extraits des tables de matières sont liées (via les liens hypertextes) aux vedettes matières de LCSH. Pour 48.000 notices enrichies par l'incorporation de termes des tables de matières, Micco a construit un réseau de plus de un million de liens. Une autre possibilité de ILSA est de mettre en contexte les vedettes matières de la LCSH et les classes de la DDC.

3.1.3. 1. L'affichage des liens de la liste d'autorité

Dans le cas d'une liste d'autorité, les vedettes représentent les nœuds et les différentes relations sémantiques entre les vedettes représentent les liens. Une telle représentation pose cependant un certain nombre de problèmes. Parmi les plus importants, citons:

Les problème d'affichage des subdivisions

Un des problèmes majeurs de l'accès sujet concerne la surcharge d'information qui se présente lorsque la visualisation consiste en un long segment de l'index matière dans lequel une vedette-matière est suivie d'un ensemble impressionnant de subdivisions. Par exemple à partir d'un mot courant comme Arts on feuillette des quantités de sous vedettes pendant des dizaines

d'écrans. Une solution à ce problème consiste à utiliser à la fois un affichage structuré de l'index et des techniques de compression qui permettent d'introduire un ordre conceptuel dans la présentation des vedettes. Il s'agit d'une part de regrouper les subdivisions de forme et de sujet en des catégories conceptuelles générales représentant les différentes facettes d'une vedette sujet, et d'autre part la suppression de certaines subdivisions (par exemple les subdivisions géographiques) à l'affichage principal. La représentation des vedettes et de leurs subdivisions devient alors manipulable et l'utilisateur peut ainsi butiner d'un niveau hiérarchique à un autre selon qu'une vedette admet un niveau de subdivision ou plus. Les tests effectués par (Allen, 1993) indiquent qu'une telle représentation réduit l'acte de feuilletage de plus de 50 % sans réduire la pertinence de recherche. Cependant, le nombre de subdivisons est encore très limité (sujet, géographique, chronologique, forme).

Un autre problème concerne la surcharge d'information "horizontale" : les vedettes matières sont longues et complexes. Exemples :

- *« Indonésie -- Histoire -- 1825-1830 (Guerre de Java) -- Récits personnels belges »*
- *« Chansons soudanaises -- Indonésie -- Java (Indonésie ; ouest) -- Histoire et critique*

Ces problèmes sont la conséquence principale de la pratique des subdivisions affranchies. Nous pensons qu 'il est nécessaire de réduire la pratique des subdivisions affranchies, et qu'au lieu d'indexer un livre par une ou deux vedettes matières construites aussi longues, il serait préférable d'augmenter le nombre de vedettes moins longues, mais qui expriment certains aspects du document. Les utilisateurs comprendront plus facilement la signification des vedettes matières moins longues :" *to increase end-user understanding, efforts must be taken to shorten subject headings by reducing the number of subdivisions per string and by reducing the number of words in individual subject heading subdivisions*" (Franz, 1994) .

Les problèmes de renvois dans la liste d'autorité.

Il s'agit essentiellement des vedettes qui n'admettent pas de renvois " les termes sans liens " qui représentent prés de 20 % dans le cas de la liste d'autorité du congrès (Library of Congress Subject Headings) (Markey,1994). Ceci signifie que dans un cas sur cinq une vedette matière sujet n'a aucun lien qui permette de cheminer dans la hiérarchie. Plus de 50 % des vedettes qui admettent une subdivision n'ont pas de liens avec d'autres vedettes.

Un autre problème de renvoi concerne les termes qui ne pointent pas vers des notices bibliographiques et qui encombrent la liste d'autorité inutilement. Une solution pragmatique à ce problème est le développement de programmes informatiques de "nettoyage" permettant d'identifier les éléments de la liste qui ne pointent pas à des notices bibliographiques. C'est la solution adoptée à l'université de LAVAL pour la liste RVM (répertoire des vedettes matières). L'usager n'est donc pas dirigé vers un sujet pour lequel la bibliothèque n'a pas encore de document. Une autre solution est d'afficher l'ensemble des renvois contenus dans la notice d'autorité qu'ils conduisent ou non au signalement de documents existants dans la bibliothèque. Ces renvois peuvent être " actifs " ou " inactifs ". Cette terminologie signifie que le renvoi a, dans le premier cas, pour fonction de renvoyer à un document, et dans le deuxième cas de proposer une orientation de recherche dépourvue de finalité immédiatement bibliographique. La mise en ligne d'une liste d'autorité n'est opérationnelle que dans le cas de bibliothèques qui n'effectuent pas de catalogage en local car les relations de la liste ne portent que sur les autorités et non sur les vedettes construites. Nous avons montré qu'actuellement, seules les grandes bibliothèques productrices des listes d'autorités (bibliothèques du congrès et la BNF) peuvent le faire. On peut aussi noter l'incohérence du réseau de renvois de la liste d'autorité. En effet depuis une dizaine d'années, la bibliothèque du congrès a modifié le réseau des renvois de la liste LCSH pour les rapprocher de celui d'un thésaurus. Le remplacement d'une façon automatique des renvois « voir » et « voir aussi » par des relations sémantiques (TG,TA,TS) n'a fait que révéler la faiblesse et l'incohérence des hiérarchies de la liste.

Enfin, nous pouvons constater que l'affichage en ligne de la liste ne règle pas le problème de surcharge d'information. Une question sur le droit aboutit à l'affichage de dix écrans dans RAMEAU. Or comme le signale Lahary (1994) : « *un bruit de vedettes est certainement pire qu'un bruit de notices* ». Nous pensons d'ailleurs que l'affichage des différentes relations aboutira fatalement à un problème de désorientation. La distinction entre ces renvois n'est pas claire pour les usagers. Sinkakas (1977) constate qu'ils se perdent en digression en suivant les renvois voir et voir aussi.

3.2. Les liens dans l’espace documentaire

Ceux qui ont écrit sur les hypercatalogues n'ont pris en compte que l’aspect navigation et la facilité d'utilisation , sans aucune référence aux objectifs du catalogues. Ce sont donc des théoriciens du catalogage et du contrôle bibliographique qui ont montré l’importance de lier les différentes éditions d'une œuvre , cela bien avant l’introduction et le développement de l’hypertexte. Dans notre partie, nous parlerons souvent de l’importance de cette technique pour faciliter la recherche et mettre en évidence sa relation avec des techniques telles que la reformulation interactive. Néanmoins, cette analyse est à placer dans un contexte plus général qui est celui d’étudier les moyens de mise en œuvre du deuxième objectif du catalogue à savoir faciliter le regroupement des œuvres bibliographiques.

3.2. 1. Les relations dans l'espace documentaire

3.2. 1.1 Les relations bibliographiques

Pour étudier et analyser comment sont définies les relations entre les documents, il est nécessaire d'abord de revenir sur les objectifs du catalogues pour déterminer le niveau où se situe le catalogage. Cette pratique (catalogage à niveau) fut introduite en France dés 1960 et était réservé aux exclusivement monographies en plusieurs volumes :

- A un premier niveau, on notait, toutes les informations communes à l’ensemble des différents volumes composant la monographie en son entier.
- A un niveau inférieur, on transcrivait les éléments propres à chaque volume.

Cette solution a le mérite d’être économique, Pour Cazabon (1993) : « en évitant les redondances, elle faisait gagner du temps et de la place dans un catalogue manuel. Un catalogueur n’avait pas à reprendre en totalité la notice chapeau dès qu’il ajoutait un nouveau volume à la suite et se limitait à un rappel chapeau abrégé ». En intégrant certaines recommandations de l’IFLA, L’AFNOR a abandonné, le catalogage à niveaux dès 1989. C'est la pratique du catalogage partagé qui est à l’origine de l'abandon du catalogage à niveaux. En effet les notices de monographies en plusieurs volumes établies selon le catalogage à niveaux étaient rejetées par les bibliothèques anglo-saxonnes.

3.2. 1.2 Définition de la notion d'œuvre:

Une œuvre (traduction de l'anglais du concept de work) est une création intellectuelle ou artistique distincte. Il s'agit d'une entité abstraite qui permet d'attribuer à une création intellectuelle ou artistique un nom et des liens relationnels. Par exemple "les misérables" de Victor Hugo. Il n'existe pas de définition précise de la notion de "work" ou d'œuvre.

Voici la définition proposée Simirglia (1999) :

> *" the intellectual content of a bibliographic entity; any work has two properties: a) the propositions expressed, which form ideational content; and b) the expression of those propositions (usually a particular set of linguistic , musical, string) which form semantic content. Any variation in the linguistic content of work is considered to result in the creation of a new work*".

Une entité bibliographique se compose de deux éléments :

- une œuvre intellectuelle
- une entité physique.

Jusqu'à maintenant, Les SRI ne prennent en comptent que le deuxième élément (item physique). L'œuvre intellectuelle est décrite implicitement au moyen du regroupement. Cette méthode a induit un outil , le catalogue, qui ne décrit pas les attributs d'une œuvre bibliographique (historique, son genre, forme intellectuelle et les relations bibliographiques avec les autres œuvres). Un groupe de spécialistes de l'Ifla (1996) , sous la responsabilité de Madisson , a effectué une étude sur les besoins fonctionnels pour les notices bibliographiques. Ils ont permis de distinguer ces quatre entités:

1. **Œuvre** : c'est une création intellectuelle ou artistique distincte. Il s'agit d'une entité abstraite qui permet d'attribuer à une création intellectuelle ou artistique un nom et des liens relationnels. *Ex: le nom de la rose d'Umberto Eco*
2. **Expression** : la réalisation intellectuelle ou artistique d'une œuvre. Cette entité comprend les termes, phrases ou paragraphes particuliers qui résultent de la réalisation ou de l'expression d'une œuvre. Ex: la traduction française de Jean-Noel Schifano
3. **Manifestation** : l'incarnation matérielle d'une expression d'une œuvre. La manifestation correspond à l'ensemble des objets matériels qui présentent les mêmes caractéristiques quant à leur contenu intellectuel et leur forme matérielle. Ex: l'édition broché parue chez Grasset en 1982

4. **Exemplaire** : Exemplaire unique d'une manifestation. Dans de nombreux cas, il s'agit d'un objet physique unique. Ex: l'exemplaire de BNF

Panizzi, cité par Leazer (1993), fut sans doute parmi les premiers « bibliothécaires » à soutenir et à défendre l'idée d'un catalogue qui contrôle des œuvres bibliographiques. Pour lui : " a reader may know the work he requires ; he cannot be expected to know all the parculiarities of differents editions … » . Il estimait fortement que le rôle d'un catalogue est d'identifier, d'afficher et de regrouper les différentes éditions d'une œuvre. Pour lui, un catalogue doit présenter ces cinq caractéristique :

- Il donne suffisamment de détails pour faciliter l'identification d'un ouvrage
- Il doit posséder une seule « entrée » pour chaque unité bibliographique.
- Cette entrée doit être normalisée
- toutes les éditions et les traductions d'une œuvre doivent être disposées ensemble
- enfin, il y a affichage des renvois (pour les auteurs et titre).

Ce sont donc les deux dernières fonctions qui constituent cet objectif de regroupement.

Quelques années plus tard et dans un contexte différent , Cutter (1904) énonçait dans son célèbre ouvrage « Rules for printed dictionnary catalog » les objectifs d'un catalogue manuel. Ce sont les principales sources et règles de catalogage pour les bibliothèques dans le monde.

Selon les objectifs de Cutter, le classement d'un catalogue manuel est basé sur l'idée qu'un usager cherche en connaissant au moins l'un des trois points d'accès (auteur, titre, sujet). Or les études montrent que les usagers arrivent au catalogue avec une information incomplète sur ces clés d'accès. Ils emploient d'autres sources externes au catalogue (bibliographies, la liste d'autorité) pour obtenir des données afin de mieux exprimer leurs besoins d'information. On peut noter que dans la version de Cutter , seul le regroupement des œuvres d'un auteur est rendu explicite. Pour lui, le moyen le plus favorable dans le choix d'un ouvrage est l'édition elle même (et les notes). Cutter ne préconise pas d'afficher l'ensemble des éditions d'une œuvre pour faciliter le choix de l'usager.

Cette fonction de regroupement, connue chez les américains sous le nom du deuxième objectif : « the second objective » est distincte de celle formulée par Panizzi. Selon Cutter le regroupement n'inclut pas les différentes éditions ou traductions d'une œuvre mais se limite à rassembler tous les ouvrages qui ont en commun soit l'auteur, soit le sujet , soit la forme.

Yee (1996) montre que bien avant Cutter, des « théoriciens » du catalogage ont recommandé de disposer ensemble les œuvres d'un même auteur, notamment Jewett, Bodleain et Panizzi. Cutter fut cependant le premier à recenser explicitement les objectifs d'un catalogue et a

indiqué les techniques pour les implémenter. Sous l'égide de la bibliothèque du Congrès, Lubetzky, cité par (Leazer,1993), a réexaminé les fonctions d'un catalogue . Son étude a servi de base de travail pour la conférence internationale sur les principes de catalogage de PARIS en 1961 (Ifla,1961). Le catalogue doit être un instrument efficace pour déterminer :

- Si la bibliothèque contient un livre particulier caractérisé par
  - Par son auteur et son titre, *ou*
  - Si l'auteur n'est pas nommé dans le livre, par son titre seul, *ou*
  - Si l'auteur et le titre ne conviennent pas ou sont insuffisants pour l'identification, par un élément de remplacement approprié.

- Quelles œuvres d'un auteur particulier
- Quelles éditions d'une œuvre particulière } figurent dans la bibliothèque.

Pour Lubetzky les deux premiers objectifs sont en conflit. En effet le premier objectif stipule que l'unité de base de la description bibliographique est l'item (ouvrage) alors que le deuxième objectif sous entend le contrôle de l'œuvre bibliographique. A la conférence[25] internationale sur les principes de catalogage de Paris en 1961 , seul le premier objectif a été entièrement approuvé; les deux autres ne le sont que partiellement. Actuellement, le niveau où se situe le catalogage est celui de l'édition du document et non celui de l'œuvre.

D'autres chercheurs ont aussi critiqué le choix de l'ouvrage "item" comme unité de catalogage. Vizine (1989) considère que les objectifs d'un catalogue comme ceux exprimés à PARIS sont inadéquats à l'ère de l 'accès à distance et des catalogues collectifs. Wilson (1989) critique pour sa part, non les objectifs mais leur priorités . Il précise que le regroupement des œuvres doit être l'objectif principal d'un catalogue. Il distingue à cet effet le « médium » d'enregistrement des documents du « médium » d'affichage. Pour lui, un catalogue doit permettre l'accès à des « copies virtuelles qui existent dans les catalogues distants et qui n'existent pas en local. Il considère qu'à l'ère des accès à distant et des catalogues collectifs,

[25] Ces principes connus sous le nom de " principe de Paris" sont à l'origine des règles de catalogage dans plusieurs pays, notamment les AACR en Etats-Unis.

c’est l’œuvre qui constitue un intérêt et non l’item physique. Ce qu'il importe de décrire, c'est le contenu du document, et non sa forme matérielle.

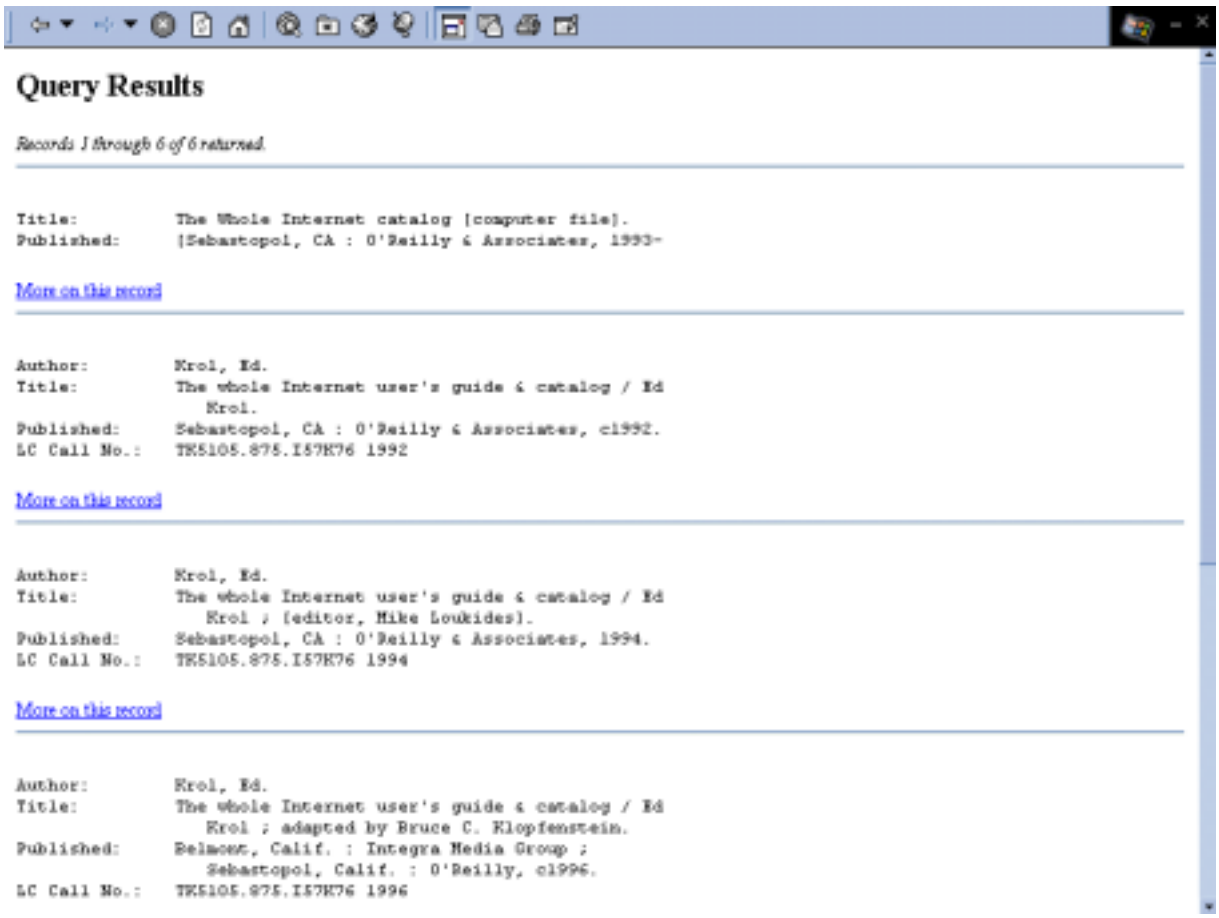
L'intérêt porté à la description physique d’un document au lieu de l’œuvre est contestable pour plusieurs raisons :

un usager porte plus d'intérêt à l’œuvre et non à une édition bien particulière. Pour preuve, la majorité des études d’usage montrent que les champs bibliographiques qui permettent de distinguer les différentes éditions ne sont jamais employés.

Les catalogues en ligne mais aussi d’autres SRI qui sont en réseau sont utilisés pour rechercher des œuvres. Lorsqu’un usager veut un document particulier (une édition) il peut, soit le vérifier dans son catalogue en local, dans les étagères d’une bibliothèque ou bien en utilisant d’autres moyens comme le PEB (prêt en bibliothèques).

Le modèle « bibliographique » n’est plus valable à l’heure d'Internet. En effet pour un même document, il existe plusieurs versions électroniques mises à jour continuellement.

Ne pas prendre en compte les compte l’œuvre comme unité de catalogage augmente le problème de la surcharge d'information. Une question portant sur le livre « the whole internet » permet d’avoir six réponses dans le catalogue de la bibliothèque du congrès (figure 12). Or un affichage d’une seule notice aurait pu suffire s’il n’y avait un lien hypertexte qui indique « les autres éditions du livre ». Ceci permet de réduire le nombre de réponse à afficher.



Query Results

Records 1 through 6 of 6 returned.

Title: The Whole Internet catalog [computer file].
Published: [Sebastopol, CA : O'Reilly & Associates, 1993-

More on this record

Author: Krol, Ed.
Title: The whole Internet user's guide & catalog / Ed Krol.
Published: Sebastopol, CA : O'Reilly & Associates, c1992.
LC Call No.: TK5105.875.I57K76 1992

More on this record

Author: Krol, Ed.
Title: The whole Internet user's guide & catalog / Ed Krol ; [editor, Mike Loukides].
Published: Sebastopol, CA : O'Reilly & Associates, 1994.
LC Call No.: TK5105.875.I57K76 1994

More on this record

Author: Krol, Ed.
Title: The whole Internet user's guide & catalog / Ed Krol ; adapted by Bruce C. Klopfenstein.
Published: Belmont, Calif. : Integra Media Group ; Sebastopol, Calif. : O'Reilly, c1996.
LC Call No.: TK5105.875.I57K76 1996

Figure 12 : exemple de l'affichage d'une oeuvre

En résumé, pour les tenants de cette approche, l'usager ne cherche pas un ouvrage particulier mais une œuvre à partir de laquelle , le catalogue en ligne doit lui proposer l'ensemble des éditions (ou manifestations) de cette œuvre et/ou les liens lui permettant d'atteindre d'autres œuvres distinctes, mais qui restent liés par un ensemble de relations bibliographiques.

Voici les arguments qui soutiennent la prise en compte de l'entité physique dans le catalogage :

- Le premier argument et sans doute le plus important est la difficulté , sinon l'imprécision, dans la définition de la notion d'œuvre.

- Pour Velluci (1997) , les résultats de la majorité des études effectuées sur l'usage des catalogues « mais aussi sur les SRI » sont inadéquats pour clarifier si les usagers préfèrent des œuvres bibliographiques ou seulement des ouvrages bien précis. La principale critique concerne la méthodologie utilisée. L'utilisation de l'analyse transactionnelle ne permet pas de savoir si un usager cherche une œuvre (représentée par une ou plusieurs notices) ou un item (représenté par une notice spécifique). De même, ces données collectées ne permettent pas de savoir comment les éléments sont utilisés pour identifier un type de relation.
- Un autre problème est d'ordre « linguistique », ils est difficile pour un usager de « discuter » du concept des relations bibliographiques (surtout lorsqu'on voit l'inconsistance dans la manière dont les liens sont exprimés dans les formats MARC).
- La majorité des études trouve que les accès par sujet , par auteur et par titre sont les mode de recherches les plus employés par les usagers. Les auteurs de ces études présument que ce sont des recherches sur des éléments « connus » : l'usager possède déjà une partie de l'information demandée. Pour Velluci (1997), personne n'a pris en compte le fait qu'un usager peut interroger le catalogue pour chercher des versions d'une œuvre par exemple.
- D'autre part avec l'analyse transactionnelle, , il est impossible de savoir comment l'usager détermine la nature d'une relation sans une interview avec lui.
- Enfin, il est parfois difficile de généraliser à partir de résultats de l'analyse transactionnelle. Ainsi Velluci note que si le champ « accompaying material » est rarement usité dans un catalogue classique , il l'est dans le cas d'un catalogue d'un fonds musical.

3.2. 2. Typologie des relations bibliographiques

Une relation bibliographique est donc une association entre deux entités bibliographiques, c'est à dire: R(œuvre, œuvre) , R (œuvre, Expression) et (manifestation, exemplaire). Très peu d'études ont été effectuées sur la définition des relations entre les documents. Dans cette partie, nous résumons principalement les travaux de Tillet (1989), Leazer (1993) et de Smirglia (1999). Dans son étude analytique, Tillet (1989) a déterminé les différents moyens et outils mis en œuvre utilisés pour décrire les liens entre les entités bibliographiques dans 24

codes de catalogage , depuis les travaux de Panizzi aux règles récentes américaines du AACR2.

1: **la relation d'équivalence:** qui lie une œuvre et ses copies exactes existants sur différents supports indépendamment du "médium" de présentation (imprimé, microfilm, électronique...). Elle permet de lier :

- les différentes copies d'une même manifestation d'œuvre
- une œuvre originale et ses reproductions tant que le contenu intellectuel est préservé.

2: **les relations dérivées**: qui lient une œuvre bibliographique et une modification de la même œuvre , elle peut être soit:

- une version d'une œuvre (édition, révisions, traduction, résumé…etc.)
- une adaptation qui devient une autre œuvre mais qui reste liée à d'autre œuvres.
- Un changement de genre : dramatisation
- Une nouvelle œuvre

3: la relation descriptive: qui lie une œuvre bibliographique et ses critiques , commentaires, les éditions annotées.

4: la relation hiérarchique: qui lie une œuvre et une partie de celle-ci. (i. e anthologie, série)

5: la relation d'accompagnement: qui lie une œuvre bibliographique et des documents auxiliaires (cassette accompagnant un livre, ...)

6: la relation séquentielle: qui lie deux œuvres bibliographiques qui se suivent, par exemple la relation qui lie une publication en série catalogue à un titre précédant.

7: la relation de caractéristiques communes "shared caracteristique" : qui lie des œuvres bibliographiques ayant par exemple un de ces éléments en commun: auteur, titre, sujet, indice Dewey.

Ces relations entre notices bibliographiques existent partiellement dans *le format UNIMARC (bloc 4XX).*

L'intérêt des données empiriques ci-dessous est de mettre en valeur le pourcentage de notices bibliographiques qui ont des liens entre elles, bien que les catalogues actuels ne le signalent pas. Tillet (1989) montre que presque 75 % de la base MARC du catalogue de la bibliothèque du congrès, admettent un ou plusieurs types de ces relations dont 3.35 % sont des relations

d'équivalence, 14.27% sont des relations dérivées , 62.36% sont des relations Tout-partie, 3.91% sont des relations d'accompagnement et enfin 16.11% sont des relations séquentielles. Smirglia (1992) estime que près de 49.9 % de son corpus a des relations dérivées. Contrairement à Tillet (1989), il trouve que prés de 40% de ces relations ne sont pas indiquées dans les notices. Il estime que les facteurs suivants (discipline, langue, date et place de publication) influent sur ce pourcentage. Enfin, Pour sa part, Vellucci (1997) montre que près de 97 % des notices d'une base (musicale) ont au moins un lien avec une autre notice .

3.2. 3. les relations dans les métadonnées

Plusieurs métadonnées incorporent un ou plusieurs champs "relations". C'est le cas de Dublin-Core (Is part, HasPart, Is version of, Has version, Is format of, Has format, References, Is referenced by); de GILS[26] (Cross reference Title , Cross reference linkage, Cross reference Type linkage , linkage type); de NEDRES (related publications, related records, requires, Is required By); EAD (description of subordinate components,related material, separated material); FGDC[27] (metadata entry level, metadata identifier, metadata serie object identifier, metadata set object identifier).

3.2. 4. Les relations non-bibliographiques dans l'espace documentaire.

Les chercheurs en informatique documentaire se sont intéressés tardivement aux relations entre les documents. Leur intérêt s'est focalisé à l'apport des liens (entre documents) dans la recherche d'information. Les travaux de Croft (1991), Turtle (1991) et de Savoy (1992) représentent ce courant de recherche. D'autres auteurs se sont attachés à montrer la catégorisation des relations pour faciliter la représentation visuelle graphique de l'espace documentaire et d'améliorer les algorithmes de filtrage et clustérisation des documents. Les documents peuvent être ici, des articles, des pièces musicales, des images. Le récent séminaire de SIGIR (Hetzler,1997) a permis de regrouper les chercheurs qui travaillent sur ce domaine. Ils ont développé huit catégories de relations dont certaines ressemblent à celles définies par Tillet (1989). Le prototype SPIRE de Hetzler (1998) permet de représenter en 3D une partie

[26] http://www.usgs.gov/gils

[27] Content Standards for digital geospatial metadata (FGDC) (http://www.fdgc.gov)

de ces relations. Savoy (1992) a démontré l'utilité des liens hypertextes dans la recherche d'information. Il propose un modèle qui prend en considération les relations entre les documents. L'idée est que si un document est « *jugé pertinent par le SRI, les documents lui faisant référence devraient également être extraits ou pour le moins voir leur degré de similarité avec la requête augmenter* » . Les travaux de (Frisse,1989) et de (Croft,1991) constituent les deux premières approches qui tiennent en compte des relations entre documents. Frisse (1989) suggère de compléter la recherche d'information vectorielle en tenant compte des liens ou similarités entre les documents. Ici les références bibliographiques sont traitées comme des liens hypertextes

Une variété de liens offre la possibilité de structurer une collection de documents sous forme d'un hypertexte, par exemple:

- les liens de référence bibliographiques
- le lien plus proche voisin "nearest neighbor link" calculé grâce à la formule de DICE .
- les liens de citation

Savoy (1992) a construit un hypertexte dans lequel les références bibliographiques sont traitées comme des liens hypertextes.

Le degré de pertinence d'un document D_i se calcule selon la formule suivante:

$$DP(D_i) = PI(D_i) + PE(D_i)$$

$$PI(D_i) = \sum_{j=1}^{q} w_{ij}$$

$$PE(D_i) = \sum_{k=1}^{r} \alpha_{ik} \, . \, DP(D_{ik})$$

α_{ik} α reflète la force du lien unissant les nœuds D_i et D_k

Savoy (1992) démontre que pour des valeurs de α=0.3 à α=0.5 , le schéma de recherche donne des résultats meilleurs que le modèle vectoriel seul (α=0.0). La précision augmente de 6,8% pour (α=0,4). Il a ensuite testé l'efficacité de la relation « plus proche voisin » comme lien hypertexte. Les résultats auxquels il aboutit ne sont pas satisfaisants. Dans l'étude de

Turtle (1991), les liens de citations apportent une augmentation de la précision de prés de 7.3 %.

Remarques

1. Nous n’avons pas trouvé d’étude qui tient compte des relations de citations dans un fonds encyclopédique, ni dans un catalogue en ligne. D 'ailleurs, nous ne pensons pas que cela puisse augmenter les performances de recherche du système. Lorsqu’on sait que près de 20 % d’une collection d’une bibliothèque est constituée d’œuvres de collaboration, dont les parties ne sont pas toujours similaires, il est fort probable que cela générerait du bruit.
2. L’intérêt des travaux sur les relations (bibliographiques ou non) , est que de plus en plus des chercheurs pensent que l’on peut exploiter ces liens dans la recherche d’information sur l’Internet. Google (www.google.com) est un exemple de moteurs de recherches qui prend en compte les liens hypertextes dans la recherche de documents.

3.3. Analyse des hypercatalogues .

3.3.1. Le projet HYPERCATlog:

La finalité de ce projet consiste à développer un système dont le mode de consultation primaire est la navigation. Rolland Hjerppe (1985) a décrit son projet HYPERCAT dès 1985 avant que le concept de l'hypertexte ou de groupeware ne soient pas vraiment popularisés. Pour lui, HYPERCAT doit être une extension et un enrichissement des catalogues traditionnels. Il doit supporter les fonctionnalités suivantes:

- La navigation non-linéaire est le mode d'accès principal à la base de donnée.
- Présenter des moyens alternatifs de présentation de la structure de l'information (nœud, relations) .
- Donner la possibilité à l'usager de créer ses propres liens
- Inclure des possibilités de recherche personnelle (annotation) et de filtrage.
- Avoir un modèle des usagers pour faciliter l'adaptation du système aux différents usagers.
- Comporter plus de relations que le catalogue classique.
- En plus des informations sur les notices, il doit présenter des informations sur la contenu de la collection sous forme graphique.

Il n'était pas possible de construire une interface avec des possibilités hypertexte, agissant directement sur le système opérationnel. Ils ont développé à la place des prototypes de démonstration de quelques centaines de notices sur des logiciels tels que Hypercard, Notecards et Guide. Ils se sont aperçus que la conception était limité par le choix des outils existants et leurs tentatives de généraliser cette solution pour les grandes bases se révélèrent un échec

Nous classons les Hypercatalogues en quatre catégories

1. Ceux basés sur les liens d'indexation :
2. les hypercatalogues basés sur la notion " d'œuvre "
3. les systèmes basés sur la métaphore de l'étagère:

4. les Hypercatalogues collaboratifs : Dans cette catégorie nous regroupons l'ensemble des travaux qui s'intéressent au filtrage collaboratif .

3.3. 2. Les hypercatalogues basés sur les liens d'indexation

3.3. 2. 1. WHIRZD: (Windewed Hypertext Interface for Zippy Retrievel and Display)

Le but de ce système développé par (Nelson,1991) est d'ajouter la possibilité de faire une recherche par navigation sur les systèmes de recherche existants. Les nœuds représentent soit un document ou un ensemble de documents alors que les liens incluent les auteurs, les classifications, les vedettes-matière, les titres, les revues, les titres de collection, les mots du thésaurus et du résumé**.** La recherche s'effectue comme dans un SRI booléen classique mais la structure de la recherche et la présentation des notices sont de type hypertexte. De plus, l'usager peut copier des enregistrements dans une autre fenêtre pour créer une bibliographie personnelle et avoir l'option d'y ajouter des commentaires. L'usager peut utiliser soit la souris, soit le clavier comme dispositifs de sélection de liens. Les aides à la navigation sont très limités (retour arrière). L'auteur n'a pas effectué d'évaluation sur son prototype.

3.3. 2. 2. La navigation dans les OPACs opérationnels.

Certains OPACs opérationnels, mais aussi des bases de données bibliographiques sur CD-ROMS (LISA, EMBASE, Medline) ont déjà employé l'approche hypertexte. On peu citer:

- TINman
- CARL : fonction " Express Search "
- Marquis: " la commande RW : related work "
- Dynix: " related work menu "
- Advance (GEAC) avec sa fonction TARZAN
- Unicorn: " commande Like "
- INNOPAC : " show items with the same SUBJECT "

Cette implémentation reste pour beaucoup de concepteurs “ marginale ”, elle n'est pas conçue pour des usagers grand public. D’ailleurs ces systèmes présentent les inconvénients suivants (Ihadjadene,1999b):

- D'utiliser le curseur pour désigner l'information textuelle ; la navigation est donc difficile.
- Les aides à la navigation sont inexistantes.
- la navigation est souvent limitée au champs auteur et au sujet.
- L'usager doit comprendre la fonction de regroupement des vedettes matières ou des auteurs pour pouvoir activer ces fonctions.
- Utilisation de commandes pour revenir en arrière

Nous n'avons pas trouvé d'études sur l'usage de ces fonctions dans les catalogues sauf celle de (Wallace,1993). Cette dernière montre que cette possibilité n’est utilisée que par 0,03% des usagers.

Avec la généralisation de systèmes intégrés basés sur une architecture client serveur , certains systèmes offrent la possibilité d'une navigation graphique. Cette introduction de l'hypertexte coïncide avec la généralisation des catalogues sur le Web : les WWW-OPACs.

Le serveur HYTELNET (http://www.insigt.com) qui recense les catalogues en ligne ayant une interface WWW dans le monde. Dans (ihadjadene99b), nous avons consulté et interrogé 66 catalogues. Le tableau suivant montre que majoritairement, ces catalogues proposent une navigation à travers les liens sujet (81,8 %) et auteurs (87,8 %).

	Total (n=66)
Auteur principale	87,8%
Sujets	81,8 %
Côte	48.5%
Collection	48.5%
Auteur secondaire	30,3%
Titre	21.2%
location	9%
Editeur	6%
Collectivité	6%
Conférence	6%
Autres	3%

Tableau 19 : les liens dans les WWW-Opacs

Dans (Ihadjadene, 1999b), nous avons recensé les problèmes que posent la navigation dans ces catalogues. Nous verrons dans les chapitres suivantes comment sont utilisés les liens hypertextes par des usagers des catalogues de l'enssib, de Lyon2 et de l'Irisa.

3.3. 3. les systèmes basés sur la métaphore de l'étagère:

Ce sont un ensemble de prototypes qui utilisent l'étagère d'un bibliothèque comme métaphore de visualisation de l'information et de navigation. En se basant sur plusieurs études qui montrent que le butinage libre dans les étagères d'une bibliothèque est l'une des stratégies de recherche les plus populaires des usagers, presque 30 à 45 % d'entre eux selon (Hancock,1987), des chercheurs ont développé des prototypes qui représentent ces étagères en deux ou trois dimensions. La navigation est "spatiale" et visuelle, l'usager n'a plus besoin de maîtriser un langage de commande, ni de formuler des requêtes. Cette solution est rendue facile par la richesse des informations contenues dans le champs 3XX (description physique de l'ouvrage) du format MARC.

Nombre de pages = épaisseur du livre

Dimension = hauteur du livre

Quatre modes de navigation sont offerts ; voir les livres qui sont situés à gauche, à droite, en haut et en bas de celui que l'usager a choisi. Parmi ces prototypes, citons PACE (Beheshti,1996), SLC (Borgman,1996), SOPAC (Sugamoto,1995). Pour faciliter la navigation dans cet espace de documents, Borgman (1996) a ajouté la possibilité de naviguer hiérarchiquement dans la classification de Dewey .

Plusieurs catalogues opérationnels permettent déjà de rechercher les livres les plus proches dans l'étagère (ceux qui ont des côtes similaires). Malheureusement les usagers ont des difficultés à comprendre le rôle de regroupement des indices de classification. Les résultats des travaux de Borgman (1996) et de Beheshti (1996) sur les prototypes SLC et PACE ne révèlent que très peu de différences avec un système classique. Cette représentation pose deux problèmes : Comment repérer une ou plusieurs notices ? comment représenter des milliers de livres dans le un dispositif de consultation comme l'écran ?. Contrairement à la navigation dans les étagères, l'usager n'a pas une véritable vue spatiale. On ne peut pas présenter plusieurs livres en même temps.

3.3. 4. Les hypercatalogues basés sur la notion "d'œuvre "

Différentes possibilités ont été identifiées pour permettre une navigation à travers les sept liens existants entre les œuvres bibliographiques. Les rares prototypes développés n'ont pas pris en compte la totalité de ces relations. DIFWICS (digital index for works in computer science) est un projet de bibliothèque numérique dont le but est de rechercher des documents (articles de périodique, congrès, séminaires, en ligne ou des rapports) informatiques sur l'Internet et de regrouper les documents similaires (Hylton,1996). Pour déterminer si les documents sont liés, Hylton (1996) a développé un algorithme regroupement " clustérisation". Celui-ci permet de retrouver tous les documents ayant le même nom d'auteur et le même titre, facilitant ainsi l'identification en général des relations d'équivalences et dérivées. Par exemple une étude peut être publiée sous forme d'un article, d'une communication lors d'une conférence ou juste sous forme d'un rapport. Au lieu de présenter les trois réponses, le système affiche une grappe (cluster) avec trois liens différents. L'intérêt de la thèse de Hylton est qu'il a mis en lumière les trois problèmes que pose cet algorithme. Le premier concerne les erreurs (typographiques, d'écriture) qui surviennent parfois dans les bases documentaires ; le second problème est dû au fait que les noms d'auteurs des articles ne sont

pas toujours normalisés . Enfin le problème des abréviations se pose aussi. Ces difficultés restreignent l'usage de cet algorithme.

Bradfort OPAC[28] est un prototype qui permet d'accéder à plusieurs catalogues en ligne distribués (via le standard Z39.50) et de présenter à l'usager les versions d'une œuvre bibliographique. A l'instar de DIFWICS, il obtient ces grappes en recherchant toutes les notices bibliographiques qui ont le même auteur et le même titre. L'intérêt principal de ce système est qu'il permet à l'usager d'interagir avec les résultats obtenus (ici des grappes) pour obtenir plusieurs vues (plusieurs façons d'organiser les résultats). L'usager peut demander de sélectionner une vue (par édition, par date, par serveur, par auteur, par titre, selon le format ou la langue). Les auteurs ont testé leur système sur 69 usagers (dont la moitié sont des bibliothécaires). Ils montrent que BradfortOPAC2 permet d'améliorer l'affichage des résultats par rapport aux catalogues traditionnels. Les usagers optent en majorité pour la présentation des résultats en sélectionnant les critères suivants: auteur, titre et serveur.

BibRelEx (Brüggemann,1999) est une base de données bibliographique comportant plus de qui gère et visualise les liens de citations existant dans plus de neuf documents (fig.13)

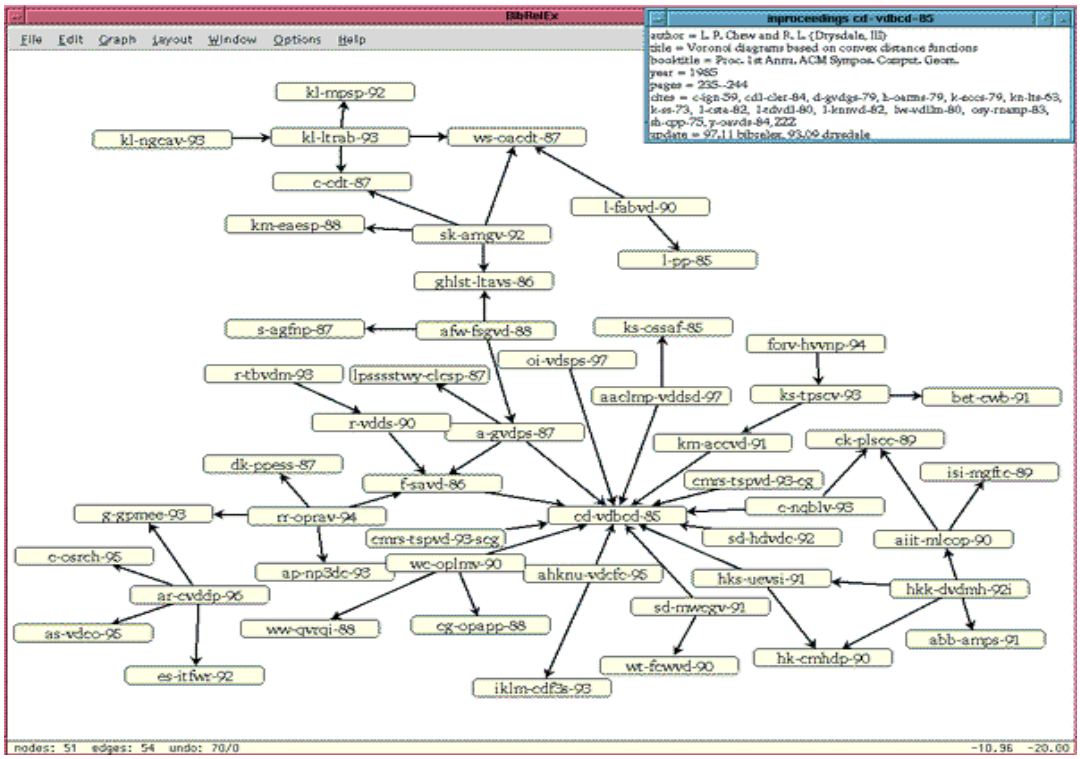


Figure 13 : les relations de citations dans Bibrelex

[28] http://www.bopac.ac.uk

Dans le cadre du projet de bibliothèque numérique de l'université de Michigan, une base de connaissances offre la possibilité de chercher toutes les œuvres écrites par, ou sur, Beethoven. Cette base de 493 notices est construite manuellement; elle permet à l'usager de naviguer dans un réseau de liens proches de ceux définies par Tillet (1989) . Les auteurs n'ont malheureusement pas indiqué comment généraliser leur solution à une large base documentaire. Il est clair qu'aucun prototype ne permet de gérer l'ensemble des relations entre les œuvres telles qu'elles sont définies dans les études de Tillet (1989), Leazer (1993) ou de l'Ifla (1996). Il sera nécessaire de préciser la notion d'œuvre, de mieux définir ses attributs mais surtout de revoir le format MARC.

3.4. limites des hypercatalogues

L'application de l'hypertexte pour un système de recherche d'information est prometteuse mais plusieurs difficultés s'y posent. Premièrement la plupart des bases de données bibliographiques sont composées de millions d'enregistrements distincts par opposition aux centaines de documents qui s'y trouve dans la majorité des systèmes hypertexte. Tous les systèmes étudiés dans les sections précédentes, en sont restés à l'étape de prototype. Les hypertextes auteurs tels que Hypercard, Guide ou Notecards sont se avérés inadaptés pour gérer et maintenir de gros volumes d'informations comme les bases de données MARC. Les autres approches restent limités, la majorité des auteurs privilégient soit la navigation dans les langages documentaires, soit dans le réseau de liens documentaires. De plus, la navigation dans le catalogue est favorisé au détriment de la recherche analytique (booléenne ou autre). Or, toutes les études effectuées sur les hypertextes ont montré l'insuffisance de la navigation , pour des tâches de recherche d'information, lorsqu'elle est employé toute seule (Wolfram,1996).

Pour, toutes ces raisons plusieurs chercheurs (Khoo,1997), (Hassoun & Roger,1994) ont vivement critiqué cette l'approche. Nous pensons qu'au contraire, il est important de concevoir des hypercatalogues qui associent aussi bien les stratégies de recherche analytique que des possibilités de navigation hypertextes (au niveau de l'espace des concepts et au niveau des documents). Les WWW-OPACs sont donc les seuls hypercatalogues opérationnels. Ils nous offrent la possibilité d'étudier leurs usages en situation réelle comme ce fut le cas pour les catalogues de deuxième génération. Il convient de constater entre autre, si les usagers arrivent à améliorer leurs recherches, à naviguer sans problèmes et à déterminer les insuffisances de ces systèmes. C'est le but des parties suivantes de la thèse.

Partie II : Etudes empiriques

Chapitre quatre : La recherche d'information médiatisée.

Les bibliothécaires mettent en œuvre des connaissances qui leur permettent de rechercher efficacement de l'information. Pour améliorer les SRI, de nombreux auteurs ont cherché à identifier les stratégies de recherche des documentalistes (Chen,1992), (Fidel ,1991). Nous recherchions au départ à adapter ces travaux pour les catalogues en ligne. Les premières analyses des verbalisations des usagers nous a conduit à élargir notre démarche à l'étude de l'interaction , et donc des échanges, effectués entre les bibliothécaires et les usagers. Notre but n'est pas de développer un « énième » système expert mais de comprendre davantage les différentes phases de dialogues et d'échanges.

L'objectif de cette section est de répondre à ces interrogations:

- Quelles sont les connaissances mises en œuvre par les usagers et les bibliothécaires ?
- Quelles sont les stratégies de recherche et tactiques utilisées ?
- Quels sont les différents types d'éclaircissements suscités par les bibliothécaires et par les usagers ?
- Comment les bibliothécaires contournent-ils les problèmes d'échecs ou de surcharge d'information ?

4.1. Les Stratégies de recherche d'information

4.1.1. Rappel des stratégies de recherches

Afin d'améliorer les résultats d'une recherche et de réduire les coûts liés au repérage de l'information, les intermédiaires ont développé des stratégies de recherche. Le développement et le choix d'une tactique ou d'une stratégie est sans doute l'aspect le plus conceptuel d'une recherche d'information. Ce choix exige souvent une grande connaissance des fonctionnalités d'un SRI, du vocabulaire contrôlé, des objectifs de la recherche en termes de bruit et de silence, et des besoins d'information de l'usager.

Le terme tactique ou heuristique désigne une action qu'on entreprend à un point donné, une décision à suivre pour atteindre un objectif immédiat comme utiliser un synonyme, un terme

associé pour améliorer le résultat (Harter, 1986) . On dénombre plusieurs sortes de tactiques. Bates (1977) en recense plus de vingt neuf.

Par contre, une stratégie de recherche est un plan général pour atteindre un but (Marchionini, 1995). Pour ce dernier**,** deux type de stratégies s'imposent:

- Les stratégies analytiques: elles sont guidées par des buts et nécessitent l'élaboration d'un plan. Elles sont déterministes, formelles ou directes.
- Les stratégies de navigation qui sont informelles, guidées par les données et qui sont donc heuristiques.

Nous allons examiner dans cette section trois sortes de stratégies de recherche analytiques et nous montrerons le lien existant entre l'une d'elles et une stratégie de navigation que l'on verra au chapitre suivant. D'autres stratégies comme « Interactive scanninig », « pairwise facettes »**,** **«** simple research », « citation indexing strategies » sont décrites dans les ouvrages de (Marchionini, 1995) , de (Harter,1986) ou de (Lancaster,1993).

4.1.1. 1. "Building Block strategy": la recherche par combinaison d'étapes

Cette stratégie de recherche est très efficace. Elle consiste à effectuer sa recherche en plusieurs étapes de façon fragmentaire. Chaque étape est numérotée et l'usager peut consulter l'historique de sa recherche (l'ensemble des étapes) . Il contrôle ainsi le processus de recherche. En fonction des résultats obtenus à chaque étape , il pourra les combiner au moyen des opérateurs booléens. A chaque phase, l'usager peut recourir à plusieurs tactiques. Bates (1990) conseille de regrouper les termes proches et synonymes avec l'opérateur OU et de combiner ensuite chaque étape avec un ET. Plusieurs CD-ROMs (comme celui de LISA) et serveurs (comme Dialog) permettent d'exécuter cette stratégie. Nous pensons que cette démarche de recherche est difficile pour des non-professionnels.

Elle peut être représentés par la figure suivante:

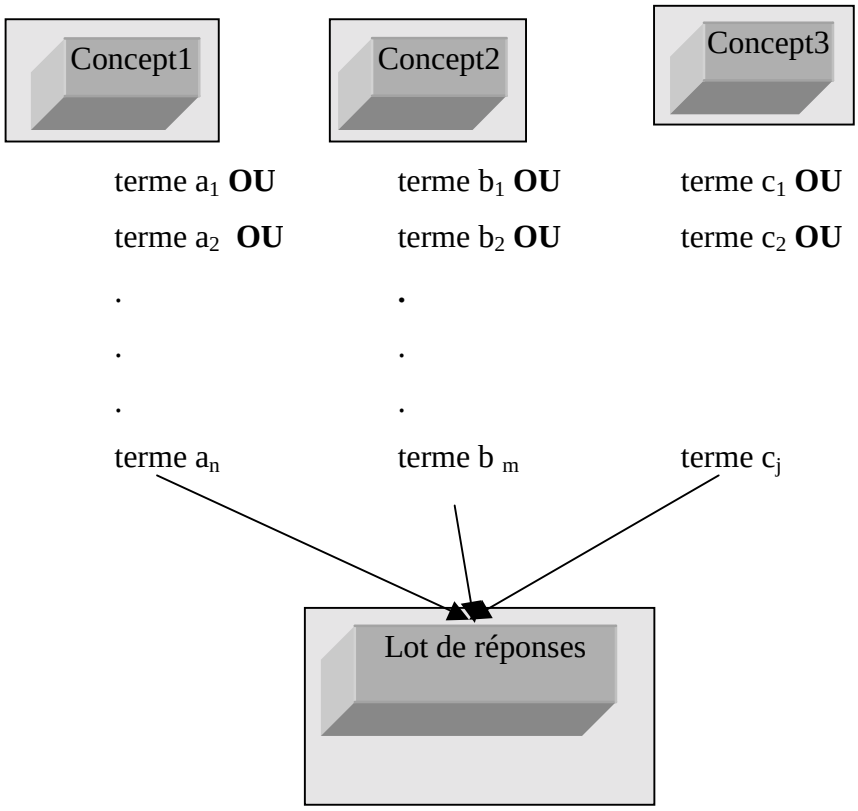


Figure 14 : building block strategy

L'équation finale peut s'écrire : (a_1 OU...a_n) ET (b_1 OU b_m) .....ET (c_1.....OU c_j)

4.1.1. 2. Successive facet strategy

L'idée principale de cette stratégie est de commencer la recherche avec un terme général, en vue de l'obtention d'un grand bruit puis successivement de réduire l'ensemble des résultats en y ajoutant d'autres mots clés.

Elle est illustrée dans la figure15.

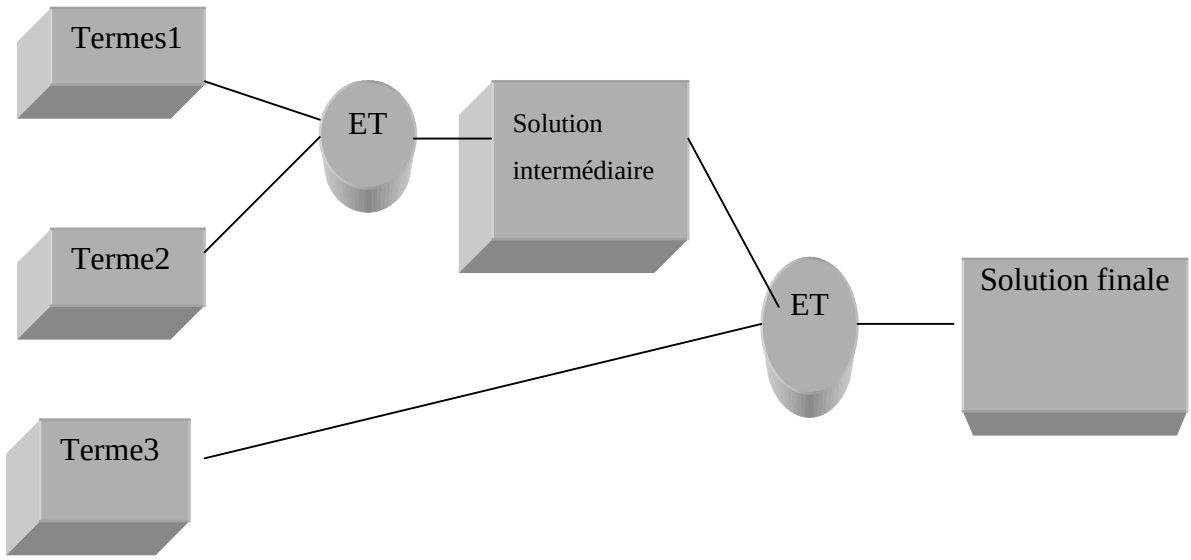


Figure 15 : facette strategy

Harter (1986) recommande d'appliquer cette stratégie lorsque le besoin d'information de l'usager est vague et, d'introduire d'autres termes comme les dates ou la langue dans la requête afin de limiter le nombre de réponses.

4.1.1.3. Citation pearl growning.

La finalité d'utilisation des deux stratégies présentées ci-dessus, est d'obtenir un grand bruit puis de restreindre le résultat pour avoir de la précision. Le but de la stratégie "citation pearl growning" est inverse: l'usager commence sa recherche avec 100% de précision. Il examine ensuite les références pour extraire les termes qui lui semblent importants (auteur, descripteur, nom du journal, indice de classification...etc), puis il reformule sa requête pour trouver des documents similaires. Cette opération est itérative et cyclique. Chen (1992) a développé un système expert basé sur cette stratégie de recherche. Il montre que les performances du système sont meilleures que dans système classique.

Walker (1989), Chen (1991), Meadows (1992) et Marchionini (1995) estiment que cette stratégie de recherche est une variante formelle du relevance feedback. Nous verrons au chapitre suivant, une autre version de cette stratégie que nous avons nommée BRF : Browsing Relevance feedback.

Remarque

En plus des stratégies de recherche formelles, il convient de citer de nombreuses autres stratégies de recherches (informelles) dont l'efficacité est parfois confirmé selon le contexte de recherche et le besoin de l'usager (Erdelez, 1995). On peut citer :

- La participation aux conférences, congrès…etc.
- La visite de librairies et de salons d'expositions
- La lecture des catalogues d'éditeurs (livre hebdo, Eyrolles,…etc.) les ouvrages récents
- Le feuilletage de périodiques et la navigation dans les étagères de bibliothèque
- Les échanges d'information à travers les listes de discussions et usage du courrier électronique
- La navigation ,au hasard, dans Internet (Erdelez, 1995).

4.2 La Recherche d'information médiatisée

4.2.1. Recherche médiatisée et interface homme-machine

Pour améliorer les performances de recherche des usagers, plusieurs chercheurs ont évoqué la nécessité de mieux comprendre au préalable le processus de communication entre l'usager et les professionnels de l'information. Dans le domaine des interfaces homme-machine (IHM), un certain nombre d'auteurs remettent en cause cette approche. Ils se demandent si les interactions humaines constituent un bon modèle pour une interaction homme machine (Shneiderman, 1992) . Il n'est pas sûr que les phénomènes et faits mis en œuvre dans les interactions se reproduisent tels quels dans les interactions homme-machine. Il est donc difficile, sinon illusoire d'extrapoler tous les résultats de cette étude pour un dialogue homme-machine. L'une des critiques que l'on peut formuler à l'égard de ces d'observations est qu'elles se préoccupent essentiellement des dialogues oraux alors que la communication

homme-machine se fait principalement par écrits et de plus en plus par des dispositifs de pointage comme la souris. De plus, les communications interpersonnels ne se limitent aux échanges oraux. Une interaction homme-machine n'est efficace que si la communication est multimodale. Par ailleurs, Polity (1993) reste critique sur la possibilité de transfert de ces résultats pour améliorer la recherche d'information en langage naturel. Enfin, d'autres auteurs comme (Osmont,1993) y sont catégoriquement opposés.

> " *on est en droit de se demander s'il y a réellement dialogue entre une machine, si interactive et conviviale soit-elle, et une personne...il n'y pas d'interprétation par l'ordinateur du faisceau causal du dire de l'utilisateur, et encore moins de variation de l'interprétation en fonction de l'utilisateur. En fait, lorsque l'on compare la relation utilisateur-machine dans une interrogation sur banque de données à la relation d'interlocution entre deux être humains la première semble à la fois plus simple et plus complexe. Plus simple, car ce qui provient de l'ordinateur ..est sinon prévisible, du moins contrôlable. Plus complexe, par ce que c'est en fait, face aux données de la machine, davantage un langage intérieur qui s'instaure chez l'usager, dont peu d'éléments de surface émergent du langage dans la pensée, qu'un réel dialogue. Et les traces d'interrogation des banques de données laissent des repères procéduraux et déclaratifs, de ce langage intérieur".*

Saracevic & SU (1991) , Marchionini (1995) et Belkin (1993) sont d'un autre avis. Ils considèrent que tout processus de recherche d'information (médiatisée ou non) est un acte de communication. Nous pensons pour notre part, qu'une partie des analyses sur l'interaction entre un usager et un professionnel peut être transférée au contexte d'une recherche d'information interactive. Ces études peuvent nous aider d'abord à mieux comprendre les insuffisances du dialogue homme-machine (dans le contexte des logiciels documentaires en situation réelle) . Cela peut aussi nous indiquer quelles sont les améliorations à apporter à l'interaction dans les systèmes notamment pour faciliter la mise en œuvre de stratégies de recherche. Plusieurs études expérimentales ou théoriques ont examiné la communication usager - professionnels de l'information. Ingwersen (1992) les a regroupé dans son livre " Information retrieval and interaction" . Dans ce qui suit, nous parlerons plutôt de récentes études (celles de SARACEVIC et de Nordlie). Signalons que toutes ces études, y compris la nôtre, ont été effectuées dans des SRI qui n'offrent pas de possibilités de navigation ou de reformulation.

4.2.2. Les travaux de Saracevic et de Spink

Les travaux de (Saracevic & SU,1991) , (Saracevic & Spink, 1997) et de (Spink,1997) portent sur une quarantaine de recherches effectuées sur le serveur DIALOG par quatre documentalistes. Ils se concentrés sur le processus de sélection des termes. L'intérêt de leurs réflexions, provient sans doute, de la définition et de l 'évaluation du rôle de la visualisation des réponses dans le processus de recherche et de sélection des termes. Ils considèrent ces visualisations comme un élargissement du concept de bouclage de pertinence. Pour Saracevic et Spink, les deux approches de sélection de termes (algorithmique et humaine) sont différentes. Leurs études ont permis de distinguer cinq sortes de feedback:

- Content relevance feedback: l'usager évalue les réponses obtenues avant de reformuler ou de soumettre une autre requête.
- Magnitude feedback : l'appréciation de l'usager concerne le nombre de réponses trouvées.
- Terminology review feedback : l'appréciation de l'usager concerne la pertinence des mots affichés dans l'index (alphabétique ou non)
- Terms relevance feedback : l'usager relance une recherche avec de nouveaux termes d'indexation ou les clés d'accès d'une référence déjà trouvée .
- Tactical review feedback: l'usager affiche l'historique des recherches avant de reformuler la question.

(Spink & Losee, 1996) s'est intéressée au processus du choix des termes ainsi qu'à l'efficacité des différentes sources (usager, bibliothécaire, thésaurus, TRF). Pour elle : « *the transition between sources of search terms at time T to those at tile T+1 is not random* ».

4.2.3 Les Travaux de Nordlie

Pour Nordlie (1996) , le dialogue entre l'usager et le bibliothécaire se déroule en trois phases**:**

1. la définition du problème
2. la consultation du catalogue
3. la navigation dans les étagères

La première phase occupe près de 65% du temps d'une interaction. Il montre que dans un tiers des cas (30%) , on passe directement de la définition du problème à la navigation des étagères alors que dans notre étude la consultation du catalogue est une phase importante (seulement 8% de cas où le dialogue s'effectue sans consultation de catalogue). Nordlie ne s'est pas intéressé à catégoriser les échanges effectués entre les bibliothécaires et les usagers. Ainsi dans son étude, la visualisation des réponses et le feedback ne jouent pas un rôle aussi important dans le repérage de l'information. L'intérêt des travaux de Nordlie (1999) se situe dans la comparaison qu'il réalise entre la pertinence d'une recherche médiatisée et un repérage de l'information dans le catalogue. Il montre que la différence majeure entre ces deux types de recherches, est liée au performances trop élevées de la recherche médiatisée (45 % d'échecs vs 10% dans une recherche médiatisée). C'est le processus de sélection des termes et une meilleure définition du besoin de l'usager qui permettent d'atteindre ce taux d'échecs très réduit.

Remarque

Le corpus de dialogues (cinquante dialogues) recueillis par Polity en situation naturelle dans plusieurs centres documentaires et bibliothèques est le seul corpus existant en français (Pollity, 1993). Leur but était d'étudier l'interaction pour améliorer l'accès aux logiciels documentaires en langage naturel. Ils montrent que les requêtes de type "matière" (recherche de documents sur un sujet) ne sont pas importantes (elle varie de 6 à 16 % de l'ensemble des dialogues). Des études n'ont pas menées sur la structuration des dialogues, ni sur l'usage des tactiques par les bibliothécaires. Si l'on compare nos résultats et ceux de Pollity, on observe une prédominance des requêtes sujets dans notre étude et l'absence de certains types de requêtes (concours et jeux, etc.).

4.2.4. La recherche médiatisée: le cas de la BPI

Nous avons réalisé nos travaux dans la bibliothèque publique d'Information (BPI) à Paris pendant le mois de Juillet 1996. Le catalogue de la BPI est de deuxième génération. Nous avons employé deux approches pour le recueil des données: un questionnaire en vue d'obtenir des données socio-démographiques sur les cents usagers et neufs bibliothécaires, la satisfaction des usagers et leurs connaissances et expériences dans l'usage du catalogue. Cette

partie fut l'objet d'un DEA (Aboud, 1997). Nous avons aussi enregistré tous leurs échanges et verbalisations que nous avons ensuite transcrits.

Dans cette étude, nous nous sommes intéressé aux sessions où les questions initiales de l'usager concernaient des recherches par sujets. Nous avons élargie notre corpus et sélectionné une partie (61) des 122 interviews des sessions pour la transcription et l'analyse des verbalisations. Dans l'annexe3, nous présentons quelques exemples de ces échanges.

*4.2.4.*1.Les différentes catégories d'éclaircissement

En plus de l'observation directe des bibliothécaires et des usagers en situation d'interrogation, nous avons enregistré sur cassettes tous les dialogues qui se passaient entre eux. La transcription des dialogues obéit au formalisme défini dans (Saracevic & SU, 1991) et (Spink,1997) puisque le but est de comprendre les catégories d'éclaircissements. L'objectif principal des premières pré-expérimentations était de vérifier et de valider la typologie des requêtes que nous avions établie à priori en se basant sur les études de Spink. Par éclaircissements, nous entendons toute question , toute explication, tout échange verbal entre l'usager et le bibliothécaire "*a verbal request for information. Any utternace by person A, which directly produces a response from person B*" (Spink & Losee,1996).

Nous avons décidé de regrouper les éclaircissement liés au choix des termes (de l'index ou de la notice affichée) en une seule catégorie.

Cette typologie comprend les catégories suivantes:

Catégories	Exemples
Questions relatives au choix des termes	B :«Pouvez-vous me donner un terme plus précis ? »
les objectifs de l'usager et ses motivations	B : « C'est pour votre mémoire ? »
connaissance de la base, aspects techniques, procédures de recherche	B : « savez vous comment faire une recherche par mot dans le catalogue? »
choix du type de la bases de donnés ou de la bibliothèque (sources)	B : "voulez vous d'autres bibliothèques plus spécialisées
Profil de l'usager (urgence, niveau de difficulté, restriction de langue, nombre de matériels)	B : « vous voulez chercher des ouvrages plus *récents* que celui de 1991? »
recherches antérieures dans le catalogue ou autres sources	B : « Il y a aussi l'annuaire de la presse et de la publicité. Est ce que vous les avez déjà consulté ? »
Eclaircissement relatifs au nombre de réponses, au format des données..	B : « c'est trop élevé ? »
Eclaircissements relatifs à la pertinence des réponses	B : « Ça vous intéresse ? »
éclaircissement relatifs aux choix des termes issus des notices ou de l'index affichés (TRF)	B : « voir le sujet : analyse des questionnaires : ça vous intéresse ?"
questions relatives au développement d'un plan de recherche (immédiat ou futur)	B: « on essaye dans le titre ? »
Autres types de questions (prêt, collègue, ...etc.)	B: « voulez vous prendre ce livre ? »
Eclaircissements et questions relatifs à l'usage des rayons de bibliothèques (explication de l'usage des côtes, ,,,)	B : « vous comprenez c'est quoi la 802 ? »

Tableau 20 : Catégories des échanges

*4.2.4.*2. Analyse des résultats

Dans toute interaction , le bibliothécaire oriente souvent les échanges (ses questions, éclaircissements, objectifs du dialogue) vers deux directions:

- La question elle-même pour déterminer le besoin réel des usagers

- La réponse pour trouver les sources potentielles de la bibliothèque qui répondent à la question posée

(Le Coadic, 1998) parle de démarche orientée " émetteur" pour la première, et démarche "récepteur" pour la seconde. Ces deux démarches sont antérieures à l'informatisation des fonds documentaires. Avec l'introduction des catalogues en ligne, une autre partie du dialogue concerne la formation des utilisateurs à l'usage du catalogue: en plus de la recherche d'information elle même, le bibliothécaire explique comment effectuer une recherche d'information, la signification des commandes, la mise en œuvre de stratégies de recherche. L'analyse des données de la BPI montre cependant que cette phase d'instruction est relativement faible (moins de 1%). Environ 7% des bibliothécaires ont opté pour une démarche orientée « émetteur», en ce sens que l'accent est mis sur les collections de la bibliothèque et les différentes ressources. Ils estiment et choisissent la ressource la plus appropriée à la question posée puis orientent l'usager à un feuilletage des étagères. Cette pratique , orientée émetteur, est fortement critiquée par (Le Coadic,1998) .

D'une façon générale, ces deux types de démarches ne sont ni opposées, ni distinctes: Un bibliothécaire peut commencer à préciser le besoin de l'usager et au cours de l'interview, il peut l'orienter à l'usage des autres ressources. Ces deux procédés s'observent aussi bien chez les nouveaux bibliothécaires que chez les anciens. Nous pensons que dans le cas des bibliothèques, cette démarche "orientée émetteur" n'est pas toujours une mauvaise solution, surtout lorsqu'elle est suivie par des explications sur les différents types de ressources et leurs usages. Il existe des situations où l'usager est pressé et a besoin d'une réponse urgente. Lorsque les bibliothécaires, orientent les lecteurs vers les étagères sans aucune explication sur l'organisation des ouvrages dans la bibliothèque, et lorsque cela se produit au début du dialogue, il est évident que cette démarche orientée "émetteur" ne répond pas aux besoins des usagers. Une grande partie des bibliothécaires a donc adopté une démarche orienté « récepteur ».

L'analyse de l'ensemble des échanges, nous a permis d'identifier deux phases. La première concerne les dialogues avant l'interrogation du catalogue, ceci constitue presque 20% des échanges verbaux. La seconde, qui est la plus riche en terme d'éclaircissement, se déroule pendant la consultation du catalogue : près de 80% (tableau 21). Ces données ne corroborent pas avec ceux de (Nordlie,1996) . Vu le nombre d'éclaircissements qui s'effectuent pendant la consultation du catalogue , on peut donc estimer que cette phase est moins importante dans notre étude.

Il n'existe pas de phase " navigation dans les étagères". Nous avons observé juste une seule fois un bibliothécaire qui a accompagné un usager afin de localiser un livre. Nordlie (1996), quand à lui, indique que les bibliothécaires danois accompagnent souvent les usagers pour discuter de la pertinence des ouvrages trouvés pendant la phase recherche en ligne. Il montre que cette phase occupe presque 75% de l'interaction. Dans notre cas, les bibliothécaires se limitent habituellement à expliquer comment fonctionnent les classifications mais surtout à mettre en lumière le rôle de regroupement des indices de classifications. Cela signifie que dans les deux contextes, la consultation des catalogues a pour principal objectif de repérer des points d'entrée (ici des indices de classification ou des côtes) pour la navigation dans les étagères. C'est à l'inverse des résultats de (Spink & Losee,1996) et de (Villame,1994). La recherche d'information dans les catalogues n'est donc qu'une étape parmi d'autres, d'un processus plus général dont le catalogue en est la partie centrale. L'objectif des bibliothécaires n'est pas toujours de trouver l'ensemble des réponses pertinentes, mais une ou deux notices et ensuite de donner à l'usager des orientations pour feuilleter dans les étagères de bibliothèques.

On s'est intéressé par la suite à étudier dans le détail, la démarche des bibliothécaires pour répondre à la formulation d'un besoin d'information. Une bonne proportion de dialogues sont à l'initiative du bibliothécaire. Ils contrôlent le dialogue dans 68% des cas. Ce résultat converge avec ceux de (White,1998) et de (Spink & Wu,1996) . L'étude des négociations entre l'usager et le bibliothécaire montre que ce sont les questions verbales directes fermées qui sont majoritaires (60%). Plusieurs auteurs ont montré les limites de ce type de questions qui n'encouragent pas l'usager à préciser ses besoins en information. Ils préconisent d'employer les questions ouvertes ou neutres (Dervin,1986).

Le tableau suivant donne un aperçu de l'importance de chaque catégorie.

	total		
	Avant la session N=41	Pendant la session (en ligne) N=163	Total n=204
choix des termes, et définition du sujet	46,3%	30,7%	33,8%
les objectifs de l'usager et motivation	17,1%	3,1%	5,9%
BDD (connaissance de la base, aspects techniques, procédures de recherche	4,9%	3,7%	3,9%
choix du type de la bases de donnés ou de la bibliothèque (sources)	0,0%	1,8%	1,5%
profil de l'usager (urgence, niveau de difficulté, restriction de langue)	9,7%	4,9%	5,8%
Expérience passée (usage du catalogue, feuilletage de la bibliothèque, ressources consultées)	17,1%	3,1%	5,9%
Réponses (nombre, format, lecture)	2,4%	6,1%	5,4%
Pertinence des réponses	0,0%	17,8%	14,2%
éclaircissement relatifs aux choix des termes issus des notices ou de l'index affichés (TRF)	0,0%	7,9%	6,4%
plan de recherche immédiat ou futur	0,0%	6,1%	4,9%
Autres (prêt, impression,...)	2,4%	5,5%	4,9%
rayons (explication de l'usage des côtes)	0,0%	9,2%	7,3%
Total	100,0%	100,0%	100,0%

Tableau 21: Catégories d'éclaircissements des usagers et des bibliothécaires

On peut constater que la phase pré-recherche est plus significative pour les usagers que pour les professionnels (30,3% vs 15,2%). Contrairement aux travaux de (Belkin,1984) , nos résultas indiquent que la modélisation de l'usager n'est pas une étape importante dans le dialogue (5,8%) . Cet auteur suggérait que les intermédiaires s'intéressent beaucoup aux objectifs de recherche de l'usager ainsi qu'à ses recherches passées.

Nous pouvons distinguer trois étapes essentielles dans le dialogue : le choix du vocabulaire, la visualisation des résultats et l'identification du contexte de la recherche .

Le choix du vocabulaire et la définition du besoin d'information.

La nature des questions du professionnel change durant l'interaction. Dans la première phase, le but du bibliothécaire est souvent de cerner le besoin de l'usager. Les renseignements et éclaircissements liés au choix des termes et à la définition du sujet représentent 31,2% pour le bibliothécaire et 39,4% pour les usagers (tableaux 22 et 23). Néanmoins, nous remarquons que les bibliothécaires contribuent plus dans la définition et la précision du besoin de l'usager ainsi que dans le choix des termes à travers leurs dialogues avec les usagers. En général ces précisions peuvent être classées en trois catégories : le choix de l'orthographe correct des termes, l'élimination des ambiguïtés et enfin la suggestion de nouveaux termes. Lorsque l'usager connaît bien le domaine de recherche concerné, il propose plus de termes.

La phase pré-recherche est primordiale pour la définition du sujet pour les usagers mais moins importante pour la proposition de nouveaux termes. Comme les professionnels, ils produisent plus de termes pendant la phase consultation du catalogue en ligne.

Comme les usagers ont tendance à exprimer leur besoins d'informations avec des termes généraux, les bibliothécaires leur demandent souvent des précisions. Ce n'est qu'après un dialogue de clarification avec un intermédiaire ou après une série d'interactions (visualisation des réponses, TRF) avec la base que les motivations réelles peuvent être précisées. Cette étape est absente des catalogues en ligne.

La visualisation des résultats

La visualisation des résultats (soit l'index des termes, soit la liste des notices abrégées, soit la liste des notices complètes) est une étape essentielle dans la recherche d'information médiatisée. Elle joue un rôle important dans la définition de la pertinence des résultats, dans la sélection des documents et dans la suggestion de nouveaux termes de recherche. Cette étape représente pour les 34,1% de l'ensemble des échanges pour les professionnels mais seulement 9,1% pour les usagers.

L'affichage des réponses a abouti à la suggestion de nouveaux termes : c'est plus le bibliothécaire qui est à l'origine des suggestions que l'usager (8,7% vs 1,5%) . Cependant, nous remarquons que cette sélection s'effectue couramment à partir de la liste des index (liste intermédiaires entre la requête et l'affichage des notices). Dans deux sessions seulement, le bibliothécaire, après examen de la notice complète, a suggéré d'employer les descripteurs pour reformuler la recherche. On peut donc affirmer que la mise en œuvre de la stratégie d'instanciation de références connues est très rare. Seul un usager (doctorant en droit) a pu affiner sa requête en demandant : " est ce que je peux avoir les documents qui ont le même auteur ? ". Ces résultats confirment les conclusions de (Chen,1992) et de (Spink,1997). Dans une recherche médiatisée, les usagers n'exploitent pas souvent les réponses du système pour reformuler leurs questions. Enfin, nous avons remarqué que les usagers ont parfois des difficultés à répondre aux questions: ils ne savent pas si la référence affiché est pertinente ou non. Dans ces cas, ils diffèrent leurs jugements.

Le contexte de la recherche d'information

Nous regroupons dans cette étape, les éclaircissements et les questions relatifs au profil de l'usager et à son expérience passée. Elle représente 11,6% des questions que posent le bibliothécaire et 12,1% des éclaircissements fournis par l'usager . Avant de consulter le catalogue, les bibliothécaires sollicitent davantage de renseignements sur l'usage antérieur des utilisateurs, sur les ressources documentaires (en général pour savoir si l'usager a consulté le catalogue en ligne, les CD-ROMs), sur la navigation dans les étagères et enfin des questions relatives à son profil *(B :" vous avez déjà vu dans les rayons ?*); (*B: avez vous consulté les dictionnaires?*) (*B: C'est pour vos études ?*). L'usager fournit aussi beaucoup d'informations concernant ses recherche précédentes aussi bien lors de la consultation du catalogue que dans la phase pré-recherche.

Ces informations constituent parfois un point de départ pour une recherche (soit au niveau du choix des termes) ou pendant l'évaluation de la pertinence des résultats (*U : j'ai déjà consulté cette notice*) ; (*U : j'ai déjà effectué une recherche mais je veux des données plus récentes*). L'analyse des interactions révèle qu'une partie des usagers (moins de 5%) ont déjà effectué des recherches sur le même thème. A l'inverse d'un logiciel documentaire, l'intermédiaire se sert de cette connaissance dans le choix des références (pertinence), dans la définition des termes et l'usage des stratégies de recherche (Item instanciation). Les usagers effectuent les mêmes requêtes pendant plusieurs sessions de recherches. Il semble donc que la variable "temps" est importante dans le processus de recherche (aussi bien pour la définition du problème que dans le choix des références).

	bibliothécaires		
	Avant la session N=21	Pendant la session N=117	Total N=138
Choix des termes, et définition du sujet	6,5%	24,6%	31,2%
Les objectifs de l'usager et motivation	2,2%	1,5%	3,6%
BDD (connaissance de la base, aspects techniques, procédures de recherche	0,7%	1,5%	2,2%
Choix du type de la bases de donnés ou de la bibliothèque (sources)	0,0%	1,5%	1,5%
Profil de l'usager (urgence, niveau de difficulté, restriction de langue)	0,7%	2,2%	2,9%
Expérience passée (usage du catalogue, feuilletage de la bibliothèque, ressources consultées ,,etc.)	5,1%	3,6%	8,7%
Réponses (nombre, format, lecture)	0,0%	5,8%	5,8%
Pertinence des réponses	0,0%	19,5%	19,5%
TRF	0,0%	8,7%	8,7%
Plan de recherche immédiat ou futur	0,0%	4,4%	4,4%
Autres (prêt, impression,...)	0%	1,5%	1,5%
Rayons (explication de l'usage des cotes, etc.)	0,0%	10,1%	10,1%
Total	15,2%	84,8%	100,0%

Tableau 22: Catégories d'éclaircissements des bibliothécaires

	usager		
	Avant N=20	En ligne N=46	Total N=66
Choix des termes, et définition du sujet	15,1%	24,2%	39,4%
Les objectifs de l'usager et motivation	6,1%	4,5%	10,6%
BDD (connaissance de la base, aspects techniques, procédures de recherche	1,5%	6,1%	7,6%
Choix du type de la base de données ou de la bibliothèque (sources)	0,0%	1,5%	1,5%
Profil de l'usager (urgence, niveau de difficulté, restriction de langue)	4,5%	7,6%	12,2%
Expérience passée (usage du catalogue, feuilletage de la bibliothèque, ressources consultées ,,etc.)	0,0%	0,0%	0,0%
Réponses (nombre, format, lecture)	1,5%	3,1%	4,5%
Pertinence des réponses	0,0%	3,1%	3,1%
TRF	0,0%	1,5%	1,5%
Plan de recherche immédiat ou futur	0,0%	6,1%	6,1%
Autres (prêt, impression,...)	1,5%	10,6%	12,1%
Rayons (explication de l'usage des cotes)	0,0%	1,5%	1,5%
Total	30,3%	69,7%	100,0%

Tableau 23: Catégories d'éclaircissements des usagers

Les questions relatives au plan de recherche concernent le choix de stratégie immédiat pour déterminer les conditions d'arrêt d'une recherche d'information ou le choix d'une tactique de recherche. Elle représente 4.3% des questions du bibliothécaire et 6,1% de cas pour l'usager.

4.2.4.3. Tactiques et stratégies de recherches observées

Lors de leurs recherches, les bibliothécaires recourent à plusieurs tactiques qui leur permettent notamment de contourner les échecs et le problème de la surabondance des réponses. Celles qui reviennent le plus souvent sont:

- Rechercher par titre lorsque la recherche par sujet n'aboutit pas.
- Employer des opérateurs booléens pour réduire ou élargir une recherche
- Utiliser la troncature pour obtenir plus de réponses.
- Limiter les recherches par date et par langue
- Saisir un terme général qui correspond à une tête de vedette, puis feuilleter les subdivisions sujets.
- Chercher et utiliser les renvois de l'index.

Nous n'avons pas observé un développement de stratégie de recherches tel que celui observé lors des documentalistes pour des SRI. La seule stratégie de recherche observée porte sur l'instanciation de références connues (tableau 21) . Nous avançons plusieurs raisons:

- les catalogues de deuxième génération, ne facilitent pas cette mise en œuvre de stratégies. Alors que les serveurs comme Dialog offrent des possibilités (en termes de commandes, ou de dialogue) qui permettent de concevoir différents types de stratégies (building block, la commande Zoom, ...etc.), les possibilités de recherche et de navigation dans les catalogues sont restreintes.
- la recherche d'information dans les catalogues n'est pas liée par des facteurs de coûts. Pour les documentalistes , la connexion à des serveurs est souvent coûteuse d'où l'importance de développer des stratégies de recherche.
- A l'inverse des documentalistes, les bibliothécaires ne consacrent pas un temps suffisant au dialogue avec les usagers. Nous avons observé que la durée moyenne d'une interview est de 3,10 minutes alors qu'elle est de 13,04 min dans l'étude de Spink. Parfois , les bibliothécaires doivent gérer plusieurs demandes en même temps.
- Les documentalistes doivent souvent répondre à des usagers (souvent des spécialistes d'un domaine) qui recherchent une information la plus complète et la plus précise possible. Ce n'est pas le cas pour les usagers d'une bibliothèque dont les recherches ne sont pas toujours exhaustives.

- Enfin, le développement du libre accès dans les bibliothèques permet d'autres types de recherche complémentaire de la consultation du catalogue comme le feuilletage dans les étagères ou la consultation des autres ressources (CD-ROM, ouvrages de références…etc.)

4.2.4.4. Changement du besoin de l'usager:

Lors d'une recherche d'information , l'intérêt et le besoin des usagers peuvent changer et évoluer. Ce phénomène est souvent observé dans une recherche médiatisée. En analysant le dialogue entre l'usager et les bibliothécaires, nous avons observé que ces derniers font graduellement progresser leurs interviews. Ces changements peuvent se manifestent de plusieurs façons :

- D'une recherche basée sur un terme générique au choix d'un terme plus spécifique.
- D'une recherche spécifique à une autre plus générique.
- L'emploi des synonymes.
- Suggestion de nouveaux termes.

Nous avons remarqué que, parfois ces changements ne sont pas seulement relatifs à la stratégie du choix des termes mais aussi à la nature même du besoin d'information. C'est le cas, lorsque l'usager passe d'une discussion thématique à une autre non thématique. Comme cet un usager qui cherchait des livres sur la "Tunisie". Au cours de l'interview, l'usager précise mieux son besoin " ce que je veux , ce sont des *statistiques récentes* sur le tourisme en Tunisie : de 1993 à 1995". Ce qu'il cherchait ce sont des données statistiques et non des livres d'où son insatisfaction par rapport aux réponses du catalogue . Dans l'approche classique des SRI, la requête initiale représente l'expression du besoin d'information statique. Pour nous, la requête initiale ne représente pas toujours ce besoin, l'interaction avec le système pour aboutir à un changement du besoin, du moins à sa précision.

Conclusion

L'expertise des professionnels ne se limite pas aux connaissances techniques liées au système telles que la manipulations de la recherche booléens ou de stratégies de recherche, ni à la connaissance d'un vocabulaire contrôlé comme RAMEAU. Ce point de vue est largement remis en cause par notre étude. D'abord, ils font largement appel à leurs connaissances du domaine pour générer des suggestions de précision et d'élargissement des

demandes qui leur sont présentées. Ensuite, ils développent une modélisation de l'usager pour déterminer son niveau, sa connaissance du fonds et de la bibliothèque, ses recherches antécédentes. Les bibliothécaires tentent de comprendre les buts des usagers avant de pouvoir leur répondre. Ils posent des questions pour aider les usagers à mieux expliciter leurs attentes.

Nous avons ensuite constaté que la majorité de l'interaction concerne deux problèmes fondamentaux en SRI, celui du choix du vocabulaire et l'examen des résultats. Nous avons observé que durant l'interaction entre l'usager et le bibliothécaire, les éclaircissements relatifs aux choix des termes sont essentiels pour un meilleur ajustement du besoin d'information de l'usager durant toute la durée de l'interaction. Il est donc important que tout système de recherche d'information puisse supporter des outils d'aide à la formulation et à la reformulation des requêtes. Ceci rejoint les études de Spink, White, Ingewersen et de Nordlie. Ces auteurs ont proposé d'inclure des interfaces intelligentes qui proposent des termes à l'usager. Ce dernier est le seul capable de confirmer la pertinence de ce choix. Il ne suffit plus de proposer une liste de termes au début de l'interaction, mais au contraire les SRI doivent supporter cette aide durant tout le processus de recherche. Enfin, nous avons remarqué que parfois, la visualisation des réponses joue aussi un rôle important dans les redéfinitions du besoin des usagers. Par conséquent, les tâches de jugements de pertinence doivent être simplifiée pour l'usager durant l'interaction avec le système. L'affichage des notices abrégées est insuffisant pour déterminer si les réponses accèdent ou non à son besoin d'information. Si les études sur l'interaction bibliothécaires-usagers est étudié, il est important d'effectuer d'autres études notamment pour comprendre la nature de la communication interpersonnelle, entre médiateurs et usagers, par la médiation de l'Internet[29] (courrier électronique, système coopératifs) pour la recherche d'information.

[29] voir l'exemple du site http://www.webhelp.com qui propose de faire des recherche sur l'Internet par un millier de personnes. Ces derniers engagent un dialogue avec l'usager lorsqu'ils estiment que les requêtes sont ambiguës.

Chapitre cinq : Usage des catalogues sur l'Internet : répartition des points d'accès.

L'analyse transactionnelle (analyse du fichier log) est de plus en plus reconnue comme étant une excellente méthode d'étude, parce qu'elle fournit des informations justes et spécifiques sur le comportement véritable des usagers. C'est actuellement la méthode la plus employée dans les études sur le commerce électronique ou sur l'usage des moteurs de recherche.
Dans une bibliothèque, l'étude des traces informatiques peut servir à plusieurs personnes : l'administrateur du système, le concepteur du catalogue, le responsable de service des références, celui du contrôle bibliographique et enfin les responsables de formation. Chacune de ces personnes peut exploiter ces données en vue d'améliorer son service :

- Accroître le développement du catalogue et l'ergonomie de son interface.
- Mieux exploiter l'allocation des ressources et l'utilisation de la collection.
- Mieux cerner les besoins de formation des usagers.
- Il est possible d'établir des statistiques sur la proportion d'usagers par pays
- Mesurer les temps de réponses
- Etc.

5.1 Analyse transactionnelle classique

La majorité des systèmes intégrés qui existe sur le marché incorpore des possibilités d'enregistrement des traces. Une transaction peut être définie comme une action initiée par un usager qui est enregistrée dans un système informatique. C'est donc un fichier qui comprend des enregistrements contenant l'ensemble des informations relatives au contexte de l'interaction de l'usager avec le système informatique.
Un fichier log classique comprend les informations suivantes :

1. Le contenu de la requête de l'usager.
2. La date et l'heure de la consultation du catalogue.
3. Le terminal utilisé pendant la consultation.
4. La base de donnée interrogée.
5. Le nombre de réponses obtenues pour chaque requête.

6. Les réponses du système.

La qualité de ces informations varie bien sûr selon les systèmes.

C'est donc un procédé intéressant pour déterminer les critères de recherche utilisés par le lecteur, les raisons pour lesquelles certaines recherches n'aboutissent pas ou au contraire conduisent à un nombre de réponses élevées.

Voici un exemple d'un fichier log qui enregistre l'ensemble des transactions concernant l'accès sujet qui n'ont pas abouti. Ce fichier "simple" n'enregistre que trois informations : le numéro du terminal utilisé, le contenu de la requête et la date (Peters, 1993).

Search	Terminal	Date
Aids	97	2 : 23 PM Jan 9
Agnets	75	7 : 51 AM Jan 6
Arts and culture	104	10 :56 PM Jan 4

Dans ce deuxième exemple, le fichier contient les informations suivantes :

La source de la requête (publique ou bibliothécaire…), le type de recherche effectuée (auteur ou sujet), le contenu de la requête (Hunter, Internet), le nombre de réponses obtenues (4 ou 120), le nombre de fois que l'usager doit afficher le contenu des notices bibliographique (5 ou 10), le nombre de fois qu'il a opté pour un affichage en format MARC (0 ou 1) et enfin la date.

Source	Type de la requête	Contenu de la requête	Nombre de réponses obtenues	Nombre de fois que l'usager affiche le contenu d'une notice	Format Marc (1) ou non '0)	Date
Publique	Auteur	Hunter	4	5	0	5 : 21 AM
Publique	Sujet	Internet	120	10	1	4 : 11 AM

Le logiciel OLIVE de (Hancock, 1997) est sans doute l'outil le plus perfectionné et le plus complet. Il permet d'enregistrer tout ce qui est saisi par l'usager mais aussi tout les éléments qui sont affichés à l'écran: la durée d'une session, le nombre de recherches dans la même session, le temps moyen de recherche, le nombre de fois où l'usager a fait une recherche par auteur, titre, sujet. etc., le nombre de références affichées durant la session et le nombre de références affichées pour chaque mode d'accès. L'analyse transactionnelle comporte malheureusement plusieurs faiblesses du fait qu'elle ne permet pas de distinguer l'identité de celui qui a effectué la recherche ; les traces sont réellement anonymes : Cela favorise un assez bon contrôle des données recueillies. En effet lors de l'élaboration des interrogations, aucun phénomène d'interaction entre l'observateur et l'usager n'est à prendre en considération. En outre, elle ne nous apprend rien sur les raisons qui poussent le lecteur à procéder de telle ou telle façon, ni sur les intentions qui l'anime. Par contre, c'est un procédé intéressant pour déterminer les critères de recherche utilisées par le lecteur, les raisons pour lesquelles certaines recherches n'aboutissent pas, ou au contraire conduisent à une surabondance de réponses. Enfin, même si cette méthode permet de recueillir facilement des informations, leur analyse et leur interprétation peuvent s'avérer longues et coûteuses. Leurs enregistrements sur plusieurs mois, voire des années nécessite beaucoup de ressources informatiques. Il est souvent nécessaire d'utiliser en complément une autre méthode de renseignements. C'est ce que nous avons fait en ayant recours en plus à des questionnaires en ligne ou à l'observation des usagers en situation de recherche.

Comme ces fichiers n'enregistrent pas les liens utilisés lors de la navigation des usagers, Il est impossible d'étudier la navigation dans un catalogue en employant cette technique. Ceci explique peut être l'absence de données sur cet aspect de la recherche[30]. C'est pour cette raison que nous utiliserons les fichiers log générés par les serveurs HTTP sur l'Internet.

5.2. Analyse des fichiers log sur le Web

Les fichiers log sont des fichiers textes qui stockent les milliers de lignes d'informations générées par le serveur.

[30] Nous avons interrogé l'ensemble des bases de données relatives à ce sujet, nous n'avons pas trouvé de réponse. Par la suite, nous avons échangé des messages avec plus de dix auteurs qui ont étudié les catalogues et envoyé des messages sur des listes de discussion. La majorité des répondants nous ont souligné l'absence de données empiriques sur la navigation (l'utilisation de l'hypertexte) dans les catalogues.

Un serveur HTTP enregistre diverses informations dans un fichier log. Le plus souvent, elles sont de deux types :

- Les informations d'accès ou de transferts (l'ensemble des requêtes[31] reçues par le serveur et le résultat de la requête) sont enregistrées souvent dans un fichier access-log
- Les erreurs[32] enregistrées dans un fichier error-log qui recense l'ensemble des problèmes rencontrés dans une connexion par le serveur.

Les fichiers log peuvent aussi contenir des informations sur l'URL d'origine des connexions ainsi que sur la page d'arrivée sur le site. On peut trouver aussi des informations qui renseignent sur le type d'équipements des utilisateurs (navigateurs couramment employés, système d'exploitation utilisé. etc.).

Les deux formats d'enregistrements les plus utilisés sont CLF et XLF.

CLF: Common Log Format. Il est de la forme suivante :

Adresse-source-compte [date : heure 0100][33] " méthode chemin protocole "code-réponse taille_réponse

Exemples :

---

[31] Le terme « requête » désigne tout envoi d'un message, il a un sens différent de celui utilisé en informatique documentaire.

[32] Le terme « erreurs » est différent de celui utilisé communément en informatique documentaire. Les fichiers log d'erreurs conservent la trace des incidents et des dysfonctionnements intervenus lors d'une transaction HTTP.

Code d'état	**description de l'erreur**
401 (non autorisé)	utilisateur non trouvé
402 (non autorisé)	erreur mot de passe
403 (interdit)	tentative d'assimiler un fichier à un script
404 (non trouvé)	le fichier demandé n'existe pas
405 (non trouvé)	script non trouvé
406 (erreur interne)	bug dans l'application

[33] Le date n'est pas exacte : c'est celle du serveur, cependant il est possible d'avoir un rapport sur les statistiques d'accès d'une façon précise sur des intervalles de temps assez significatifs, exemple sur chaque heure, chaque jour, sur une semaine, sur un mois…etc.

- ENSSIBpc12.ENSSIB.fr – [09/mar/1998 :10 :23 :11 + 0100]"GET /cgi-bin-ever/DORIS_WEB_LIVRES_SIMPLE ? *mots = centre de documentation* HTTP/1.0" 200 21407
- ENSSIBpc12.ENSSIB.fr – [09/mar/1998 :10 :23 :39 + 0100]"GET /cgi-bin-ever/DORIS_DOC_WEB_LIVRES ? *NOTICES_W3=19931* HTTP/1.0" 200 2046.
- ENSSIBpc12.ENSSIB.fr – [09/mar/1998 :10 :23 :51 + 0100]"GET /cgi-bin-ever/DORIS_REF_WEB_LIVRES ? *AUT_MATIERE :4765* HTTP/1.0" 200 439

Le premier exemple montre que le 9 mars 1998 à 10h23 la machine ENSSIBpc12 a demandé la ressource /cgi-bin-ever/DORIS_WEB_LIVRES_SIMPLE ? mots = centre de documentation en utilisant la méthode GET du protocole HTTP/1.0, que le serveur a répondu normalement (code=200, donc pas d'erreur) et que la taille de la réponse est de 21407 octets.

XLF: Extended log format :

XLF est plus riche que le format CLF. En plus des informations classiques inscrites dans le format CLF, il permet d'enregistrer l'URL du dernier document html chargé (c'est à dire celui où était présent le lien qui a permis de parvenir au document courant) et le type de navigateur utilisé lors de la navigation.

Avec XLF on peut savoir comment un visiteur a pu parvenir à certains documents :

- On peut mesurer la proportion des usagers qui ont explicitement demandé l'URL du site.
- Ceux qui sont passés par un autre serveur.
- Il est possible d'obtenir des informations sur la navigation des usagers et leurs parcours favoris.

La présence des caches et des proxy sur le réseau est un obstacle à une bonne analyse des statistiques. Avant d'envoyer une requête sur le Web, le navigateur vérifie s'il ne l'a pas conservée. S'il dispose de l'information demandée, il envoie l'information au client sans aller chercher le fichier sur le serveur. Ceci veut dire qu'une partie du trafic n'est pas enregistrée dans les fichiers log. Les caches ne sont pas un obstacle à une requête utilisant des scripts CGI[34] (méthode de programmation qui permet de créer des pages dynamiques et de les relier à

[34] Common General Interface

des bases de données). C'est le cas de notre étude. Il est possible maintenant d'obtenir des informations sur la navigation[35] des usagers et leurs parcours favoris.

Une panoplie d'outils permet d'effectuer des statistiques à partir des fichiers log du serveur. On peut citer :

Webtrends (www.webtrends.com)

Accrue Insight (www.accrue.com)

Net Analysis (www.netgen.com)

Remarque :

La gestions des transactions n'est pas intégré dans HTTP1.0. Le protocole est sans session. On dit qu'il est sans état. Il existe cependant plusieurs techniques (URL longues, cookies, etc.) pour simuler des transactions. Pour répondre aux besoins du commerce électronique, le futur protocole HTTP2.0 devrait incorporer un véritable gestion des transactions.

5.3. L'utilisation de l'analyse transactionnelle dans notre étude

Pour déterminer les points d'accès, nous avons utilisé un ensemble de programmes informatiques sur plusieurs années. Nous avons vérifié les résultats obtenus plus de trois fois. Lorsqu'un doute survient , ou une légère perte d'information, nous les avons retirés de notre corpus. Ceci explique pourquoi nous n'avons pas pris les années 1996, 1998 et 1999 de l'ENSSIB. Dans notre analyse, nous pouvons décrire la navigation et le parcours d'un usager en réécrivant les requêtes initiales. Ceci nécessite par contre la connaissance précise de la signification de chaque information (numéro de la clé d'un notice, d'un lien hypertexte auteur, sujet, etc.). Pour l'étude de la navigation des usagers, il fallait non seulement ressaisir les données, mais interroger en même temps le catalogue de l'ENSSIB en mode caractère (telnet) pour pouvoir effectuer une recherche par le numéro de clé de la notice ou par les liens. Enfin, Pour l'analyse des taux d'échecs ou de la surcharge d'information comme pour l'étude des erreurs, nous avons imprimé l'ensemble des traces et interrogé le catalogue avec les mêmes termes que les usagers.

[35] Dans notre analyse, nous pouvons décrire la navigation et le parcours d'un usager en réécrivant les requêtes initiales. Ceci nécessite par contre la connaissance précise de la signification de chaque information (numéro de la clé d'un notice, d'un lien hypertexte auteur, sujet, série, ….etc.)

Le tableau 24 synthétise les méthodes utilisés pour le recueil des données selon les objectifs poursuivis.

Objectifs de l'étude	bibliothèques concernéss	Méthode utilisé
Répartition des points d'accès	Enssib, Lyon2, Irisa	Fichier log
Usage des liens hypertexte	Enssib, lyon2, irisa	Fichier log
Typologie des liens hypertextes	Enssib, lyon2, irisa	Fichier log
Analyse des échecs	Enssib, Lyon2	Fichier log + réécriture des requêtes
Analyse de la surcharge d'information	Enssib, Lyon2	Fichier log + réécriture des requêtes
Usage des opérateurs booléens et de la troncature	Enssib, Lyon2	Fichier log
Catégorie des accès par auteur	enssib	Examen et lecture du fichier log
Analyse des tactiques	Enssib, Lyon2	Fichier log
Analyse qualitative de la navigation	Lyon2	Observation + fichier log + interview (questionnaires)
Analyse de l'accès et de la navigation à distance	enssib	Fichier log + questionnaire en ligne

Tableau 24 : méthodes utilisées pour le recueil des données.

Dans cette partie de la thèse , l'objectif est de :

- Etudier la distribution des points d'accès selon les mois et les années.
- Etudier la distribution des points d'accès selon les établissements suivants (enssib, Lyon2, irisa)
- Avoir des données quantitatifs sur l'usage des liens lors de la consultations des catalogues.
- Etudier l'usage des opérateurs booléens et de la troncature.

5.3.1 Etablissements étudiés

5.3.1.1 Présentation des catalogues étudiés :

Le logiciel de la base est un logiciel documentaire LORIS conçu par la société Ever (http://www.ever.fr) et utilisé en France par plusieurs centres documentaires et quelques bibliothèques universitaires.

LORIS est une application de gestion intégrée des bibliothèques et centres de documentation. LORIS est développée sur le moteur de base de données documentaire DORIS.

DORIS[36] comprend un système unique de gestion de modèles de données pour structurer les références et les contenus des documents sous une forme orientée objet. Ce modèle de données, souple et original, est idéal pour la gestion des standards catalographiques tels que le MARC (Machine Readable Cataloging) et l'ISBD (International Standard Bibliographic Description). Il est également parfait pour des standards de contenus tels que le SGML (Standard Generalized Markup Language) et le XML (eXtended Markup Language). DORIS est basé sur SQL. Il est disponible sur les systèmes d'exploitation UNIX et WINDOWS NT. DORIS utilise les SGBD ORACLE, SYBASE ou SQL SERVER. DORIS complète la puissance du relationnel par des fonctions très riches de structuration de l'information parfaitement adaptées à la manipulation et l'indexation des documents. LORIS bénéficie de toutes les fonctions de recherche documentaire du moteur DORIS, incluant les données de type numérique, date, texte court et texte intégral. Il est basé sur le modèle booléen et inclut des possibilités de troncature. LORIS est fourni avec des interfaces standardisées, de type client-serveur Windows et Internet (Web) offrant ainsi la possibilité de naviguer dans une base documentaire.

Deux modes de recherches sont présentés aux usagers :

- La recherche simple
- La recherche experte

[36] nous avons recueillis les informations sur Doris à partir de la documentation et des manuels d'utilisation de la société Ever et aussi à partir du serveur (http://www.ever.fr).

Le mode simple permet de rechercher les mots dans le titre et le sujet pour le catalogue de l'enssib (figure 16) . Il permet d'effectuer une rechercher dans toute la notice dans le cas du catalogue de Lyon2.

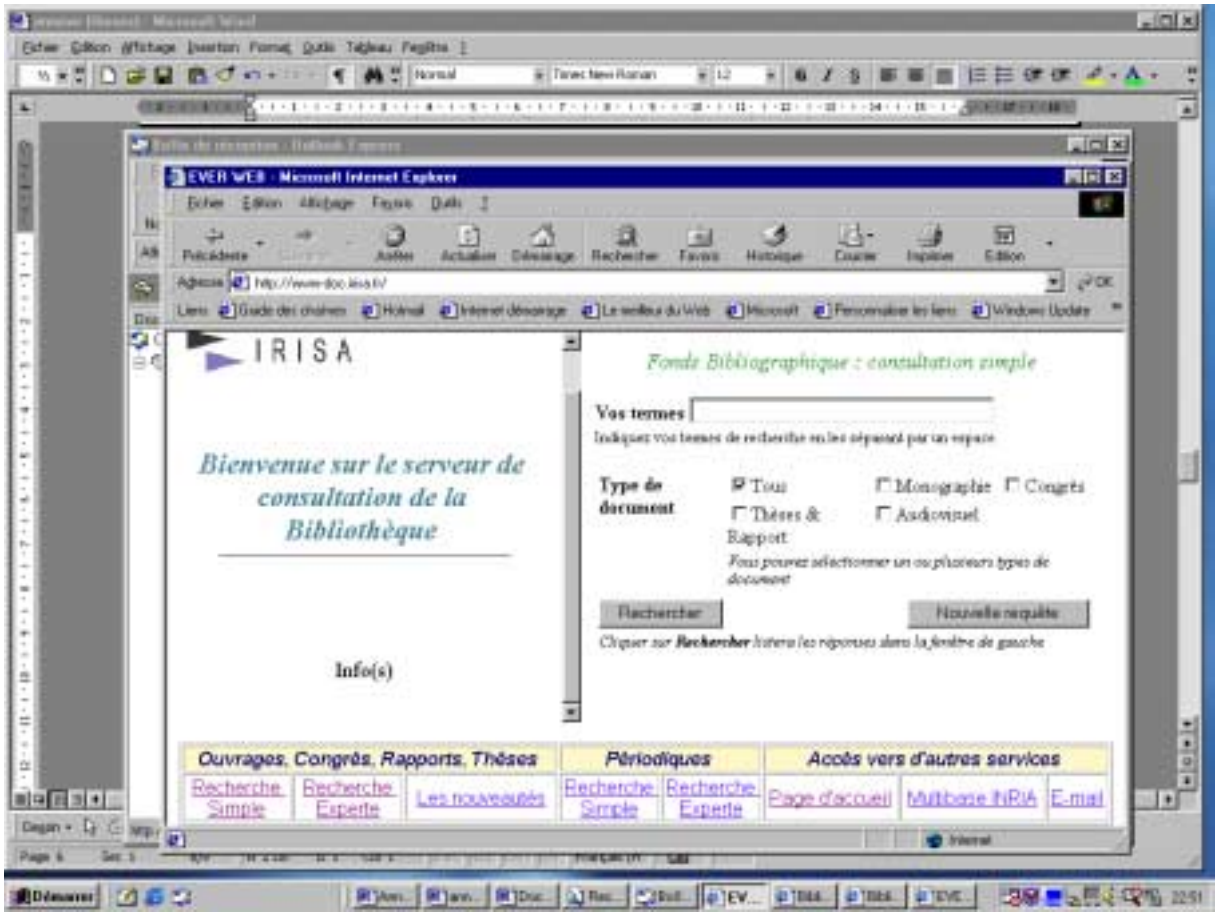


Figure 16 : recherche en mode simple dans le catalogue de l'Irisa

La figure 17 présente l'interface à partir duquelle les usagers peuvent faire des recherches expertes. L'usager peut soit remplir un champ (auteur, sujet, titre, éditeur, ..etc.) , soit faire une recherche multicritère en utilisant les opérateurs booléens (ET, OU, Sauf).

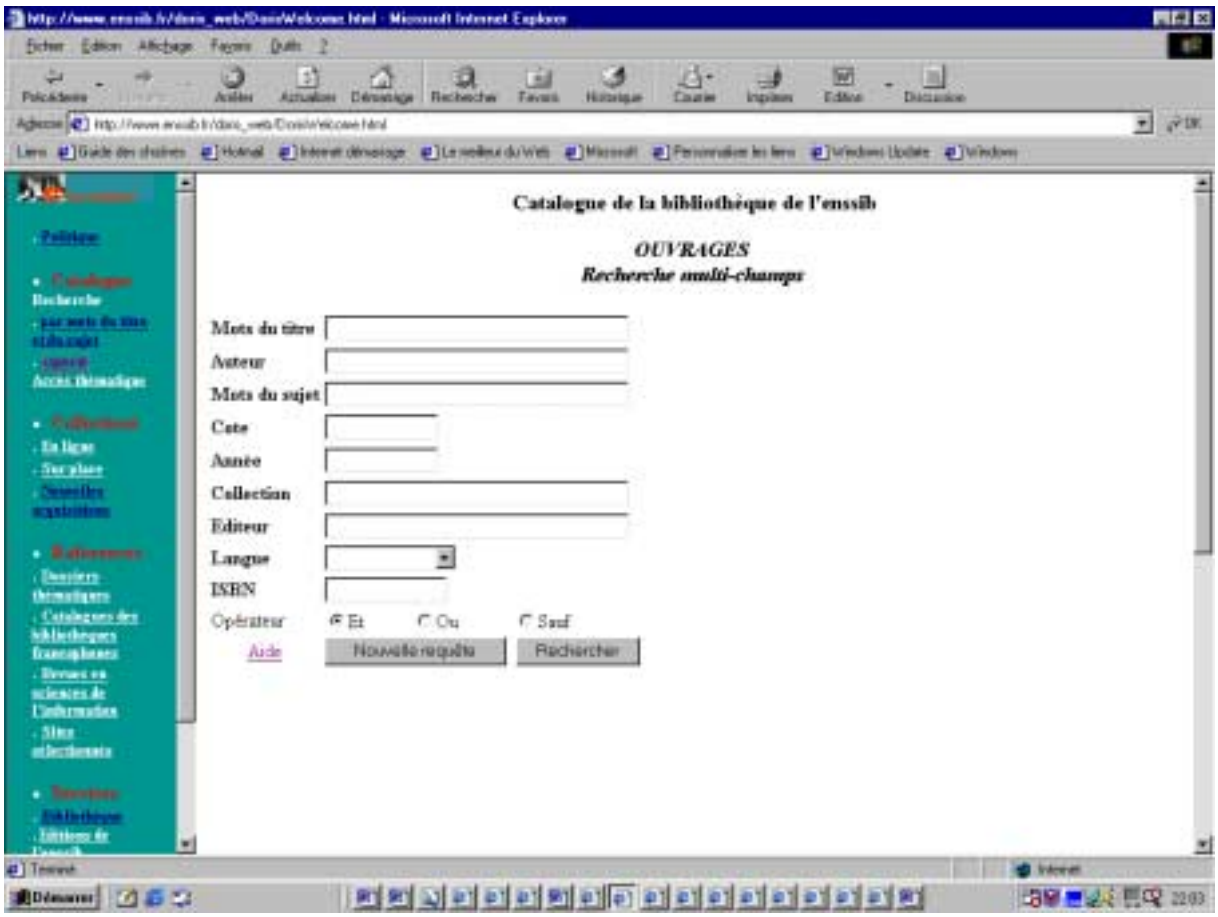


Figure 17 : recherche experte dans le catalogue de l'enssib

Après une requête, LORIS présente une liste de titres abrégés sous formes de liens hypertextes ainsi que le nombre de réponses trouvées (figure 18).

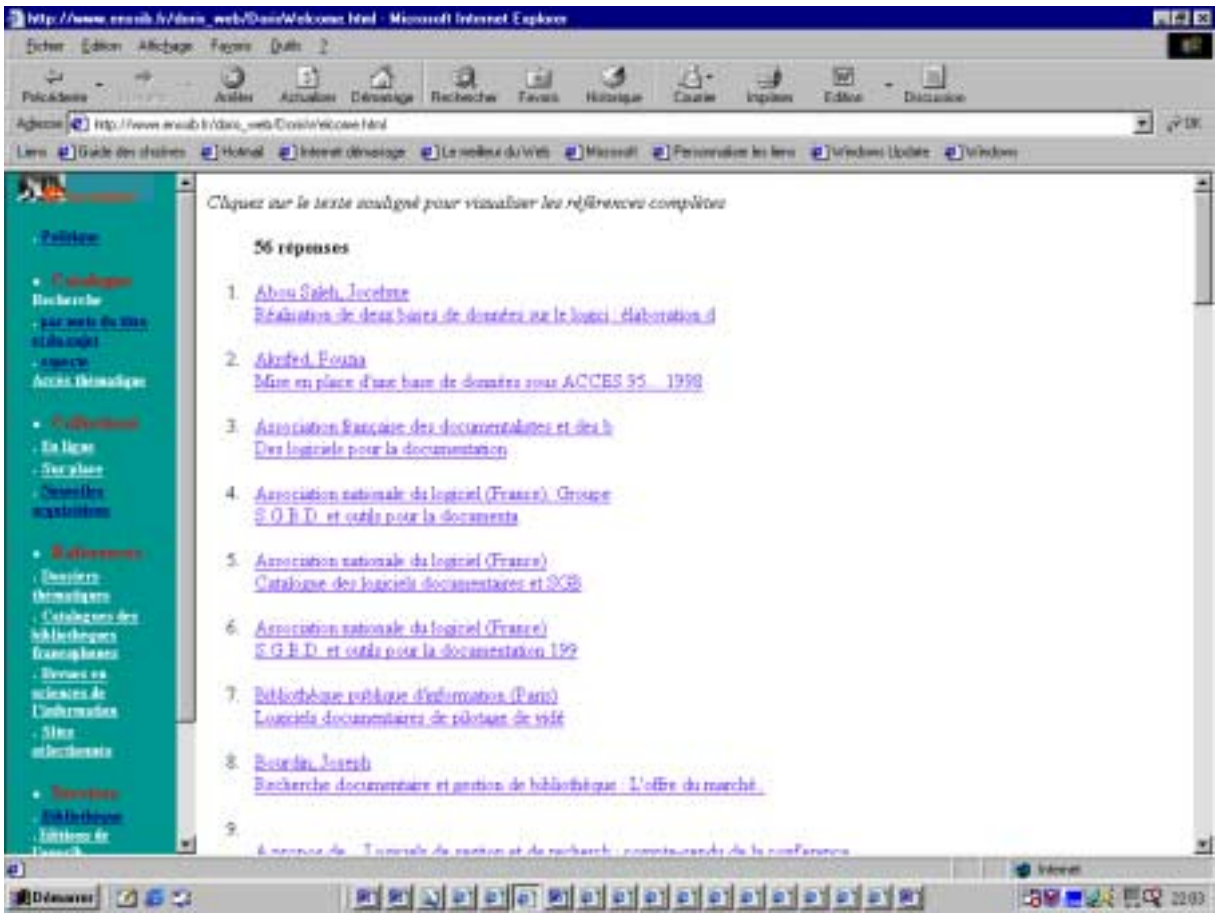


Figure 18 : affichage des réponses (notices abrégées) dans le catalogue de l'enssib

En cliquant sur l'un des liens, LORIS présente la notice bibliographique détaillé qui présente plus d'informations sur le contenu du document . Cet affichage propose à l'usager d'autres liens hypertextes (auteur, sujet, …) permettant ainsi à l'usager de naviguer d'une façon non linéaire dans la base. La figure 19 présente un exemple de l'affichage d'une notice complète pour le catalogue de l'IRISA.

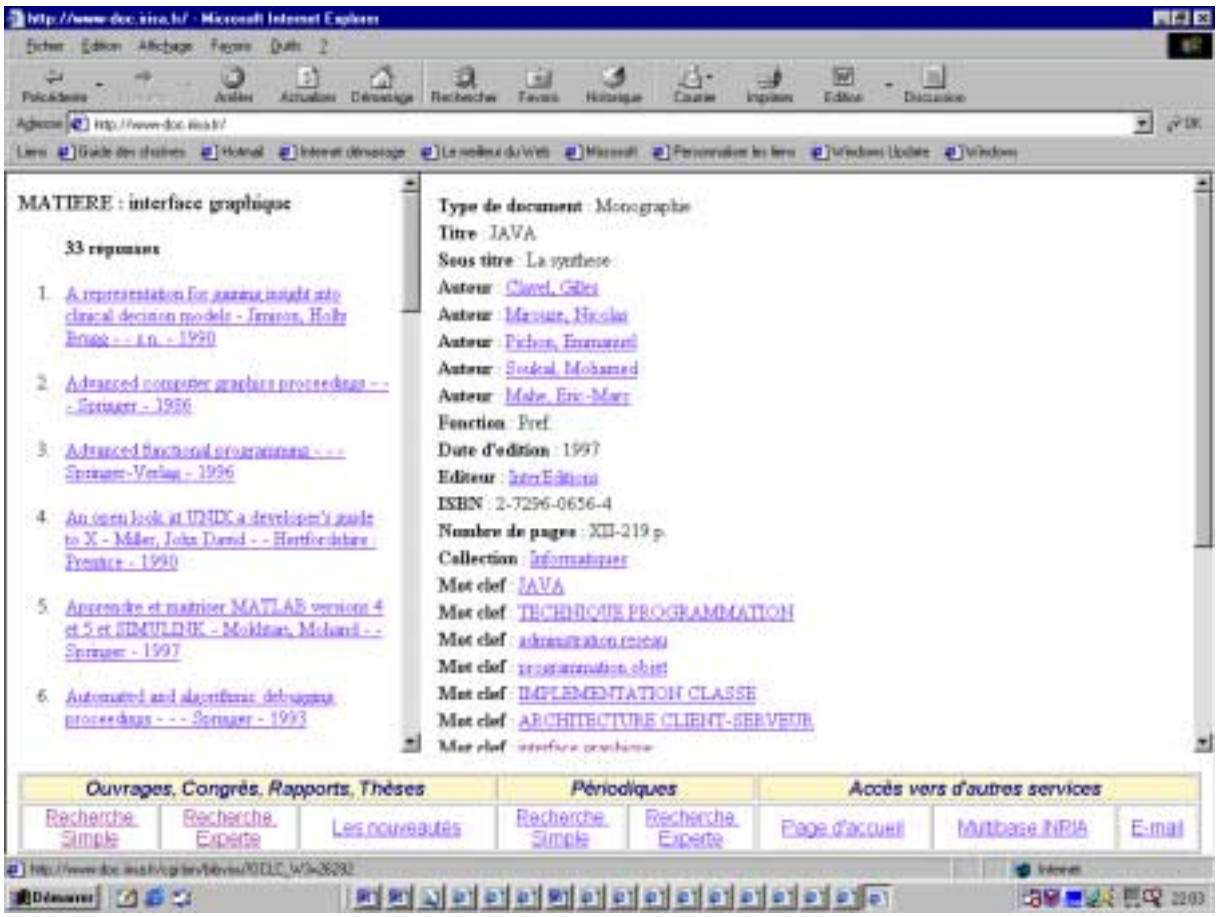


Figure 19 : affichage d'une notice complète dans le catalogue de l'Irisa

5.3.1.2. Les publics

Les catalogues sont libres d'accès sur place et sont accessible par Internet. Le catalogue de l'enssib (http://www.enssib.fr) est destinée essentiellement aux spécialistes de l'informations (étudiants et enseignants des SIC, documentalistes, bibliothécaires, etc.). Celui de l'Université Lumière (http://www.univ-lyon2.fr) aux universitaires intéressés par les sciences humaines, économiques et juridiques. Enfin, le catalogue de l'IRISA (http://www-doc.irisa.fr) est souvent utilisé par des chercheurs spécialisés en informatique. En ce qui concerne la qualification de ces publics, on peut affirmer que la majorité des usagers des catalogues de l'enssib et de l'irisa possèdent des connaissances sur l'usage de la logique booléenne et des ordinateurs. Rappelons toutefois que les étudiants de l'enssib recevoivent en plus une formation spécifiques à l'interrogation des catalogues en ligne.

Même s'ils ne reçoivent de formation à la recherche documentaire, les usagers du catalogue de Lyon2 maîtrisent l'outil informatique.

Le fonds documentaire de l'enssib comme celui de l'irisa est spécialisé. Le catalogue de l'enssib contient plus de 20.000 documents (ouvrages, rapports, articles, thèses, mémoires) en science de l'information et de la communication. Le contenu du catalogue de Lyon2 est multidisciplinaire. Il est riche de plus de 250.000 références.

5.3.2 Recueil des données

Nous avons d'abord observé la répartition des interrogations selon les mois de l'année pour repérer une éventuelle différence d'usage. Il ressort de cette étude que le nombre d'interrogations varie considérablement selon les mois. On peut penser que hormis les mois de juillet, août et septembre, qui correspondent à la période des vacances, tous les autres mois sont représentatifs, et que l'analyse des catalogues peut être faite de façon indifférente quel que soit le mois retenu pour le recueil de données. Nos données corroborent avec celles de Larson (1991).

L'objectif de cette deuxième partie de l'étude consistait à analyser l'existence de variations quantitativement dans le comportement des usagers suivant, les heures et les jours de la semaine. Pour cela, nous avons interprété le contenu des enregistrements des catalogues de l'ENSSIB et celui de Lyon2 pendant six mois. On constate que 80, 7% des recherches ont lieu entre 10 et 18 heures. Les usagers consultent chaque jour dans des proportions très proches le catalogue. Le nombre d'interrogations décroît au cours du week-end (8, 8%), ainsi que la fréquentation des bibliothèques. Il en résulte qu'on peut aussi étudier indifféremment le comportement des usagers pendant toute la semaine. Lorsqu'on examine la répartition des interrogations au cours d'une semaine, nous ne remarquons pas un usage important les lundis et les vendredis. Les données correspondant aux journées du samedi et de dimanche correspondent en général au accès à distance.

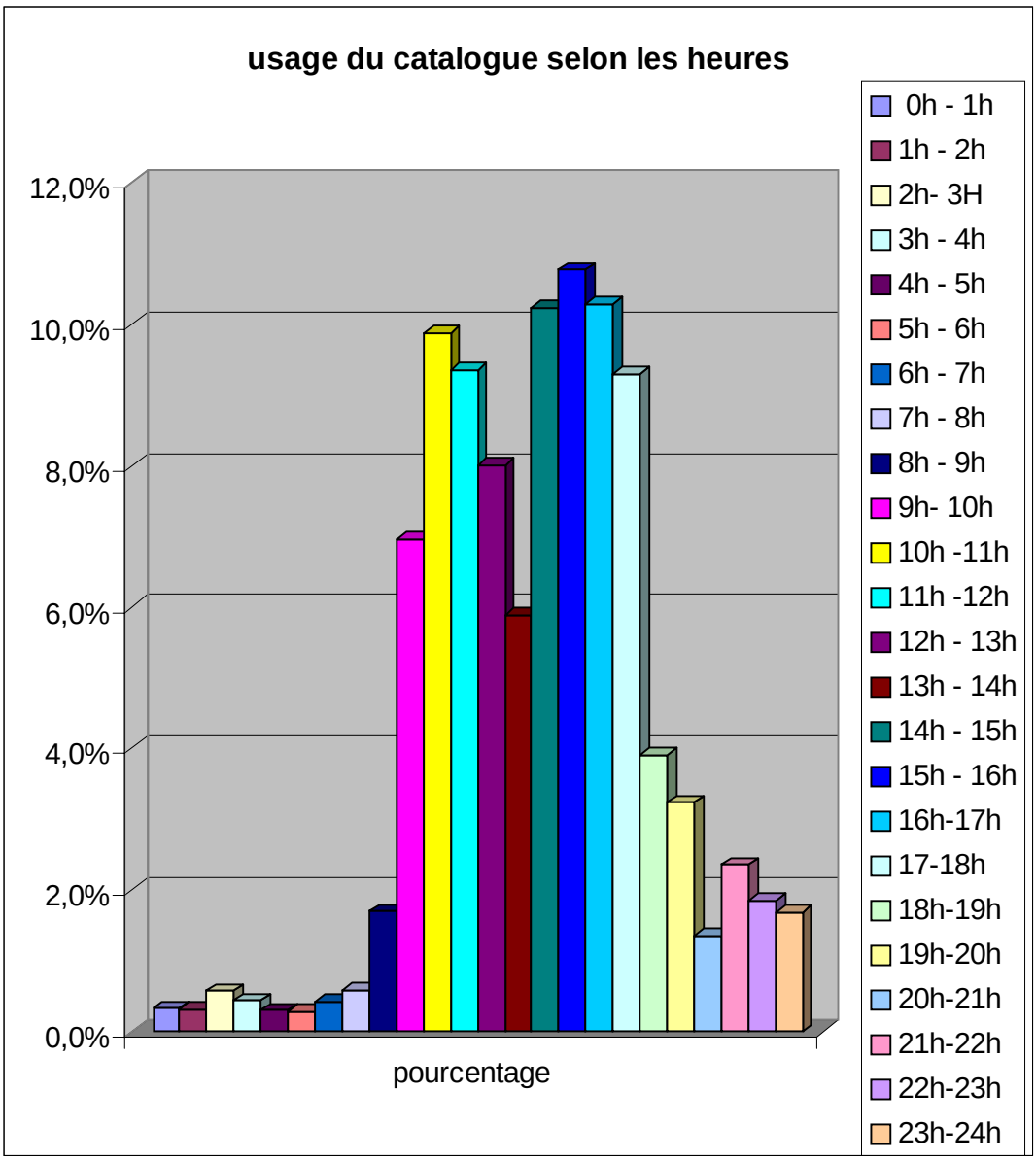


Figure 20 : répartition des accès selon les heures : le cas de Lyon2

Pour l'analyse des taux d'échecs et de surabondance, nous n'avons pas pris en compte les recherches effectuées pendant le mois d'août, les week-end et pendant les tranches horaires suivantes : 0h-8h et 18h-24H.

samedi	4%
dimanche	5%
lundi	19%
mardi	18,50%
mercredi	17,60%
jeudi	16,50%
vendredi	19,60%

Tableau 25 : répartition des accès selon les jours (enssib)

5.3.3. La répartition des points d'accès

Le premier facteur étudié concerne la proportion des points d'accès et leur évolution selon les mois et selon le type de catalogue. A partir du tableau 26, on peut observer que les recherches simples (recherche par mots du titre et par mots de sujet) sont de loin les plus nombreuses pour trois catégories d'utilisateurs (ENSSIB et Lyon2, IRISA). On peut considérer ce mode recherche comme un accès sujet. Ceci est en conformité avec plusieurs autres études de deuxième génération (voir 2.2. 3.) qui montrent la prédominance de l'accès sujet dans les catalogues en ligne. Les interrogations se font davantage à partir du mot de sujet qu'à partir de noms propres d'auteur ou de la désignation exacte du document recherché. Dans la majorité des cas, les usagers qui consultent le catalogue en mode"expert" se contentent d'un seul critère de recherche. La recherche booléenne (combiner plusieurs index de recherche) est rarement usitée. Elle ne représente qu'un dixième des transactions. Ceci est en conformité avec les 9.8% de (Ferl, 1996). Lorsque plusieurs critères de recherche sont croisés, on retrouve le plus souvent l'une des combinaisons suivantes:

- auteur + titre
- auteur + date de publication
- auteur + mots-clé
- titre + date de publication.

Les autres combinaisons sont très hétérogènes et beaucoup plus rares. Les recherches avec trois clés ou plus ne représentent que 1% de l'ensemble des interrogations.

On remarque d'ailleurs que certains points d'accès ne sont presque jamais utilisés; l'exemple le plus approprié est la recherche par cote dans le catalogue de l'ENSSIB qui ne représente que 1, 3% des recherches. Ce qui différencie aussi ces résultats de ceux obtenus dans l'étude des catalogues de deuxième génération est la forte proportion de l'usage des liens hypertextes (plus de 12% des accès).

L'examen du tableau 26, nous révèle une nette différence entre l'usage des points d'accès selon les bibliothèques. La moitié des accès de l'IRISA concerne la recherche simple alors qu'ils sont de l'ordre de 30% pour Lyon2. De même, les recherches sur des éléments connus (titre et auteur) sont plus importantes chez les étudiants de Lyon2. Enfin, on peut aussi remarquer le faible emploi des recherches multicritères par les utilisateurs du catalogue de l'IRISA. Cependant, à partir de ces données, il s'avère difficile de dégager les raisons pour lesquelles les usagers choisissent les différents modes d'accès. Il faudrait une autre méthode de recueil de données (verbalisation ou interview par exemple). Nous n'avons pas effectué cette analyse dans cette thèse.

La prépondérance des recherches mono-index laisse penser que les usagers des catalogues en ligne sont mieux servis quand on leur fournit dans un premier temps un nombre limité de clés de recherche, le sujet, soit l'auteur, le titre. On peut leur donner accès dans un deuxième temps aux types de recherches plus pointus.

	Établissement (nombre d'accès)		
Type d'accès	**Lyon2 (33829)**	**Irisa (5274)**	**ENSSIB (31451)**
Simple	30,0%	50,8%	40,1%
Auteur	21,0%	10,7%	16,2%
Liens hypertextes	12,0%	9,8%	15,1%
Titre	12,0%	9,5%	8,8%
Multicritère	12,7%	3,8%	9,4%
Mots clé	12,1%	13%	8,8%
Autres	0,1%	2,4%	1,5%
total	100%	100%	100,0%

Tableau 26 : Répartition des points d'accès selon les trois catalogues

5.3.4. Les opérateurs booléens et la troncature

Nous avons compté le nombre de fois où les usagers des trois catalogues introduisent un opérateur booléen ou la troncature. Seul, l'opérateur ET est utilisé et ce dans des proportions proches pour les catalogues de l'ENSSIB et de Lyon2. L'opérateur OU est employé dans 1% des cas pour les trois catalogues alors que l'opérateur SAUF est pratiquement absent des requêtes (moins de 1%). Si l'on peut conclure facilement sur la rareté d'usage des opérateurs booléens (OU, SAUF), il est, par contre, difficile d'affirmer cela pour l'opérateur (ET) ; en effet les trois catalogues permettent de lier implicitement les mots entre eux. Enfin, nous avons noté que la troncature est très faiblement utilisée. Ces résultats sont aussi conformes à ceux observés dans les études des catalogues de deuxième génération. Seule l'étude de (Nielsen, 1993) contredit nos résultats. Elle montre que dans le cas de l'Europe du Nord, les lecteurs emploient souvent la troncature et l'opérateur ET. Elle pense que cela est dû à des particularités linguistiques de ces pays (morphologie riche).

. Voici un exemple de l'usage de ces opérateurs pour le catalogue de l'ENSSIB :

	auteur	**Mots clé**	**cote**	**simple**	**date**	**titre**
ET	0,1%	7%	2%	5%	0%	3%
OU	0,1%	0%	0,1%	0,1%	0%	0,1%
SAUF	0%	0%	0%	0%	0%	0%
Troncature	3%	4%	4%	2%	2%	1%

Tableau 27 : Utilisation des opérateurs booléens dans le catalogue de l'ENSSIB

Nous avons aussi observé que les usagers introduisent parfois les opérateurs directement en anglais (AND, OR)

On peut donc affirmer qu'il n'existe pas un usage intensif des opérateurs booléens même pour les lecteurs qui ont suivi une formation en documentation et/ou à l'outil d'interrogation. Ce qui est le cas des usagers de l'ENSSIB.

Nous avons par la suite examiné et compté le nombre de mots saisis par les usagers de l'ENSSIB dans les champs suivants : recherches simples, sujets et mots clé. L'analyse nous a montré que les requêtes des usagers sont très courtes. Le nombre de mots saisis, en moyenne, par les usagers en local est de l'ordre de 1,8 alors que ceux qui accèdent à distance tapent en moyenne 1,6 mots. Nos données confirment les observations de Markey (1994). Le nombre de mots dans une requête est en général inférieur à deux termes.

Chapitre six : Les problèmes de l'échec et de la surabondance d'information dans la consultation des WWW-OPACs.

La gestion des erreurs est un facteur important dans le dialogue homme-machine. Elle concerne les moyens permettant d'une part de réduire ou d'éviter les erreurs et d'autre part de les corriger lorsqu'elles surviennent. Les identifier, les catégoriser est une tâche importante pour améliorer l'interaction dans les catalogues en ligne. Cela constitue une source d'information dans la prévention des erreurs, leur correction et la rédaction des messages. L'un des aspects les plus frustrants de la recherche dans les catalogues est le message d'inexistence de réponses sans aucune autre indication sur les sources de l'erreur.

6.1. Qu'est-ce qu'un"échec" dans une recherche d'information

Aucun consensus sur la définition de l'échec dans un processus de recherche d'information n'est établi. Dans sa thèse de doctorat, Tonta (1992) a regroupé l'ensemble des travaux sur les échecs. Elle distingue trois approches différentes:

- Les mesures de performances (précision/ bruit).
- Les mesures de satisfaction : il y a échec lorsque le document trouvé ne correspond pas aux besoins de l'usager.
- Inexistence de notices : les recherches sont infructueuses lorsqu'elles ne permettent pas d'identifier au moins une notice. L'analyse transactionnelle, comme méthode de recueil de données, apparaît comme la plus apte à analyser ce type d'échec..

Nous avons appliqué deux méthodes pour le recueil de données : l'analyse transactionnelle et l'interview d'étudiants. Toutes les recherches des usagers ont été reconduites dans les deux catalogues (ENSSIB et Lyon 2). Afin de déterminer quelles sont les raisons qui expliquent les insuccès des lecteurs, nous avons effectué des corrections orthographiques, troncature, passage d'une recherche sujet par titre, suppressions des mots, consultation du fichier d'autorité RAMEAU, consultation d'autres catalogues[37], feuilletage dans les rayons des deux bibliothèques.

[37] Notamment celui de la bibliothèque du congrès, de la BNF.

L’interview des étudiants s'est déroulée en deux étapes. La première à l’ENSSIB entre le 15.03.97 et 30.03.97 sur une population de 20 usagers. La seconde, effectuée à Lyon2, du 20.02.98 au 24.02.98 (sur une population de 30 usagers). Nous avons examiné les 50 étudiants en situation réelle d'interrogation, puis nous les avons ensuite questionnés à la fin de chaque session pour mieux comprendre les problèmes qu'ils ont rencontrés lors de leurs recherches. Ces observations nous ont servi de base pour effectuer une typologie d'erreurs ; elles nous aussi aidé dans la compréhension des tactiques mises en œuvres par les usagers lorsqu'ils ne trouvent pas de réponses. Cette méthodologie est très différente de celles habituellement utilisés dans les études sur les catalogues de deuxième génération (Peters 1993). Souvent, ces auteurs se contentent de la lecture d'un petit échantillon des données pour fixer les causes des échecs (Annexe1).

Dans ce qui suit, nous présentons les résultats de l'analyse des échecs et les catégories de ces échecs pour deux catalogues (ENSSIB et Lyon2). Dans la section suivante, nous analyserons les résultats de la deuxième étude sur les tactiques

6.2. Typologies des échecs

Une partie des erreurs observées dans les travaux de Peter (1993), Hunter (1991) et de Markey (1994) a été analysée, affinée et regroupée en 10 catégories principales.

1. **BDD (base de données) :** La requête est bien formée mais les documents demandés ne sont pas dans la base
2. **Booléen :** L'échec est dû à l'emploi d'un opérateur booléen
    - par exemple, Holm ET Nelson
3. **Index** : Cette catégorie regroupe l'ensemble des erreurs où l'usager s'est trompé de clé d'accès. Au lieu de saisir les termes dans un champs bien déterminé (par exemple, l'index titre), il saisit les mots dans le champ auteur.
4. **Écriture:** nous avons rassemblé dans cette catégorie, l'ensemble des insuccès engendrés par des fautes orthographiques ou typographiques.
5. **Langue étrangère:** cette catégorie est composés de l'ensemble des échecs occasionnés par l'emploi d'un ou plusieurs termes en langue étrangère dans la requête.
6. **Troncature**: le choix de l’opérateur de la troncature ne correspond pas à celui du catalogue.
    - Par exemple $ au lieu de *.

- Plusieurs lettres saisies en même temps : **
- Troncature au milieu alors que cela n'est pas possible.

7. **Précision:** les termes de la question sont trop précis par rapport à l'indexation des ouvrages.
8. **Un mot en plus**: l'échec est dû à un nombre important de termes dans la question. Si on dégrade la question d'un mot, il n y a plus d'échec.
9. **Format**: les lecteurs ne savent pas quel est le format exact des données. C'est le cas pour les accès par date et par cote (par exemple, mémoire 1998 au lieu de M1998)
10. **Multiples**: l'usager a commis plusieurs erreurs en même temps (par exemple faute d'index + faute d'écriture)
11. **Autres**: dans cette catégorie nous regroupons toutes les erreurs qui ne correspondent pas à celles citées au-dessus
    - La requête est vide.
    - Le contenu de la question est inintelligible (par exemple, aaaa, jhfdff.etc.)
    - Lorsqu'on n'arrivait pas à un choix de la catégorie
    - Utilisation d'abréviation : Ce type d'erreurs correspond fréquemment à des accès auteur collectivité (cne vs commission nationale d'évaluation)

6.3. Raisons des échecs

Nous avons recherché les causes des échecs pour les quatre points d'accès les plus employés par les usagers. Elles sont distinctes selon les catalogues et selon les points d'accès. On peut affirmer que globalement, le taux d'échec s'explique essentiellement par les deux raisons suivantes (figures 20 et 21) :

- Les usagers commettent des erreurs (souvent typographiques et orthographiques) lors de l'écriture des requêtes. C'est d'ailleurs la principale cause des échecs pour les usagers de Lyon2.
- La bibliothèque dont on consulte le catalogue ne possède pas d'ouvrages sur le thème recherché. Dans le cas de l'ENSSIB, c'est la cause essentielle des échecs selon les quatre points d'accès. L'analyse de corpus peut être un élément de réflexion pour déterminer la clarté de l'information sur le contenu de la base et sur les politiques d'acquisition. Nous n'avons pas étudié cet aspect.

Dans ce qui suit, nous examinons plus en détail ces raisons et selon chaque point d'accès, nous proposons les solutions appropriées.

Lorsqu'on analyse le contenu des requêtes concernant le champ auteur, on observe une nette préférence des recherches par auteur pour les personnes (87% pour les accès en local et 79,1% pour les consultations à distance). Les usagers ne saisissent pas en totalité le nom et le prénom de l'auteur: la proportion de requêtes composées conjointement d'un nom et d'un prénom sont faibles. Il faut constater, que contrairement à la plupart des autres études (Peters, 1993) (Hunter, 1991), la recherche par auteur n'a pas connu un taux d'échecs élevé du fait que le système permet la recherche des noms d'auteurs dans n'importe quel ordre. Grâce aux possibilités offertes par les recherches par mots de l'auteur, les catalogues en ligne n'ont plus besoin d'inclure des programmes qui permettent automatiquement d'inverser les prénoms et les noms des personnes. C'est l'une des premières raisons qui justifie le faible pourcentage des échecs. Dans les études précédentes, ce fût l'une des principales sources d'erreurs.

Très peu de recherches effectuées par auteur concernent les collectivités (moins de 2% pour les usagers qui accèdent à distance). Ce pourcentage corrobore avec les conclusions de Markey (1994) et confirme ainsi la très faible utilisation des noms de collectivités dans les recherches en ligne. Ainsi on peu s'interroger s'il est toujours nécessaire de normaliser ce champs ? Nous pensons que la recherche par mots rend inutile cette normalisation.

On trouvera ci-dessous, les résultats des analyses des contenus des requêtes au catalogue de l‘ENSSIB.

	Accès local (n=991)	**Accès distant (n=822)**
Nom	79,1%	87,6%
Collectivité	3,54%	1,4%
Prénom	3,83%	1,8%
(Nom, prénom)	11,8%	9,01%
Autre	1,73%	0,01%

Tableau 28 :Catégories des accès par auteur pour le catalogue de l’ENSSIB.

De prime abord, on constate que les causes des échecs ne sont pas les mêmes pour les deux catégories de lecteurs. Plus de la moitié des échecs (57.17%) est due à l’absence d’ouvrages dans le catalogue de l’ENSSIB. L'examen détaillé de la liste des auteurs saisis montre qu'une partie (6%) des insuccès résulte de la confusion qui est liée au contenu de la base de

l'ENSSIB. Précisément les lecteurs recherchent des auteurs qui ont écrit des articles[38]. Le faible pourcentage de ce genre d'échecs chez les lecteurs de Lyon2 peut s'expliquer par la richesse du fonds documentaire. Ces derniers font davantage d'erreurs lors de la saisie des termes.

Voici des exemples de fautes d'écriture :

- Les fautes de frappe.
- Les erreurs de segmentation (**sal un vs saläun**).
- Des espaces manquants entre deux ou plusieurs mots (**michelsalun**).
- Des erreurs dactylographiques : lettres manquantes (**yves-francois** ; **bordieu vs bourdieu**), lettres en trop (mmarkey vs markey).
- L'oubli de la troncature.

Contrairement à notre attente, peu d'échecs sont la conséquence d'une utilisation des opérateurs booléens dans l'accès par auteur.

[38] le catalogue de l'ENSSIB comprend quelques articles et le dépouillement de quelques revues.

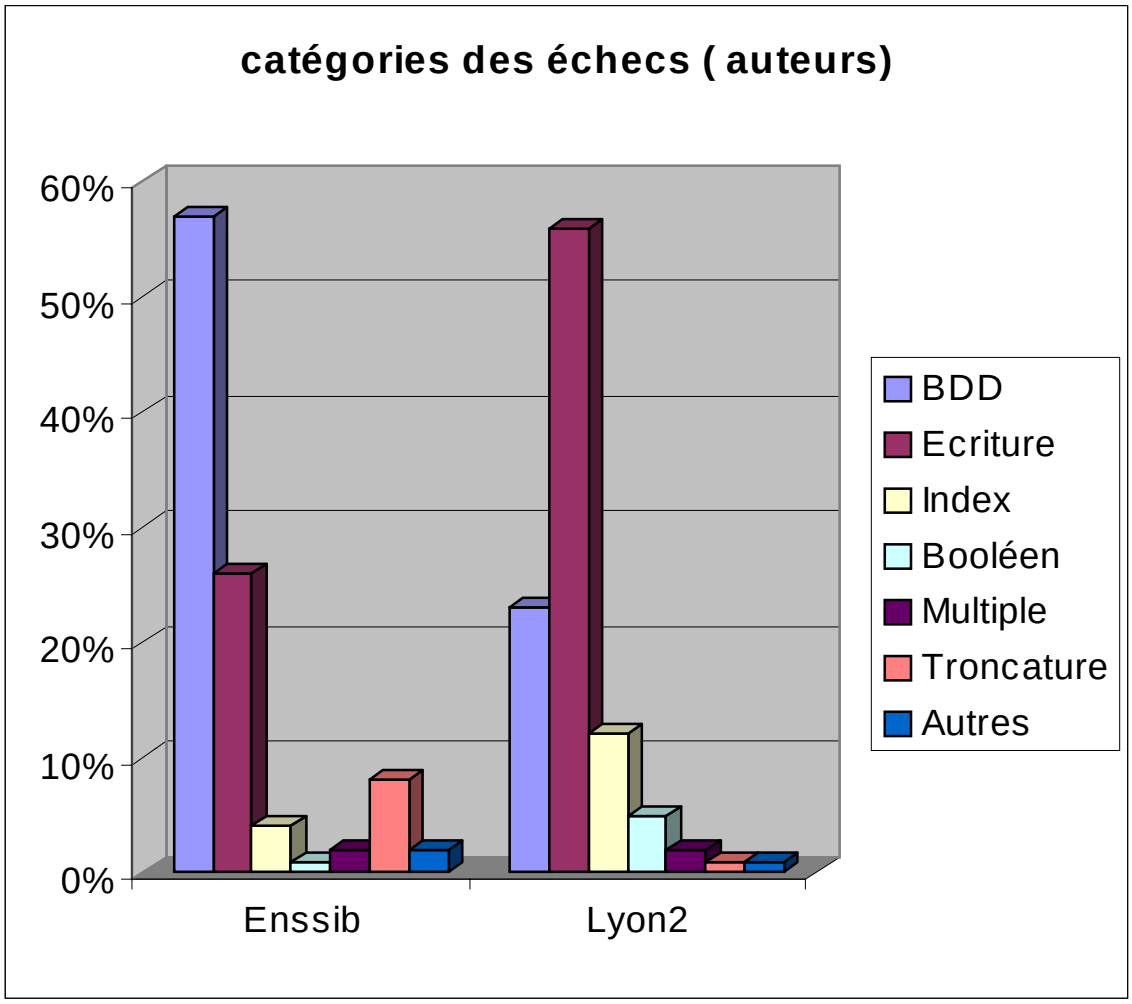


Figure 21 : catégorie des échecs pour l'accès par auteur

La troisième source d'erreurs regroupe les recherches où, par exemple, un titre d'ouvrage a servi de requête en mode auteur. Cela représente 12% chez les lecteurs de Lyon2 contre 4 % pour ceux de l'ENSSIB. Les questions suivantes n'ont aucune chance d'aboutir:

Auteur = asis proccedings

Auteur = diriger une bibliothèque, image en bibliothèque

Le résultat le plus surprenant sans doute concernent les erreurs effectuées par les lecteurs de l'ENSSIB lorsqu'ils emploient la troncature. Cela représente 8% du total des échecs.

Pour les deux catégories d'usagers, à propos des trois autres types de consultation (par titre, simple, mots de sujets), les résultats révèlent que les échecs sont dus à l'absence d'ouvrages et aux erreurs de saisie (figures 21, 22 et 23). On peut toutefois observer que les erreurs sont relativement plus nombreuses dans la recherche par mots du titre que dans la recherche

simple. Ce résultat est étonnant à plus d'un titre. Si l'on considère que l'usager recherche dans le champs titre des documents dont il possède une partie de l'information[39], il est censé faire moins d'erreurs. Par conséquent, on peut estimer qu'une partie des recherches par mots du titre sont en fait des recherches" sujets". Larson (1991) fut le premier à avancer cette hypothèse. Dans son enquête effectuée à l'Université de Californie, il a démontré qu'on assistait à un déclin constant de la recherche par sujet et que la recherche par les mots-clés dans les titres constituait la méthode de remplacement de la recherche par sujet.

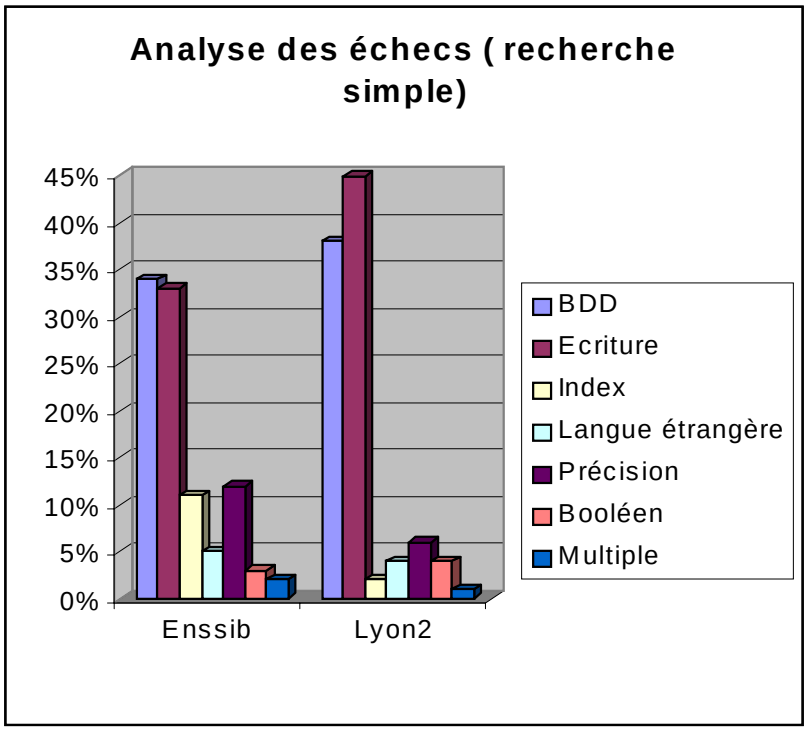


Figure 22 : catégorie des échecs pour l'accès simple

Les difficultés de recherche simple, proviennent en partie de l'insuffisance de l'information (en termes d'aide et de messages) sur les pratiques de catalogage. Il est en effet difficile pour les lecteurs de savoir quelles sont les options de recherche dans le cas d'une recherche par mots du titre. A-t-on besoin juste du premier mot, ou deux mots du titre, ou est-il nécessaire d'introduire tout le titre (le, la compris) ?

Sans surprise, ce sont les lecteurs de l'ENSSIB qui commettent principalement des erreurs de la catégorie" langue étrangère". Comme la bibliothèque est riche en ouvrages publiés en

[39] Les anglo-saxons parlent de "know search".

anglais, mais aussi parce qu'ils effectuent souvent des recherches bibliographiques sur des thèmes pointus[40].

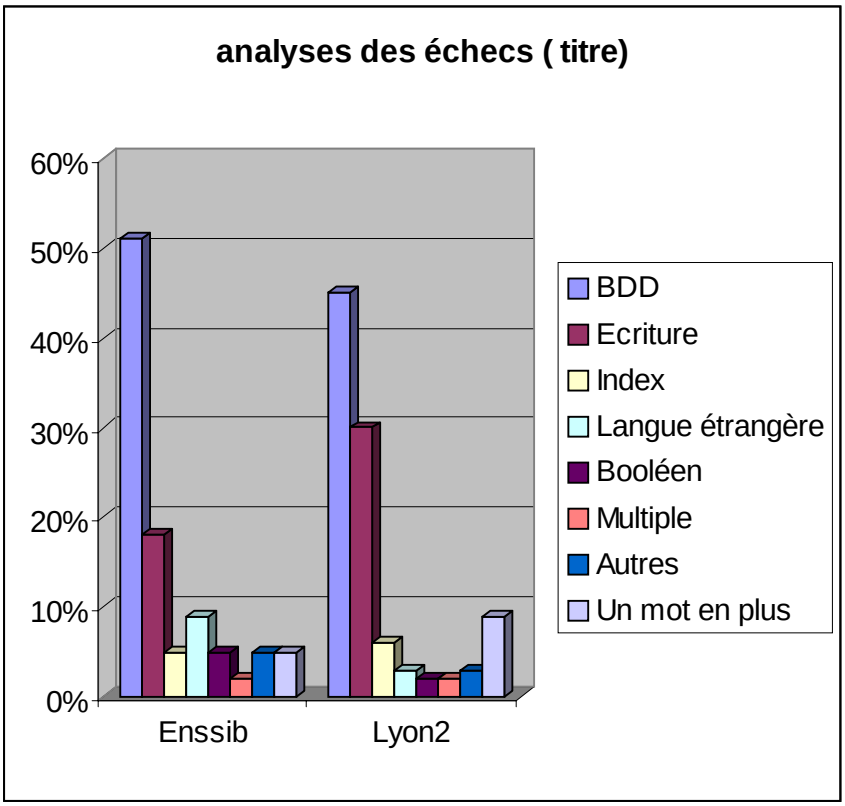


Figure 23 : catégorie des échecs pour l'accès par titre.

Nous avons trouvé peu d'études [41] établissant les causes des échecs d'un accès par cote. Peters (1993) montre que 100 % des accès par une cote aboutit à zéro réponse. La méconnaissance des règles du système (catalogue) est la cause essentielle (92.5%) des échecs. Les problèmes de formats représentent plus d'un tiers des incorrections. Ainsi la requête suivante "cote = mémoire 1996" ne peut aboutir à cause du mot" mémoire". Les problèmes d'écriture sont aussi importants dans un accès par cote. Voici les deux erreurs qui reviennent le plus :

- Manque d'un espace pour les requêtes suivantes, par exemple, m1995 au lieu de lieu de M 1995
- Emploi d'une lettre non acceptée par le catalogue, par exemple, m_1995 au lieu de lieu M 1995

[40] Data mining, datawhahouse..etc.

[41] On ne peut pas prendre en compte l'étude de Hunter qui se base sur un échantillon de 9 termes

Parfois c'est l'ajout d'un terme (par exemple, *374* hen) ou une mauvaise troncature *(*par exemple, cote = 4**) qui occasionne l'échec.

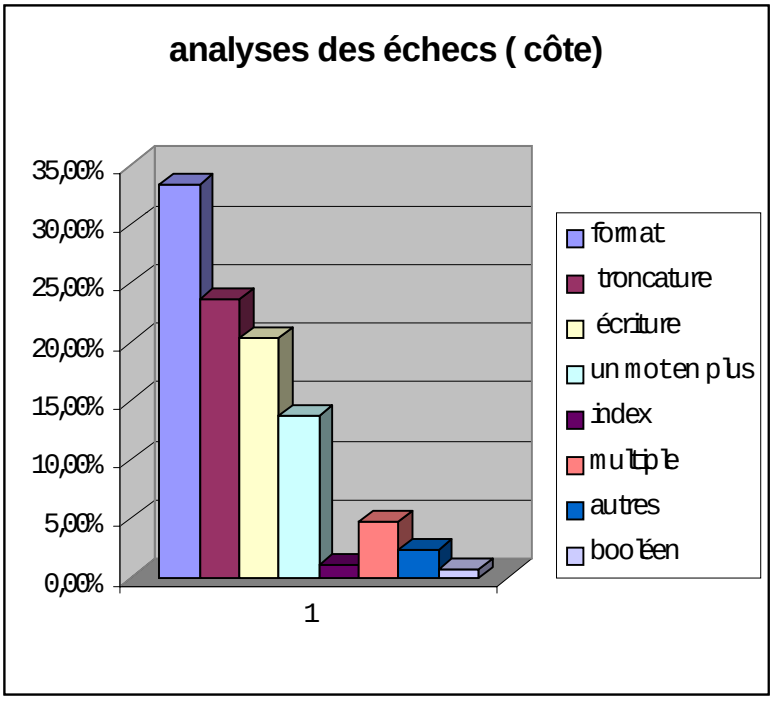


Figure 24 : catégorie des échecs pour l'accès par cote

Quelques améliorations au niveau de l'interface suffiraient pour résoudre une partie de ces problèmes (en donnant des exemples ou en améliorant l'aide en ligne). Dans le cas de l'ENSSIB préciser l'existence des libellés suivants (art pour désigner une recherche par article, DCB pour rechercher un mémoire des élèves du DCB, D.E.A. pour les mémoire de DEA, etc.). Le problème avec l'accès par cote provient de l'inefficacité des indices de classification. L'usager ne pourra jamais deviner ce qui se dissimule derrière ces chiffres énigmatiques.

Concernant l'accès par date, ce sont les erreurs de format (par exemple, après 1998) et d'écriture (19977 vs 1997) qui sont à l'origine des échecs.

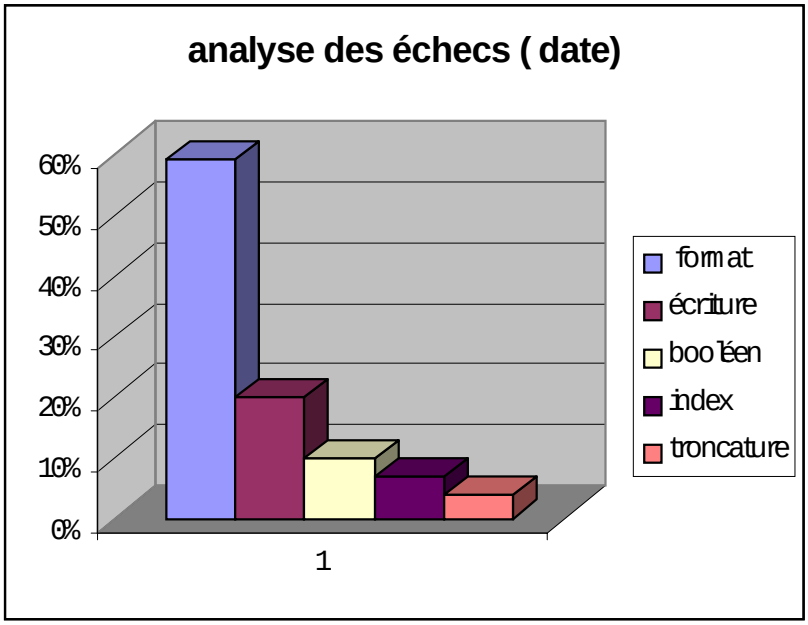


Figure 25 : catégorie des échecs pour l'accès par date

6.4. La surcharge d'information

Peu de recherches se sont intéressées au problème de la surcharge ou du trop grand nombre de réponses résultant d'une requête. Par conséquent, très peu de données empiriques portent sur cette question. Seuls Larson (1991) et Wiberley (1988) ont manifesté un intérêt à ce problème. De plus, ils n'ont examiné ce problème que pour l'accès sujet.

Quand peut-on parler de surcharge d'information ?

La réponse à cette question varie certainement selon les usagers, et pour chacun d'eux suivant ce qu'il cherche et selon le contexte de recherche (urgence, faire une bibliographie, etc.). On considérait dans les catalogues de première génération, que le nombre maximum de réponses à afficher se situait entre 30 et 35 réponses (Wiberley, 1988).

Pour notre part, le seuil à partir duquel nous considérons qu'il y a surabondance de réponses est fixé à 60: A partir de l'observation d'usagers en situation réelle, nous avons remarqué qu'ils ne consultent pas plus de quatre écrans (avec en moyenne 15 réponses par écran, d'où le chiffre de 60). Si les usagers lisent attentivement les résultats du premier écran, cette lecture se transforme en feuilletage de listes dans les autres écran. La lecture est plus attentive dans le premier écran et parfois à la fin de la liste. Ils s'arrêtent en moyenne après le défilement du quatrième écran.

L'observation des lecteurs (50) de l'ENSSIB et de Lyon2 a confirmé les analyses de Wiberley (1988) : trois types d'attitudes se dessinent : l'abandon immédiate de la recherche; la reformulation de la question et le feuilletage de la liste des références. Les lecteurs ne lisent en général que les deux premiers écrans. Nous avons examiné deux milles sessions de recherches et trouvé que près de 70% des notices sont affichées à partir du premier écran.
Cette abondance des réponses peut produire un effet de saturation. Wiberley (1988) indique que le défilement d'un index ou d'un ensemble de réponses lasse les usagers et que la sélection de références nécessite un grand effort de leur part.

L'analyse des données (tableau 29) montre que dans le cas du catalogue de Lyon2, le taux de surcharge est malheureusement très élevé pour les points d'accès suivants : mots du titre (30%), mots du sujets (39,4%), interrogation simple (21,6%). Dans le cas de l'ENSSIB, c'est plutôt la recherche simple (24%) et l'accès par mots clé (42,7%).

	ENSSIB - Nombre de réponses pour un type d'accès donné					
	0	1 à 30	30 à 60	60 à 120	> 120	Total
Mots du titre (n = 8600)	38%	44%	6%	3%	9%	100%
Auteur (n = 7800)	25,4 %	70,1%	1,7%	1,9%	1%	100%
Simple (n = 10000)	25 %	33,5%	17,5%	4%	20%	100%
Date (n=1350)	6,7%	0%	0%	0%	83,3%	00%
Mots clé (n=9800)	24,1%	23,2%	10 %	7,9%	34,8%	100%
Cote (n=1418)	34,3%	11%	16,1%	10,2%	28,4%	100%
Multicritère (n =4000)	58%	41%	1%	0%	0%	100%

Tableau 29: taux d'échecs et de surcharge pour le catalogue de l'ENSSIB

Contrairement à l'analyse des échecs, nous n'avons pas trouvé de catégorisation des origines de la surcharge d'information. Deux raisons peuvent expliquer cette surabondance de l'information:

- les requêtes sont trop générales par rapport au contenu de la base. Des équations telles que droit, information, bibliothèque aboutissent forcément à un nombre grand de réponses. C'est la cause principale de surcharge pour les accès suivants (simple, mots clé, mots du titre, sujet)
- L'utilisation de la troncature à droite.

	Lyon 2 – Nombre de réponses pour un accès donné					
	0	**1 à 30**	**30 à 60**	**60 à 120**	**> 120**	**Total**
Mots du titre (n = 4440)	18 %	38, 4 %	13, 6 %	9, 6 %	20, 4 %	100%
Auteur (n = 7850)	21, 5%	66, 7%	2, 7%	4, 5%	4, 6%	100%
Simple (n = 5000)	24, 2%	44, 8%	9, 4%	6, 6%	15%	100%
Date (n = 800)	19%	0%	0%	0%	81%	100%
Mots du sujets (n = 5700)	26%	19, 7%	14, 9%	8%	31, 4%	100%
Multicritère (n = 3200)	65 %	35 %	0 %	0 %	0%	100%

Tableau 30 : taux d'échecs et de surcharge pour le catalogue de Lyon2

La cause essentielle d'une surabondance d'information lors d'un accès par auteur est l'utilisation d'un prénom seul. Toutes les requêtes comportant des prénoms comme Michel, Bertrand ou Jean aboutissent toutes à des réponses en nombre plus grand que 60.

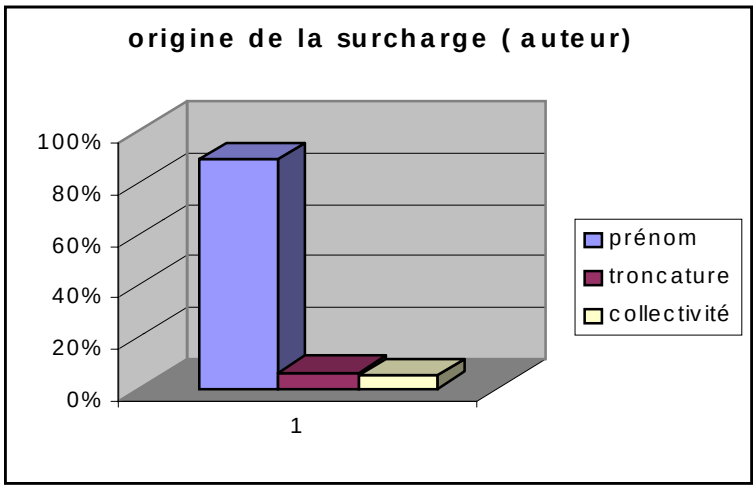


Figure 26 : Origine de la surabondance de l'information à propos de l'accès par auteur.

La deuxième raison concerne l'utilisation du début d'un nom de collectivité comme par exemple : association, ministère…etc. Enfin la troncature à droite aboutit souvent à une augmentation du nombre de réponses.

Accès par la cote

Le nombre élevé de réponses lors d'un accès par cote est souvent le résultat d'un manque d'information (> 95 % de cas). En effet, les usagers n'inscrivent que le début de la cote. Ainsi une requête qui ne contient que le mot "dcb" aboutit à l'affichage de tous les mémoires des étudiants du DCB.

La troncature à droite influe aussi négativement sur le nombre élevé de réponses (5% de cas).

Conclusion

On ne peut pas considérer les insuccès résultants d'une absence d'ouvrages dans la base comme des échecs. Donc le pourcentage d'ouvrages qui n'est pas repéré à cause des erreurs commises par le lecteur est très faible par rapport aux études précédentes. Nous avons réexaminé les taux des échecs des points d'accès les plus utilisés, en prenant en compte les données sur les catégories d'erreurs. Nos résultats indiquent qu'en vérité le pourcentage d'échecs, aussi bien pour l'ENSSIB que pour Lyon2, est compris entre 10 et 19%.

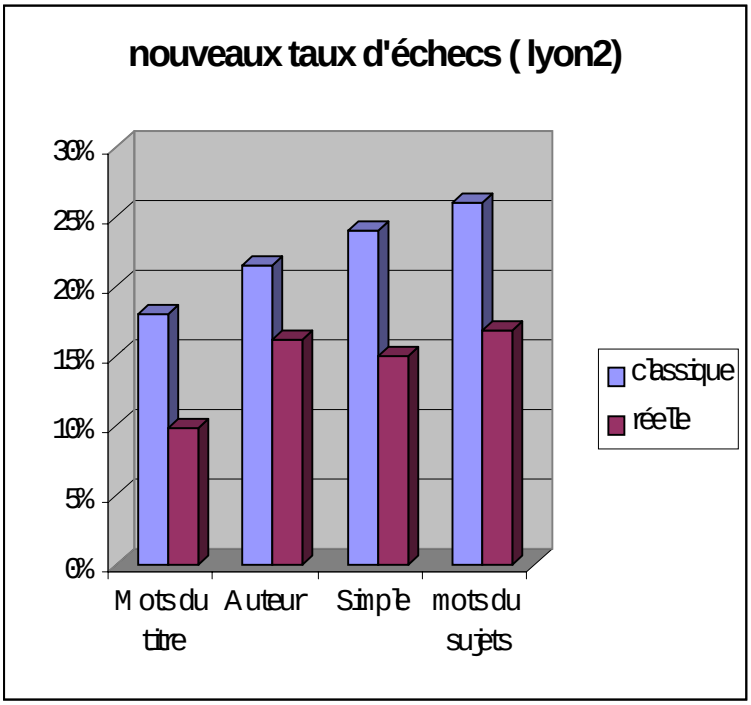


Figure 27 : Nouveau taux d'échecs (Lyon2)

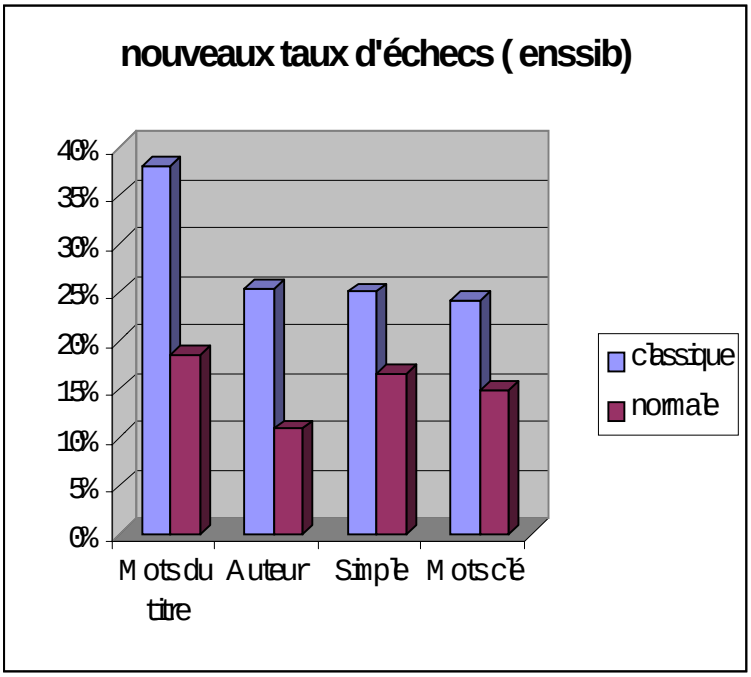


Figure 28 : Nouveau taux d'échecs (ENSSIB)

Ces résultats se révèlent importants pour plusieurs raisons. En premier lieu, cela signifie que l'usage des catalogues et le comportements des usagers a complètement évolué. Pendant des

années[42], pratiquement, toute la littérature scientifique sur les catalogues se limitait à l'examen des difficultés d'accès et au problème du "zéro réponse". On trouve encore des écrits chez des spécialistes sur "l'opacité" des catalogues, se référant à des analyses faites sur des systèmes de première génération. Nous sommes loin des "50%" d'échecs qui caractérisaient désavantageusement les anciens catalogues[43]. L'usage de plus en plus fréquent des traitements de textes, de tableurs, des jeux vidéos et récemment de l'Internet rend l'accès aux catalogues WWW plus habituel et réduit ainsi le nombre et le type d'erreurs. Dans son livre sur les usages, Perriault (1989) constate que *"les néophytes d'aujourd'hui ne sont pas ceux d'hier, ce qui modifie sensiblement le contenu des représentations d'usage et les modèles de références*". En outre, avec les possibilités d'enrichissement des catalogues par les tables de matières, ce problème tend encore à diminuer. Nous pouvons donc affirmer que l'absence d'un document apparaît comme le facteur le plus déterminant d'échec des consultations.

[42] durant la décennie 80 et le début des années 90.

[43] Ne dit-on pas qu'un OPAC est "opaque".

Chapitre sept : L'analyse des échecs et la surcharge d'informations dans un SRI : examen des tactiques mises en œuvre.

Généralement, les études sur les échecs ou sur la surabondance de l'information se limitent à trouver le taux d'échecs et éventuellement les causes des erreurs. Par contre, rares sont les auteurs qui ont examiné les tactiques employées par les usagers dans leurs processus de recherche. Nous n'avons pas trouvé de données empiriques concernant par exemple le type de tactiques mises en œuvre par les usagers lorsqu'ils ne trouvent pas de réponses ou lorsqu'il y a une surabondance d'information. Par tactique, nous entendons l'ensemble des actions conceptuelles et/ou les mouvements opératoires de l'usager pour améliorer le résultat d'une recherche.

Cette partie de notre thèse a pour but de répondre à ces questions:

- quels sont les taux d'échecs et de surcharge d'informations pour chacun des points d'accès ?
- quelles sont les causes des échecs pour chacun des deux catalogues ?
- quelles sont les tactiques employées par les usagers ?
- y a-t-il une différence entre les tactiques employées lors d'un échec et lorsque la recherche aboutit à une abondance de réponses ?
- existe-t-il une grande différence entre les usages du catalogue de l'ENSSIB et celui de Lyon2 ?
- quelles sont les conclusions à tirer pour l'amélioration des catalogues ?

7.1. Catégorisation des tactiques utilisées

Pour déterminer les tactiques, nous avons imprimé et analysé[44] plus de 2000 sessions pour chacun des deux catalogues (1000 sessions où il y a un échec et 1000 sessions où il y a une surabondance de réponses). Nous n'avons pas pris en compte les sessions dans lesquelles les usagers activent les liens hypertextes.

Notre typologie est basée principalement sur une analyse des travaux de Bates (1990), Wildemuth (1991), Fidel (1985), Shenouda (1990) et Bruza (1997). Bates distingue vingt et

[44] nous avons interrogé les deux catalogues avec les mêmes termes que ceux des usagers, visualisé les mêmes notices pour déterminer les différents tactiques utilisées.

neuf catégories de tactiques selon que la recherche d'information est automatisée ou non. Sa catégorisation n'est malheureusement pas dérivée à partir de données empiriques. En se basant sur les travaux de Bates et de Fidel, Wildemuth (1991) a défini douze catégories qu'elle a appliqué à l'analyse des traces de 53 usagers d'une base de données factuelle médicale. Bruza (1997) a déterminé et utilisé douze tactiques[45] pour étudier le processus de reformulation effectué par des usagers d'un moteur de recherche. Ingrid Yee (1998) montre aussi que les tactiques employées lors d'une recherche d'informations dans le Web (moteur de recherche) par des usagers sont simples et limitées. Lorsqu'ils trouvent trop de réponses, les usagers ont tendance soit à ajouter un terme, à feuilleter l'ensemble des résultats ou à changer de moteur de recherche.

Bates, Fidel et Wildemuth ont mis en évidence, qu'à l'inverse des usagers novices, les professionnels de l'information emploient plusieurs tactiques qui leur permettent d'améliorer leurs recherches d'informations. Les études précédentes sont limitées car elles n'ont pas déterminé une possible différence entre les tactiques utilisées lors de l'échec ou de la surabondance d'information, Or, nous pensons que cette distinction, nous donnerait plus d'éléments de réflexion pour comprendre le comportement des usagers et pour améliorer les catalogues existants.

Après la lecture d'une centaine de sessions, nous avons décidé d'ajouter deux autres tactiques : le glissement de champs à champs (INDEX) et la traduction des termes d'une requête (TRADUCTION). Parfois, à partir d'une même requête initiale, les usagers effectuent successivement la recherche dans un champs puis dans un autre. La seconde tactique (TRADUCTION) concerne la situation où un usager effectue une recherche dans une langue (en l'occurrence le français) et reproduit la même requête en traduisant les termes en langue étrangère (en anglais). C'est le cas par exemple dans certaines sessions de recherches observées à l'ENSSIB. Comme le fonds documentaire est riche en ouvrages écrits en anglais, certains usagers effectuent cette tactique. En l'absence de possibilités de recherches multilingues, cette tactique s'avère efficace pour obtenir plus de documents. Notre liste comprend donc les treize tactiques suivantes:

1. Suppression d'un ou plusieurs termes pour rendre une requête moins spécifique (SUPPRESSION).
2. Ajout d'un ou plusieurs termes pour affiner une recherche (ADJONCTION)

[45] Similaires à celles de Wildemuth

- 159.84.68.121 - - [06/Jun/1997:16:03:31 +]"GET /cgi-bin-ever/bibinterro?titre=&auteur=&dated=&sujet=**acquisition**&operat=2 HTTP/1.0" 200 19497
- 159.84.68.121 - - [06/Jun/1997:16:03:59 +]"GET /cgi-bin-ever/bibinterro?titre=&auteur=&dated=&sujet=**acquisition + société** &operat=2 HTTP/1.0" 200 717.

3. Répétition d'une même requête (REITERATION)
4. Substitution d'un ou plusieurs termes d'une requête (SUBSTITUTION)
5. Dérivation de formes (racine, ajouts ou retraits de suffixe, préfixe mais aussi usage de la troncature : DERIVATION)
   - [06/Jun/1997:17:56:05 +]"GET /cgi-bin-ever/bibinterro?titre=verbe&auteur=&dated=&sujet=**ling**&operat=2 HTTP/1.0" 200 743
   - [06/Jun/1997:17:56:05 +]"GET /cgi-bin-ever/bibinterro?titre=verbe&auteur=&dated=&sujet=**ling*** &operat=2 HTTP/1.0" 200 744
6. Correction (orthographique, changement dans la ponctuation, remplacement de chiffre par des nombres etc.) : CORRECTION
   - schutz.[10/Jun/1997:17:38:25 +]"GET /cgi-bin-ever/bibinterro?titre=&auteur=**Perreti**&dated=&sujet=&operat=2 HTTP/1.0" 200 256
   - schutz.[10/Jun/1997:17:38:39 +]"GET /cgi-bin-ever/bibinterro?titre=&auteur=**Peretti**&dated=&sujet=&operat=2 HTTP/1.0" 200 3565
7. Élargissement ou contraction d'une abréviation (ABBREVIATION)
   - 159.84.68.121 - - [06/Jun/1997:16:02:34 +]"GET /cgi-bin-ever/bibinterro?titre=&auteur=&dated=&sujet=**btp**&operat=2 HTTP/1.0" 200 252
   - 159.84.68.121 - - [06/Jun/1997:16:02:54 +]"GET /cgi-bin-ever/bibinterro?titre=&auteur=&dated=&sujet=**batiment + travaux + publics**&operat=2 HTTP/1.0" 200 1657
8. Usage d'un opérateur booléen (BOOLEEN)
   - scdinfp3 - - [06/Jun/1997:16:35:20 +]"GET /cgi-bin-ever/bibinterro?titre=**ecole et societes**&auteur=&dated=&sujet=&operat=2 HTTP/1.0" 200 2298

- scdinfp3 - - [06/Jun/1997:16:37:27 +]"GET /cgi-bin-ever/bibinterro?titre=**ecole + societes**&auteur=&dated=&sujet=&operat=2 HTTP/1.0" 200 2295

9. Index : changement de clés d'accès (INDEX)
   - 159.84.68.121-- [06/Jun/1997:15:53:36 +]"GET /cgi-ever/bibinterro?titre=**mondialisation**&auteur=&dated=&sujet=&operat=2 HTTP/1.0" 200 2106
   - 159.84.68.121 - - [06/Jun/1997:15:58:46 +]"GET /cgi-bin-ever/bibinterro?titre=&auteur=&dated=&sujet=**mondialisation**&operat=2 HTTP/1.0" 200 2106
10. Traduction d'un terme dans une autre langue (TRADUCTION)
    - "GET /cgi-ever/bibinterro?titre=**bibliothèques numériques**&auteur=&dated=&sujet=&operat=2 HTTP/1.0"
    - "GET /cgi-ever/bibinterro?titre=**digital libraries** &auteur=&dated=&sujet=&operat=2 HTTP/1.0"
11. Union : remplacement de l'opérateur ET par le OU : (UNION)
12. Ensemble: remplacement d'un OU par un ET (ENSEMBLE)
13. Utilisation d'une ou plusieurs tactiques en même temps (COMBINAISON)
    - [06/Jun/1997:17:51:49 +]"GET /cgi-bin-ever/bibinterro?titre=&auteur=**calvino**&dated=&sujet=langues&operat=2 HTTP/1.0" 200 263
    - [06/Jun/1997:17:52:16 +]"GET /cgi-bin-ever/bibinterro?mots=**calvino** HTTP/1.0" 200 1571

7.2. Résultats et analyse

Plusieurs travaux ont montré que dans les catalogues de première et deuxième génération, les ressources sont sous-utilisées: les recherches des usagers sont simples, ils n'emploient pas les possibilités fournies par le système et modifient rarement leurs stratégies de recherche. C'est inexact dans la mesure où ils reformulent leurs questions, mais leurs reformulations restent souvent primaires. Elles portent sur la structure logique de la requête (modification des opérateurs booléens) et/ou sur les concepts (changer le contenu de la requête). L'analyse des traces, nous montre que les usagers optent de préférence pour cette dernière. Elle se matérialise par plusieurs tactiques (remplacement d'un concept par un autre, adjonction,

suppression, . etc.). Plusieurs catégories de tactiques n'ont pas été observées, notamment celles qui requièrent la maîtrise des opérateurs booléens (ET, OU et le SAUF). Ce résultat confirme les travaux de Wildemuth (1991).

Les tactiques utilisées par les usagers sont différentes selon le type de bibliothèque mais aussi selon l'échec ou la surcharge d'informations (tableau 31). Dans les deux cas, nous pouvons remarquer que les phénomènes de reprise sont importants. Osmont (1995) constate que:*"une palinèrèse n'est pas un retour à la case de départ; elle est constitutive du processus de la recherche."* Ce qui semble intéressant, c'est que la réitération est aussi une tactique souvent utilisée dans les moteurs de recherche (Bruza, 1997). Par ailleurs, nous avons constaté qu'il peut y avoir alternance entre les tactiques au sein d'une même recherche. Ceci survient surtout lorsque l'utilisation d'une tactique ne donne pas de résultats satisfaisants.

	normal		**Zéro réponse**		**surcharge**	
	ENSSIB (n =1500)	Lyon2 (n =1500)	ENSSIB (n =2774)	Lyon2 (n =2436)	ENSSIB (n =2090)	Lyon2 (n =2016)
Réitération	29%	15%	14%	15%	9%	3%
Suppression	21%	16%	24%	26%	0%	7%
Adjonction	16%	17%	0%	13%	61%	38%
Index	11%	15%	15%	5%	10%	30%
Substitution	4%	19%	19%	23%	8%	10%
Dérivation	4%	3%	7%	5%	0%	0%
Correction	4%	5%	9%	10%	1%	7%
Traduction	2%	1%	5%	0%	0%	0%
Booléen	1%	1%	1%	0%	3%	0%
Combinaison	5%	8%	4%	0%	5%	6%
Union	1%	0%	1%	0%	0%	0%
Ensemble	1%	0%	0%	0%	3%	0%
Abréviation	1%	1%	1%	3%	0%	0%
Total	100%	100%	100%	100%	100%	100%

Tableau 31 :analyse des tactiques (n= nombre de tactiques)

7.2. 1. Tactiques liées aux échecs

Lors d'un échec, les usagers procèdent souvent à une autre alternative de recherche (73% de cas pour les usagers de l'ENSSIB et 58 % pour les usagers de Lyon2). Cette différence peut s'expliquer par le niveau de formation suivie par les deux catégories d'usagers. En plus de leurs formations aux techniques documentaires, les étudiants de l'ENSSIB reçoivent une formation à l'usage du catalogue. On constate que:

- Le taux de réitération est assez élevé (14% vs 15%). Lorsque les usagers ne trouvent pas de réponses, ils ont tendance à formuler la même requête.
- Lorsqu'aucune notice n'est trouvée en appariement, l'usager procède souvent à la dégradation de sa requête, c'est à dire à la suppression d'une des notions figurant dans la requête (24% vs 26%).
- Les usagers ont tendance soit à supprimer, soit à remplacer un terme par un autre plus proche (19% vs 23%).
- l'usage de certaines tactiques nous semble non seulement improductif mais illogique comme l'usage de la tactique d'adjonction par les utilisateurs de Lyon2.
- Les usagers n'arrivent à employer ni les opérateurs booléens, ni la troncature dans leurs stratégies de recherche. Les tactiques suivantes (Booléen, union, ensemble, dérivation) qui font appel à une maîtrise de la logique booléenne sont peu sélectionnées.
- Contrairement à plusieurs études d'usage des catalogues de première et de deuxième génération (Hildreth, 1989) et (Markey, 1994) (Hunter, 1991), nous constatons que parfois les usagers changent leur mode d'accès lorsqu'ils ne trouvent pas de réponses (tactique INDEX). Ce sont toutefois, les étudiants de l'ENSSIB qui font le plus usage de cette tactique (15% pour les usagers de l'ENSSIB et 5% pour ceux de Lyon2). Ce type d'usage a été rarement observé dans les études des usagers grand public. Il serait intéressant de voir si des lecteurs de bibliothèques publiques peuvent mettre en œuvre ce type de tactiques, d'autant que lors d'un échec par sujet, l'usage de la tactique INDEX est plus efficace qu'une troncature automatique ou la suppression d'un ou plusieurs mots de la requête.
- Contrairement à notre attente, les usagers ne corrigent pas souvent leurs requêtes (9% vs 10%). Les possibilités de correction automatique (typographiques et/ou orthographiques) sont donc toujours d'actualité.

- Les utilisateurs de l'ENSSIB traduisent parfois leurs concepts dans une autre langue. L'analyse des sessions, nous a montré que cette tactique est souvent efficace.

7.2. 2. Tactiques liées à la surabondance des réponses

La surcharge d'informations est par contre, un problème plus urgent: il concerne plus d'un tiers des requêtes pour les accès "sujets" et "mots du titre". Ce problème est aggravé puisque seulement un usager sur deux tente de réduire le nombre de réponses en ajoutant un terme à l'équation d'origine. Cette stratégie n'est malheureusement pas toujours concluante. Pour eux, il s'agit souvent de deviner les termes (vedettes matières) qui ont servi à l'indexation: le choix des nouveaux termes se fait par essai /erreur. Le catalogue ne leur offre aucune aide à cette étape. L'introduction du "feedback" aurait pu faciliter cette tâche pour les usagers. Malheureusement, la majorité des catalogues en ligne opérationnels n'intègrent pas cette fonctionnalité. Nonobstant, nous avons montré ailleurs (Ihadjadene, 1998b) que la navigation à travers les liens hypertextes permet d'atténuer ce défaut. Lorsqu'on examine les sessions où les usagers activent des liens hypertextes, on remarque un faible pourcentage de tactiques. Lorsque celles-ci existent, elles ne concernent que le début des recherches. L'usager préfère plutôt naviguer que redresser ses requêtes.

Les usagers se contentent donc d'afficher quelques notices ou abandonnent la recherche. On constate que:

- Les usagers ont tendance à employer massivement les tactique adjonction (61% vs 38%) et INDEX (10% vs 30%).
- De nombreuses recherches restent partielles. Ainsi dans notre analyse des traces, si une recherche par mots du sujet (ou mots du titre) aboutit à une surcharge d'informations, elle est rarement (voir tactique de combinaison) suivie de recherches sur l'auteur, sur les indices de classifications ou sur les dates pour préciser la recherche. L'usage de cette tactique (combinaison) n'est cependant pas toujours très probant : en ajoutant un nouveau mot pour limiter sa recherche, il risque d'aboutir à un échec. Plus de 50% des usagers qui l'ont essayé, n'ont pas obtenu de réponses.
- En outre, ils emploient rarement la tactique ensemble.

Conclusion

Avec un corpus, plus large que celui d'Osmont, et Wildemuth, nous avons mis en évidence l'importance de certaines tactiques pour résoudre le problème des échecs mais surtout

l’incapacité des usagers à affiner leurs recherches. La mise en application des stratégies essai-erreurs peut expliquer le taux élevé de révision. Si l’usager ne trouve pas de documents, il revient en arrière pour explorer une piste avec un autre terme. Nous pouvons avancer, comme OSMONT, que les phénomènes de reprise sont importants dans tous processus de recherche d'information. Elle constate :" *la reprise d'une interrogation semble très rarement pouvoir se prolonger sur plus de trois termes sans la reprise partielle ou totale d'un élément déjà employé*" (Osmont, 1995). Ce qui semble plus intéressant, c'est que la réitération est aussi une tactique souvent utilisée dans les moteurs de recherches (Bruza, 1997). Nous allons donner un support graphique à cette tactique dans notre prototype. L’usager pourra en plus consulter la progression de sa recherche et revenir à une étape précédente facilement.

Si la méthodologie choisie (analyse des traces de milliers de sessions) se montre adéquate pour déterminer l'usage de ces tactiques, elle requiert des méthodes complémentaires (comme les verbalisations et les interviews) pour déterminer les causes du choix (ou du non usage de ces tactiques), les éléments de situation qui influencent le choix de chaque type de tactique. Ce travail laborieux n'a pas été entrepris dans cette thèse, ni ailleurs.

Les résultats de cette étude remet en cause plusieurs travaux antérieurs: il est évident que le problème du "zéro réponses" n'est pas aussi important que cela. Nous avons montré qu'en général (73% pour les usagers de l’ENSSIB et 63% cas pour ceux de Lyon2), les usagers ont essayé une autre alternative de recherche pour surmonter cet échec. La principale raison de cette diminution du taux d'échec est liée à l'introduction des recherches par mots clé et aussi à l'interface graphique. Dans les catalogues de première génération, les erreurs dues à l'écriture de commandes étaient importantes ; de plus, l'usager était obligé de formuler sa requête dans les mêmes termes que ceux de la base. Cette diminution est malheureusement accompagnée d'un taux d'échecs inquiétant, d'autant plus que les usagers ne disposent d'aucune d'aide pour reformuler leurs requêtes. Ce problème est aggravé puisque seulement un usager sur deux tente de réduire le nombre de réponses en ajoutant un terme à l'équation d'origine. Cet aspect de la recherche fut longtemps occulté par celui des échecs. Nous aborderons dans les chapitres suivants des solutions pour le réduire et nous proposerons un prototype expérimental CATHIE (CATalogue Hypertextuel, Interactif, Enrichi) que nous avons développé.

Chapitre huit : La navigation dans les WWW-OPACs

Dans les catalogues traditionnels, les concepteurs présumaient qu'une fois la notice bibliographique trouvée, la recherche était terminée. Or, une notice comporte des données très utiles qui peuvent être des points de départ pour créer des liens avec d'autres notices. La lecture de la notice bibliographique permet de suggérer à l'utilisateur l'emploi de nouveaux termes auxquels il n'avait pas pensé au préalable et de trouver des documents similaires.

Offrir une interface WWW permettra aux usagers de ne plus se limiter à des stratégies de recherche faibles que sont l'essai-erreur ou le feuilletage alphabétique (Chen, 1991). Au contraire, ils peuvent mettre en œuvre certaines stratégies des professionnels, notamment l'instanciation de références connues (c'est à dire permettre de relancer une recherche avec les termes d'indexation ou les clés d'accès d'une référence déjà trouvée). C'est une forme de reformulation itérative dirigée par l'utilisateur final. Nous avons appelé ce processus de sélection de termes BRF: Browsing relevance feedback (Ihadjadene, 1998b)

Après avoir affiché la notice ci-dessous (figure 29) , l'usager peut sélectionner un lien (auteur, sujet ou collection) pour trouver des documents similaires c'est à dire indexé par le même sujet ou écrit par le même auteur.

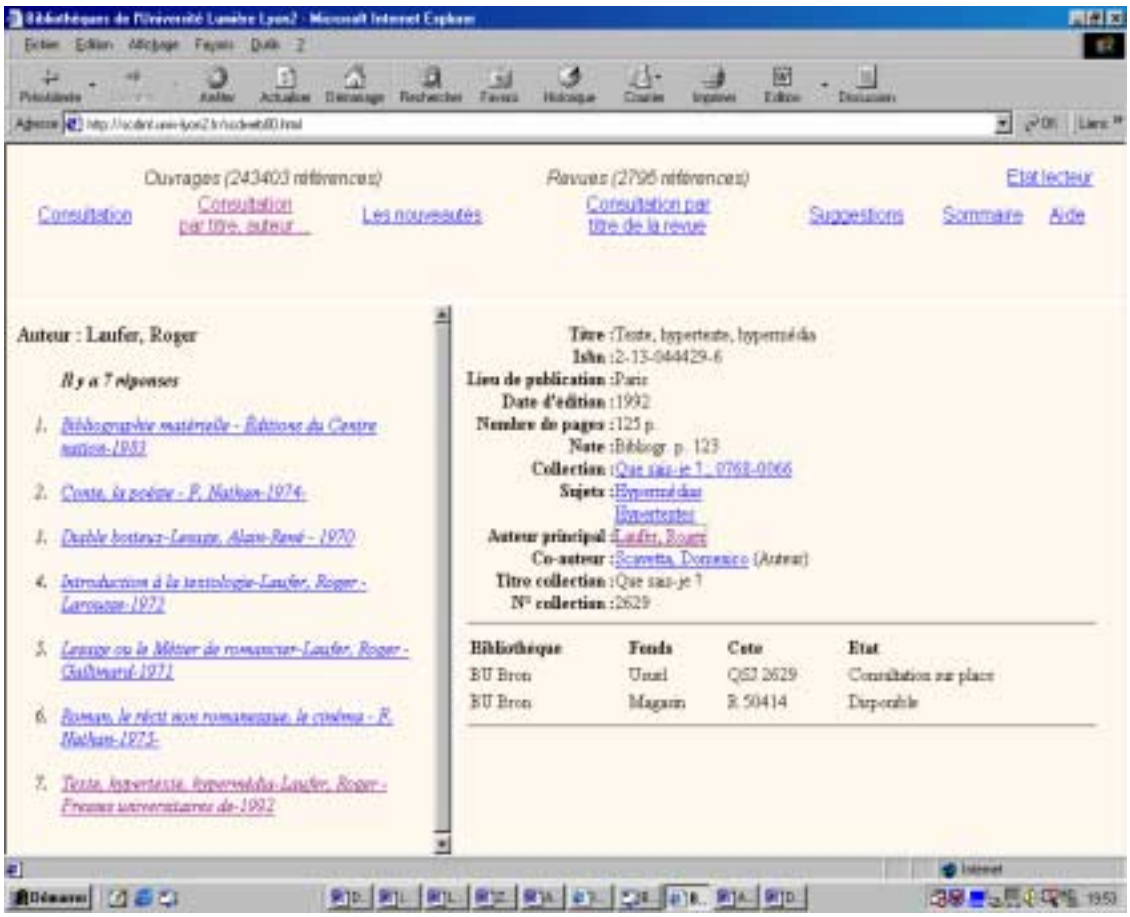


Figure 29 : exemple de navigation par auteur

Par le passé, des chercheurs ont identifié des stratégies de recherche proches de BRF, dans la mesure ou elles exploitaient deux sous - processus de recherche:

- La visualisation des références trouvées.
- L'extraction d'éléments d'informations (descripteurs, mots du titre, auteur, .etc.) pour poursuivre la recherche.

Il s'agit principalement des stratégies telles que : citation pearl growing (Markey, 1978), item instanciation strategy (Chen, 1991), term relevance feedback (Spink, 1997), la reformulation (Walker, 1989) (Meadows, 1992) et (Marchionini, 1995).

- *"the citation pearl growing approach can be viewed in principle as the "manual" equivalent of the relevance feedback search" (*Meadows, 1992)
- *"retrieval by instantiation search strategy is a type of relevance feedback"*

*(*Chen, 1991).

- *"relevance feedback is like browsing in that it depend on feedback as the user explores and probes the system interactively" (Marchionini, 1995).*
- *"the use of a retrieved record for reformulation is much like relevance feedback, but the user does note rate each individual record"* (Meadows, 1992)

Jusqu'à maintenant ces stratégies de recherche (TRF, Pearl growing, reformulations) étaient considérées par plusieurs auteurs comme une stratégie de recherche secondaire, juste utile pour les bibliothécaires et les documentalistes.

Parmi les cinq catégories de bouclage de pertinence définies par (Spink, 1997) et (Saracevic, 1998), c'est le bouclage de pertinence effectué à partir des clés de réponses affichées (TRF: term relevance feedback) qui correspond le mieux à la stratégie BRF. Bien qu'ils ne se réfèrent pas aux anciens travaux de Markey ou de Chen, cette stratégie de recherche (TRF) est la même que celle décrite par Markey, c'est à dire :"pearl growing strategy".

Trois facteurs différencient BRF des autres stratégies de recherches :

- Le mode de sélection des termes s'effectue par la navigation.
- Le choix des termes est limité aux seuls liens proposés par le WWW-OPAC. Dans notre étude, ce sont les liens sujet, collection, collectivité, éditeur, congrès et auteur.
- A la différence du feedback automatique, la recherche BRF s'effectue ici sur un seul terme (soit la vedette matière, soit l'auteur, soit etc.).

La stratégie "the citation pearl growig" suppose en principe que la recherche initiale " ou principale" ait donné des résultats jugés pertinents pour l'usager. C'est en quelque sorte, une recherche planifiée. Ce n'est pas le cas de la stratégie BRF au cours de laquelle l'usager n'a pas besoin de commencer sa recherche par un élément connu pour trouver des documents similaires. Enfin, nous pensons qu'il est nécessaire que l'usager intervienne lui même pour modifier sa requête: dés lors qu'il s'agit de préciser ou d'élargir la requête, le critère qu'il ajoute ou celui qu'il substitue ne peuvent être choisis que par lui. L'usager doit être actif durant tout le processus de recherche.

8.1. Efficacité et usages de ces stratégies de recherches

Toutes les études éffectuées en laboratoire [46] sur l'efficacité de ces trois stratégies (TRF, bouclage de pertinence et la reformulation) montrent qu'elles améliorent de façon significative les performances des recherches. Salton (1983) assure que le bouclage de pertinence enrichie de 50 % le taux de précision. Dans le cas d'une recherche médiatisée, Spink & Saracevic (1997) affirment que les termes extraits des documents affichés (TRF) permettent d'améliorer la réponse dans 70% de cas.

Si les tests expérimentaux, effectués en laboratoire ont tous montré l'efficacité du bouclage de pertinence et de la reformulation, il n'en est pas de même pour leurs usages en situation réelle d'interrogation. Lorsqu'on examine les résultats des études effectuées en situation réelle, on s'aperçoit que cette aide n'est pas toujours sélectionnée. Hancock (1994) a mis à la disposition des usagers d'une bibliothèque universitaire, une version de OKAPI qui offre la reformulation interactive pendant plusieurs mois. L'analyse des traces informatiques a révélé que seuls 11% des usagers a utilisé cette fonctionnalité. Non seulement la reformulation ne fut pas beaucoup employée, mais 70% des usagers ne sont pas satisfaits des réponses obtenues. Pour les moteurs de recherche, la situation est plus grave. Seul 6% des utilisateurs ont exploité le feedback pour le moteur de recherche Excite (Spink & Jansen, 1998).

Type de reformulation	Pourcentage des utilisateurs qui l'ont sélectionné	Efficacité
Interactive (Okapi)	11%	31%
Automatique (Okapi en VT100)	31%	50%
Avec un thesaurus (Okapi)	21%	56%
Médiatisée (TRF)	11%	70%
Médiatisée (thesaurus)	19%	46%
Automatique (moteur de recherche)	6%	nsp

Tableau 32 : Usage du feedback et de la reformulation en situation réelle.

Dans le cas des catalogues classiques, Wallace (1993) a étudié l'usage de la fonction "Express Search" qui permet de retrouver des notices similaires. Elle montre que seul 0,3% du total des

[46] Voir les résultats de TREC (http://www.nist.org/trec)

transactions concernent cette option. Deux raisons expliquent ce faible pourcentage. La première est liée au fait que les usagers n'affichent pas les références complètes, la seconde est que cette option est dissimulée dans l'ensemble des commandes du catalogue.
Dans leurs analyses de plus de 800 sessions d'usagers, Hassoun et Roger (1994) ont constaté l'absence d'utilisation de commandes de renvoi et des possibilités de reformulation. L'instanciation d'une référence connue ne semble jamais être mise en œuvre, même pour les usagers experts. A la suite de leurs études, plusieurs auteurs (Kolmayer, 1997) (Hassoun et Roger, 1994), (Chen, 1991), (Marchionini, 1995) et (Hildreth, 1993) ont observé que l'instanciation d'une référence connue est l'une des principales limites posées à l'usager grand public. Sa méconnaissance est un frein à un usage intelligent et fiable du catalogue car le lecteur ne peut pas tirer profit de l'identification antérieure d'un document portant sur un même thème.

A partir de ces études, on peut faire le constat suivant :

- Bien que efficace, la reformulation (automatique ou non, médiatisée ou non) est une option qui n'est pas souvent sélectionnée par les usagers.
- Dans les catalogues traditionnels, les usagers ne reformulent pas leurs questions.

Les travaux de Zizzi (1996) ou de Khorlfage (1998) sur la visualisation de l'information, ont montré que les stratégies de présentation de l'information conditionnent les performances de recherche des usagers. Le critère de présentation des résultats détermine la facilité avec laquelle un usager peut exploiter les résultats de sa recherche. Nous faisons l'hypothèse qu'avec la clarté des interfaces Web des catalogues, les usagers auront plus de facilité à utiliser la stratégie BRF. L'analyse transactionelle nous offre une opportunité de déterminer si les usagers grand public l'emploient. De plus, nous avons choisi trois corpus différents pour cette analyse.

8.2. Méthodologie et résultats

Nous avons eu recours à trois sortes de données pour examiner l'usage de la navigation dans les WWW-OPACs. Nous avons examiné l'ensemble des traces informatiques des catalogues de l'ENSSIB (année 1997), de LYON2 (1996 et 1997) et enfin celui de l'IRISA (1997)

8.2.1. La navigation : données quantitatives

Nous avons vu précédemment que la navigation concerne plus d'un dixième des accès (tableau 33).

	ENSSIB (N = 31420)	**Lyon2** (N = 33829)	**IRISA** (N = 5274)
Recherche analytique	84, 95%	87, 96%	91, 5%
(BRF)	15, 05%	12, 04 %	9, 5%
Total	100%	100%	100%

Tableau 33 : Utilisation de la stratégie BRF

On peut remarquer qu'il existe une différence significative entre les lecteurs des trois bibliothèques. Ceux de l'ENSSIB ont tendance à utiliser davantage les liens hypertextes que les autres usagers. En outre, on peut remarquer que l'importance des liens varie selon les bibliothèques. Les lecteurs des trois bibliothèques préfèrent naviguer fréquemment en activant les liens sujets et auteurs (Tableau 34). Cela veut dire que, non seulement ils interrogent par sujets mais qu'ils préfèrent aussi naviguer par sujet. Aucune différence n'est à noter dans le choix des liens selon les mois de l'année.

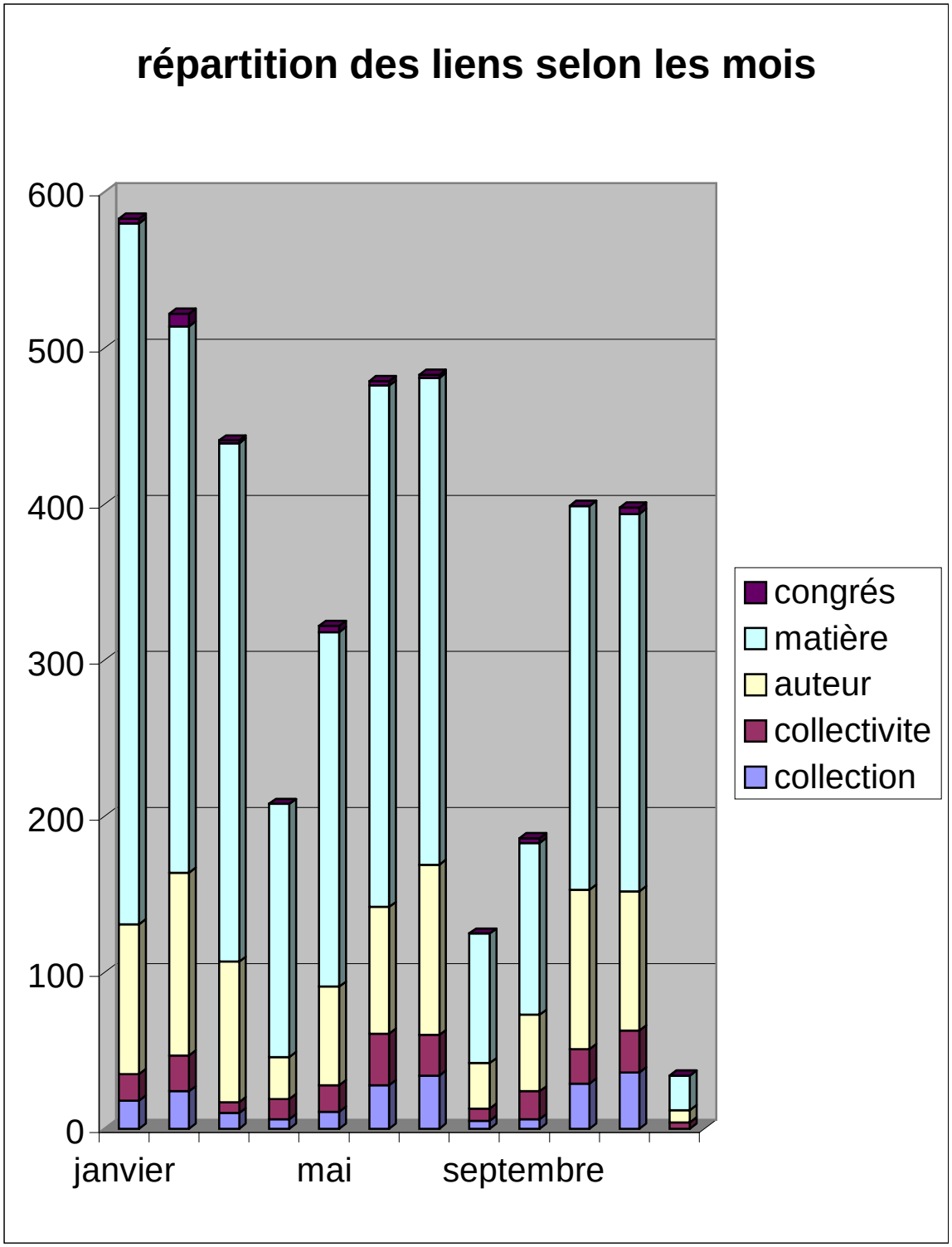


Figure 30 : répartition des liens selon les mois (enssib)

La deuxième remarque porte sur l'usage du feedback dans les SRI traditionnels. Dans toutes les études effectuées sur le bouclage de pertinence, les concepteurs assument que l'usager n'a

besoin que de documents qui ont des descripteurs en commun. Dans son étude sur la reformulation dans un environnement médiatisé, Spink (1997) constate que les bibliothécaires utilisaient souvent les mots du titre (27%), les descripteurs (58%) et les mots extraits du résumé (3%). Nos données montrent l'opportunité d'utiliser d'autres sources de termes pour la reformulation, notamment le champ auteur.

	ENSSIB (4727)	**Lyon2 (4073)**	**IRISA (498)**
Matière	70, 32%	43, 88%	53, 8%
Auteur	19, 65%	30, 98%	23, 9%
Collectivité	4, 89%	5, 14%	10, 8%
Collection	5, 14%	9, 60%	4, 6%
Éditeur	NA	10, 90%	1, 4%
Congrès	0, 7%	Nsp	5, 4%
Total (Liens)	100%	100%	100, 0%

Tableau 34 : Typologie des liens utilisés

8.2.2. Navigation et buts de recherches : analyse qualitatives

Nous avons voulu déterminer quels sont les facteurs qui influencent le choix des liens et quelles sont les raisons qui conduisent l'usager à se contenter de la notice affichée sans voir les autres documents qui y sont liés ? Pour cela, nous avons observé, puis interviewé quarante quatre étudiants en sciences humaines de l'université de Lyon2. Trente d'entre eux ont un niveau DEUG et les quatorze autres évoluent soit en licence, soit en maîtrise. La majorité de ces lecteurs (90%) ne sont pas formés à l'usage des navigateurs, ni à la consultation des catalogues en ligne sur l'Internet. On peut aussi faire les observations suivantes :

- 86% d'entre eux ont l'habitude de fréquenter la bibliothèques plus de deux fois par semaine
- Presque deux tiers d'entre eux utilisent plus de deux fois par semaine le catalogue de la bibliothèque.
- Ils consultent le catalogue en général pour rechercher des ouvrages dans 68% de cas.

- Plus de la moitié (52%) d'entre eux ont connaissent des ouvrages portant sur le même sujet ou le même auteur.
- Dans 75% de cas, ils trouvent ce qu'ils cherchent. Seul 18% d'entre eux se plaignent du problème de surcharge d'information
- Une fois le document trouvé, ils l'empruntent dans 54% de cas et/ou ils le consultent d'abords au rayon pour voir s'il est vraiment pertinent (27% de cas).

En plus de l'observation, nous avons récupéré l'ensemble des traces de leurs consultations.

Environ, un tiers (29, 5%) d'entre eux a employé la stratégie BRF avec une moyenne de 2 liens par session. Ils favorisent en général la navigation par les liens sujets (tableau 35)

Nombre d'usagers	44
Nombre d'usagers qui ont utilisé BRF	13
Liens sujets	20
Liens auteurs	6
Lien collectivité	1
Nombre moyen de lien par session	2.07

Tableau 35 : Analyse des 44 sessions

Les étudiants fournissent quatre raisons principales pour motiver leurs choix. La première est de trouver des ouvrages similaires. La seconde est de préciser leurs besoins, en identifiant un terme plus précis par rapport à leur première requête ou juste pour réduire le nombre de réponses. Seul un usager a élargi sa requête en utilisant le lien sujet. Enfin deux étudiants ont employé les liens pour revenir à la page précédente (liste des notices). Ils ont utilisé les liens hypertextes comme un outil d'aide à la navigation.

Contrairement aux études de Spink (1997), Hancock (1997), Kolmayer (1997) et de Chen (1991) nos résultats indiquent que le pourcentage de termes par la méthode BRF est assez important (33 %). Ceci peut s'expliquer par le mode navigationnel de l'interface Web du catalogue. La seconde différence est liée au choix des termes. Dans (Saracevic & Spink 1998), les sources des termes proviennent successivement de titres ou des descripteurs (les usagers et les intermédiaires n'utilisent pas les résumés pour le choix des termes). Nos résultats indiquent au contraire que les usagers optent pour les sources suivantes:

sujet, auteur mais aussi éditeur, collectivité et le lien congrès. Concernant l'efficacité de BRF, 67% des usagers sont satisfaits des nouvelles réponses obtenues. Ceci corrobore avec les données de Spink (1997).

Voici les principaux facteurs qui peuvent expliquer le non usage des liens .

- Satisfaction devant les résultats obtenus : les réponses fournies par le catalogue correspondent à leurs attentes, ils n'ont pas besoin de voir d'autres notices. "J'ai trouvé ce que je cherche, cela me suffit"
- Localisation : Ils ont noté une ou deux cotes pour poursuivre leurs recherches et naviguer dans la bibliothèque. Ainsi pour cette étudiant en Deug" je fais toujours comme ça, je prends la cote et je flâne dans le rayon".
- Incertitude : Pour eux, la navigation ne leur apporterait pas de nouvelles informations.
    - "Peut être qu'il (cet auteur) n'a pas écrit sur le même thème."
    - "Un auteur peut écrire plusieurs ouvrages sur différents thèmes."
    - "Ca ne m'intéresse pas de savoir ce qu'il a écrit d'autre."
    - "Je ne pense pas trouver de choses intéressantes."
    - "Il va me donner les mêmes réponses, n'est ce pas ?"
- Le manque d'information: Ils ne peuvent pas déterminer si la référence obtenue est pertinente ou non. Cela est dû à deux facteurs, l'absence de tables de matière et la difficulté qu'ils ont eu à comprendre le sens de certaines vedettes (vedettes inversées).
    - "Je ne sais pas si c'est bien."
    - "Il n y pas de mots clé, je ne pourrai pas savoir."
- facteurs contextuels: ils n'ont pas le temps de naviguer "J'ai un examen d'histoire demain, je veux juste un ou deux livres en urgence."
- habitude: ils ne sont pas familiarisés avec le navigateur.
- Disponibilité de l'ouvrage: Ils interrogent le catalogue juste pour voir si le livre est disponible ou non.
- Expérience passée : "Je l'ai déjà essayé, ce n'est pas toujours efficace"

Une étude relative aux souhaits des usagers donne une idée de ce que le public attend ou souhaiter trouver. Pour les usagers de Lyon2, le manque d'index auteur est une source de problèmes; 80% des usagers regrettent son absence. Pour eux c'est un moyen très utile lorsqu'on a déjà une bibliographie. Dans la version précédente de LORIS (sur Windows), au fur et à mesure que l'usager saisit son terme, l'index défile. Ils ne sont pas obligés de saisir le nom en complet. Les usagers expriment principalement une forte demande en matière d'aide à

la formulation (on afficher des termes proches) et du contenu (plus de sujets ou tables de matières).

8.3. La désorientation dans les hypercatalogues

Plusieurs auteurs ont critiqué l'approche hypertexte pour les catalogues en ligne en raison du risque de désorientation inhérent à cette technique, car la navigation peut entraîner l'usager vers une navigation compliquée ou l'éloigner rapidement de son sujet (Hassoun et Roger, 1994) (Khoo, 1997). Contrairement aux premiers prototypes d'hypercatalogues où les aides à la navigation étaient insuffisantes (seule la possibilité de revenir en arrière était disponible), les catalogues sur l'Internet exploitent les différentes aides des navigateurs. Étudier la désorientation des usagers, c'est avant tout, évaluer l'usage de ces aides. Nous avons observé cinquante cinq étudiants de l'enssib en situation réelle d'interrogation pour analyser les problèmes qu'ils rencontrent ainsi que les erreurs commises. Cette étude est réalisée entre le 15 mai 1996 et 10 juin 1996. Les usagers concernés sont des étudiants de troisième cycle (DEA, DCB, DESS, Doctorat) en science de l'information. ils ont tous reçu une formation et acquis une maîtrise des techniques documentaires et de l'Internet. Le navigateur utilisé est Netscape version 2. Les innovations apportées par les nouvelles versions des navigateurs, induisent que certains problèmes soulevés par cette étude ne sont plus d'actualité, notamment la confusion entre les options "précédent" et le "précédent du cadre" du navigateur.

Voici l'ensemble des observations et les remarques sur l'usage de ces aides:

- Près de 70% des usagers n'ont pas suivi de formation à l'usage des aides. C'est en sollicitant une aide auprès des autres utilisateurs qu'ils apprennent le fonctionnement du navigateur.
- Plus de 90% des usagers connaissent le rôle des aides fournies par le navigateur, sauf pour : Find "rechercher" et adresse d'URL en bas.
- La majorité des usagers (90%) ne savaient pas comment revenir à la page précédente : comment utiliser le cadre (Frame back). C'est une fois que nous leur avons montré son fonctionnement que les usagers l'emploient. Bien qu'il existe une brochure et un MOT sur le bouton droit de la souris (retour arrière) personne ne l'a lu. Ils éprouvent tous des difficultés lorsqu'ils essayent cette fonction.

- L’adresse en bas : elle n’est d’aucune utilité dans le cas d’une recherche d’information dans ce catalogue. Voici un exemple d’adresse :

  GET /cgi-bin-ever/DORIS_DOC_WEB_LIVRES ? NOTICES_W3=19931

  Dans le cas d’un site WWW, cette aide sert à indiquer l’adresse URL d’un lien. Dans notre cas l’adresse donnée n’est souvent pas compréhensible. D’ailleurs, les usagers soit ne la connaissent pas (plus de 80% usagers), soit ils ne font pas attention puisqu’elle ne leur sert pas

- Rares sont les usagers qui utilisent les signets (bookmarks). Ils préfèrent prendre des notes. Ceci peut s’expliquer par deux raisons :

  - la première est que dans le cas du catalogue Web de l’ENSSIB, l’adresse du signet est inintelligible (*GET /cgi-bin-ever/DORIS_DOC_WEB_LIVRES ? NOTICES_W3=19931).*

  - La deuxième est que le microordinateur d’où ils interrogent est public, d’où de nombreux signets (difficulté de personnaliser le PC) et la possibilité de les perdre (un autre usager peut les effacer).

- La fonction historique n’est guère utilisée (5%). Les informations qui sont données sont, comme dans les signets, incompréhensibles.

- Cinq usagers ont des problèmes avec l’usage de l’environnement Windows : exemple il clique dans le hors-cadre, l’application courante (ici Doris-Web) disparaît : ils ne savent comment revenir à l’application.

- L’utilisation du bouton "arrêter" est souvent le signe d’une frustration, d’une impatience surtout lorsque les temps de réponses sont importants.

- Au lieu d'utiliser les boutons Frame back pour revenir aux étapes précédantes, 10 % des usagers préfèrent cliquer sur "nouvelle recherche" ou sur un lien BRF et recommencer la recherche.

- L'usage des "liens déjà passés" pose cependant des problèmes

  - Le premier est ergonomique, certains usagers ont des difficultés avec certaines couleurs (le rouge)

- Le second est liée à l'usage d'un PC public. Parfois en affichant une notice (les références) un certain nombre de liens sont déjà en "rouge" et ceci déroute les usagers : ils ne savent plus s'ils ont déjà activé le lien ou non ?

- Comme les usagers dans un PC destiné spécialement pour la recherche de l'information, l'utilisation des URL n'est d'aucune utilité.

Une autre problème majeur de la navigation dans les hypertextes est le phénomène de "noyade en disgression" (Foss, 1989) qui se manifeste pour certains utilisateurs par la nécessite de retourner à des liens déjà consultés. Pour étudier ce problème, nous avons analysé les traces de cinq cents sessions des utilisateurs du catalogue de l'enssib (dont deux cents concernent des usagers qui accèdent à distance). Nous avons constaté que le taux de répétition est de 20 % pour les liens. Nous pensons qu'une part importante des usagers utilise les liens comme un outil d'aide à la navigation. En effet, la difficulté de revenir en arrière dans les interfaces Web basées sur les "cadres" peut expliquer cette réitération. Il est plus facile de revenir en arrière en activant un lien déjà visité que d'utiliser le bouton droit de la souris. Toutefois, cela nécessite plus d'études. Il serait intéressant de voir si ce taux est inférieur dans les catalogues qui n'utilisent pas les cadres (frames). Une analyse plus fine permettrait de mieux cerner les problèmes de désorientation que posent les WWW-OPACs.

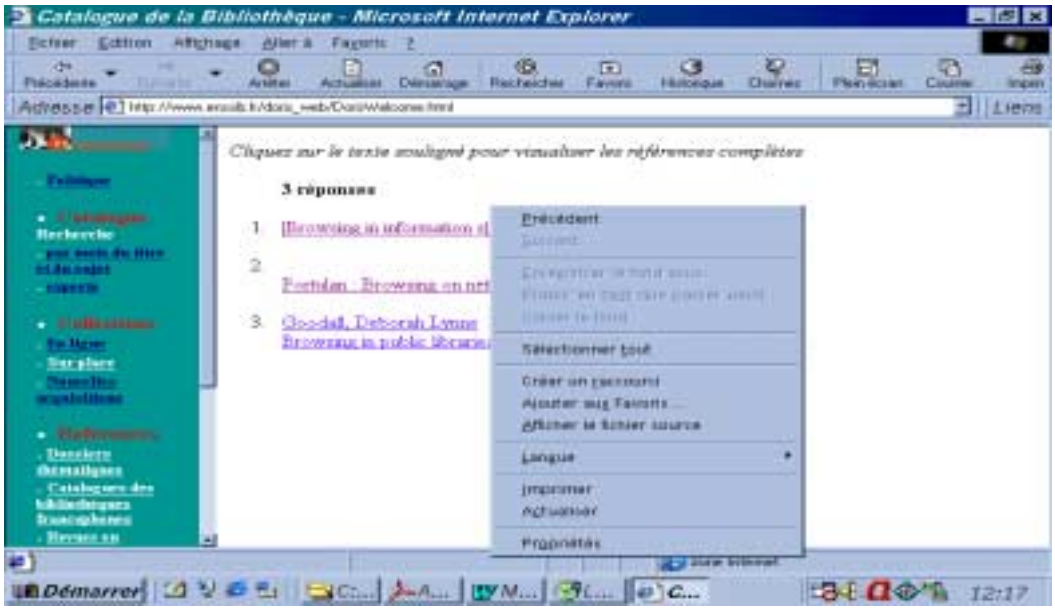


Figure 31 : Les aides à la navigation dans Netscape

Conclusion

Barthes et Glize, cités par Villeme (1994), considèrent la recherche d'information dans les bases de documentaires comme un processus planifié qui se limite à résoudre un problème. Or dans les systèmes interactifs, les usagers grand public ne planifient pas leurs recherches. L'analyse des traces nous a montré que près d'un tiers d'entre eux, modifient sa recherche, non seulement pour avoir des documents similaires mais aussi pour préciser leurs besoins et les redéfinir. La visualisation des réponses joue un rôle important dans les redéfinitions du besoin qu'élaborent les usagers. Elle leur donnent des pistes de recherche, de nouveaux termes qui peuvent être à l'origine d'un changement dans le besoin d'information. Il existe d'autres facteurs qui aident à l'élaboration de ce processus exploratoire de la recherche, notamment l'utilisation des organisateurs paralinguistiques (couleur, typographie, espace, dispositifs dynamiques). Selon, Cori (1996), le rôle des organisateurs dans le domaine des SRI est double. D'une part, ils permettent de distinguer les informations en fonction de leur importance, D'autre part, ils établissent une différence entre les informations selon nature. Ils permettent aussi à l'usager d'identifier facilement les liens et les possibilités de navigation du catalogue. Nous pensons que c'est un des éléments qui expliquent pourquoi les usagers naviguent facilement dans les catalogues en ligne sur l'Internet alors que ce n'est pas le cas dans les catalogues traditionnels.

Nous avons observé que près d'un tiers des utilisateurs passent d'un mode de recherche à un autre. Ceci montre que le processus de recherche n'est pas déterminé à l'avance, mais qu'il est au contraire évolutif en fonction des circonstances de la recherche et des besoins d'information des différents utilisateurs. Le besoin d'information ne reste pas constant au cours de la consultation mais évolue au grès des visualisations. On peut donc émettre l'hypothèse que ce sont les réponses du système qui sont à l'origine de l'évolution de la demande. C'est la base du modèle "graspillage" proposée par Bates (figure 32) :

- Le but et les objectifs des usagers, qui ne sont pas prédéterminés dés le départ de la recherche, changent et évoluent en fonction des réponses du système.
- Le résultat de la recherche n’est pas constitué par une seule notice mais par un ensemble de références.
- La recherche d’information n’est pas linéaire mais plutôt exploratoire et interactive.

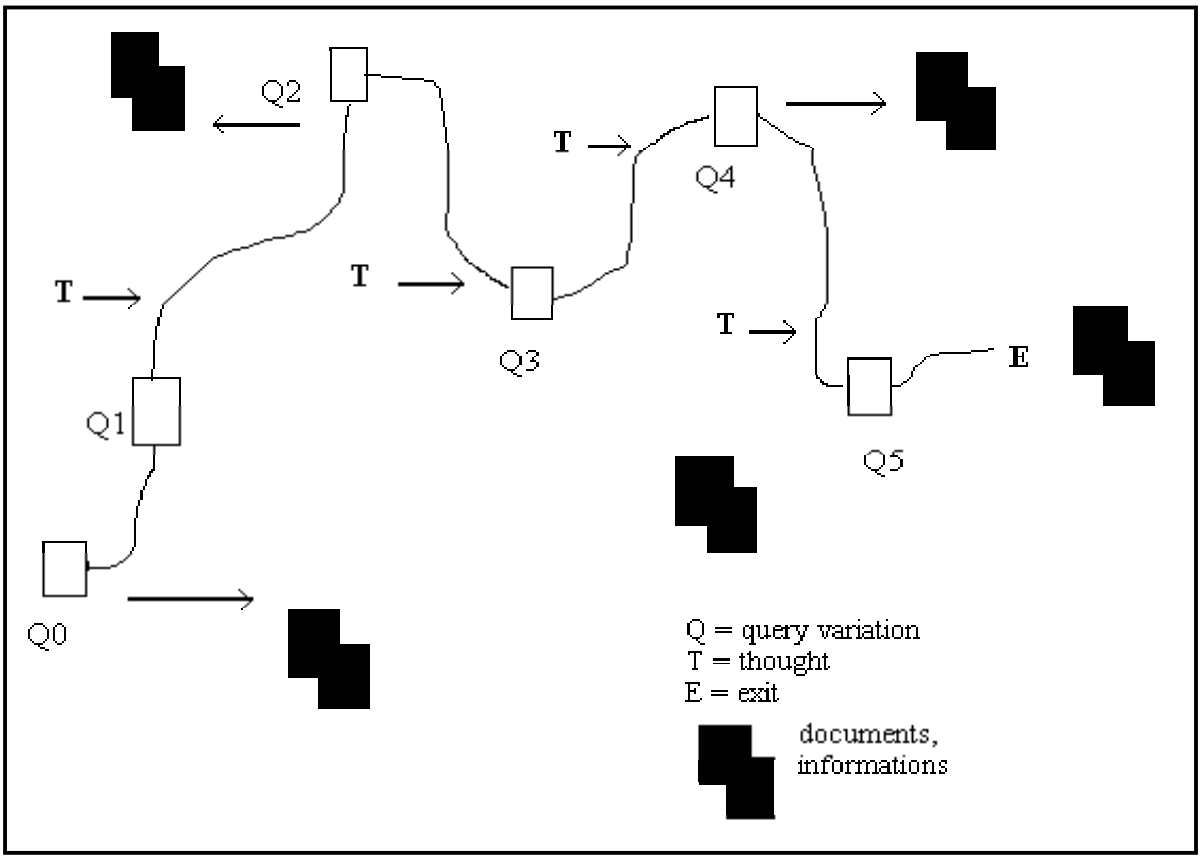


Figure 32. Le modèle de "berrypicking" de Bates

C'est en se référant à ce modèle, que plusieurs auteurs justifient l'importance donnée à l'exploration dans les bases de données. Néanmoins, le modèle de Bates, ne permet pas d'expliquer tout le processus de recherche. Si les réponses jouent un rôle important dans la sélection des termes , il existe cependant d'autres facteurs qui favorisent ce choix comme la maîtrise des vocabulaires contrôlés, la consultation d'autres personnes (bibliothécaires) ou son expérience passée. En outre, une part importante des usagers (70%) continue à privilégier les recherches analytiques au dépens de la navigation.

En résumé, on peut soutenir qu'un usager formule rarement sa requête de manière optimale la première fois et qu'il lui est souvent nécessaire de préciser son besoin d'information. L'usage de la stratégie BRF est une aide à cette reformulation. Nous avons montré qu'avec WWW-OPACs, le pourcentage est passé de 1% à prés de 30%. Ce qui est un saut qualitatif déterminant dans le mode de consultation des catalogues en ligne. Nonobstant, cette contribution est réduite, car elle ne survient pas au début du processus de recherche. Si l'usager ne trouve pas de réponses qui répondent à ses besoins et/ou s'il n'affiche pas la notice complète, il ne pourra pas bénéficier de cette aide. Le prototype CATHIE que nous

présenterons au chapitre suivant permet de contourner cette insuffisance.

Chapitre neuf : L'accès à distance aux WWW-OPACs.

L'accès à distance aux ressources d'informations sur l'Internet devient de plus en plus important. Avec la généralisation de l'enseignement à distance et des possibilités d'autoformation, il devient important d'étudier le comportement des usagers qui accèdent à distance pour mieux cerner leurs besoins d'informations. Snelson (1993) rappelle à juste titre que :

> « *our perceptions of patrons' behaviors is still limited by our experience with patrons within the traditional library context* ».

Toutes les études sur l'accès distant que nous avons consultées portent sur des catalogues classiques (sans interface hypertexte) (Ferl,1992) , (Ferl,1996), (Kalin,1991), (Millsap,1993), (Sloan,1991), (Snelson,1993). Hormis l'étude de (Snelson,1993), rares sont ceux qui ont étudié le processus navigationel des usagers distants. Nous nous intéresserons dans cette étude à la navigation comme un processus de reformulation manuelle et interactive.

Cette étude, exploratoire par nature, a comme objectifs d'analyser les points suivants :

- les caractéristiques de l'accès à distance
- la stratégie BRF est-elle souvent employée par les utilisateurs distants ?
- quels sont les liens invoqués lors de cette stratégie ?
- quel est le nombre moyen de liens invoqués par session ?
- analyse socio-démographique des utilisateurs qui accèdent à distance.

9.1 Méthodologie

Cette étude a été réalisée en deux phases :

- lors de la première phase, l'objectif consiste à récupérer l'ensemble des transactions des interrogations et de la navigation des usagers (internes et à distance) sous forme d'un fichier log. La collecte des données s'est effectuée du 1.12.96 au 31.12.97 .

- lors de la seconde , nous avons traité 300 sessions de recherche sur une période de 45 jours. Il était proposé aux utilisateurs de répondre à un questionnaire électronique (voir annexe 5). Après analyse, seules 200 sessions d'utilisateurs à distance ont été retenues.

9.2 *Analyse des résultats*

9.2.1 Analyse des transactions (fichier log)

Elle nous a permis de dégager les points suivants :

- L'accès à distance représente 42.9 % des accès.
- Les différences entre les modes de recherche des deux catégories d'utilisateurs ne sont pas significatives (tableau 36).

	Accès en local	Accès à distance
Simple	37.2%	44%
Auteur	17.5%	14.5%
BRF (navigation)	14%	16.4%
Titre	10.5%	6.6%
Recherche multicritères	9.9%	8.9%
Mots clé	8.9%	8.7%
Côte	1.8%	0.4%
Date	0.14%	0.4%

Tableau 36 : répartition des modes d'accès(local vs distant)

La recherche simple (accès par mots du sujet et par mots du titre) est le mode qui prédomine dans les deux catégories d'utilisateurs. Nous avons par la suite imprimé 4000 transactions et interrogé le catalogue avec les mêmes termes des usagers. Nous avons noté que les usagers internes emploient davantage d'expressions que les usagers à distance (1.8 vs 1.6) ; leurs requêtes aboutissent à un taux d'échec moins élevé que ceux de l'extérieur (19.36% vs 36.24%). Par contre, le taux de surcharge (lorsqu'il y a plus de 60 réponses) est relativement supérieur (20.31 % vs 18.6%).

Le taux d'échec s'explique essentiellement par les deux raisons suivantes (tableau 37) :

- la bibliothèque ne possède pas d’ouvrages sur le thème recherché (30.3% vs 37.8%). Contrairement aux usagers locaux, ceux de l’extérieur emploient des termes qui ne correspondent pas au degré de couverture thématique du catalogue.
- les usagers effectuent des erreurs typographiques et orthographiques et leurs requêtes sont trop précises

	Accès en local	Accès à distance
La bibliothèque ne possède pas d'ouvrage sur le thème	30,30%	37,80%
Erreurs typographique et orthographique	37,70%	28,90%
Se tromper de clés d'accès	9,45%	13,37%
Requêtes trop précises	12,90%	10,70%
Emploi de termes d'une langue étrangère	4,50%	5,40%
Autres types d'erreurs	5,15%	3,83%
Total	100,00%	100,00%

Tableau 37 : Raisons des échecs lors d'un accès à distance

- La stratégie BRF est très employée par les deux catégories d'usagers (14.05% vs 16.37%). Lorsqu'ils naviguent, ils ont tendance à activer les liens sujet et auteur (tableau 38).

	Accès en local	Accès à distance
Lien sujet	72.98%	67.29%
Lien auteur	18.06%	21.48%
Lien collectivité	4.76%	5.03%
Lien collection	4.21%	6.21%
Total	100%	100%

Tableau 38: liens utilisés dans la stratégie BRF

9.2.2. Analyses des sessions de recherche et du questionnaire

Cet examen ne concerne que les usagers qui accèdent à distance et les réponses font apparaître une diversité d'éléments :

- 56.34% de la population étudiée sont des femmes et 43.66 % sont des hommes
- Les étudiants représentent 41.5% de notre corpus (tableau 39) .

Etudiants	41.5%
Bibliothécaires/ Documentalistes	31.5%
Enseignants	12%
Autres	12%
Sans réponse	3%

Tableau 39: répartition par activité

- presque 31.5% des usagers sont des spécialistes du traitement de l'information (bibliothécaires ou documentalistes) et 12 % sont des enseignants. Enfin 12 % des usagers appartiennent à des catégories socioprofessionnelles très variées (retraité, chômeur, technicien, magasinier,...etc.).

- la majorité des étudiants ont un niveau licence/maîtrise de troisième cycle (tableau 40).

Licence/maîtrise	40.38%
3°cycle (dess, dea, doctorat)	36.54%
Deug	23.08%

Tableau 40 : répartition par niveau

- Plus d'un quart d'entre eux , sont issus de disciplines scientifiques (tableau 41)

Sciences humaines	36.14%
Sciences de l'information	34.94
Sciences exactes, technologie	26.51%
Autres (économie, médecine...)	2.41%

Tableau 41 : répartition par discipline

- En ce qui concerne la fréquence d'usage du catalogue, nous observons que la majorité (68.25 %) des usagers l'utilise pour la première fois, 9.5% d'entre eux l'utilisent une à deux fois par semaine et 9.5 l'utilisent une à deux fois par mois.
- Les bibliothécaires et les enseignants effectuent des recherches dans le catalogue pour un usage professionnel alors que les étudiants s'en servent en général pour leurs études.

- La majorité des usagers (65.8%) a l'habitude de naviguer dans un WWW-OPAC. Pour 8.73% d'entre eux, c'est la première fois.

Pour analyser les stratégies de navigation, (Snelson,1993) a identifié quatre variables:

- *Scrol : moving forward or backward alphabetically through an index file*
- *Select: specification of which retrieved records will be scanned or examined*
- *Display: asking for further bibliographic information for a record*
- *Refine: combining sets with boolean operators, limiting results by date or language.*

Elle trouve que le comportement des usagers qui accèdent le catalogue à distance n'est pas différent de ceux qui le consultent en local.

Pour notre part, nous avons étudié la navigation comme un processus de reformulation. L'analyse des 200 sessions indique que 83 usagers soit 37.89% du corpus emploient la stratégie BRF. Lorsqu'ils naviguent, ils activent en moyenne 2.42 liens par session. Sur les 158 sessions ou il y a affichage, les utilisateurs visualisent en moyenne 4.04 notices bibliographiques. L'analyse manuelle des transactions des 83 sessions montre qu'environ 58.6 % des usagers passent d'un mode de recherche (ou de navigation) à un autre durant une même session. Le parcours le plus fréquent est le passage d'une recherche simple à la navigation par sujet. Nous avons observé qu'une partie importante des 83 utilisateurs effectue plusieurs fois la même opération, soit en visualisant une référence plus d'une fois (32.9 % de cas), soit en activant le même lien (20.8 % de cas).

Conclusion

Le peu d'études qui existent sur le comportement des usagers qui consultent le catalogue à distance rend difficile la comparaison de leurs résultats. (Sloan, 1991) trouve que les usagers effectuent plus d'erreurs que les usagers internes alors que (Kalin, 1991) soutient le contraire. En ce qui concerne les modes d'accès, les résultats sont contradictoires. Les modes d'accès les plus populaires sont l'accès par titre (62.2%) pour (Millsap,1993), l'accès par auteur (38.9%) pour (Kalin,1991). Le taux d'échecs varie entre 37.4 % pour (Kalin,1991) et 57 % pour (Millsap,1993). Les erreurs typographiques et orthographiques représentent la cause essentielle des échecs pour (Millsap,1993). Pour (Kalin,1991), c'est plutôt l'absence d'ouvrages sur le thème recherché. Elle constate que les usagers internes effectuent plus d'erreurs que ceux de l'extérieur. Dans notre étude, la recherche simple est le mode d'accès le

plus utilisé par les deux catégories d'usagers. Le faible pourcentage des échecs (zéro réponse) est dû principalement à la possibilité de recherche combinée dans les champs titre et sujet.

Nous avons observé que les utilisateurs qui consultent le catalogue en local tendent à effectuer plus d'erreurs orthographiques que ceux de l'extérieur. Ceci est conformité avec les résultats de (Kalin,1991).

Dans le cas de l'enssib, il n'existe pas une grande différence entres les pratiques des usagers qui consultent le catalogue en local et ceux de l'extérieur. Toutefois , il ne faut pas généraliser ces observations à d'autres bibliothèques. Nous pensons qu'il serait nécessaire de faire des études similaires sur le catalogue de l'enssib (et ailleurs) sur plusieurs années. Cette dimension temporelle de l'usage des catalogues à distance, ouvre la voie à la prise en compte des phénomènes générationels dans l'acculturation techniques.

Partie III : Conception et développement

Chapitre dix : Conception et réalisation d'un catalogue de troisième génération : le prototype CATHIE.

CATHIE (CAT Hypertextuel, Interactif, Enrichi) est un prototype de catalogue de troisième génération, développé avec Microsoft Visual-Basic entreprise version 6. Pour le tester, nous avons importé l'ensemble des notices du catalogue Loris de l'ENSSIB.

10.1. La conception de CATHIE

Trois principes généraux nous ont guidé dans la conception de CATHIE. Le premier, sans doute le plus important, est que la recherche d'information suit le principe du moindre effort. Cela veut dire que contrairement aux professionnels de l'information, les usagers ne font pas de recherches exhaustives. Ils n'emploient pas toutes les capacités du logiciel documentaire.

La deuxième règle que nous avons suivie consiste à faciliter les processus de reconnaissance et d'identification de l'information et de minimiser toute demande de spécification de l'information. En effet, il est plus facile aux usagers de découvrir ou de repérer quelques chose qui peut les intéresser, que de produire des descriptions formelles. Après plusieurs études sur les usages des catalogues en ligne de deuxième génération, Yee (1996) recommandait fortement pour les concepteurs des systèmes de recherches d'information d'adopter ce principe :

> *" one job of the catalog is to facilite recognition on the part of the user rather demanding exact specification"*.

Ceci a pour conséquence de concevoir une interface où la navigation est le mode principal pour feuilleter des listes, pour le filtrage des réponses, pour affiner les recherches ou pour trouver des documents similaires. L'approche que nous avons développée dans CATHIE reproduit ainsi, intuitivement un des processus cognitifs du raisonnement humain: l'association d'idées. Enfin, nous considérons que les usagers sont les "seuls" qui puissent décider quels sont les ouvrages qui répondent à leurs besoins en informations. Par conséquent, il est impératif de développer des interfaces qui proposent des choix et c'est à l'usager d'opter pour la meilleur stratégie convenant à ses besoins.

En plus de ces trois principes, nous avons montré, dans les chapitres cinq et six, que le problème de "zéro réponse" n'est plus aussi important que dans le passé. Il le deviendrait de

moins en moins avec l'enrichissement des tables de matières et la généralisation des accès par mot (sur un champ ou dans toute la notice). Par contre, le problème de surcharge d'information demeure et nous pensons qu'il s'intensifierait de plus en plus avec l'interconnexion des catalogues entre eux. Nous privilégions donc dans notre conception les solutions qui permettraient de réduire cet épineux problème. Au chapitre quatre, l'analyse des interactions entre les bibliothécaires et les usagers, nous a montré que deux phases du dialogue sont capitales dans la recherche d'information: celle de la sélection des termes et l'exploitation des résultats pour continuer la recherche. Il ne s'agit donc pas d'adapter toutes les dernières "trouvailles" issues de la recherche en informatique documentaire ou des solutions sophistiquées mais plutôt des techniques qui permettent d'améliorer le dialogue entre l'usager et le catalogue en ligne, en l'aidant à définir son sujet, à comprendre les réponses du système et en facilitant la navigation dans le catalogue.

Pour mettre en application ces trois principes, nous avons utilisé ces quatre techniques :

- Enrichir le vocabulaire d'entrée.
- La navigation.
- Le filtrage d'information.
- La mise en œuvre de deux stratégies de recherche.

10.1.1 Enrichir le vocabulaire d' entrée.

Les listes d'autorités font souvent l'objet de critiques. On leur reproche un vocabulaire désuet et une syntaxe incompréhensible pour un non bibliothécaire. Avant d'interroger le catalogue, l'usager doit déterminer quels sont les mots retenus comme vedettes matières ainsi que la syntaxe de ces vedettes. Il doit effectuer la fonction inverse de celle du bibliothécaire sans avoir ni le manuel de l'indexation ni de formation au catalogage. On s'étonne par la suite de trouver plus de 50% d'échecs dans les études d'usages. L'accès par mot clé permet à l'usager d'éviter cette difficulté et d'avoir plus de chances dans l'appariement entre ses termes et ceux de la liste RAMEAU. Bates (1986) considère que le catalogage matière repose sur le principe que les usagers peuvent penser à un petit nombre de termes sélectionnés et les utiliser. Le postulat fondamental du catalogue matière affirme qu'il est possible en trouvant une ou deux vedettes matières pour définir un sujet, de permettre un bon accès à n'importe quel domaine. Toutes les études sur l'usage des catalogues en ligne (de première et de deuxième génération) en situation réelle ont montré que ce postulat est insoutenable. Bates montre que la probabilité moyenne pour que deux usagers utilisent un même terme pour désigner un concept

se répartit entre 10 et 20%. Cette grande diversité s'oppose au principe de choix d'une ou deux vedettes matières pour décrire un ouvrage. Nous pensons comme Bates , que pour résoudre ce problème, il serait important de distinguer l'analyse du sujet (qui peut être maintenue) de l'accès par sujet. Le vocabulaire d'entrée des listes d'autorité (RAMEAU et LCSH) est encore très restreint par rapport à la richesse sémantique des termes qui sont quotidiennement entrés dans les catalogues. Il faudra donc enrichir le vocabulaire d'entrée aux catalogues en concevant des interfaces qui dirigent l'usager vers les vedettes à partir des termes de la requête de l'usager. Cela revient à développer fortement les relations de synonymie et de pseudo-synonymie. Pour enrichir le vocabulaire d'entrée , il faut spécifier quelles sont les sources de ces termes et quel est le moyen le plus adéquat pour établir ces liens. Il existe trois sources possibles:

- les termes du vocabulaire contrôlé (thesaurus, liste d'autorité, classification).
- les mots entrés par les usagers et qui ne donnent pas de résultas (d'après le fichier des transactions).
- les termes que l'on peut extraire des sources suivantes (titre, tables de matières, résumé...etc.).

Markey (1990) est sans doute la première à explorer cette approche; elle a ajouté les termes extraits des classes de la DEWEY comme source supplémentaire à l'index du catalogue. Micco (1991) et Khoo (1998) ont opté pour des techniques purement statistiques. Le prototype ILSA de Micco intègre l'hypertexte comme une interface frontale aux interrogations en langage naturel: Les mots clés extraits des tables de matières sont liés (via les liens hypertextes) aux vedettes matières de LCSH. Pour 48.000 notices enrichies par l'incorporation de termes des tables de matières, Micco a construit un réseau de plus de un million de liens. Khoo (1998) présente une autre approche. Pour chaque mot d'un titre, les auteurs recherchent les titres contenant ce mot et extraient toutes les vedettes matières indexant ces documents (titres). Enfin, Buckland[47] (1999) conjugue les techniques statistiques et des traitements linguistiques.

Pour Buckland : *« Searching is likely to be effective and efficient only if the searcher is familiar with the terms used in the classification, categorizing, and indexing schemes (metadata vocabularies) being searched. Therefore, it is obviously beneficial to provide a*

[47] Le project « Unfamiliar Metadata Vocabularies" du Professeur Buckland est disponible à l'adresse suivante : http://www.sims.berkeley.edu/research/metadata

mapping between the user's ordinary language and the metadata vocabularies of the unfamiliar database in order to diminish any lack of familiarity. An "Entry Vocabulary Modules" provides associations between the user's ordinary language to domain-specific technical metadata vocabulary with which the user would begin a search. The process of creating an entry vocabulary module is one of Bayesian inference, wherein sufficient training data (consisting of document texts) are downloaded (using the Z39.50 protocol for efficiency) from a document database to provide a probabilistic matching between ordinary language terms and the specific metadata vocabulary which have been used to organize the data. Developing the entry vocabulary utilizes both natural language processing modules as well as statistical techniques to identify noun phrases (e.g. `color laser printer') and individual words to map to specialized classification".

Notre approche plus simple, a l'avantage d'être économique. L'usager peut poser sa requête en un langage pseudo-naturel. Le système effectue la recherche aussi bien sur les mots du titre que sur les mots du sujet augmentant ainsi les chances d'appariement avec le contenu de l'index. Le système propose alors comme réponses un ensemble de vedettes matières sous forme de liens hypertextes. Nous n'utilisons aucune base de connaissance, ni aucun traitement linguistique. L'inconvénient de cette solution est que le risque d'ambiguïté est plus élevé par rapport à la solution proposée par Buckland. En effet, pour un terme comme « java », CATHIE ne différencie pas si le terme relève de l'histoire (Indonésie) ou de l'informatique. Pour cela, nous avons ajouté un autre processus : le filtrage de l'information.

10.1.2. Filtrage de l'information

Il reste encore un grand effort à faire pour trouver des moyens efficaces pour l'organisation des sources d'informations. L'information peut :

- soit être organisée au moment où on la saisit dans le SRI : c'est l'approche métadonnées ou du catalogage, ce qui peut simplifier sa récupération ultérieure.
- soit elle est organisée à la sortie et c'est l'usager qui a la tâche de trier cette masse d'information: c'est l'approche "filtrage d'information".

Pour Fondin (1999) : « *le principal problème documentaire qui se pose n'est pas réellement celui de la procédure ou la méthode utilisée par l'usager mais bien davantage, celui posé, une fois encore, par le choix et le façonnage préalables - à faire ou non- des éléments informationnels qui seront utilisés pour retrouver les objets mis en mémoire* ».

Victorri (1999) présente un autre regard sur le futur de l'informatique documentaire. Il considère que: « *l'informatique documentaire doit renoncer à intervenir en amont, du côté du fournisseur, et elle doit plutôt se centrer sur le seul pôle qui garde un minimum de stabilité, à savoir l'utilisateur … Ainsi, ce serait en aval, à partir des besoins bien identifiés de groupes d'utilisateurs, que pourrait agir l'informatique documentaire, en offrant à chacun une grille de lecture appropriée de la masse foisonnante et exubérante de documents qui peuplent l'univers documentaire* ». Cela revient à développer largement des outils de filtrage.

Nous ne pensons pas qu'il faut opposer ces deux approches. La structuration de l'information (catalogage ou métadonnée) n'est pas suffisante, il faut donc inclure des possibilités de filtrage d'information. De même que la qualité des techniques linguistiques ou automatiques adoptées pour le filtrage ne suffisent pas et gagneraient à être enrichies par une meilleure structuration de l'information, d'où l'engouement actuel pour XML.

Jusqu'à maintenant les bibliothécaires ont privilégié la première approche. Nous suggérons d'inclure les fonctions de filtrage dans les système de recherche d'information bibliographiques destinées au grand public. La solution que nous proposons permet d'inclure les outils de classification dans ce processus de filtrage. Avec l'interconnexion des SRI, on ne peut plus se satisfaire des réponses du système. Dans ce cas, il faudrait développer des outils facilitant le filtrage et/ou former les usagers à mettre en œuvre ce processus.

10.1.3. La navigation dans la base documentaire

Le principal inconvénient dans une recherche d'information réside dans la formulation d'une requête. En effet, les performances attendues sont étroitement liées à la concordance entre le vocabulaire utilisé pour l'indexation et celui employé par l'usager. Lorsqu'ils ne parviennent pas à décrire ou préciser leur besoin d'information, le catalogue doit leur fournir des indications de recherches. L'usager a besoin d'aide dans sa démarche d'exploration. Faciliter le choix des vedettes grâce à des aides sémantiques, faciliter l'orientation contextuelle constitue pour les catalogues un objectif vital du fait de son caractère encyclopédique.

Nous pensons que l'ajustement aux besoins de l'usager peut être réalisé en les laissant explorer à leur guise:

- Un espace de concepts.
- l'ensemble des documents retrouvés

Ces deux aides (navigation dans l'espace des concepts, navigation dans la base des documents) permettent à l'usager d'organiser sa recherche et d'adapter la meilleure stratégie pour trouver les termes des questions et les documents qui répondent au mieux à ses besoins.
Les deux aides sont intégrées dans CATHIE d'une façon transparente. L'usager peut passer facilement d'un niveau à un autre en pointant sur les différents liens hypertextes. Jusqu'à maintenant, les SRI séparent souvent ces deux aspects en présentant les vocabulaires contrôlés dans un module et la base documentaire dans un autre. Nous pensons qu'il est nécessaire d'établir des liaisons entre ces deux modules. L'architecture que nous avons adoptée rend claire la navigation dans les deux espaces.
A l'heure actuelle, le seul mode de navigation des termes qui existe dans les catalogues est l'affichage alphabétique. Ces listes alphabétiques sont rarement employées dans une recherche sujet et, lorsqu'elles sont utilisées, une majorité des usagers n'assimilent pas son fonctionnement ; plus particulièrement ils ne comprennent pas l'ordre alphabétique des sous vedettes (Kolmayer, 1997) dont le but est de permettre à l'usager d'identifier la forme correcte d'un terme dans la liste des termes. Cette visualisation alphabétique des termes est d'une grande importance dans le cas d'une recherche par auteur, mais son efficacité est restreinte pour une recherche par sujet. Markey (1994) a pu établir qu'un très petit nombre des termes entrés par les usagers correspondent au vocabulaire contrôlé. Elle remarque qu'il est très important d'assister l'usager dans son choix des termes, et que les listes alphabétiques des vedettes matières ne font qu'effleurer le problème. Elle préconise un affichage conceptuel des termes sans indiquer en quoi cela consiste. Nous avons vu qu'actuellement, seuls deux catalogues offrent à l'usager la possibilité de naviguer d'un concept à un autre d'une façon non linéaire. Ce sont les bibliothèques de Fresnes et celle de la BNF.

10.1.4. Stratégies de recherches

Les usagers ont besoin d'être guidés dans la formulation de leurs recherches, mais aussi dans l'élaboration des stratégies de recherche. Actuellement, les catalogues laissent aux usagers le soin d'élargir et d'affiner leurs recherches sans aucune assistance. La majorité des usagers n'a pas les moyens d'accomplir cette tâche. L'analyse des tactiques des usagers de l'ENSSIB et de Lyon 2 nous a montré que leurs tactiques sont inefficaces lorsque le nombre d'information affiché est trop important.

Nous avons vu que lors d'une recherche médiatisée, la définition et le choix des termes à soumettre au logiciel constitue l'une des étapes principales dans une recherche d'information.

Elle dépend des capacités d'expression de chaque usager. De plus, nous avons constaté que le rôle du bibliothécaire consistait à mieux élucider ce besoin, à aider l'usager à mieux cerner ses termes et parfois à lui proposer des pistes (termes) de recherche. Toute cette "négociation" du choix des termes est absente dans la majorité des SRI grand public. Au lieu d'automatiser ce processus par le biais d'une expansion de requête, nous avons opté pour une approche interactive. L'usager aura ainsi une meilleure représentation de la façon dont sont structurées les informations du catalogue.

10.2. Description du prototype CATHIE.

10.2.1. La reformulation dans CATHIE

Dans son étude sur les techniques d'expansion de requêtes, Efthimidiadis (1996) distingue trois niveaux qui nous donnent les possibilités de les différencier entre elles. Le premier niveau, concerne la source des termes utilisés dans la reformulation. Proviennent- ils des résultats de recherches ou d'une base de connaissance (thésaurus, classification, réseau sémantique, etc.) ?. Le deuxième facteur concerne le choix de la méthode ou de l'algorithme qui permet de selectionner des termes. Enfin, le dernier élément est relatif au rôle actif ou passif de l'usager dans le processus de sélection des termes.

10.2.1.1. Typologie des reformulation

On peut donc considérer qu'il y a trois façons de reformuler une question. La première est manuelle, la seconde est automatique et la dernière est intéractive.

10.2.1.1. 1. La reformulation manuelle

Ce type de reformulation est surtout associé aux systèmes de recherche booléens. On peut procéder à la reformulation de la requête, en utilisant un vocabulaire contrôlé (RAMEAU, thesaurus ou classification). Cet outil permet à l'utilisateur de trouver les bons termes pour compléter sa requête. Il s'agit principalement de mettre en œuvre une ou plusieurs tactiques associées aux différentes stratégies de recherche que nous avons examinées au chapitre quatre (building block, successive search strategy, etc.). L'approche système expert incorpore substantiellement ces tactiques avec plus ou moins de succès. C'était l'approche dominante en informatique documentaire durant les années 80.

10.2.1.1. 2. La reformulation automatique

Lorsque le feedback de pertinence s'accompagne d'une adjonction (et/ou de suppression) de termes, on parle de reformulation automatique. La requête de l'usager est remaniée automatiquement pour intégrer les descripteurs des documents jugés pertinents ou rejetés. On trouve plusieurs variantes de cette technique; celles qui sont basées sur le modèle vectoriel, les réseaux de neurones, les algorithmes génétique ou probabilistes. (Efdimiadis, 1997) (Salton, 1983) (Korfhage, 1998). Dans les systèmes statistiques, le feedback est effectuée de cette façon : les résultats d'une première recherche sont utilisés automatiquement pour reformuler la requête en augmentant les poids des termes de la requête présents dans les documents jugée pertinents par l'usager, et à l'inverse, en diminuant les poids des termes qui sont également présents dans les documents non pertinents restitués. La modification de la requête peut s'accompagner d'une extension de la requête en ajoutant les termes caractérisant les documents pertinents. Le problème avec la reformulation automatique est d'estimer les bons termes qui peuvent effectivement améliorer le processus de recherche, car l'introduction des termes inappropriés peut entraîner un silence, ou au contraire une augmentation du bruit.
Les systèmes SPIRIT, SMART utilisent des techniques de reformulation de ce type dans leurs processus de recherche.

10.2.1.1. 3. La reformulation interactive

Dans une reformulation interactive, l'usager joue un role actif. A l'inverse de la méthode automatique; ici ce sont le systeme et l'usager qui sont responsables dans la détemination et le choix des termes candidats à la reformulation. Le système joue un grand rôle dans la suggestion des termes; le calcul des poids des termes; et l'affichage à l'écran de la liste ordonnée de ces termes. L'usager examine cette liste et décide du choix des termes à ajouter dans la requête. C'est donc l'usager final qui prend la décision ultime dans la selection des termes.

Les techniques de reformulations interactives peuvent être aussi subdivisées selon que les termes sont extraits des résultats de recherches, ou d'une base de connaissances. Dans ce dernier cas, on peut encore distinguer le type de reformulation selon que la base de connaissance soit liée ou non à un corpus (collection) (figure 33)

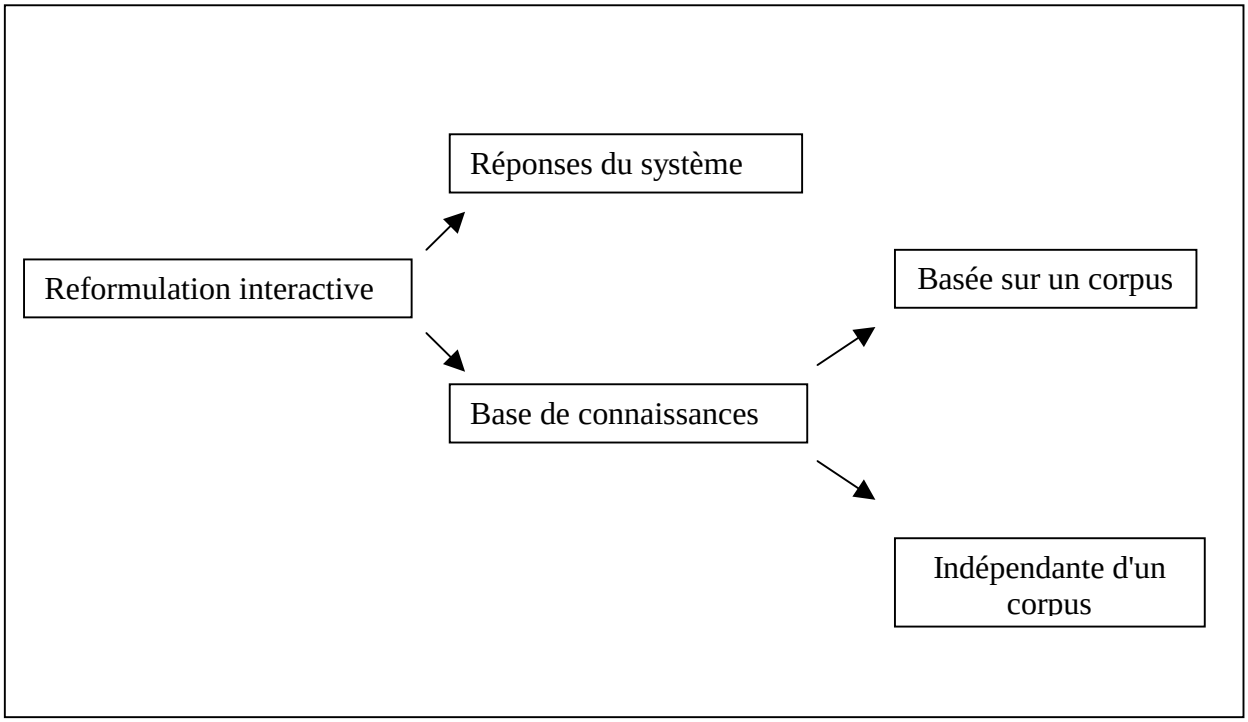


Figure 33 : Les différentes types de reformulation

Parmi les SRI qui utilisent cette technique; on peut citer CITE; CANSEARCH; MenUSE mais surtout les serveurs opérationnels tels que ZOOM (ESA/IRS), IT (fonction EXPLOR), ORBIT (GET), MEMSORT (QUESTEL) ; RANK (DIALOG) . La fonction SUMMARIZE du prototype OASIS (Buckland, 1992) associe par exemple à un mot une liste ordonnée de vedettes matières proches, afin de donner à l'usager une idée de mots réellement utilisés pour indexer la base documantaire.

10.2.1.2. Algorithmes utilisés pour le feedback et l'expansion de requêtes

Efthimidiadis (1993) a évalué la performance de six algorithmes probabilistes de classements des termes selon le point de vue de l'usager. Il montre que les six algorithmes produisent des résultats similaires avec une légère préférence pour les termes produits par les algorithmes EMIM et w (p-q). Voici les principaux algorithmes étudiés par Efthimidiadis (1993). Ce sont ceux qui sont à la base de plusieurs prototypes de recherche.

- L'algorithme de Porter

  Poids (t) = $\frac{r}{R} - \frac{n}{N}$

- ZOOM **:** calcul de fréquences
- F4
  - P (t) = $\log \frac{(r+c)(N-n-R+r+1-c)}{(n-r+c)(R-r+1-c)}$
- F4 point 5
  - P (t) = $\log \frac{(r+c)(N-n-R+r+1-c)}{(n-r+c)(R-r+1-c)}$ avec c=0, 5
- W (p-q).
  - P(t) = $\left(\log \frac{(r+0.5)(N-n-R+r+0.5}{(n-r+0.5)(R-r+0.5)} (\frac{r}{R} - \frac{n-r}{N-R} \right)$
- EMIM: (expected mutual information measure)

P (t) = $\left(\log \frac{rN}{Rn}.r - \log \frac{(n-r)N}{(N-R)n}.(n-r) - \log \frac{(R-r)N}{(N-n)R}.(R-r) + \log \frac{(N-n-R+r)N}{(N-n)(N-R)}.(N-n-R+r) \right)$

Avec :

N = nombre de documents dans la collection ;

n= fréquence du terme t ;

R = nombre de documents pertinents ;

r =nombre de documents indexés par le terme t ;

P (t) : le poids du terme t ;

Q : nombre de mots dans la question.

L'intérêt des travaux de (Efthimiadis, 1993) est de montrer que :

- ❖ L'algorithme de PORTER fournit le même ensemble de termes que celui de ZOOM.
- ❖ Bien que moins complexe que les cinq autres algorithmes, la fonction ZOOM permet d'obtenir des résultats satisfaisants. Nous allons donc étudier ces fonctions de tri, puisqu'elles vont servir de base à l'algorithme de reformulation de CATHIE. L'intérêt du

calcul de fréquence est qu'il est facile à implémenter sur des systèmes courants (SRI booléens)

10.2.2. Les fonctions d'analyse statistiques et de tri.

Il est possible de pouvoir établir des statistiques sur les résultats d'une recherche. Cette analyse consiste à classer selon un ordre de fréquence d'occurrence les termes extraits d'un ou plusieurs champs. La fonction ZOOM permet par exemple de repérer puis de classer les mots clés les plus fréquemment utilisés ou alors de classer les auteurs travaillant sur un sujet donné, etc. Ingwersen (1984) et Belkin (1985) considèrent que ce mécanisme est proche de celui du bouclage de pertinence.

Récemment dans son article sur l'avenir des SRI, le professeur Meadows (1999) estime que l'analyse statistique des résultats est l'une des fonctions les plus importantes à développer dans les futurs SRI. :

> *«I have mentioned the difficulty of users not being able to evaluate retrieved information in terms of its value, and to extract information from such records to be used in modifying a query. One way to help is to offer an analysis of a set of retrieved records. What terms occur often ?what query terms never occurred? What terms tend to occur, elsewhere in the database, togther with query terms ? ESA-IRS had a version of this called their ZOOM command, in the 1980s but it never became popular with other services"*

Actuellement, la plupart des serveurs opérationnels ont une fonction statistique similaire à ZOOM (tableau 42).

Serveurs	commandes
DIALOG	Rank
ESA/IRS	ZOOM, SuperZOOM
ORBIT	Get
QUESTEL	Memsort
NASA RECON	Frequency
DIMDI	Extract
STN	Select, Sort

Tableau 42 :Reformulation interactive dans les serveurs

Ainsi, dans le cas du serveur de Dialog, une requête sur le terme (**fat and (substitute? or replacement?))** aboutit à la visualisation de cette liste :

• To use the prompted version of RANK:

?**select fat and (substitute? or replacement?)**

7441 FAT

118447 SUBSTITUTE?

18205 REPLACEMENT?

S1 646 FAT AND (SUBSTITUTE? OR REPLACEMENT?)

?**rank pa**

Started processing RANK

.Ranking 100 of 646 records

.Ranking 200 of 646 records

.Ranking 300 of 646 records

.Ranking 400 of 646 records

.Ranking 500 of 646 records

.Ranking 600 of 646 records

Completed Ranking 646 records

DIALOG RANK Results

RANK: S1/1-646 Field: PA= File (s) : 340

(Rank fields found in 646 records -- 261 unique terms) Page 1 of 33

RANK No. Items Term

-------- ----- ----

1 67 UNASSIGNED OR ASSIGNED TO INDIVIDUAL

2 24 PROCTER & GAMBLE CO THE

3 23 ARCO CHEMICAL TECHNOLOGY INC

4 22 NABISCO BRANDS INC

5 17 NABISCO INC

6 16 KRAFT FOODS INC

7 12 CIBA-GEIGY CORP

8 12 CPC INTERNATIONAL INC

P = next page Pn = Jump to page n

P- = previous page M = More Options Exit = Leave RANK

On peut noter qu'une fonction proche de la commande ZOOM est présentée au sein de l'anté-serveur développé par (Thomazo, 1997). Après chaque requête, ce système propose à l'usager une fonction intitulée" pistes" qui présente la liste des termes associés à un lot de documents obtenus. Belkin (1993) a aussi introduit cette fonction dans son prototype BRAQUE qui permet d'interroger le serveur ESA. Ces commandes sont destinées exclusivement à des professionnels formés aux techniques documentaires et à l'usage de chacun des serveurs. Les premières études bibliométriques employaient cette fonction d'analyse statistique (Dou, 1993).

Elles présentent, malheureusement, les limites et inconvénients suivants :

- Elles nécessitent la connaissance et la maîtrise d'une syntaxe particulière
- L'usager doit être initié à l'analyse statistique.
- Cette fonction est employée sur les documents extraits par la recherche, mais pas sur les documents pertinents.
- Il faut maîtriser l'usage du clavier (c'est souvent un langage de commande qui reste complexe pour des usagers grand public).

- Il est indispensable de mieux comprendre la structure des données des bases documentaires interrogées pour pouvoir les utiliser.
- Ces outils ont été introduites pour des chercheurs de haut niveau, souvent des bibliothécaires et des documentalistes qui les utilisent fréquemment, et non pour des novices qui font leur propre recherche d'une façon occasionnelle.
- Ces documents ne sont pas classés et le nombre de termes à visualiser est souvent trop important ; d'où un temps de traitement qui peut être long.
- Comme le contrôle d'autorité n'est pas toujours effectué sur les termes (auteur, mots clés, .etc.) on rencontre parfois une certaine redondance au niveau de l'affichage des termes.
- Ce n'est qu'une fois le résultat obtenu, que le système permet de mettre en œuvre cette fonction de reformulation. Dans CATHIE on anticipe ce traitement et on présentera les termes et les réponses en même temps.

Enfin, ces reformulations ne sont pas interactives. Elles présentent le même problème que les vedettes matières dans une notice. Nous avons montré dans le chapitre sept que c'est grâce à la généralisation des interfaces hypertextes que la mise en œuvre d'un stratégie de recherche comme BRF était possible. Il est donc important de présenter ces pistes de recherches sous forme de liens hypertextes. Ce type de représentation présente des avantages par rapport aux formes textuelles: la représentation des informations est plus explicite; la démarche de consultation est plus intuitive

L'intérêt de cette fonction dépend essentiellement de la qualité du référentiel utilisé (liste de termes, thesaurus, classification, liste d'autorité, etc.). Nous pensons que contrairement aux anciens développements, l'usage d'une liste d'autorité, basée sur la précoordination, apporte plusieurs avantages :

- Le thesaurus par exemple oblige l'usager à lier les termes entre eux et à utiliser les relations pour atténuer, entre autres, les problèmes de polysémie alors que dans le cas d'une liste d'autorité ce travail intellectuel est déjà réalisé par les indexeurs.
- Un usager doit générer autant de diversité dans sa recherche que le domaine interrogé en comporte dans son indexation. Il est donc nécessaire de fournir à l'usager un outil d'orientation dans le domaine interrogé.
- Les requêtes des usagers sont souvent très courtes, elles comportent un seul terme, alors que les vedettes matières (grâce à la précoordination) sont plus longues. Nous pensons que l'une des raisons qui expliquent, le faible usage des

reformulations est lié au mode d'indexation choisie. En effet, l'indexation par des termes simples (unitermes) ne permet pas une description fine et précise du contenu d'un document. Le recours à des termes composés tels que « informatique documentaire » représente une indexation plus précise que l'emploi des termes simples « informatique » et « documentaire ». De même, une affichage de termes composés apporte plus de pertinence qu'une liste de termes simples.

Notre projet est d'ajouter des outils permettant ce genre d'analyse pour des usagers qui ne sont pas formés, ni aux techniques documentaires, ni à l'usage de ces commandes et qui n'entreprennent pas de recherches exhaustives.

10.2.3. Les choix de CATHIE

Après une analyse des résultats obtenus, CATHIE permet de charger toutes les vedettes qui indexent les notices retrouvées, et de les classer en nombre décroissant de leurs occurrences. Cet affichage permet aux utilisateurs d'étendre leurs recherches sur un même sujet. L'idée sous entendue dans CATHIE est qu'il existe une relation sémantique indirecte entre les vedettes indexant un même livre. Plutôt que de lui proposer un ensemble de référence, le système lui proposera une liste de termes sous forme de liens hypertextes. L'usager peut, soit activer un lien pour trouver les ouvrages indexés par cette vedette matière, soit activer le lien pour trouver d'autres vedettes matières. Cette outil l'aiderait certainement dans la formulation et reformulation de requêtes.

Le premier intérêt de cette solution est de proposer une liste de termes à l'usager : de nombreux catalogues n'ont pas d'index sujet, c'est le cas des catalogues en lignes de l'ENSSIB et celui de Lyon2. L'ensemble des vedettes matières générées par CATHIE constitue une solution à ce problème.

10.2.3. 1. Le choix de l'algorithme pour CATHIE

L'existence d'un catalogue en ligne, basé sur le modèle booléen, est un élément important dans notre choix. Nous considérons qu'il est plus simple de greffer sur ce système existant les outils nécessaires afin d'améliorer les performances plutôt que de concevoir un prototype entièrement nouveau basé sur un autre modèle (vectoriel ou non). Il nous a semblé que, parmi tous les algorithmes de reformulation, l'incorporation d'une fonction similaire à ZOOM était plus facile.. Ce choix s'explique par d'autres facteurs :

- Une architecture à deux niveaux est indispensable. Devant l'impossibilité d'utiliser RAMEAU en ligne, il nous a semblé pertinent de proposer une autre liste de termes plus faciles à incorporer en local ou en réseau.
- La majorité de ces reformulations présentées par (Efthimidiadis , 1996) sont très efficaces lorsque le domaine est très spécifié. Or le caractère encyclopédique des catalogues et la diversité des usagers rendent ces techniques de reformulation difficiles à appliquer. Nous pensons donc qu'il est important d'employer les vedettes matières comme source de reformulations. Cette procédure combine les avantage d'une recherche postcoordonnée et une précoordonnée. La recherche simple (recherche par mots du sujet et mots du titre) permet d'avoir le plus de documents et réduit le taux d'échecs alors que l'usage des vedettes matières aboutit à plus de précision. Une fois que l'usager a saisi sa question, l'interface le guidera vers les vedettes indexant les documents qui l'intéressent. Ainsi le nombre de vedette matières peut rester stable alors que les possibilités de recherche réussies augmenteront.
- Cette manière d'enrichir l'accès et l'accessibilité du catalogue n'impose pas de modifier l'indexation des documents déjà catalogués. Par conséquent, elle n'est pas coûteuse contrairement à une ré-indexation.

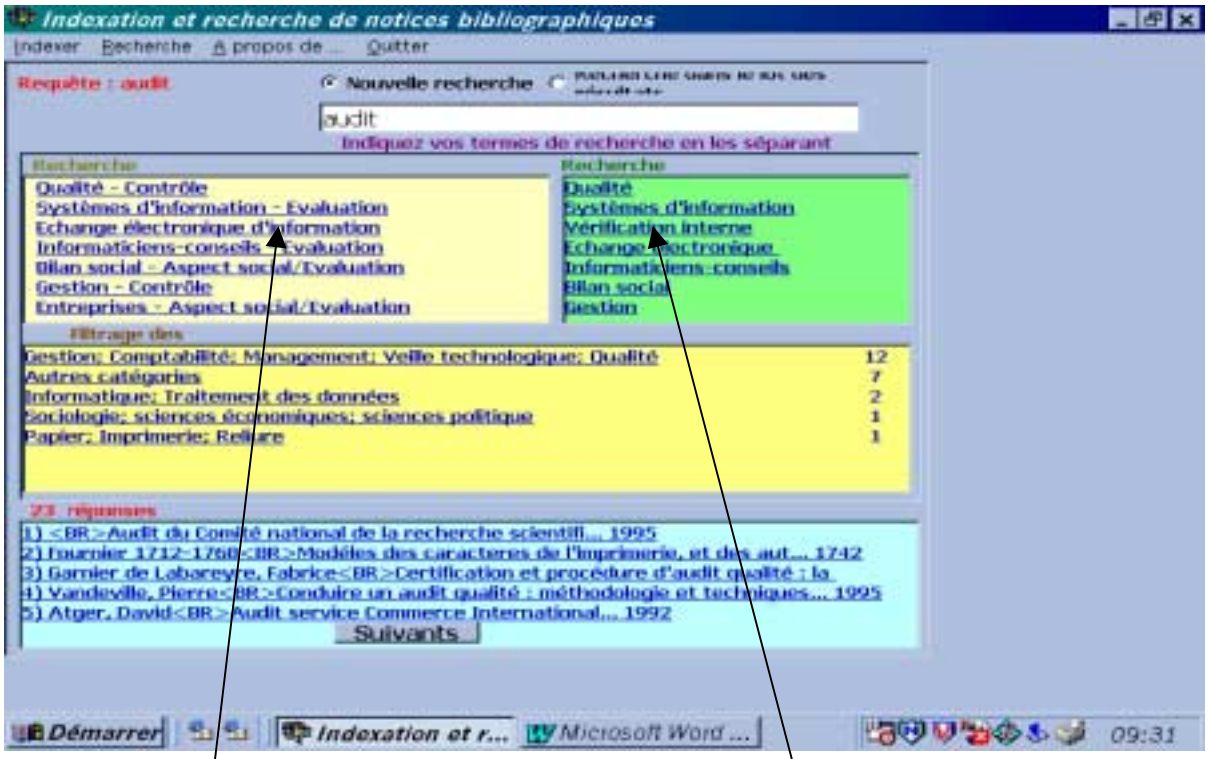
Enfin, les travaux de (Efthimidiadis , 1996) ont montré que d'un point de vue des usagers, les résultats obtenus par la commande ZOOM sont aussi importants que les autres algorithmes de reformulation. Il est d'ailleurs intéressant de questionner l'opportunité d'utiliser ces algorithmes pour la recherche dans un SRI grand public. Toutes les études sur la reformulation et/ou sur le feedback présupposent que l'usager ou le système emploient plusieurs dizaines de termes dans la reformulation. En accord avec Spink (1998), nous pensons qu'il est aussi important d'examiner la façon de rendre cette reformulation plus aisée à l'usager, que d'étudier la pertinence des termes générés par ces algorithmes.

10.2.3.2. Extraction et traitement des données

La source des termes candidats à la reformulation est un facteur déterminant dans notre solution. En effet dans la notice, plusieurs sources de termes se côtoient (auteur, vedettes matières, indices de classifications, mots du titres, etc.). Le type et la variété de termes prévus à la conception conditionne plus ou moins le type de reformulation que les usagers pourront utiliser.

L’algorithme que nous avons développé permet d’extraire l’ensemble des vedettes matières construites à partir des cent cinquante (150) premières réponses. Ces vedettes matières sont ordonnées selon un ordre de fréquence. Lorsque deux vedettes ont la même fréquence, nous effectuons un classement par ordre alphabétique. Seules les dix premières vedettes sont affichées. Nous avons opté pour un nombre limité puisque nos observations ont montré que les usagers ne consultent guère plus d’une page. Ceci évite ainsi le problème de surcharge d’information liée à la lecture de l’index. L’usager a cependant la possibilité de fixer lui même les deux seuils (le nombre de documents à traiter, le nombre de vedettes à trouver, le nombre de documents à afficher et celui des vedettes à afficher). Par défaut, CATHIE examine donc les 150 premières notices et les 50 vedettes (pour chacune des listes).

Pour la source des termes, nous avons décidé d’utiliser les champs MARC suivants $650 xy$z pour la liste des termes spécifiques et simplement le champ $650 pour la liste des termes généraux. La première liste permet de présenter les têtes de vedettes et les différentes subdivisions alors que la seconde se limite aux têtes de vedettes. Nous avons pensé qu’il serait plus approprié de suggérer en même temps, les deux listes à l’usager. Nous avons donc développé un prototype qui proposait la liste des têtes de vedettes pour élargir la recherche et celle des vedettes matières construites pour réduire le nombre de réponses. Voici un exemple (figure 34)



Affiner la recherche élargir la recherche

(tête de vedettes + subdivision sujet) (tête de vedette)

Figure 34 : CATHIE (exemple1)

Chaque fois que l'usager décide de cliquer sur un lien, le système reconstruit les liens : c'est une opération qui se fait en temps réel. L'extraction et le traitement des données est

dynamique et se fait en temps réel. Contrairement à la commande ZOOM, la reformulation dans CATHIE est réellement interactive. Chaque fois que l'usager relance sa recherche en sélectionnant sur un lien, CATHIE relance la recherche et affiche de nouveaux termes et d'autres notices bibliographiques.

10.2.3.3 Pertinence des termes extraits par CATHIE

Il existe plusieurs façons de mettre en œuvre cette fonction. On peut décider de ne prendre en compte que les têtes de vedettes (champs MARCc 650), ou que les subdivisions de formes et de chronologie. Ainsi pour une recherche portant sur l'expression "système d'information", le nombre de liens est différent selon qu'on prend en compte les têtes de vedette seules ou l'ensemble des vedettes matières (voir tableau 43)

Requête	Avec subdivisions	Sans subdivisions
Système d'information	*Système d'information* Système d'information – congrès ■ automatisation ■ guides, manuels ■ informatique ■ législation ■ répertoires *système, conception de* gestion—informatique merise, méthode—informatique	*système d'information* systèmes experts système en ligne services de documentation ré ingénierie organisationnelle archéologie merise, méthode données gestion *système, conception de*
Droit	*Droit d'auteur* *Droit administratif* Droits d'auteur – droits voisins Communication audiovisuelle—droit Informatique—droit Bibliothèque—droit Audiovisuel—droit Presse – droit	*Droit d'auteur* *Droit administratif* Communication audiovisuelle Informatique Bibliothèques presse Droit

Tableau 43 : termes extraits à partir de CATHIE

Nous avons défini et utilisé deux paramètres pour contrôler l'efficacité de notre système de reformulation.

Le taux de cohérence

Pour déterminer s'il y a redondance entre les vedettes retrouvées par les deux méthodes, nous avons adapté l'indicateur de cohérence externe qui est utilisé pour mesurer l'efficacité de l'indexation. Cet indicateur est donné par la formule suivante (Lancaster, 1993) :

$Cab = Ea \cup Eb / Ea \cap Eb$

Ea représente l'ensemble des concepts utilisés pour décrire un document par l'analyste A, et Eb l'ensemble des concepts utilisés pour décrire ce même document par l'analyste B. La mesure du taux de cohérence permet un suivi de la qualité de l'indexation.

Dans notre cas, cet indicateur est produit par la formule suivante :

C = M1 ∪ M2 / M1 ∩ M2

Dans laquelle M1 représente le nombre de termes trouvés par la première méthode (têtes de vedettes et subdivisions), et M2 le nombre de termes total repéré par la seconde méthode (juste les têtes de vedettes). Ainsi si 20 termes différents ont été retrouvés au total et que sur ces 20 termes, deux l'ont été par les deux méthodes, on obtiendra un taux de cohérence de 2/20 =0, 1. Ce taux peut varier de 0 (pas de vedettes en communs) à un (les deux méthodes retrouvent les mêmes vedettes)

Nous avons pris au hasard 40 termes extraits du fichier log (des requêtes des usagers) de l'ENSSIB. Nous les avons employés pour interroger CATHIE. Après analyse des termes retenus par chacune des deux méthodes, nous avons trouvé qu'en moyenne ce ratio **C** est de l'ordre de 16%.

Ceci implique que dans 16% de cas, la liste de termes proposées en utilisant les têtes de vedettes est similaire à celle qui est produite par l'ensemble de la vedette matière construite (tête de vedettes et subdivision). Si l'on examine le tableau 43, nous remarquons que pour le terme"droit", les deux listes ont en commun les deux vedettes matières suivantes : Droit d'auteur et droit administratif. L'analyse de plusieurs listes produites par les quarante termes de notre échantillon et les notices bibliographiques qui leurs sont liées, nous a permis de comprendre qu'il s'agissait fréquemment de références bibliographiques indexées seulement par des têtes de vedettes. Si l'on se réfère à notre étude (Ihadjadene, 1998c) sur le contenu des notices de l'ENSSIB et aux travaux de Markey (1994), on peut estimer que le taux de cohérence se situerait entre 13,7% et 37%. En effet, nous avons montré que le nombre de notices qui n'admettent pas de subdivisions ne dépassait pas les 37% dans une bibliothèque. Il est évident que plus le fonds documentaire d'une bibliothèque est important, plus ce pourcentage tend à diminuer. En plus de taille de la base, deux autres facteurs influent sur le taux de cohérence. Le premier concerne le nombre maximal de termes affichés, le second est le nombre de documents à partir duquel, on établit l'analyse statistique. Dans notre cas, nous avons analysé les 150 premières notices pour extraire les 50 vedettes les plus récurrentes.

Le taux du bruit

Il s'agit de déterminer quel est le pourcentage de termes de la liste affichée qui n'a aucune relation sémantique avec la requête de départ. Nous avons effectué cette analyse sur les quarante termes, aussi bien sur la première liste (avec subdivisions) que sur la deuxième liste. Par exemple dans le tableau 43, on peut estimer que pour la question "droit", la deuxième liste comporte quatre termes (Communication audiovisuelle, Informatique, Bibliothèques, presse) qui ne sont pas associés à "droit"

Taux de bruit = le nombre de termes inappropriés/ l'ensemble des termes de la liste

Ce taux est de l'ordre de 7,8% pour la première liste et de 27,6% pour la deuxième liste. Par conséquent, il y a une forte probabilité que les termes extraits à partir des têtes de vedettes ne soient pas liés sémantiquement à la requête initiale de l'usager.

Nous avons par la suite montré à sept usagers de l'ENSSIB les deux listes qui sont proposées par CATHIE. Tous les sept ont fortement apprécié cette possibilité de reformulation. Cependant, ils ne comprenaient pas pourquoi le système fournit, parfois, des termes similaires dans les deux listes. Nous avons, par la suite, modifié les stratégies d'affichage de CATHIE. Ce dernier ne présente qu'une seule liste de termes. Le système permet de proposer la liste des termes extraits à partir des têtes de vedettes lorsque le nombre de réponses est inférieur à 30. Ces termes sont moins précis que ceux extraits par la première méthode. Lorsque le nombre de réponses dépasse 30, le système propose une liste de termes ordonnés extraits à partir des vedettes matières construites (têtes de vedettes et les différentes subdivisions). Le seuil de trente réponses est choisi car il correspond pour Bates (1996) au nombre maximum de réponses, à partir duquel, il y a effectivement un problème de surcharge de réponse.

10.2.3.4.Utilisation des mots du titre pour la reformulation

Les titres ne sont pas toujours évocateurs du contenu d'un document et ne fournissent peu ou pas d'éléments d'information concernant le sujet traité dans le document. Cette situation est surtout valable dans les ouvrages de sciences sociales et humaines. Fernandez (1991) a comparé d'une point de vue linguistique la terminologie employée par les auteurs dans les titres et celle utilisée par les indexeurs dans la description sujet. La nature de cette relation est de type hiérarchique, à savoir que les titres sont plus précis que les vedettes matières dans presque 39% des cas.

Malgré cela, la recherche par mots du titre présente plusieurs avantages :

- La terminologie employée par l'auteur est plus à jour que celle des vedettes matières

- ❖ Les usagers peuvent oublier l'ordre exact des mots du titre mais se rappellent facilement de l'existence de ces mots dans les titres
- ❖ Dans les domaines scientifiques, les termes employés par les usagers correspondent mieux à la terminologie de l'auteur.

Nous pensons que les recherches par mots du titre complètent le repérage de l'information à l'aide des vedettes matières. Nous avons montré que lorsqu'on effectue une recherche simple (par mots du titre et sujets) le taux d'échecs diminue.

Si les mots du titres sont importants pour l'accès aux documents, ils ne sont pas souvent pertinents pour la reformulation des requêtes. Dans sa thèse, Hildreth (1993) a montré que la navigation à travers les liens "mots du titres" n'apportent pas d'amélioration pour la recherche. La méthode employée pour générer ces liens est très simple. Pour chaque affichage de référence bibliographique, il extrait automatiquement les mots clés du titre de la notice. Il n'a aucune analyse statistique sur ces mots. Pour tester cette hypothèse, nous avons élargi notre prototype aux mots titres. Après chaque requête de l'usager, CATHIE analyse les réponses et classe les mots du titres selon leurs fréquence. Nous avons aussi étudié le taux du bruit pour les mêmes termes, nous avons abouti à un taux de 38%.

10.2.3.5. Utilisation du champ auteur

S'il est aisé de représenter une liste de termes organisée d'un point de vue conceptuel pour les usagers, il est par contre difficile d'ajouter cette fonction pour les auteurs dans un catalogue pour deux raisons. La première, est que l'affichage alphabétique des auteurs pose moins de problèmes que l'affichage des sujets. Les usagers n'ont pas beaucoup de difficultés à se retrouver dans cette liste. Leur proposer une liste d'auteurs proches peut rendre la recherche confuse. Nous avons opté pour un choix ergonomique plus simple. Après chaque requête d'un usager, le système lui propose sous forme d'un lien hypertexte le message suivant : "liste des auteurs travaillant sur le même thème que X"

Si l'usager choisit de cliquer sur ce lien, le système lui propose une liste d'auteurs extraits des réponses. Dans ce cas, une fonction de classement des résultats peut améliorer le calcul.

Comme pour la recherche sujet, nous avons décidé d'y ajouter des mécanismes de filtrage par domaine. Cela est particulièrement utile pour traiter le problème des homonymes qui existent dans les grandes bases documentaires.

Cette solution présente trois avantages :

- Dans un catalogue comme celui de l'ENSSIB qui est dépourvu d'un index alphabétique des auteurs, cela peut constituer une alternative
- Dans le cas où il n'y a pas de contrôle d'autorité sur les auteurs (ce qui n'est pas rare), on peut trouver les variantes d'écriture du nom d'un auteur.
- Pour un usager novice dans un domaine, cela peut lui servir de piste de départ pour se rendre compte des auteurs qui travaillent sur un domaine de recherche.

Cette fonctionnalité nous permet une première lecture recherche d'auteurs sur un domaine émergeant. Ainsi sur un thème nouveau comme les bibliothèques numériques ou les métadonnées, CATHIE permet certainement d'avoir une vue d'ensemble sur les auteurs qui ont écrit sur ce thème.

10.2.4. Le rôle de l'usager dans la reformulation

La réflexion sur le rôle de l'usager dans la reformulation interactive ne se limite pas seulement sur son statut actif ou passif dans la recherche. Nous pensons qu'il est nécessaire de penser aussi aux moyens mis à sa disposition pour lui faciliter le choix des termes, sur les possibilités offertes pour contrôler ce processus (le nombre de documents à examiner, le nombre de termes à visualiser, etc.) et surtout sur les problèmes de charge cognitive que peuvent engendrer ces outils.

On peut donc conclure, qu'à l'inverse des fonctions statistiques tels que ZOOM, notre approche est sensiblement différente. Voici un ensemble d'éléments qui distinguent notre approche de celles des grands serveurs de banques de données comme ESA ou Dialog :

	Fonctions statistiques	CATHIE
Sources des termes	Mots clés, titre, résumé,	Vedettes matières construite et auteur
Niveau de contrôle des termes	aucun	Important (liste d'autorité)
Présentation de la fonction	Sous forme d'une commande	Sous forme de liens hypertextes
Résultats de la fonction	Des références	De nouveaux liens et des références bibliographiques
Utilisation de la reformulation	A la fin du processus de recherche	Durant tout le processus
Filtrage des termes	inexistant	Thématique
Niveau du bruit	Non étudié	Très faible
Algorithme de sélection des termes	booléen	Probabiliste
Base documentaire	spécialisée	encyclopédique

Tableau 44 : comparaison entre les fonction statistique et la reformulation dans CATHIE

Dans CATHIE, les reformulations s'inscrivent dans la logique de navigation et sont donc complètement graphiques et visualisées sous forme de liens hypertextes. L'originalité de notre approche consiste à donner à ce type de reformulation un support graphique (liens hypertextes). Ce processus peut être répété itérativement de manière à étendre ou restreindre progressivement le domaine concerné par la requête. Le choix d'un lien (vedette matière ou auteur) revient à relancer la recherche et par conséquent à recalculer l'ensemble des liens.

10.2.5. Critique de la solution adoptée

Lors de la récupération des documents, le catalogue adopte un ordre par défaut qui est en général l'ordre chronologique. Les documents sont donc numérotés du plus récemment entré

dans l'ensemble documentaire au plus ancien. Parfois c'est juste un ordre alphabétique. Cette situation nous a posé des problèmes de traitement des données. En effet, pour certaines requêtes qui aboutissent à la surabondance de réponses, le temps de traitement est très long (parfois plus d'une minute). De plus, en limitant l'analyse aux cent premières réponses, la qualité et la richesse des termes extraits à partir de ces références sont amoindris. Nous avons donc opté pour un classement des réponses par ordre de pertinence.

10.3. La classification automatique des résultats dans CATHIE

Pour résoudre le problème de la surabondance des réponses, plusieurs auteurs ont proposé d'utiliser un algorithme de calcul de pertinence pour présenter à l'utilisateur quelques dizaines de documents jugés les plus pertinents parmi tous ceux correspondant à sa requête. C'est la solution adoptée dans la conception des moteurs de recherche. On peut considérer que c'est la première fois que cette technique est réellement utilisée pour les grandes bases de données. Les premières études d'usages des moteurs de recherches ont malheureusement, montré les limites de cette technique. Dans le cas du moteur de recherche ALTAVISTA (Silverstein, 1998) , 85% des usagers se contentent des dix premiers résultats fournis sur la première page et 78% des requêtes ne sont pas modifiées dans le but de les améliorer.

- Tout d'abord, les algorithmes utilisés pour le calcul de la pertinence ne sont pas connus des utilisateurs. Or, il s'avère parfois important pour les usagers de comprendre ce mécanisme. On a souvent reproché au modèle booléen son opacité alors que cette lacune subsiste avec les autres modèles. Les usagers ne savent pas comment sont classées les réponses. En outre, ils sont incapables de savoir comment agir pour obtenir un nouveau classement. Notre étude effectuée à l'INIST (Ihadjadene, 1998d), nous a révélé les réticences des usagers à employer la recherche pondérée. Comment donner des poids aux termes ? Quelle est la signification réelle de ces poids ? Quelle est la différence entre des réponses qui ont un poids proche (90% vs 76%) ?
- Ces algorithmes ont été développés dans une optique où le nombre de termes utilisés dans la requête est très important. Dans les expérimentations de TREC par exemple, ce nombre est compris entre 10 à 20 termes. Notre étude sur l'usage du catalogue de

l'ENSSIB, comme celles de Markey ont prouvé que le nombre moyen de termes saisis est souvent inférieur à deux termes.

Wilkinson (1997) souligne que les algorithmes actuels de calcul de pertinence ne sont pas adaptés aux requêtes qui sont courtes (moins de trois mots). Il constate que les usagers sont souvent insatisfaits du classement des résultats. Cet auteur conclut sur l'importance d'offrir une forme de coordination des termes. Clark (1996) a développé un prototype qui effectue automatiquement la reconnaissance des phrases dans les requêtes des usagers. Son étude a montré que cette technique a amélioré la qualité de recherche. Gregory Grefenstette (1997) a proposé un prototype SQLET basé sur une analyse linguistique qui offre à l'usager un ensemble de phrases (groupes nominaux) pour reformuler sa recherche.

L'utilisation des vedettes matières construites et donc de la précoordination, pour la reformulation des requêtes se justifie essentiellement par les récents travaux de (Bruza, 1997) (Grefenstette, 1997), de (Clark, 1996). Ces derniers constatent qu'en réponse à une requête courte, le système doit impérativement permettre de visualiser des phrases et non des unitermes. Ceci a pour but de contourner les problèmes d'ambiguïté et d'augmenter la précision des réponses (les questions et les reformulations avec des requêtes courtes aboutissent souvent à l'affichage de plusieurs réponses). Cependant, si les approches linguistiques peuvent se justifier sur de petits corpus, elles sont difficilement applicables à l'analyse de contenu d'une grande base documentaire. Ce qui distingue notre approche de celles de Bruza ou de Grefenstette, est l'utilisation d'un langage documentaire encyclopédique conjugué avec une analyse statistique.

Notre propos, n'est pas de critiquer les modèles vectoriels ou probabilistes, encore moins de privilégier le modèle booléen, mais de pointer sur les imperfections de ces modèles pour traiter des requêtes courtes dans un SRI grand public et encyclopédiques (moteurs de recherches, catalogues en ligne, bibliothèques numériques). Il est donc urgent d'adapter ces techniques pour ce type de système. Notre approche permet d'utiliser simultanément l'analyse statistique, des algorithmes probabilistes et deux langages documentaires (RAMEAU et la classification de Dewey).

Il existe en effet plusieurs manières d'inclure un algorithme de calcul de pertinence dans un SRI booléen. Le choix de notre algorithme d'appariement est basé essentiellement sur les travaux de Smith (1995).

Voici les six algorithmes de pondération de termes testés par Smith :

- **Inverse document fréquency : IDF**
  - P (t) = (log N/n) + 1
- **F4**
  - P (t) = log $\frac{(r+c)(N-n-R+r+1-c)}{(n-r+c)(R-r+1-c)}$ avec c=0, 5
- **F4 point 5**
  - P (t) = log $\frac{(r+c)(N-n-R+r+1-c)}{(n-r+c)(R-r+1-c)}$ avec c=0, 5
- **F4 point 50**
  - P (t) = log $^{(N-n+0.5)}$ /n + 0.5
- **L'algorithme de Croft**:
  - P (t) = log N-n /n
- EMIM:

$$P(t) = \left(\log\frac{rN}{Rn}.r - \log\frac{(n-r)N}{(N-R)n}.(n-r) - \log\frac{(R-r)N}{(N-n)R}.(R-r) + \log\frac{(N-n-R+r)N}{(N-n)(N-R)}.(N-n-R+r) \right)$$

- **Coordination :** tous les termes ont un poids égal à un.

Avec :

N = nombre de documents dans la collection ; n= fréquence du terme t

R = nombre de documents pertinents ; r =nombre de documents indexés par le terme t

P (t) : le poids du terme t

Q : nombre de mots dans la question.

P(d) : poids du document D

Le poids du document peut être déterminé par la somme des poids des termes qui composent la question. De cette façon les documents peuvent être classés par ordre de pertinence.

Pour prendre en compte la longueur du document (tables de matière, texte intégral), on peut aussi utiliser par exemple la formule de Harman (1990) ou de Robertson (1994).

Les travaux de SMITH ont montré que si le nombre de termes est inférieur ou égal à quatre, les fonction de calcul ci-dessus donnent souvent les mêmes résultats. Nous avons donc opté pour la méthode suivante : F4 point 50

L'accès, comme la reformulation, sont basé donc sur un nouveau système d'appariement : le modèle probabiliste. L'extraction des termes se fait à partir des 100 notices les plus

pertinentes. Contrairement à OKAPI, cet algorithme se fonde essentiellement sur les vedettes matières construites.

10.4. Le filtrage des réponses dans CATHIE

Lorsqu'on parle de filtrage d'information, il faut toujours distinguer le groupements des termes d'indexation de celui des documents. Dans ce qui suit nous décrivons d'abord quelques prototypes qui offrent ce procédé. Nous exposerons par la suite, une autre approche basée le modèle probabiliste aussi bien pour le classement des résultats que pour le regroupement des termes.

10.4.1. La Catégorisation dans les SRI

10.4.1. 1. Le regroupement des termes

L'idée de regrouper les termes d'indexation n'est pas apparue avec les SRI. Il se doit d'être signalé qu'on trouve souvent ce genre de regroupement par thème pour les thesaurus ayant adopté la représentation graphique.

- le regroupement par facettes **:** dans de nombreux thesaurus, le vocabulaire a été regroupé selon un découpage en facettes.
- Le regroupement par thème **:** Il suit le découpage en domaine de la discipline concernée.

Le prototype Cansearch (Pollit, 1986) qui permet de rechercher de l'information sur le cancer dans Medline est un exemple intéressant qui propose la structuration des descripteurs d'un corpus de données. Pollit (1986) ne se contente pas des listes disponibles de mots clés. Il propose de structurer ces dernières en rapport avec le sujet traité, pour améliorer l'adéquation entre la question et les réponses et permettre un accès guidé aux données bibliographiques.

La clustérisation est une approche automatique qui permet d'identifier des agrégats rassemblant les mots clés qui sont fréquemment associés les uns aux autres. En France, les chercheurs du Centre de sociologie de l'innovation, ont amplement utilisé cette méthode pour concevoir des outils comme Leximap ou LiveTopic pour le moteur de recherche Altavista.

Quotidiennement, les professionnels de l'information (surtout ceux qui travaillent en veille technologique), se servent d'outils qui leurs permettent d'analyser statistiquement des volumes importants de données dans le but d'obtenir une information plus synthétique. Parmi ces outils, on peut citer Semiomap (http://www.semio.com) développé par Claude Vogel. Les cartes graphiques créées par Sémiomap sont des cartes lexicales sont constituées sur la base des cooccurrences constatées dans les pages WWW. Les clusters (agrégats statistiques des mots ou expressions) sont représentés sous forme graphique en temps réel. Le professeur Claude Vogel a mis en place un outil qui permet d'obtenir une carte graphique sur laquelle s'affichent les thèmes les plus significatifs évoluant autour d'une requête. Il est possible de "zoomer" sur chaque thème pour visualiser des termes proches et ensuite d'obtenir les références correspondantes.

On peut aussi inclure dans cette partie, les travaux et prototypes suivants :

- Shadocs (Zizzi, 1996)
- Concepts Maps (Gaines, 1995)
- AI ET Map (Chen, 1998)

Dans ces trois prototypes, les résultats sont présentés sous forme de carte montrant les liens entre des nœuds (les termes). Pour Chen (1998), ces relations sont associatives. L'importance des sujets se traduit visuellement par l'espace alloué sur une carte. Voici un exemple de carte conceptuelle issus de ETMap :

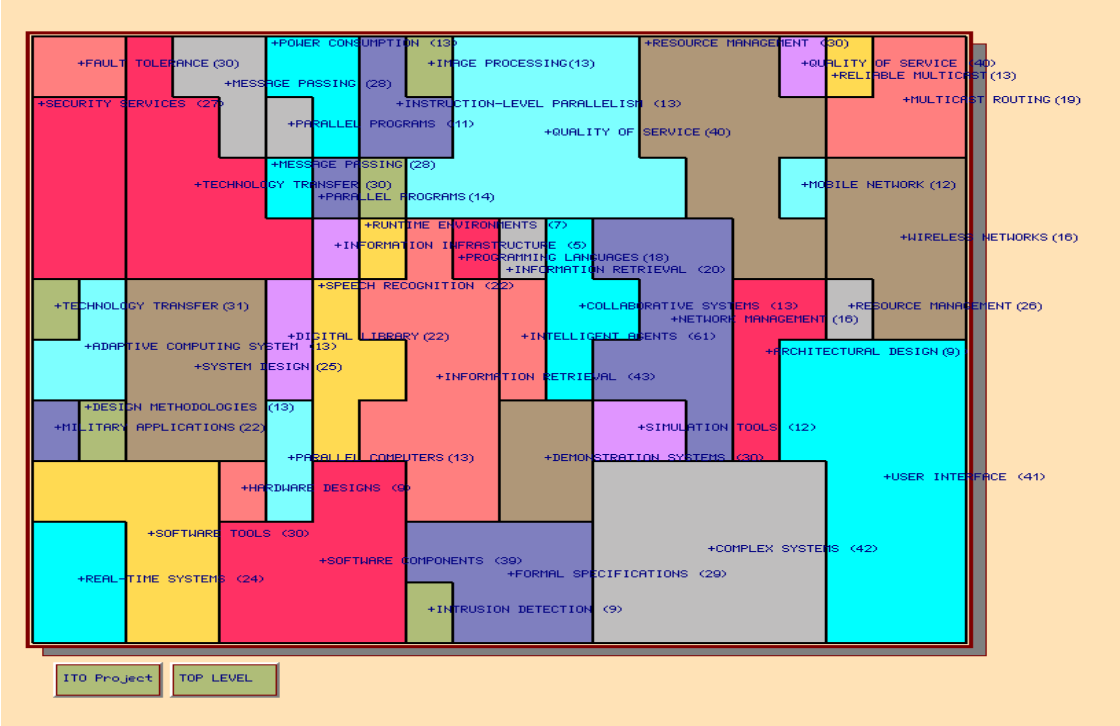


Figure 35 : exemple d'une carte conceptuelle ETMap

Ces méthodes nécessitent souvent l'avis d'un expert pour contrôler cette catégorisation. De plus, elles ne sont efficaces que dans un domaine spécialisé. Nous n'avons pas trouvé de prototype qui puisse effectuer une catégorisation des mots clés dans une base encyclopédique.

L'inconvénient des méthodes statistiques telles que Live Topic est qu'elles engendrent du bruit an niveau des termes proposés qui n'ont souvent aucun lien sémantique avec les termes initiaux de la requête**.** Plus spécifiquement, ces méthodes statistiques n'arrivent pas à résoudre les problèmes d'ambiguïté lors de l'analyse, puisqu'elles ne peuvent rendre compte des relations syntaxiques et des phénomènes de synonymie, de polysémie et d'anaphore. De nouvelles approches sont apparues comme celles basées sur une analyse linguistique comme (Bruza, 1997) ou celles qui s'appuient conjointement sur une analyse statistique et les connaissances propres d'un domaine (Buckland, 1999).

10.4.1. 2. La clustérisation des documents

Une des solutions au problème de surcharge d'information, consiste à construire des interfaces utilisateur qui ordonnent les informations trouvées. Certains algorithmes regroupent automatiquement les résultats en catégories.

Cette méthode (clustérisation) est apparue avec l'introduction du modèle vectoriel. Elle a été constamment améliorée. Plusieurs travaux récents ont permis de donner un support visuel à cette catégorisation des documents. Le professeur Khorflag (1998) a effectué, à l'Université de Pittsburg, un ensemble de recherches pour mettre au point des interfaces facilitant la visualisation de l'information, et par conséquent, le repérage Il a conçu trois prototypes : VIBE (*Visual Information Browsing Environment*), GUIDO (*Graphical User Interface for Document Organization*) et BIRD (*Browsing Interface for Retrieving Documents*). On peut aussi citer les prototypes suivants :

- TileBars (Hearst, 1995)
- Scatter/Gather (Cutting, 1992)
- InfoCrystal (Spoerri, 1993)

Nous pensons que ces interfaces de visualisation sont encore difficile à utiliser pour le grand public. L'interface de recherche d'information restera encore, une interface textuelle. Pollit (2000) critique[48] les possibilités de représentation des connaissances par des interfaces 3D.
C'est la raison pour laquelle nous pensons que les classifications et/ou les ontologies peuvent jouer un rôle dans le filtrage d'information.
Les concepteurs du moteur de recherche NorthernLight (http://www.northernlight.com) ont développé une technique dite Custom Search Folders qui permet de classer des résultats des recherches dans des dossiers représentant des catégories d'une ontologie construite à la main par des experts. Ces concepteurs, ont toutefois introduit de nouveaux facteurs, particulièrement pertinents, dans la composition des dossiers. Ce sont des critères suivant les thèmes, le type des documents, la source et la langue des pages Web.

[48] "*our world of three dimensions provides the interface designer with ideas for information retrieval systems, however, the effect of increasing the number of dimensions can only be realised effectively in the mind. When we conceptualise an information space we can attribute dimensions to facets of a data object an can mentally locate in it n-dimensional space, where each facet provides a compoment for the co-ordinate. The document object, on the other hand, with subject attributes that defy location to a point on the axis of a dimension, is more difficult to conceptualise. Nevertheless wa can image the caracteristics of the space and the relative locations of documents in it*"

10.4.2 Le filtrage de l'information dans CATHIE

Nous avons décidé que les vedettes matières issues de la liste RAMEAU doivent être regroupées selon le champs sémantique. Nous effectuons donc deux types de classification. La première concerne un calcul de fréquence des vedettes matières construites (VMC) sur un lot de documents. La seconde consiste à structurer ces VMC selon les domaines. De ce fait, nous faisons intervenir plus de sémantique en établissant une classification des vedettes par domaine. Le catalogue en ligne, perd ainsi une part importante de son opacité. Dans le cas d'une base encyclopédique, il est plus économique d'utiliser les classifications comme filtre à ce découpage.

Chan (1995) considère que les classification peuvent jouer un rôle important dans le filtrage d'information. La classification hiérarchique de DEWEY exemplifie deux fonction de la classification traditionnelle : la collacation (inclusion) et la partition (exclusion). L'inclusion rapproche les objets et les idées semblables. Mais dans un domaine d'information très vaste, il est tout aussi important d'exclure l'information non désirée que d'inclure ce qui est recherché. La partition peut être opérée en divisant une grande quantité d'information en parties plus petites comme moyen d'isoler la partie qui a la plus grande probabilité d'être pertinente.

Comme toute analyse statistique de données, CATHIE permet d'atteindre un ensemble d'objectifs:

1. Gagner du temps de recherche
2. Évaluer quantitativement le fonds documentaire et les réponses
3. Extraire une partie de l'information et faciliter la lecture d'un ensemble de données qui restent difficiles à lire avec une navigation linéaire des notices ou des termes.
4. Améliorer la recherche (en permettant de mieux choisir les termes)
5. Effectuer un filtrage sur les vedettes matières et sur les documents.

Un autre intérêt de notre approche est d'ordre pédagogique, . L'affichage des classifications montre à l'usager que pour un thème il existe différentes approches d'études (sociologique, informatique, etc.).

Pour aboutir à cette catégorisation, nous appliquons l'algorithme de regroupement sur les indices de la classification de Dewey. Cette méthode nous permet d'adapter la construction des liens hypertextes aux différents types de besoins des usagers. L'intérêt de l'analyse des résultats varie selon la richesse du fonds documentaire de la bibliothèque et

selon le type de bibliothèque (spécialisée ou non). En effet une requête sur le terme "bibliothèque" est moins générale dans une bibliothèque spécialisée en mathématique alors qu'elle aboutirait à une surcharge d'information pour le cas de l'ENSSIB. Comme l'évolution des collections n'est pas prévisible, il est donc important de donner des moyens de contrôle aux bibliothécaires selon le volume d'information entré et selon l'usage du catalogue (quels sont les thèmes les plus sollicités par les usagers). Selon la qualité et la nature du fonds documentaire de chaque bibliothèque, une fonction de filtrage peut être mise à disposition des bibliothécaires. Il suffirait de repérer les options qui correspondent le plus à leurs fonds. Exemple, pour une bibliothèque spécialisée comme celle de l'ENSSIB, où les requêtes sur la physique ou la médecine sont rares, il suffirait juste d'opter pour la classe correspondant à la médecine. Par contre, il est primordial de développer toutes les sous-classes liées aux sciences de l'information et de la communication. L'algorithme permet d'établir la correspondance entre les indices Dewey et les libellés. Dans une bibliothèque qui ne possède pas beaucoup d'ouvrages en sciences de l'information et de la communication, il suffira juste d'indiquer la classe correspondant à ce domaine.

Le tableau 45 illustre certaines catégories et leurs indices respectifs pris en compte pour le filtrage d'information dans le cas de la bibliothèque de l'ENSSIB. Par exemple la catégorie "informatique documentaire" rassemble l'ensemble des notices dont l'indice de classification est compris entre 025.04 et 025.06.

Indices de Dewey	Libellés
001 à 002	Le savoir; histoire du livre
003 à 006	Informatique
025.04; 025.06	Informatique documentaire
010 à 018	Bibliographie
20	Science de l'information
021 à 028	Bibliothéconomie; types de bibliothèques; lecture
25	Techniques documentaires
030 à 069	Encyclopédie et dictionnaires
070 à 079	Média et journalisme
090 à 098	Manuscrits et livres rares
100 à 149; 160 à 196; 200 à 220	Philosophie et religion
150 à 158	Psychologie
300 à 338	Sociologie; sociologie de la lecture; sciences économique; politique
340 à 363	Science juridiques; administration publique
370 à 379	Éducation ; enseignements
380 à 383	Commerce
384	Communication et télécommunications
390 à 398.2	Folklore; littérature populaire orale
400 à 490	Linguistique
500 à 539	Mathématiques et statistiques
601 à 621	Technologie et électronique
650 à 658	Gestion; comptabilité; managements
658.403	Veille technologique
676 à 686	Papier, imprimerie, reliure
700 à 900	Arts et architecture
801 à 890	Littérature
901 à 944	Histoire et géographie

Tableau 45: les différentes catégories du filtrage dans CATHIE

Comme les thèses et les mémoires des étudiants n'ont pas d'indices de classification, nous avons ajouté deux autres catégories "mémoires des étudiants" et "thèses de doctorats". Puisque la bibliothèque de l'ENSSIB est spécialisée en science de l'information et communication, nous avons privilégié la précision des catégories qui sont liées à l'informatique documentaire, à la bibliothéconomie et aux technique documentaires.

L'usager aura le choix entre trois tâches différentes. La première consiste à activer le lien pour n'afficher que les documents relatifs à une classe (exemple : afficher les documents en informatique documentaire) ; la seconde consiste à affiner une recherche en choisissant un autre terme plus précis et en cochant sur un thème bien particulier (exemple : économie de l'information et veille stratégique). Enfin la dernière qui nous semble la plus importante et innovante consiste pour l'usager à comprendre la relation qui existe les vedettes matières et la classification. Cette dernière fonctionnalité offre en plus la possibilité de réduire sensiblement les problèmes de polysémie. Ces trois procédés sont itératifs. Notre approche associe les avantages des analyses de contenu employées dans le filtrage d'information et dans la veille avec la richesse de l'analyse documentaire.

A la suite des travaux de Shamber (1994) et de Wang (1995) on sait maintenant**,** que dans une situation réelle de recherche, les usagers utilisent plusieurs facteurs pour déterminer la pertinence des réponses ainsi que pour le filtrage de l'information. Notre approche comme celles de tous les logiciels de veille, n'emploie que des critères thématiques liés au contenu du document. Il est donc important d'ajouter de nouveaux éléments dans le classement et la clusterisation des résultats. C'est la raison pour laquelle, nous avons ajouté ces quatre facteurs dans le classement des résultats et des termes. Nous avons introduit en plus un autre facteur "la date" pour pouvoir visualiser les ouvrages les plus récents. Voici les attributs de cinq critères :

- **Thèmes** (classification Dewey)
- **Type** (thèses, rapports, texte intégral, archive, ouvrages, articles, Cédéroms)
- **Source** (ENSSIB, Lyon2, etc.). Ce champ est particulièrement utile dans un accès en réseau.
- **Langue** (français, anglais, allemand, italien, autres langues)
- **Date** (documents les plus récents)

Ce n’est qu’à partir du deuxième filtrage que nous utilisons les autres critères. Après une recherche sur la question "access" et si l’usager décide de spécifier les termes et les réponses relatifs au domaine informatique, CATHIE affiche ces nouveaux dossiers (figure 36). On observe ici, la possibilité de filtrer les réponses soit par domaine, soit par langue ou par support.

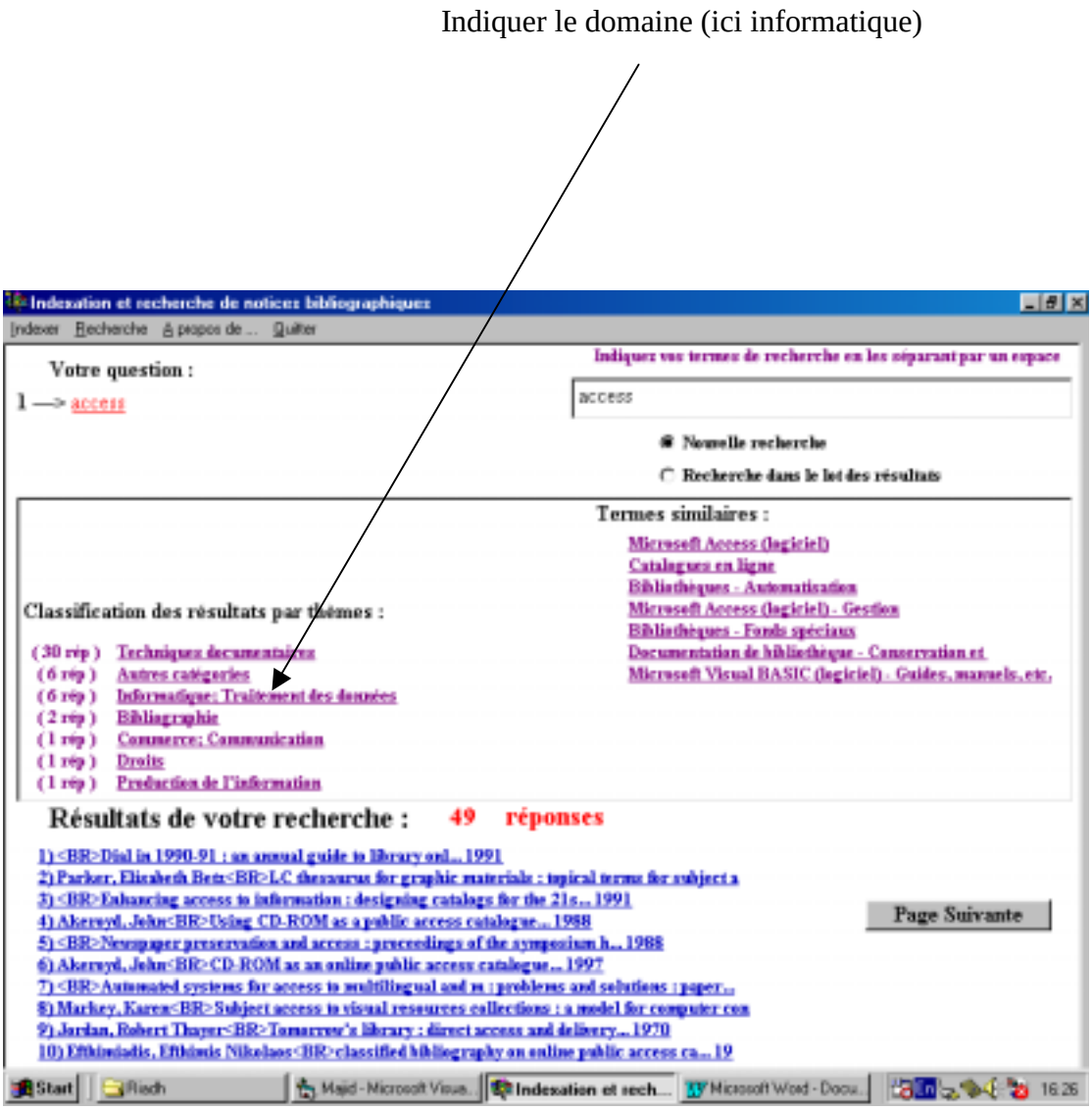


Figure 36: filtrage d’information dans CATHIE (exemple 1)

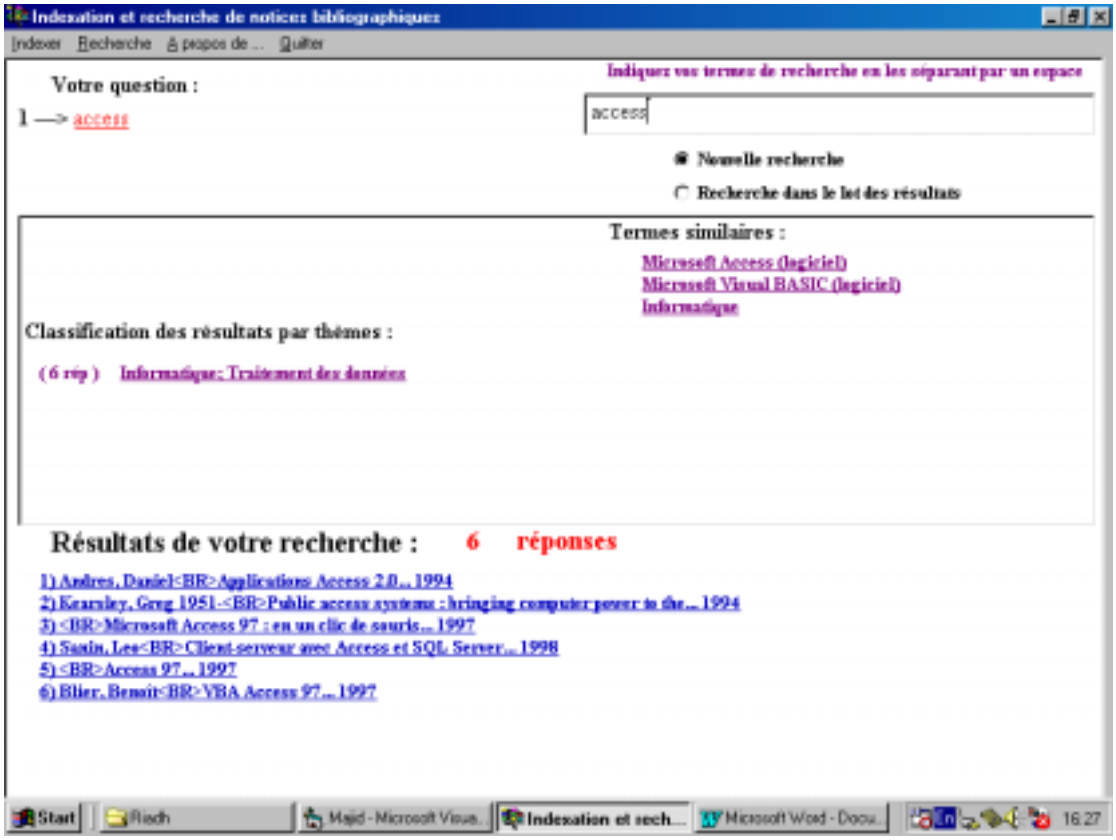


Figure 37: filtrage d'information dans CATHIE (exemple2)

Remarque 1: filtrage d'information et interdisciplinarité

Les classifications ont toujours eu à résoudre deux problèmes classiques mais qui sont toujours d'actualité: comment éviter le fractionnement d'un thème et comment représenter l'interdisciplinarité ? On ne peut pas donner deux cotes à un livre. Un ouvrage est soit dans une classe X, soit dans une autre classe Y.

Le savoir actuel est de plus en plus multidisciplinaire. Or, les grandes classification sont encyclopédiques et présentent chaque discipline académique. Cette structuration s'accommode mal des besoins de transversalité entre les sciences.

Concernant le filtrage, le problème se pose d'une manière différente. Rien n'interdit de séparer l'indexation systématique de la cotation. Un ouvrage peut avoir par exemple deux ou trois

indices de classification mais seulement une cote. On peut retrouver un ouvrage dans le sous ensemble "informatique" mais aussi dans le sous ensemble "science cognitives". Ce sont des copies virtuelles. A travers les liens hypertextes, on pourra lier ces copies à la cote.
En ce qui concerne, les documents électroniques, le problème du classement physique ne se pose pas. Il est donc possible de donner deux ou plus indices à ce document, facilitant ainsi une recherche pluridisciplinaire. Tinker (1999) et Beghtol (1999) préconisent aussi d'assigner plusieurs indices de classification à un document . Cette solution est déjà mise en œuvre dans la base bibliographique ITER (http://iter.library.utoronto.ca/iter).

Remarque2 :la recherche multilingue dans CATHIE
Nous avons remarqué qu'une partie non négligeable des usagers emploient des termes en anglais pour interroger le catalogue. Avec CATHIE, l'usager peut effectuer une recherche sur le mots du titre et le système lui proposera un ensemble de vedette matières en français. Une recherche sur "subject headings" aboutira à l'affichage des ces vedettes matières (figure 38.)
Cette fonctionnalité pourrait être sensiblement améliorée en si l'on disposait de la version en ligne de RAMEAU ou des autres traductions de la classification DEWEY. En effet, le langage RAMEAU comprend souvent une traduction anglaise extraite de la liste LCSH de la vedette matière. Voici des exemples de vedettes avec leurs traduction :

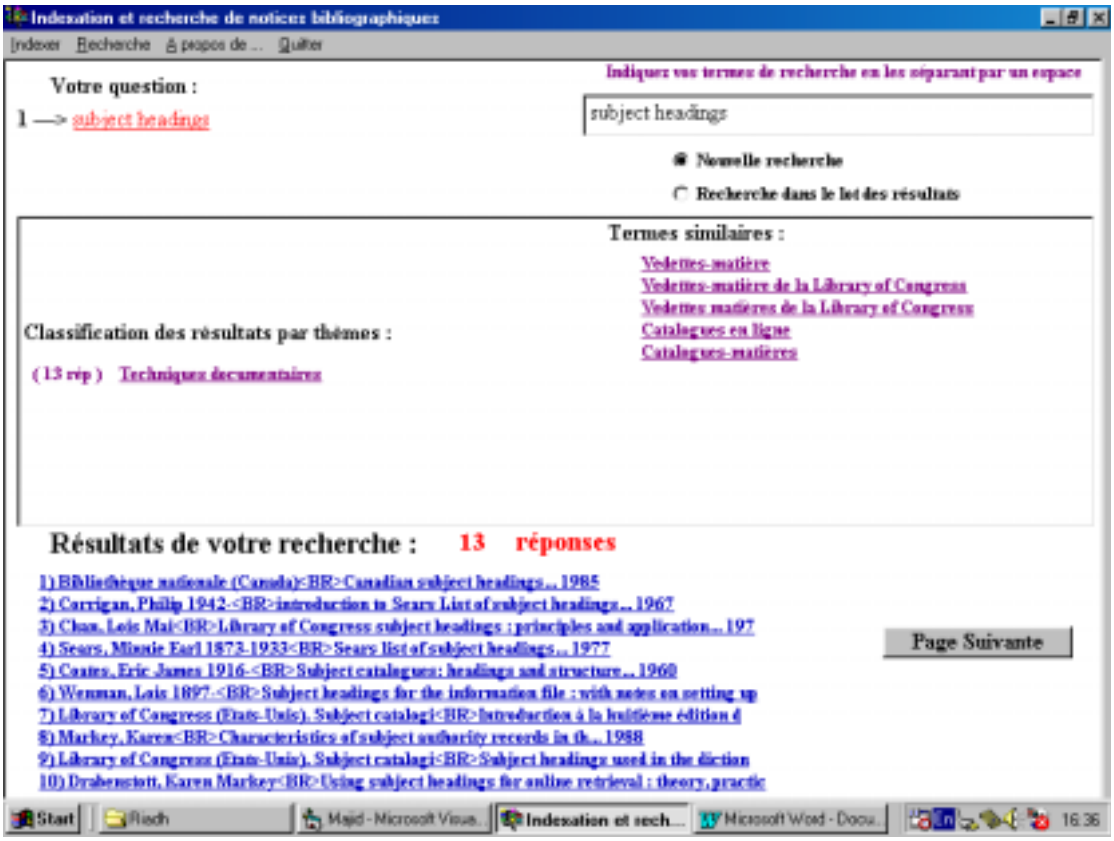


Figure 38: recherche multilingue dans CATHIE

10.5. Les stratégies de recherche dans CATHIE

Nous avons vu, qu'à l'inverse des usagers occasionnels qui formulent leur requête en saisissant tous les concepts composant la recherche, les professionnels ont tendance à chercher d'abord un lot de documents importants pour un concept général, et ensuite à filtrer les réponses en y ajoutant un autre concept plus précis. Pour faciliter la mise en œuvre de cette stratégie de recherche, nous avons ajouté au niveau de l'interface la possibilité de limiter la recherche à un lot particulier de documents.

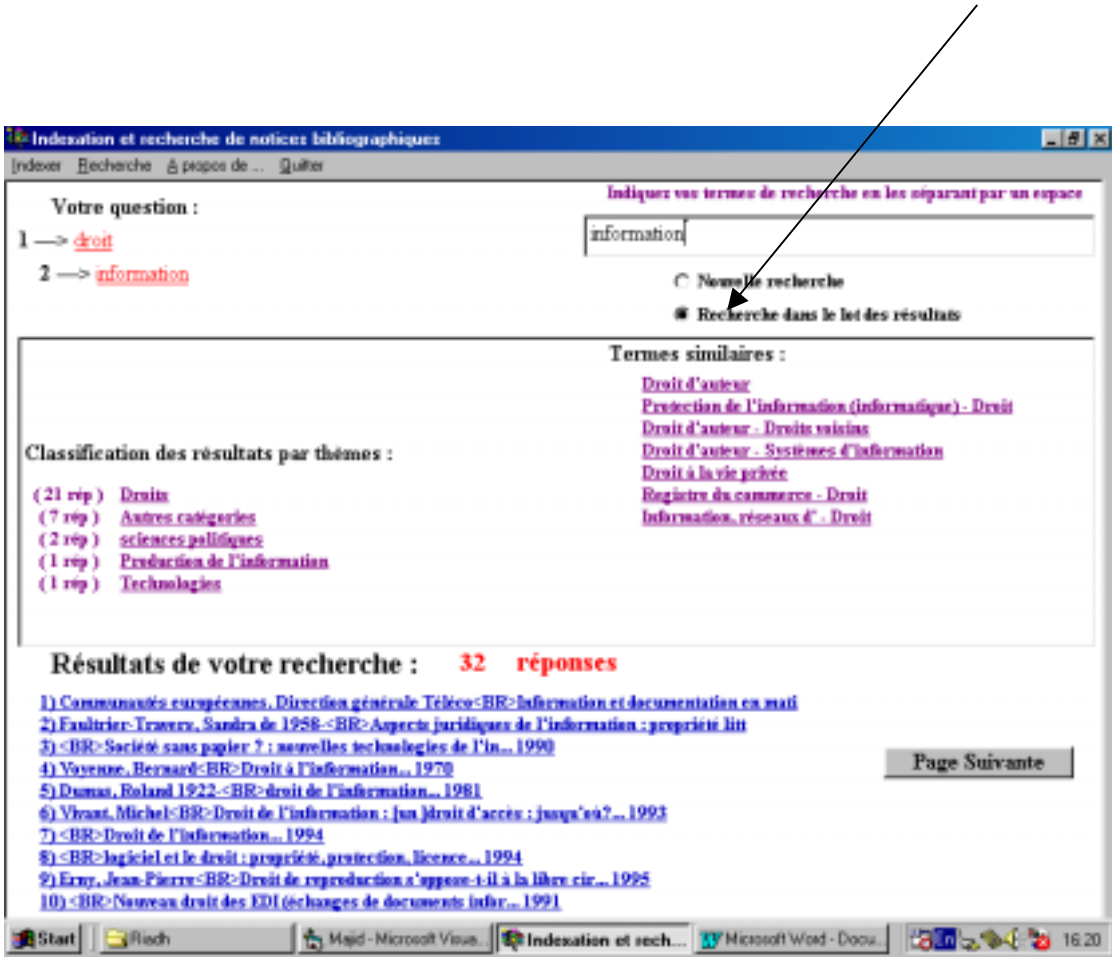


Figure 39: stratégie de recherche dans CATHIE

Cette stratégie est d'autant plus importante que nous offrons un accès intégral au contenu de toute la notice. L'usager peut ainsi développer des équations de recherche très complexes. L'usage du booléen "OU" permettra d'améliorer encore plus cette possibilité.

Concept1 : terme1 OU terme2 : pour rechercher les mots qui sont proches (synonymes)

Concept2 : termes 3 OU termes 4

Requête: concept 1 ET Concept 2

Exemple : pour une recherche sur l'usage des vocabulaires contrôlés dans les catalogues en ligne.

Première requête : concept 1 = "vocabulaires contrôlés OU langages documentaires"

Deuxième requête : Concept 2 = "catalogues en ligne OU OPACs"

Question finale : concept 1 ET concept 2 :"vocabulaires contrôlés OU langages documentaires ET (catalogues en ligne OU OPACs). Cette fonctionnalité est inexistante dans les catalogues en ligne actuels. On peut toutefois, signaler que la recherche dans Infoseek

"http//www.infoseek.com" est basée particulièrement sur cette stratégie. Ce moteur n'offre cependant aucune aide pour le choix des termes.

L'inconvénient majeur de cette stratégie de recherche est qu'elle exige de l'usager de réfléchir à d'autres termes pour réduire le nombre de réponses. Nous l'aidons en lui offrant une liste de vedettes classées par ordre de pertinence, il lui suffit alors de pointer avec la souris sur un lien et le système relance la recherche. Cette fonctionnalité est essentielle lorsque l'usager connaît peu le sujet de sa recherche. Il peut ainsi compléter sa recherche sur le catalogue en ligne en complétant sa requête avec d'autres termes issus directement du catalogue, assurant ainsi de meilleurs résultats.

Parmi les informations présentées à l'usager, on a le rappel de sa requête et la progression de celle ci (figure 40). Dans le cas où les termes qu'il a saisi n'aboutissent pas à l'affichage de réponse, l'usager n'a qu'à cliquer sur un des liens pour revenir à l'étape antérieure.

Progression de la recherche (historique)

Figure 40 : historique de la recherche

Nous avons bien sûr inclus la stratégie d'instanciation d'une référence connue à travers l'implémentation des liens hypertextes. Chaque fois que l'usager visualise une notice, on lui offre la possibilité de naviguer à travers les liens auteurs et sujets. Les liens sujets portent sur la totalité de la vedette matières construite. (Figure 41).

Figure 41: stratégie de navigation (BRF) dans CATHIE

Une reformulation sur le lien "système d'information—évaluation", aboutirait à l'affichage de ces notices (figure 42)

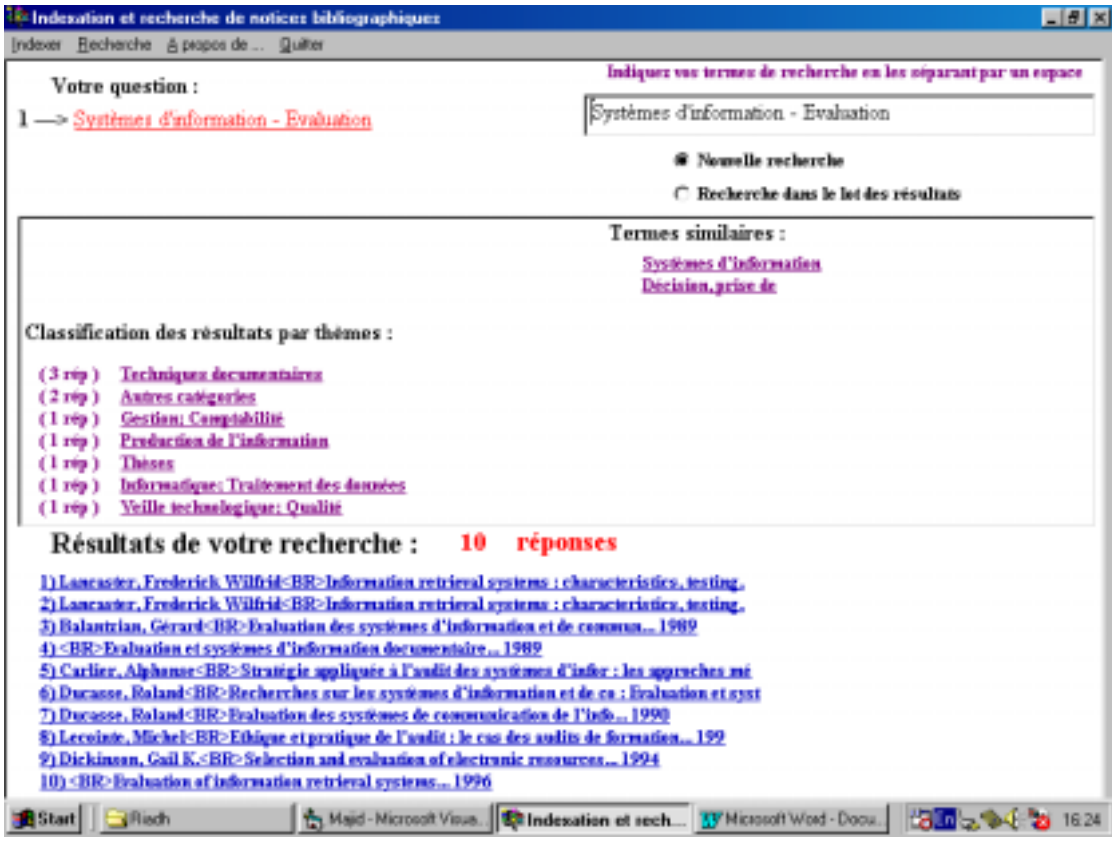


Figure 42: documentes similaires

10.6. Le regroupement des œuvres dans CATHIE

Il existe deux façons pour résoudre le problème du regroupement dans CATHIE ; la première est automatique. Elle consiste à utiliser quelques algorithmes qui permettent de détecter automatiquement toutes les duplications au niveau des réponses. Par duplication, nous entendons, les notices qui ont le même auteur et le même titre. Le système CATHIE affiche alors les réponses sous forme d'un dossier. En cliquant, l'usager peut consulter les différentes éditions d'une œuvre. Nous avons développé un algorithme similaire à Hylton (1996) : les temps de réponses sont très importants. Nous avons décidé d'abandonner cette solution. De plus, cette solution automatique a pour inconvénient de ne regrouper que les différentes éditions (dont le nom et le titre sont identiques) d'une œuvre.

La seconde solution est structurelle, c'est à dire qu'elle est basée sur la description des champs du format MARC. Ce sont les liens hypertextes qui jouent ce rôle de regroupement. Actuellement, c'est la solution la plus économique. Ainsi, lorsqu'on clique sur le lien auteur, le système affiche tous les titres écrits par l'auteur. Dans le cas où l'auteur est prolifique, l'affichage des réponses peut aboutir à des ouvrages qui ne sont pas toujours liés (d'un point de vue descriptif : ce n'est pas une édition d'un autre ouvrage).

10.7. CATHIE et le problème d'économie d'échelle

La solution que nous proposons est valable pour des petites et moyennes bibliothèque. La plupart des bibliothèques en France sont de ce type. Néanmoins notre solution peut facilement être adaptée pour les grands catalogues volumineux comme le futur système universitaire, celui de la BnF ou Melvyl aux États Unis. Bien entendu, lorsqu'on est en présence d'un volume important, on peut combiner plusieurs liens pour restreindre la recherche. Dans le cas de l'ENSSIB, nous n'avons pas besoin de cette fonctionnalité. Il reste à déterminer les temps de réponses et les stratégies de choix de ces deux seuils :

- Le nombre de documents pertinents à partir duquel il faudra extraire les termes ?
- Le nombre de termes candidats à la reformulation ?

Conclusion et Perspectives

Placer l'usager final au centre des études est devenu l'une des évolution les plus marquantes ces quinze dernières années en informatique documentaire. Notre objectif était d'étudier le comportement des usagers d'un catalogue en ligne afin de proposer de nouvelles techniques permettant d'améliorer la consultation d'une large base bibliographique. La première partie de la thèse consacrée à la présentation des principaux modèles de recherches et à la description des hypercatalogues a permis de dégager les choix réalisés par les concepteurs de ces systèmes, mais aussi leurs limites. L'analyse des travaux sur les catalogues opérationnels nous a permis de constater qu'un des obstacles à toute recherche documentaire (médiatisée ou non) est de trouver, à partir d'une requête donnée, une formulation dans les termes du catalogue : l'efficacité d'une recherche d'information nécessite la connaissance de la façon dont la base documentaire a été indexée. Les conséquences de cet écart de vocabulaire sont une des causes majeures du taux d'échec et de surcharge d'information. En effet, le langage ne permet pas toujours aux usagers d'extérioriser et de bien exprimer leurs besoins d'informations, du fait de l'économie de langage qu'ils pratiquent. Nous avons vu au chapitre quatre que l'un des rôles du bibliothécaire est de dialoguer avec les utilisateurs pour qu'à partir des réponses partielles , clarifier leurs besoins d'informations.

Plusieurs études (Chen, Hildreth, Kolmayer) ont montré que certaines stratégies d'information comme l'instanciation d'une référence connue sont hors de portée des usagers grand public puisqu'ils n'arrivent pas à raisonner à partir des termes d'indexation. Notre hypothèse de travail consiste à dire que ces problèmes ne sont pas toujours liés aux capacités cognitives des usagers, mais aux difficultés procédurales de l'accès aux catalogues de deuxième génération. Dés son apparition, l'hypertexte en général et le Web en particulier a apporté une nouvelle dimension à la recherche d'information bibliographique en offrant un accès navigationnel au contenu de la base. L'analyse des traces informatiques pendant plusieurs mois et dans trois contextes différents (enssib, irisa, lyon2), a confirmé notre hypothèse. Les études que nous avons présentées dans la deuxième partie de la thèse (chapitres cinq, six et sept) ont montré que non seulement les taux d'échec sont faibles mais qu'aussi les usagers élaborent plusieurs tactiques pour contourner l'absence d'une réponse. Cette diminution est malheureusement accompagnée d'un taux de surcharge inquiétant, d'autant plus que les usagers ne disposent d'aucune aide pour reformuler leurs requêtes. Ce problème est aggravé puisque seulement un usager sur deux tente de réduire le nombre de réponses en ajoutant

souvent un terme à l'équation d'origine. Les tactiques élaborées pour réduire ce problème de surinformation sont rudimentaires. Le fait que les usagers préfèrent changer le contenu de la requête plutôt que de modifier sa structuration logique pose le problème de la pertinence des outils logico-analytiques mis à leur disposition et suggère d'autres modalités d'exploration.

Nous avons montré dans le chapitre sept , que dans un tiers des cas, la navigation apporte une aide importante dans la sélection des termes et de documents similaires. Les usagers privilégient les liens sujets et auteurs. Il est clair cependant que la seule navigation par les liens hypertextes ne peut suffire comme mécanisme de repérage d'information bibliographique. C'est la raison pour laquelle, nous pensons qu'une solution hybride (Système de Recherche d'Information+hypertexte) apporte plus de facilité dans la consultation d'une base bibliographique.

Les récents travaux de (Jones,1998) , (Spink, 1998) , (Silverstein, 1998) et (Bruza,1997) sur l'usage des moteurs de recherches et des bibliothèques numériques ont aboutit à des résultas similaires aux nôtres à savoir que les ressources du système sont sous-utilisés et que les outils mis à la disposition de l'usager final pour explorer le nombre élevé de réponses sont insuffisantes et inadaptées. Nous avons dans cette thèse montré d'un point empirique la variété et les limites des tactiques utilisées par l'usager pour reformuler ses questions. Nous avons tenté d'apporter notre contribution à ces problèmes de surcharge en utilisant un certain nombre d'approches permettant d'assister l'usager dans le choix des termes et dans l'élaboration de stratégies de recherches. Le prototype CATHIE tend à remédier en partie à divers déficits que nos analyses d'usage ont mis en évidence. Il associe la richesse des vocabulaires contrôlés (RAMEAU et la classification décimale de Dewey), les possibilités de visualisation et de navigation de l'hypertexte et la puissance du modèle probabiliste. Après une requête, l'usager peut exploiter quatre stratégies :

- Effectuer une recherche dans le lot de documents trouvés
- Reformuler la question à travers les vedettes matières proposées
- Établir un filtrage thématique des vedettes et/ou documents
- Afficher la notice et voir les documents similaires.

Le prototype CATHIE est pour l'instant encore expérimental et de nombreuses améliorations d'ordre techniques pourraient lui être apportées. Ainsi nous souhaiterions :

1. Tester et implémenter d'autres algorithmes pour la reformulation interactive. Le traitement automatique des champs textuels de la notice bibliographique (résumé,

table des matières) en permettant une indexation automatique, donne d'autres descripteurs que ceux employés par les bibliothécaires (vedettes matières). Dans ce cas là, il serait intéressant notamment d'utiliser l'algorithme de Robertson (1994). De plus, l'efficacité des stratégies de recherche dépend en partie de la compréhension des usagers. Il est nécessaire d'effectuer des études d'utilisabilité du prototype pour rendre encore plus claires les options du système. Quel est le nombre de termes à afficher ? Quel est le seuil à partir duquel l'analyse est nécessaire ? Parmi les options proposées par CATHIE , quelles sont celles réellement utilisées ?

2. Adapter et tester CATHIE dans des domaines spécialisés (médical, brevets, droit …etc.). Comme nous avons pu le constater tout au long de nos développements, CATHIE reste largement tributaire du système d'indexation inhérent aux SRI bibliographiques. Nous voudrions utiliser simultanément des thésaurus spécialisés (Mesh, Inspec, etc.) et d'autres systèmes de classification pour faciliter le filtrage des descripteurs et des documents.
3. Développer une version Web de CATHIE. Nous souhaiterions développer un moteur de recherche facilitant le filtrage et la reformulation interactive de l'information bibliographique. Le prototype W-Cathie serait développé en Java et exploiterait la richesse de la norme Z39.50 pour accéder à des catalogues distants . Cette solution ne sera pas possible sans une réflexion et le développement d'un outil de correspondance entre quelques langages documentaires. L'accès simultané à plusieurs millions de notices bibliographiques pose le problème des doublons. Nous pensons prendre en compte les récents travaux de l'Ifla sur les œuvres bibliographiques pour faciliter leurs recherches .
4. Thompson (1997) et Dolin (1998) ont proposé des techniques (linguistiques et probabilistes) permettant d'assigner automatiquement un ou plusieurs indices de classifications à des documents non-bibliographiques (émails, groupes de discussions, sites Web, etc.). Ils ont développé deux prototypes : Scorpion (Thompson,1977) et Pharos (Dolin,1998). Nous pensons se baser sur leurs études et utiliser la classification de Dewey pour le filtrage des sites Web et/ou des listes de discussions.
5. Ajouter un accès thématique (par le biais d'une représentation hypertextuelle) des classifications en se basant sur les études de Pollitt, de Cochrane et de Markey.

Ces extensions devraient contribuer à la validation de notre modèle sur des bases importantes. L'étude de ces nouveaux modèles d'interaction (visualisation de l'information, catégorisation thématique, reformulation interactive, etc.) est encore récent. Il convient maintenant d'examiner leur usages en situation réelle d'utilisation.

Bibliographie

Aboud A (1997). *Recherche d'information médiatisée*. DEA SIC, enssib 1997**.**

Alberico R, Micco M (1990). *Expert systems for reference and information retrieval*. Meckler (eds)

Allen, B. (1996). *Information tasks : toward a user-centered approach to information systems*. San Diego, CA : Academic Press.WSU Purdy/Kresge 1996

Allen, R. (1994). *Navigating and Searching in Hierarchical Digital Library Catalogs*, Digital Libraries '94 Proceedings, College Station, TX June 19-21, 1994, pp. 95-100

Balpe J.P ; Lelu A ; Papy F ; Saleh I (1996). *Techniques avancées pour l'hypertexte*. Hermes (eds) 1996 ; 288p

Bates, M. (1977). *System meets user: Problems in matching subject terms.* Information Processing and Management *13,* 367--75.

Bates, M. (1984). *The fallacy of the perfect 30-item online search*. Reference Quarterly (RQ) 24, 43--50.

Bates, M (1986). *Subject access in online catalogs: a design model*. Journal of the American Society for Information Science , 37 (6). 357 – 376.

Bates, M. (1989). *The design of browsing and berrypicking techniques in online search interfaces*. Online review 13 : 407-423.

Bates, M.J. (1990). *Where should the person stop and the information search interface start?* Information Processing and Management 26, 575--591.

Bates, M (1996). *Indexing and Access for Digital Libraries and the Internet: Human, Database, and Domain Factors*. Journal of the American Society for Information Science 49 (November 1998): 1185 - 1205.

Ballard, T., and Lifshin, A. (1992). *Prediction of OPAC spelling errors through a Keyword Inventory*. Information Technology and Libraries 11, 139--145.

Beaudouin-Lafon M (2000). *Ceci n'est pas un ordinateur*. Techniques et science informatiques. 19 (1/2/3) . (à paraître)

Beghtol C (1999). *Knowledge domains: multidisciplinary and bibliographical classification systems*. Knowledge organisation, 25 (1/2), 1-12.

Beheshti, J; Large, V and M Bialek (1996). *PACE : a browsable graphical interface.*. Information technologies and libraries 15(4) , pp 231-240.

Belkin N (1984). *Cognitive models and information transfer*. Social science information studies, Num 4 , pp 111-129.

Belkin, N.J. & Vickery, A. (1985). *Interaction in information systems*. London: British Library, 1985.

Belkin, N. J., Marchetti, P. G., & Cool, C. (1993). *BRAQUE: Design of an interface to support user interaction in information retrieval*. Information Processing & Management, 29, 325-344

Belkin, N. J., Cool, C. (1995). *Cases, scripts and information seeking strategies*. Expert Systems with Application, 9(3),pp 379-395.

Berger, M.G. (1992). *The MELVYL system: The next five years and beyond*. Information Technology and Libraries 11, 146-157.

Blair, D., Maron, M.(1984). - *An evaluation of retrieval effectiveness for a full-text document-retrieval system*. Communications of the ACM, vol. 28, n°3, 1984, pp. 281-299.

Borgman, C.L. (1986). *Why are online catalogs hard to use? Lessons learned from information retrieval studies*. Journal of the Society for Information Science 40, 153--157.

Borgman, C. L., Hirsh, S. G., Walter, V. A., & Gallagher, A. L. (1995). *Children's searching behavior on browsing and keyword online catalogs: The science library catalog project*. Journal of the American Society for Information Science, 46, 663-684

Borgman, C.L. (1996). *Why are online catalogs still hard to use?* Journal of the American society for information science, 47(1996):7, 493-503.

Boughanem M (1992). - *Les systèmes de recherche d'informations : d'un modèle classique à un modèle connexionniste*. - Thèse de doctorat en informatique, Toulouse, Université Paul Sabatier, 1992.

Brüggemann, A K (1999). *BibRelEx Exploring Bibliographic Databases by Visualization of Annotated Content-Based Relations.* D-Lib Magazine. November 1999 (http://www.dlib.org)

Bruza, P.D. and Dennis, S. (1997). *Query re-formulation on the Internet: Empirical Data and the Hyperindex Search Engine*. In Proceedings of the RIAO97 Conference - Computer-Assisted Information Searching on Internet, 488-499, Centre de Hautes Etudes Internationales d'Informatique Documentaires.

Buckland M K, M.H. Butler, B.A. Norgard & C. Plaunt (1992). *OASIS: A front-end for prototyping catalog enhancements.* Library Hi Tech 40 (1992):7-22

Buckland, M. K.(1999). *Vocabulary as a Central Concept in Library and Information Science*. *In* Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities. Proceedings of the CoLIS3, Dubrovnik, Croatia, 23-26 May 1999, pp 3-12.

Buckland, M., Chen, A., Chen, H., Gey, F., Kim, Y., and Larson, R. (1999). *Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies*. D-Lib Magazine. Vol.5 No.1 January 1999

Butkovich, N.J., Taylor, K.L., Dent, S.H., and Moore, A.S. (1989). *An expert system at the reference desk: Impressions from users.* Reference Librarian: Expert Systems in Reference Services 23, 61--74.

Caro S (1997). *Ergonomie des documents techniques informatisés*. In les hypermédias, approches cognitives et ergonomiques, sous la direction de Tricot A et Rouet JF, Paris, éditions Hermès, 1998, pp 123-138.

Cazabon M (1993). *Un catalogage allégé*. Bulletin des bibliothèques de France, 38 (5) 1993 , pp. 42-43.

Chan L M (1995). *Classification, present and future*. Cataloging and classification quarterley. 21(2), 5-17.

Chen H , V Dhar (1991). *Congnitive process as a basic for intelligent retrieval systems design* .Information processing and management 27 : 405-432

Chen H (1992). Knowledge-based document retrieval. Journal of Information Science, 18,pp 293-314.

Chen H, A. Houston, R. Sewell, and B. Schatz (1998). *Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques. JASIS* 49 (7), 582-603

Cherry J (1994). *Improving subject access in OPACs: An exploratory study of conversion of users'queries.*The journal of Academic Librarianship, Vol 18(2)1994- p. 95-99

Clark C (1997). *Relevance ranking for one to three term queries.* In Proceedings of the RIAO97 Conference - Computer-Assisted Information Searching on Internet, 388-400, Centre de Hautes Etudes Internationales d'Informatique Documentaires.

Cleverdon, C. W., and E. M. Keen (1966) . *Factors determining the performance of indexing systems*, volume 1: design, volume 2: test results. Cranfield, England: Aslib Cranfield Research Project.

Conein B (1994). *Action située et cognition*. Sociologie du Travail .Vol 4, 475-500

Crawford J (1993). *A survey of subject access to academic library catalogues in Great Britain*. Journal of librarianship and information science- vol 25 , 1993. pp 85-93.

Croft, W.B., and Thompson, R.H. (1987). *I3R: A new approach to the design of document retrieval systems.* Journal of the American Society for Information Science 38, 389--404.

Cutter CA (1904). *Rules for a dictionary catalog*.- 4th edition, Washington DC, 1904.

Cutting, D.R., Karger, D.R., & Pedersen, J.O. (1993). *Constant Interaction-Time Scatter/Gather Browsing of Very Large Document Collections. In* Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, *Pittsburgh, 126-131*

Denos N (1997). *Modélisation de la pertinence en recherche d'information: modèle conceptuel, formalisation et application*. Thèse de doctorat, Université de Grenoble 1, IMAG,1997.

De Grolier E (1989). *Taxologie et classification*. BBF. 33 (6), 469-489.

Dervin B and Michael Nilan (1986). *Information Needs and Uses.* Annual Review of Information Science and Technology 21 (1986), pp 3-33.

Dervin, B. (1983). *An overview of sense-making research: Concepts, methods and results*. Annual meeting of the International Communication Association, Dallas, TX, May. Disponible à: http://communication.sbs.ohio-state.edu/sense-making/art/artdervin83.html

Dewdney, P. & Michell, G. (1997). *Asking "why" questions in the reference interview: a theoretical justification*. Library quarterly, 67(1997):1, 50-77

Dou H., Desvals H (1992*). La Veille Technologique* . Edition Dunod, Paris, 1992

Dolin R, D. Agrawal, A. El Abbadi, and L. Dillon.(1998). *Using Automated Classification for Summarizing and Selecting Heterogeneous Information Sources.* D-Lib Magazine January 1998. Disponible à (http://www.dlib.org)

Dujol A (1986). *Le clair et l'obscur*. BBF , 31(3), 232-237.

Efthimidiadis, E.N. (1996). *Query expansion*. In: Williams, M. ed. Annual Review of Information Science and Technology, *31*, 121-187.

Efthimidiadis, E.N. (1993). *A user centered evaluation of ranking algorithms for interactive query expansion*. In R Korfhage (eds), Proceedings of the 16th ACM-SIGIR, June 1993, N. Y, pp 146-159.

Ensor, P. (1992). *User practices in keyword and boolean searching on an online public access catalog*. Information Technology and Libraries 11, 210--219.

Ercegovac, Z. (1989). *Augmented assistance in online catalog subject searching*. Reference Librarian: Expert Systems in Reference Services 23, 21--40.

Erdelez S (1995). *Information encountering: an exploration beyond information seeking*. Unpublished dissertation, Syracuse University, New York 1995.

Ferl E T & Millsap L (1992). *Remote use of the University of california MELVYL Library System: An online Survey*. Information Technology and Libraries, September 11, 285-303.

Ferl , E T (1996). *The Knuckle-Cracker's Dilemma: A Transaction Log Study of Opac subject searching*. Information Technology and Libraries, june : pp 81-98.

Fidel, R. (1985). *Moves in online searching*. Online Review 9, 61--74

Fidel, R. (1991). *Searchers' selection of search keys: 2. Controlled vocabulary or free-text searching*. Journal of the American Society for Information Science, 42(7): 501-514.

Fondin H (1999). *La recherche d'information dans les mémoires électroniques*. Documentaliste-science de l'information, vol. 36 (4-5), 242-249.

Foss, C L (1989). *Detecting lost users : empirical studies on browsing hypertext*. Rapport de recherche INRIA n°972. Sophia-Antipolis France

Fox, E, France R and Ben E. Cline (1993). *Development of a modern OPAC: from REVTOLC to MARIAN*. SIGIR-93: Proceedings of the Sixteenth ACM SIGIR Conference (Pittsburgh, PA: June 27 - July 1, 1993) ACM, 1993. pp. 248-259.

Franz Lori (1994). End-user *understanding of subdivided subject headings*, LRTS, 38(3)1994, 213 - 226

Frisse, M and Cousins S (1989). *Information Retrieval From Hypertext*: Update on the Dynamic Medical Handbook Project, Proceedings of Hypertext '89, ACM Press.

Frost C (1994). *Next-generation online public access catalogs*. In Advances in library automation and Networking, vol 5 (1994)., 1-41

Gaines, B.R. and Shaw, M.G.L (1995). *Concept Maps as Hypermedia Components*. International Journal of Human-Computer Studies. 43(3), 1995, p. 323-61.

Genest D (1999) , *Indexation de documents en graphes conceptuels*, Deuxième conférence du chapitre français de l'ISKO, Lyon, 1999 (à paraître)

Grefenstette G (1997) . *SQLET : Short query linguistic expansion techniques*. In Proceedings of the RIAO97 Conference - Computer-Assisted Information Searching on Internet, pp 500-509, Centre de Hautes Etudes Internationales d'Informatique Documentaires

Hancock, B.M (1987). *Subject searching behaviour at the library catalogue and at the shelves*. Journal of Documentation 43, 303--321.

Hancock, B. M (1992). *Query expansion: advances in research in online catalogues*. Journal of Information Science 18: 99-103.

Hancock M.B (1994) . *A graphical interface for OKAPI*. British library research and development department Report 6144, 1994.

Hancock, B. M. (1997). *Experiments on interfaces to support query expansion*. Journal of Documentation 53(1): 8-19.

Harman D K (1990). *Retrieving records from a Gigabyte of text on a minicomputer using statistical ranking*. Jasis, 41(8), 581-589.

Harter, S.P. (1986). *Online Information Retrieval: Concepts, Principles, and Techniques*. Library and Information Science Series. Academic Press, Orlando..

Harter,S.P & Hert C (1997). *Evaluation of information retrieval systems*. ARIST, vol 37, 1997, pp 3-92.

Hassoun M & Roger D (1994). *Les catalogues en lignes,* ENSSIB 1994

Hearst M (1995) . *TileBars: Visualization of Term Distribution Information in Full Text Information Access*, Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), Denver, CO, 1995.

Hert, C.A (1995). .*Exploring a New Model for the Understanding of Information Retrieval Interactions*. unpublished dissertation. School of Information Studies. Syracuse University, 1995

Hetzler B (1997). *Beyond word relations,* SIGIR Forum, Vol 31 (2) , ACM press.

Hetzler B (1999).*Visualizing the full spectrum of document relationships*. . In: W Mustafa el Hadi, J Maniez & A S Pollitt (eds.), Stuctures and relations in knowledge organisation. Proceedings of the fifth international ISKO conference 25-29 August 1998, Lille, France. Wüzburg: Ergon Verlag

Hjerppe R (1985). *Projet HYPERCATalog: visions and preliminary conceptions of an extended and enhaced catalog*. Sixth conference on Intelligent information Systems for the Information Society. Frascati Italy 1985.

Hildreth, C.R. (1989). *Intelligent Interfaces and Retrieval Methods: For Subject Searching in Bibliographical Retrieval Systems*. Cataloging Distribution Service, Library of Congress, Washington, D.C.

Hildreth C (1993). *An evaluation of structured navigation for subject searching in Online catalogs*. Phd dissertation, departement of information science, London.

Hudon, M (1995). *Le thésaurus: conception, élaboration, gestion.* Montréal: ASTED

Hudon, M (1998). *Le catalogage à l'heure des nouvelles technologies. Education et francophonie*. [En ligne]. Adresse Internet : http://www.acelf.ca/revue/XXVI-1/articles/02-hudon.html

Hunter, R. (1991). *Successes and failures of patrons searching the online catalog at a large academic library: A transaction log analysis*. Reference Quarterly 30, 395--402.

Hylton J (1996). *Identifying and Merging Related Bibliographic Records.* Master of Engineering thesis, M. I. T. Department of EECS, June, 1996.

Ifla (1961). *Conférence Internationale sur les principes de catalogage*- Bulletin des bibliothèques de France, 1961

Ifla (1996). *Functional requirements for bibliographic records: draft report for world wide review*, [En ligne]. Adresse Internet: http://www.nlc-bnc.ca/ifla/VII/s13/frbr/frbr-toc.htm

Ihadjadene M (1998a). *L'accès sujet dans les catalogues en lignes: le cas des bibliothèques universitaires en France*. Bulletin des Bibliothèques de France , vol 34 n°4, 1998. PP. 104-110

Ihadjadene M (1998b). *Searching and surfing a WWW-OPAC: the case of remote users* - In 26th annual Canadian Association of Information Science conference (CAIS/ACSI), Elaine Toms (eds), Ottawa juin 1998, pp 303-318.

Ihadjadene M (1998c) . *Les tables de matières et les OPACs*, Bulletin de l'ABF, Mars 1998.

Ihadjadene M & Roger D (1998d). E*valuation ergonomique de ARTICLE-INIST*, Rapport d'étude (CNRS, INIST 1998).

Ihadjadene M (1999a). *Les échecs et la surcharge d'information dans un SRI: examen des tactiques mises en œuvre par les usagers*. Proceedings of the 27th Annual CAIS/ACSI Conference , Turner (eds), University of Sherbroke, 9-11 june, 1999, Canada.

Ihadjadene M (1999b). *The examination of the display of hyperlinks in the WWW-OPACs* EOCONSID'99 , 22-24 april, University of Grenade , Spain. Pp 225-229.

Ingwersen, P. (1984). *A cognitive view of three selected online search facilities*. Online Review 8, 465--492.

Ingwersen, P. (1992). *Information retrieval interaction*. London, England: Taylor Graham.

Ingwersen, P. (1996). *Cognitive perspectives of information retrieval interaction. Elements of a cognitive IR theory*. Journal of documentation, 52 (1) pp 3-50.

Iyer H (1995). *Classificatory structures*, Indeks Verlag, 1995

Johnson, P. Cochrane (1995). *A Hypertextual Interface for a Searcher's Thesa*urus, Digital Libraries '95 Proceedings, Austin, TX June 11-13, 1995, pp. 77-86.

Jones, S. Gatford, M., Hancock-Beaulieu, M., Robertson, S.E., Walker, and Secker, J. (1995) . *Interactive thesaurus navigation with intelligence rules*. Journal of the American Society for Information Science 46, 1 (January 1995), 52-59.

Jones, S. Cunningham SJ (1998). *An analysis of usage of a digital library*. Proceedings of the 2th european conference for digital library, Crete, pp 261-277.

Kalin, S (1991). *The searching behavior of remote users: a study of one OPAC*. Proceedings of the ASIS annual meeting, vol 28, pp 178-185.

Kantor, P.B.(1993). *Development of An Adaptive Network Library Interface: Progress Report and System Design Issues*. In Proceedings of the 56th ASIS Annual Meeting, Bonzi, Katzer, Kwasnik, Eds., 30, 211-216

Kern-Simerenko, C. (1983). *OPAC user logs: Implications for bibliographic instruction*. Library Hi Tech 1, 27--35.

Khoo, C (1997). *Subject access in online catalogs. Encyclopedia of library and information science*, Vol 60 (1997), p. 324-340

Khoo, C (1998). *Development of search strategies for E-referancer, an Expert system Web interface to Online catalogs.* In 26th annual Canadian Association of Information Science conference (CAIS/ACSI), Elaine Toms (eds), Ottawa juin 1998; pp 319-338.

Kolmayer E N (1997). *Contribution à l'analyse des processus cognitifs mis en jeu dans l'interrogation d'une base de données documentaires*. Thèse de doctorat. Université de Paris5.

Korfhage, R (1997). *Information Storage and Retrieval*. New York: Wiley, 1997.

Kuhlthau C C (1993). Seeking meaning: a process approach to library and information services. Norwood NJ:Ablex 1993

Lahary D (1997). *Que faire de RAMEAU ?.* Bull de l'ABF. 174. 60-63

Lancaster, F, W (1993). *Information Retrieval Systems*, 2nd ed. New York: Wiley, 1993.

Larson, R.R. (1986). *Workload characteristics and computer system utilization in online library catalogs*. Doctoral dissertation, University of California at Berkeley

Larson, R.R. (1991). *The decline of subject searching: Long-term trends and patterns of index use in an online catalog*. Journal of the American Society for Information Science 42, 197--215.

Larson, R.R. (1992). *Evaluation of retrieval techniques in an experimental online catalogs*. Journal of the American Society for Information Science 43(1), 34-53.

Larson R. (1996). Cheshire II: designing a next generation online catalog. Journal of the american society for information science, Vol 47 (7) 1996 - p.555-567

Leazer G (1993). *A conceptual plan for the description and control of bibliographic works.* DLS dissertation, Columbia University ; 1993.

Le Coadic Y F (1993) . *Histoire des sciences et histoire de la science de l'information*. Documentaliste et science de l'information ; Vol 30, n°4-5, pp205-209.

Le Coadic Y F (1998) .*Le besoin d'information*. Adbs édition, 1998.

Le Loarer P (1993). *OPAC: Opaque or open, public, Accessible and Co-operative? some developments in natural language processing*. Program - Vol 27 (1993)- p.251-268

Le Marec, J (1989). *Dialogue ou labyrinthe? : la consultation des catalogues informatisés par les usagers*, Paris : Bibliothèque publique d'information.

Le Plat J (1997). *Regards sur l'activité en situation de travail*. Puf, 1997.

Liddy E d (1993). *Reality check! book index characteristics that facilitate information access*, Dans N C Mulvany (Ed) Indexing provinding access to information: Looking back, looking ahead: Proceeding of the 25th Annual meeting of the american Society of Indexers. 1993 Port Aransas Tx: ASI, 125-138

Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge: Cambridge univ. press, 1995. 224p.

Markey K (1978). *Online training and pratical manual for eric data base searchers*. Syracuse University, 1978.

Markey, K. (1983). *The Process of Subject Searching in the Online Catalog*: Final Report of the Subject Access Research Project (OCLC Research Report Number OCLC/OPR/RR-83/1). OCLC, Inc., Dublin, Ohio.

Markey, K. (1984). *Subject Searching in Library Catalogs: Before and After the Introduction of Online Catalogs*. OCLC, Inc., Dublin, Ohio, 1984.

Markey, K. (1989). *Integrating the machine-readable LCSH into online catalogs.* Information Technology and Libraries 33, 299-312.

Markey K (1990). *Analysis of a bibliographic database enhanced with a library classification.* Library Resources and Technical Services 34, 179--198.

Markey K.B; Vizine G (1994). *Using subject heading for online retrieval:theory, practice and potential*- Edition du San Diego:Academic Press. 1994.

Markey K ; Weller M.S (1996) . *Failure analysis of subject searches in a test of a new design for subject access to online catalogs.* Jasis 47(7)1996, pp 519-537.

Matthews, J.(1982). *A Study of Six Online Public Access Catalogs: A Review of Findings.* Final report for the Council on Library Resources. Washington, D.C.: Council on Library Resources. (ED 231 389).

MeadowsC T (1992) . *Text information retrieval systems*. Academic press Toronto.

Meadows C T (1999). *Information retrieval: a view of its past, present, and future*. In 27th annual Canadian Association of Information Science conference (CAIS/ACSI), Turner (eds), Sherbrooke, 1999.pp 190-197.

Meunier J G (1997). La lecture et l'analyse du texte assistée par ordinateur comme système de traitement de l'information. BULAG. Année 1996-1997, Num 22, 211-232.

Micco, M. (1991). *Dealing with the problem of very large retrieved sets. Alternatives to `Brute Force' keyword searching.* ASIS '91. Proceedings of the 54th Annual Meeting of the American Society for Information Science, October 27--31, 1991. Learned Information, Medford, New Jersey.

Millsap, L & Ferl E T (1993). *Search pattern of remote users: an analysis of opac transaction logs.* Information Technology and Libraries 54 : 81-98.

Moss M (1997): "reference services for remote users" , Katharine Sharp Review, N°5 1997 (disponible à l'adresse: http://edfu.lis.uiuc.edu/review/5/moss.html).

Mucchielli (1995) . *Les Sciences de l'Information et de la Communication*, éd. Hachette, 1995

Nelson M (1991). *The design of a hypertext interface for information retrieval*. The canadian journal of information science. Vol 16, n 2, 1991. p. 1-12

Nielsen I H (1993). *Monitoring OPACs in the Nordic Technological University libraries*. Nordinfo Publication N°23, Danemark

Nordlie R (1996). *Unmediated and Mediated Information Searching in the Public Library***.** [En ligne]. Adresse Internet : http://www.asis.org

Nordlie R (1999). *User revealment - a comparison of initial queries and ensuing question development in online searching and in human reference interactions*. In SIGIR'99, University of California (à parraitre)

Osmont B (1995). *Dynamiques cognitives et stratégies d'utilisateurs*. Edition Masson Paris 1995

Perriault J (1989). *La logique de l'usage*. Edition Flammarion, 1989

Peters, T.A. (1989). *When Smart People Fail: An Analysis of the Transaction Log of an Online Public Access Catalog.* Journal of Academic Librarianship 15, 267--273.

Peters T (1993). *The history and development of transaction log analysis.* In Library Hi Tech - vol 42, 1993. PP 41-66.

Pollard R (1993). *A Hypertext-Based Thesaurus as a Subject Browsing Aid for Bibliographic Databases*. IPM vol 29 (3)pp 345-356.

Polity Y (1990). Recueil de dialogue homme-machine en langue naturelle écrite. Cahiers du CRISS, n°17.

Puget D (1993). *Aspects sémantique dans les SRI*. Thèse de doctorat en informatique , Université Paul Sabatier Toulouse 1993.

Pollitt A S (1986). *CANSEARCH: An expert systems approach to document retrieval*. Information Processing and Management 23, 2, 1987, 119-138.

Pollitt, A.S (1988). *A common query interface using MenUSE – A Menu-Based User Search Engine*. Proceedings of the 12th International Online Information Meeting, 445-457.

Pollitt, A. S. (1997*).* *The key role of classification and indexing in view-based searching IFLA '97* Copenhagen Aug 31 - Sept 3 1997. 63rd IFLA General Conference Booklet 4, Section on Classification and Indexing Session 95 Paper 009-CLASS-1-E.

Pollit, A.S & Tinker A (2000). *Navigating N-dimensional information space with data and documents through view-based searching*. BCS-IRSG 2000 Meeting , Cambridge (à paraître)

Provensal A (1997) in Silem A .*Dictionnaire encyclopédique de l'information et de la documentation*. Nathan (eds), 1997.

Renoult D (1994). *Les bibliothèques dans l'université*. Cercles de la librairie (eds), 1994

Robertson SE, Walker S (1994). *OKAPI at TREC 2*.In Harman (eds) TREC2, 21-34.

Ryon V (1997*).* *Le dialogue homme-machine. De l'utilisateur à l'ordinateur*. BULAG. Année 1996-1997, Num 22, 199-210.

Salton G & McGill M (1983). *Introduction to modern Information Retrieval*. – New York : McGraw-Hill Book Company, 1983. – 448 p

Savoy J (1992). *Filtrage de l'information dans les hypertextes*. Publication N°818, Université de Montréal, département d'informatique, 21pages.

Saracevic, T. (1996). *Relevance reconsidered*. In Ingwersen P,eds. Colis2; 1996 October 13-16; Copenhagen, Danmark; PP 201-218

Saracevic, T., Spink, A. & Wu, M.M. (1997). *Users and intermediaries in information retrieval: what are they talking about?* In Jameson, A., Paris, C. & Tasso, C. (Eds.). User modeling: Proceedings of the Sixth International conference, UM97 (pp. 43-54). Wien: Springer, 1997

Saracevic T, H. Mokros, L. Su (1991). *Interaction Between Users and Intermediaries During Online Searching*. Proceedings of the 12th Annual National Online Meeting, 12:329-341, 1991

Saracevic, T (1997a). *Extension and Application of the Stratified Model of Information Retrieval Interaction*. Proceedings of the Annual Meeting of the American Society for Information Science, 34:313-327, 1997.

Saracevic, T. (1997b). *Users lost: Reflections on the past, future, and limits of information science.* SIGIR FORUM, 31 (2) 16-27.

Seamana, S. (1992). *Online catalog failure as reflected through interlibrary loan error requests.* College and Research Libraries 53, 113--120.

Schamber, L. (1994). *Relevance and information behavior.* In M. E. Williams (Ed.), Annual Review of Information Science and Technology. 29 (pp. 3-48). Medford, NJ: Learned Information.

Schamber, L., Eisenberg, M. B., & Nilan, M. S. (1990). *A re-examination of relevance: Toward a dynamic, situational definition*. Information Processing and Management, 26 (6), 755-776.

Shneiderman, B. (1992). *Designing the user interface: strategies for effective human-computer interaction*. 2nd ed. New York: Addison-Wesley, 1992. 560p

Shenouda, W. (1990). *Online bibliographic searching: How end-users modify their search strategies*. ASIS '90. Proceedings of the 53rd Annual Meeting of the American Society for Information Science 27, 117--128.

Silverstein C, Marais H (1998). *Analysis of a very large Altavista query log*. SRC technical Note 1998-014. Digital, Palo Alto.

Sinkakas G (1977). *A study of the syndetic structure of the LCSH*, Phd thesis, Pittsburg University, 1977.

Sloan B G (1991). *Remote Access : design implication for the online catalog* . Cataloging and classification quarterly : 133-140.

Smail M (1998). *Vers des systèmes évolutifs de recherche d'informations : un état de l'art*. Techniques et Sciences informatiques, vol.17, n°10, 1998, pages 1193 - 1222.

Smiraglia R (1992). *Authority control and the extent of derivative bibliographic relationships*.PHD thesis. University of Chicago.

Smiraglia R (1999). *Derivative biliographic relationship: the work relationship in a global bibliographic database*. JASIS 50 (6) 1999 pp 493-504.

Smith M P (1995). *The Effectiveness of Document Ranking and Relevance Feedback Techniques in a Thesaurus-based Search Intermediary System*. PhD Thesis, University of Huddersfield, November 1995.

Snelson Pamela (1993). " Relationships between access and use in information systems: remote access to and browsing of online catalogs". Proceeding of the ASIS Annual Meeting 30 : 73-80.

Spink and T. Saracevic (1997). *Interaction in Information Retrieval: Selection and Effectiveness of Search Terms*. Journal of the American Society for Information Science, 48(5): 382-394.

Spink, A (1997). *Term Relevance Feedback and mediated database searching: implication for information retrieval practice and systems design*. Information processing and management. 31 (2) : 161-171.

Spink, A. & Losee, R.M. (1996). *Feedback in information retrieval*. In: Williams, M. ed. Annual Review of Information Science and Technology, 31, 33-78.

Spink, A., Goodrum, A. & Robins, D. & Wu, M.M. (1996). *Search intermediary elicitations during mediated online searching*. Proceedings of the 19th Annual ACM/SIGIR Conference on Research and Development in Information Retrieval. 120-127.

Spink, J. Bateman, and B. J. Jansen (1998). *Searching Heterogeneous Collections on the Web: Behavior of EXCITE Users*. Proceedings of the 1998 National Online Meeting, May, New York, 1998

Spoerri, A. (1995). *InfoCrystal: A Visual Information Retrieval Interface*. *In* Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, *Seattle, 367.*

Sugamoto S (1995). *Enhacing usability of networked-based library information system. ACM DL'95*. [En ligne]. Adresse Internet http://www.csdl.tamu.edu/DL95/papers/sugamoto/

Theureau J (1994). *Cours d'action et conception d'un système d'aide*. in Sociologie du Travail. 4, 547-585.

Thomazo L (1997). *L'anté-serveur documentaire*. Thèse, université de Caen, 1997.

Thompson, R., Shafer, K., and Vizine-Goetz, D (1997). *Evaluating Dewey concepts as a knowledge base for automatic subject assignment*. 2nd ACM International Conference on Digital Libraries. Disponible à http://purl.oclc.org/scorpion/eval_dc.html.

Tillet B (1989). *Bibliographic relationships: towards a conceptual structure of information used in cataloguing*- 1987. PHD thesis. University of california.

Tinker A J , Pollit A & Braekevelt (1999). *The Dewey Decimal Classification and the transition from physical to electronic knowledge organisation*. Knowledge organisation, 26 (2): 80-96.

Tolle, J.E., and Hah, S. (1985). *Online Search Patterns: NLM CATLINE Database*. Journal of the American Society for Information Science 36, 82--93.

Tonta, Y.A. (1992). *An analysis of search failures in online catalogs*. Dissertation, School of Library and Information Studies, University of California, Berkeley.

Turtle H.(1991). *Evaluation of an Inference Network-Based Retrieval Model*. ACM Transactions on Information Systems, 9(3), 187-222.

Twidale M, D.M. Nichols(1998). C*omputer Supported Cooperative Work in Information Search and Retrieval* (1998) *In* Annual Review of Information Science and Technology, 33, 259-319,(ed.) Williams, M.E., Medford, NJ: Information Today Inc., ISBN: 1-57387-065-X.

Van Rijsbergen C J (1979). *Information Retrieval*, 2nd ed. London.

Vickery H.M. Brooks (1987). *PLEXUS : the expert system for referral*. Information Precessing and Management 23, 2, 1987, 99-117

Villame (1994). *Modélisation des activités de recherche d'information dans les bases de données et conception d'une aide informatique*. Thèse, Université de Paris 13, 1994.

Vellucci S (1997). *Bibliographic relationships*. International Conference on the Principles and future development of AACR, Toronto, Canada, October 23-25, 1997; pp 1-41.

Victorri B (1999). *Traitement automatique des langues et recherche documentaire*. Revue d'Interaction homme-machine. Vol (1-2), 25-36.

Vizine-Goetz D. 1989. *Bibliographic Relationships: Implications for the Function of the Catalog*.In The Conceptual Foundations of Descriptive Cataloging, ed. Elaine Svenonius, 167-179. San Diego, Ca.: Academic Press, Inc.

Walker, S. (1990). *Interactional aspect of a reference retrieval system using semi-automatic query expansion***.** In Informatics 10: Prospects for Intelligent Retrieval, pp 119--136. Aslib, London.

Wallace, P (1993). *How Do Patrons Search the Online Catalog When No One's Looking? Transaction Log Analysis and Implications for Bibliographic Instruction and System Design*" RQ, v. 33, no2 (Winter 1993), pp. 239-252.

Wang, P. (1994). *A Cognitive Model of Document Selection of Real Users of Information Retrieval Systems*. Unpublished doctoral dissertation, University of Maryland.

Wang, P., & White, M. D. (1995). *Document use during a research project: A longitudinal study*. Proceedings of the 58th ASIS Annual Meeting (pp. 181-188) Medford, NJ: Learned Information, Inc.

Wiberly, S.E., and Dougherty, R.A. (1988). *Users' persistence in scanning lists of references.* College and Research Libraries 49, 149--156.

Wiberley, Stephen E., R A Daugherty (1990). *User Persistence in Scanning Postings of a Computer-Driven Information System : LCS*." Library & Information Science Research, v. 12, (341-353), (1990), pp. 341-353

White, Howard D., and Katherine W. McCain (1998). Visualizing a Discipline: An Author Co-citation Analysis of Information Science, 1972-1995. Journal of the American Society for Information Science , 49: 327-355.

Wildemuth, B.(991). *A detailed analysis of end-user search behaviors.* American Society for Information Science. Proceedings of the ASIS Annual Meeting, *1991* 29, 302--312.

Wilson T D (1999). *Models in information behavior research*. Journal of documentation, 55 (3), pp 249-270.

Wilkinson R (1997). *Similarity measures for short queries*. In TREC-4, PP 277-285, 1995.

Wolfram D (1996). *The effect of linkage structure on retrieval performance in a hypertext-based bibliographic information retrieval system*. Information processing and management 32 (5) : 529-541.

Yee I (1998). *Search tactics of Webs users in searching for texts, graphics, known items and subject.* In , Elecronic ressources: use ans user behavior; Hemalata Iyer(eds), The haworth press 1998, pp 61-83.

Yee, M. Shatford L S (1998) .*Improving online public access catalogs*, Chicago : American Library Association, 1998.

Zink, S (1991). Monitoring User Search Success Through Transaction Log Analysis : the WolfPac Example.Reference Services Review, v. 19, no 1 (Spring 1991), pp. 49-56.

Zizzi M (1996). *Cartes dynamiques interactives : une métaphore spatiale pour l'exploration des espaces informationnels*. Thèse, Université de Paris-Sud, Orsay 1996

ANNEXES

Annexe 1 : répartition des points d'accès dans les catalogues de deuxième génération

	Titre (en %)	Auteur (en %)	Sujet (en %)	mot-clé (en %)	Côte (en %)	booléen (multi-critère)	autres points d'accès
(Peter, 1989) N= 13,258	34,3	23,2	38,8	Nsp	1,1	1	0,9
(Hunter, 1991) n=3,707	25,5	21,40%	51,8	nsp	0,3	Nsp	1
(Zink, 1991) n=6,118	19,32	13,35	49,33	Nsp	Nsp	Nsp	18
(Wallace, 1993) n=4,134	24,2	21,7	Nsp	53,1	0,6	Nsp	0,4
(Ballard, 1994) n=88, 659	31,1	21	35,5	8,3	Nsp	Nsp	4,1
(Larson, 1996) n=42, 668	39,48	18,72	17,16	21,42	Nsp	Nsp	Nsp

Annexe 2: questionnaire sur les bibliothèques universitaires

SECTION I: Questions d'ordre général

1) Nom de la bibliothèque:

Adresse:

Tel:

Responsable:

2) Quel est le type de votre catalogue?

a) Catalogue papier (accès manuel) ☐

b) Catalogue en ligne (OPAC) ☐

c) Les deux ☐

Si c'est un catalogue en ligne (OPAC), reportez vous à la question 6 et suivantes

Si c'est un catalogue avec accès manuel répondez aux questions 3,4,5.

SECTION II: Système Manuel

3) Quel genre de catalogue avez vous?

a) Catalogue dictionnaire (titre) ☐

b) Catalogue dictionnaire (auteur) ☐

c) Catalogue dictionnaire (matière) ☐

d) Catalogue systématique (code de classification) ☐

e) Autres ☐

Précisez.....

4) Avez-vous un fichier matière (index sujet) ? oui ☐ non ☐

5) Si oui, quel langage documentaire utilisez vous?

a) Rameau ☐

c) LCSH ☐

b) Mesh ☐

d) Autres ☐

Précisez.....

SECTION III: Veuillez compléter les sections suivantes si vous avez un système automatisé.

6) Votre système est -il

a) Commercial ☐

b) Interne (maison) ☐

7) Si c'est un logiciel commercial, lequel:.....

8) Depuis quelle date votre OPAC est il opérationnel ?.....

9) Quel est le volume du fonds de la bibliothèque ?

10) Quel est le pourcentage de notices bibliographique informatisées de votre fonds ?

11) Quel type de documents couvre votre OPAC? (choix multiple)

a) Monographies ☐

b) Conférences ☐

c) Thèses et mémoires ☐

d) Périodiques ☐

e) Microfiches ☐

f) Logiciels ☐

g) Images animés (films, vidéo...) ☐

h) Documents sonores (K7, CD ROM...) ☐

i) Documents électroniques (numérisées) ☐

j) Autres ☐

Lesquels .....

12) Quel est le format de vos notices bibliographiques ?

a) Intermarc ☐
b) Unimarc ☐
c) LC Marc ☐
d) Autres ☐

Lesquels.....

13) Quels sont les points d‘ accès disponibles dans votre OPAC?

a) Titre ☐
b) Auteur (Nom personne) ☐
c) Auteur (Collectivité) ☐
c) Auteur/Titre ☐
d) Sujet (feuilletage alphabétique) ☐
e) Mots du titre ☐
f) Mots du sujet ☐
g) Mots des table des matières ☐
h) Indice de classifications ☐
i) Autres ☐

Précisez.....
.....

14) Quel est le mode de diffusion de votre catalogue à l’extérieur ?

a) WWW ☐
b) Gopher ☐
c) Wais ☐
d) Telnet ☐
e) Minitel ☐
f) CD-Rom ☐

15) Votre OPAC est-il en réseau avec d’autres OPACs ? oui ☐ non ☐

16) Si oui

Précisez.....

.....

SECTION IV: Stratégies d'indexation sujet

17) Comment votre base est-elle alimentée?

a) Par dérivation des notices à partir d'une source extérieure ☐

b) Par un catalogage direct (catalogage en saisie local) ☐

c) Les deux ☐

18) Dans le cas d'un catalogage dérivée quelles sont les sources que vous utilisez:

OCLC ☐

SIBIL ☐

BN OPALE ☐

Autres.....

18) Dans le cas d'un catalogage dérivée, comment obtenez-vous l'index sujet ?

a) Utilisation des champs Marc 6xx des notices acquises (dérivées)? ☐

b) Autres moyens ☐

Précisez.....

.....

.....

19) Dans le cas d'un catalogage direct, indiquer la méthode appropriée (choix multiple):

a) Ajouter des termes de Rameau aux notices bibliographiques ☐

b) Ajouter des termes de thesaurus aux notices ☐

Lequel:.....

c) Ajouter des termes libres aux notices ☐

d) Ajouter des termes à l'index « maison » qui est maintenu dans le système? ☐

e)Autres.....
.....
.....

20) Prévoyez-vous d'améliorer l'accès sujet de votre OPAC?

oui ☐ non ☐

21) Pourquoi

.....
.....
.....
.....

SECTION V: Ces questions concernent l'exécution d'une recherche sujet dans votre OPAC.

22) la recherche par indice de classification est-elle disponible dans votre OPAC?

oui ☐ non ☐

23) Si oui, spécifier le système de classification utilisé:

.....

24) Est -il nécessaire pour l'usager de consulter des listes ou index imprimés afin d'effectuer une recherche sujet?

oui ☐ non ☐

25) Si oui, lequel (choix multiples) :

a) Rameau ☐

b) Schéma de classification ☐

lequel.....

c) LCSH ☐

d) Thesaurus ☐

lequel .....

e) Autres ☐

Précisez.....

26) Votre OPAC permet -il une recherche sujet par des opérateurs booléens (ET, OU, SAUF)

oui ☐ non ☐

27) Si oui, veuillez spécifier les caractéristiques existantes:

	ET	OU	SAUF	ET implicite	Autres*:
a Auteur					
b Titre					
c Sujet					
d Ind Classe					
e Mot clé					
f Autres*:					

* autres points d'accès ou autres opérateurs booléens

28) Votre OPAC permet-il la recherche , en utilisant des caractéristiques comme la recherche floue, la troncature, l 'hypertexte ...etc. ?

oui ☐ non ☐

29) Si oui, veuillez cocher les cases ou ces caractéristiques se manifestent?

	Troncature[1]	Termes adjacents[2]	Recherche Pondérée[3]	Hypertexte	Langage naturel	Recherche phonétique	Autres[4]
a Auteur							
b Titre							
c Sujet							
d Indice classe							
e Mot clé							
f Autres*:							

(1) : troncature à gauche, à droite, interne.

(2) : termes adjacents (opérateurs de proximité) présence simultanée et consécutive des termes de recherche

(3): recherche pondérée (pondération des termes de recherche et classement des résultats)

(4) : recherche floue (identification des termes similaires correspondant le plus aux termes de recherche) , recherche par système expert...etc.

*** :** autres points d'accès.

30) Lorsqu'une recherche sujet (alphabétique) échoue, le système permet -t il une conversion en recherche:

a) Mots du titre ☐

b) Titre ☐

c) Mots sujet ☐

d) Mot des tables de matières , notes ☐

SECTION VI: Affichage de l'index et des subdivisions.

31) Votre système permet -il un « feuilletage » de l'index? oui ☐ non ☐

32) Si oui, quel type d'affichage présente votre index? (voir annexe)

a) Affichage alphabétique ☐

b) Affichage structuré [(1)] ☐

(1) affichage des têtes de vedettes, puis l 'affichage groupé des subdivisions puis l'affichage groupé des vedettes avec qualicatifs puis les vedettes avec inversion, enfin l'affichage des phrases longues

33) Votre logiciel permet-il l'affichage des renvois entre les vedettes ? oui ☐ non ☐

34) Si oui, quels types de renvois

a) Voir ☐

b) Voir aussi ☐

c) Terme générique (TG) ☐

d) Terme spécifique (TS) ☐

e) Terme associés (TA) ☐

f) Terme équivalents (TE) ☐

35) Si non, pour quelles raisons, il n'y a pas de renvois dans votre OPAC?

a) Raison budgétaires ☐

b) Manque de personnels ☐

c) Le système ne le permet pas ☐

d) Les différents types de recherches existantes

(par mot du titre, par mot de sujet, feuilletage...)

rendent la provision de renvois redondant ☐

e)Autres.....

.....

.....

.....

36) Pensez-vous qu'il est important d'afficher les renvois pour vos utilisateurs ?

oui ☐ non ☐

SECTION VII: Outils d' aides

37) quelles sont les différentes aides dont dispose l'utilisateur:

a)Panneau d'affichage prés des portes ☐

b)Aide en ligne sur écran ☐

c)Aide mémoire, brochure explicative sur papier ☐

d)Didacticiels ☐

e)Session de formation ☐

f)Disponibilité des bibliothécaires ☐

38)Quelle appréciation portez-vous sur l'accès sujet (matière) de votre catalogue ?

Annexe 3: exemple d'échanges entre usagers et bibliothécaires de la BPI

U : usager

B : bibliothécaire

Exemple N°1

U : « je cherche des livres récents sur la Tunisie »

B : « qu'est ce que vous cherchez ? sur le tourisme ? l'histoire ?

U : « non, plutôt des données économiques mais des livres récents , 82, 85 sont des données dépassées »

B : « faisons une recherche par économie ET Tunisie » (*lecture des résultats*)

B : « 1991, vous avez-vu ? »

U : « c'est déjà plus récent .. cela n'est pas en bas (bibliothèque) »

B : « vous voulez voir s'il y a d'autres ? »

U : « oui, SVP »

B : « vous avez tout ce qui est histoire »

U : « c'est bon »

B : « politique économique, oui ? »

B : « vous avez aussi des revues sur la Tunisie »

U : « oui »

B : « c'est une revue économique en anglais, oui ? »

U : « oui, heu ! relations extérieures, commerce extérieur, ça peut m'intéresser »

B : « relations internationales, c'est politique ? »

U : « alors non, commerce extérieur alors »

B : « vous voulez la consultez ? »

B : « c'est la cote 613, c'est un rayon sur la Tunisie »

U : « et sur le tourisme en Tunisie ? s'il est récent ? »

B : « sur le tourisme en Tunisie ? »

U : « hein ! »

B : « qu'est ce que vous cherchez exactement sur la Tunisie ? un guide ? »

U : « non, mais des données économique sur le tourisme en Tunisie, les entrées, le classement des hôtels, les catégories d'hôtels, etc. »

B : « la comparaison entre pays ? »

U : « non, non ! »

(recherche et lecture de l'index alphabétique sujet »

B : « Afrique ? Afrique- aspects économique ? »

U : « hein ! »

B : « plutôt le Maghreb ? »

B : « le terme afrique-tourisme-tunisie , ça vous intéresse ? » (*affichage des réponses*)

U : « oui, ça mais la date c'est quand ? »

B : « ce livre est un annuaire, il apparaît chaque année, 1994 c'est un peu trop vieux ? »

U : « non, je vais voir dans les rayons »

B : « consultez, ce qui est à cote des rayons aussi »

Exemple N°2

U : « est ce qu'ici que je trouve des informations sur les métiers ?

B : « il y a plusieurs sources d'informations » *(le bibliothécaire les cite avec les cotes*)

B : « maintenant si vous dites métier, ça veut dire recherche d'emploi ?

U : « oui, hein non, je vais préciser ma question. Il existe un fichier des entreprises informatisés qui permettraient de retrouver à partir de l'activité du service, le type d'entreprise. Quelles sont les types d'entreprises qui distribuent des revues de presse sur mesure (économie, éducation,…). J'ai cherche dans presse, et pourtant je l'ai trouvée une fois dans les pages jaunes. »

B : « Regarder le Compass (produit et services). Il y a aussi l'annuaire de la presse et de la Pub. Est ce que vous l'avez déjà consulté ? »

U : « non, …il donne tous. »

Consultation du catalogue

B : « les adresses de journaux ? »

U : « c'est ce que recherche. »

Requête= annuaire de la presse

U : « c'est des annuaires papiers… y'a pas une version informatisée ?

B : « oui, compasse sur CD-Rom »

Le bibliothécaire indique différents cotes et lui indique autre centres sucpetibles de répondre à sa demande.

Exemple N°3

U : « Bonjour, je cherche des livres sur la communication interne ?

B : « dans quel domaine ? l'entreprise ? »

U : « oui »

Recherche sujet = communication entreprise

B : « si vous voulez, essayez de voir l'indication du rayon, la cote ? »

U : « oui, ça me permettra de chercher »

B : « un rayon qui concerne la communication et l'information dans l'entreprise ? »

U : « la communication interne, fait-elle partie ? »

B : « 658.8, je vois dans la liste des titres proposés est ce qu'il correspond à la cote. C'est bien ça. »

Le bibliothécaire, lui des indications comment trouver le rayons.

Exemple N°5

U : "je fais une étude sur les jeunes et la société"

B : " quels types d'informations?"

U : " des résultas de l'Insée, des choses comme ça, sur la peur des jeunes?

B : " vous voulez des sondages?

U : « oui, des résultats de sondages … anciens si c'est possible ? »

B : « c'est un peu difficile car nous avons tout ce qui est nouveau. »

U : « il y'aura peut être des comparaisons avec les anciens chiffres sous formes de tableaux. »

Recherche dans le catalogue

B : « j'essaie avec sujet= sondage et jeunes. Nous avons aussi un rayon sociologie ou on peut trouver ce genre d'informations.

(pas de réponse)

B : « j'essaie avec statistique et jeunes . Vous voulez les (*résultats*) regarder ? »

U : « oui »

B : « C'est en 308 , rayon sociologie »

(examen et visualisation des réponses)

B : « vous avez des titres en anglais, ça vous intéresse ?

U : « non, juste en France. »

Le bibliothécaire lui explique comment lire une cote et lui indique le rayon correspondant.

Annexe 4: questionnaire utilisé dans l'analyse des stratégies de navigation des étudiants de Lyon2

Pré-questionnaire

1. Utilisez régulièrement un micro-ordinateur ou un minitel (au moins 10 fois par an) ?
    a. OUI
    b. NON
2. Avez-vous l'habitude de fréquenter la bibliothèque ?
    a. Plus de trois fois par semaine
    b. Une à deux fois par mois
    c. Moins d'une fois par moins
    d. Pratiquement jamais
    e. C'est la première fois
3. Utilisez-vous fréquemment le catalogue de la BU ?
    a. Plus de trois fois par semaine
    b. Une à deux fois par semaine
    c. Une à deux fois par mois
    d. Moins d'une fois par moins
    e. Pratiquement jamais
    f. C'est la première fois
4. Utilisez-vous fréquemment les catalogues d'autres bibliothèques ?
    a. Plus de trois fois par semaine
    b. Une à deux fois par semaine
    c. Une à deux fois par mois
    d. Moins d'une fois par moins
    e. Pratiquement jamais
    f. C'est la première fois
5. Utilisez-vous régulièrement des interfaces WWW ?
    a. Oui
    b. Non
6. Utilisez-vous des WWW-OPACs.

a. Oui
b. Non

7. Quel est votre niveau de formation
    a. Niveau Deug
    b. Licence ou maîtrise
    c. 3°cycle
    d. enseignant
    e. bibliothécaire
    f. autres
8. Dans quelle discipline ?
9. Vous consultez le catalogue pour :
    a. Rechercher des documents
    b. Juste pour voir comment fonctionne le catalogue
    c. Autres cas . Préciser
10. Quel type de documents cherchez-vous ?
    a. Tous type de documents
    b. Livres
    c. Articles
    d. Logiciels
    e. Autres cas
11. Recherchez-vous dans le catalogue :
    a. Un ou plusieurs documents précis
    b. Un ou plusieurs auteurs précis
    c. Un ou plusieurs sujets qui vous intéressent
    d. Autres cas
12. Connaissez-vous déjà des livres, revues…sur le même sujet ou auteur ?
    a. Oui
    b. Non
13. Décrivez assez précisément ce que vous voulez rechercher ?
14. est ce que votre recherche porte sur un aspect
    a. très spécifique
    b. très large
    c. autre

Post-questionnaire :

15. En utilisant la stratégie BRF, votre intention était de :
    a. Elargir votre question
    b. Affiner votre question
    c. Autre cas
16. Avez-vous trouvé ce que vous cherchiez ?
    a. Rien du tout
    b. Peu de chose
    c. Ce que j'attendais
    d. Plus que ce que j'attendais
    e. Je ne sais pas
17. Etes-vous satisfait ?

Pas vraiment 1 2 3 4 5 très satisfait

18. quelles informations avez-vous retenues ?
    a. une référence
    b. plusieurs référence
    c. localisation des documents
    d. autre cas
19. Qu'allez-vous faire maintenant ?
    a. Consulter et emprunter un ou plusieurs documents
    b. Jeter un coup d'œil en rayon
    c. Demander de l'aide à un bibliothécaire
    d. Faire une autre recherche dans le catalogue
    e. Quitter la bibliothèque
    f. autre

Annexe 5 : questionnaire relatif à l'accès à distance

1) Etes-vous ?

Une femme

Un homme

2) Vous êtes actuellement:

Etudiant

Documentaliste

Bibliothécaire

Enseignant (e)

Autre (préciser)

3) Si vous êtes étudiant, quel est votre niveau de formation:

Deug

Licence/maîtrise

Troisième cycle

4) Dans quelle discipline ?

5) Utilisez-vous le catalogue de l'ENSSIB :

Pour la première fois

Tout les jours

Toutes les semaines

Tous les mois

Rarement

6) Vous effectuez cette recherche :

Pour un usage professionnel

Pour vos études

7) Vous consultez le catalogue de l'ENSSIB pour :

Rechercher des documents

Juste pour voir comment fonctionne le catalogue

Autres (préciser)

8) Recherchez-vous dans le catalogue :

Un ou plusieurs documents précis

Un ou plusieurs auteurs précis

Un ou plusieurs sujets qui vous intéressent

Autres (préciser)

9) Connaissez-vous déjà des documents sur le même sujet ou même auteur ?

Oui

Non

10) Quel type de documents cherchez-vous dans le catalogue ?

Tous types de documents

Livres

Titres de périodiques

Articles de revues

Thèses/mémoires

Rapports

11) Décrivez assez précisément ce que vous voulez rechercher:

12) Avez vous l'habitude d'interroger des catalogues qui ont une interface WWW ?

Oui

Non

C'est la première fois

13) Vos commentaires:

14) Votre e-mail (facultatif) :