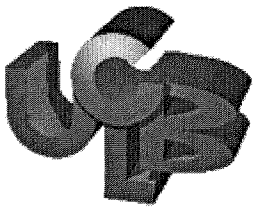


Ecole Nationale Supérieure
des Sciences de l'Information
et des Bibliothèques

Université Claude Bernard Lyon 1
43, boulevard du 11 Novembre 1918
69622 VILLEURBANNE CEDEX

DESS Ingénierie Documentaire

Rapport de recherche bibliographique

Stratégies de collecte et d'indexation de pages Web par les moteurs de recherche : conséquences sur le comportement des utilisateurs

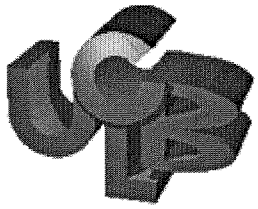
Nicolas Bayon

Sous la direction de

Claude Boisseau

Pôle Bio-Informatique Université Rennes 1
U.F.R. S.V.E
Campus de Beaulieu
35 042 RENNES cedex

Année 1999-2000



Ecole Nationale Supérieure
des Sciences de l'Information
et des Bibliothèques

Université Claude Bernard Lyon 1
43, boulevard du 11 Novembre 1918
69622 VILLEURBANNE CEDEX

DESS Ingénierie Documentaire

Rapport de recherche bibliographique

Stratégies de collecte et d'indexation de pages Web par les moteurs de recherche : conséquences sur le comportement des utilisateurs

Nicolas Bayon

Sous la direction de

Claude Boisseau

Pôle Bio-Informatique Université Rennes 1
U.F.R. S.V.E
Campus de Beaulieu
35 042 RENNES cedex

Année 1999-2000

Stratégies de collecte et d'indexation des pages Web par les moteurs de recherche : conséquences sur le comportement des utilisateurs

Nicolas Bayon

RESUME :

La recherche d'informations sur le Web amorce un tournant dans son mode de fonctionnement. La taille du réseau, la nécessité d'atteindre un niveau de pertinence dans les réponses, les regroupements, la mise en place de portails vont faire évoluer les techniques d'indexation des moteurs de recherche qui sont très disparates. Les utilisateurs de ces outils doivent avoir connaissance de ces différences pour optimiser leurs recherches ou leurs référencements.

DESCRIPTEURS :

Moteur de recherche
Indexation
Collecte
Pages Web
Internet

ABSTRACT :

Information research on the Web prepare a change. Network height, necessity of relevancy, assemblings will make an update in methods for indexing and in Web search engines which are very different. Search engines users have to know these differences in order to optimize their queries and their purchases.

KEYWORDS :

Search engine
Indexing
Collection
Web pages
Internet

TABLE DES MATIERES

I - INTRODUCTION

En raison de son développement rapide et anarchique, Internet est un réseau d'informations sans organisation ni structure, de sorte que la recherche efficace d'information parmi les centaines de millions de pages Web* est possible mais n'est pas chose aisée. Mais pour cela, il faut utiliser les outils appropriés. L'internaute a l'embarras du choix parmi tous les outils de recherche que sont les annuaires, les moteurs et les logiciels. Même si aucune de ces solutions ne peut prétendre à la perfection, elles offrent chacune des possibilités distinctes dues à des modes de fonctionnement très différents.

On s'intéressera dans cette étude aux méthodes de collecte et d'indexation des pages Web par les moteurs de recherche. Dans un premier temps, on verra quelle stratégie de recherche et quelles méthodes ont été utilisées pour mener à bien cette étude. Puis, on mettra en évidence l'état actuel de la situation des moteurs de recherche concernant l'indexation aisi que les perspectives d'avenir dans une note de synthèse. Enfin, la bibliographie ponctuera ce rapport. Les différentes techniques utilisées auront des conséquences sur les résultats exprimés par ces moteurs et par là même modifieront le comportement des utilisateurs dans leur appréhension des outils de recherche.

* On estime aujourd'hui que le Web s'enrichit d'environ un million de pages par jour.

Première Partie
METHODOLOGIE

II – METHODOLOGIE

La recherche s'est décomposée en plusieurs phases :

- La détermination des mots-clés
- La définition des critères de sélection ou de rejet des références
- L'interrogation des différentes sources d'information
- La sélection des références selon les critères définis
- L'élaboration d'une bibliographie la plus cohérente possible
- La sélection des références les plus pertinentes pour la synthèse

II-1 Recherche manuelle

C'est en utilisant les ressources de la bibliothèque que cette première étape a pu être réalisée. Dans les collections des ouvrages de références plusieurs répertoires ont été consultés pour sélectionner des descripteurs et sélectionner des bases de données.

II-1-1 Sélection des descripteurs

L'utilisation de thésaurus pour définir les descripteurs n'a pas eu des résultats très fructueux. En effet, les termes recherchés sont d'une utilisation très récente et sont pour la plupart absents des anciens thésaurus : on pense en particulier à la locution « moteur de recherche » ou aux notions de « pages Web » et « Internet ». Malgré cela, il a été intéressant de définir les autres termes à partir des thésaurus suivants :

- *Thésaurus des Sciences de l'Information : Informascience*. Paris : CNRS ; Bureau de l'IST. 1977.
- *Vocabulaire de la Documentation*. 2e ed. Paris : AFNOR. 1987.
- *Thesaurus of Information Science Terminology.../* sous la dir. de Claire K. Shulz. London : Scarecrow Press. 1978.

Descripteurs retenus

Français	Anglais
Indexation	Indexing
Collecte	Collection
Stratégie	Strategy

Le descripteur "collecte" désigne l'action réalisée par le moteur pur aller chercher de nouvelles pages. Le terme "indexation" est là pour nommer la méthode de classement utilisée par ces mêmes moteurs pour alimenter leurs bases de données.

"Stratégie" est un terme plus équivoque que les deux premiers . En effet, on y retrouve une connotation pouvant induire la présence de documents non pertinents dans les résultats. Mais on peut dire que l'utilisation simultanée de ces différents descripteurs permet d'obtenir des résultats conformes aux attentes.

Pour élargir la recherche, nous avons sélectionné des termes plus génériques qui traduisent les outils ou locutions spécifiques des nouvelles technologies et de l'Internet en particulier.

Termes associés

Français	Anglais
Moteur de recherche	Search engine
Pages Web	Web pages
Internet	Internet

De la même façon que les descripteurs, les termes associés sont très spécifiques et ne portent pas à confusion. Le terme « moteur de recherche » se traduit littéralement et donne « search engine » en anglais.

Mais à l'usage, il faut ajouter quelques bémols à ce premier constat. On a pu remarque dans un premier temps que le terme Web n'était pas forcément le bon descripteur chez certaines bases de données. Il est remplacé parfois par « WWW » ou bien encore « World Wide Web » ou « W3 ».

De plus, le mot « stratégie » n'apparaît que très rarement dans les descripteurs. Ce terme est souvent remplacé par « techniques » ou « méthodes d'indexation ». On parle peu de « stratégie d'indexation » dans les articles.

La stratégie de recherche de documents s'est affinée au fur et à mesure de l'interrogation des bases de données et des moteurs de recherche. Celle-ci consistait au départ à consulter premièrement les bases de données pour ensuite aller vers les autres types d'interrogations. Cette stratégie a ceci de positif qu'elle permet de se procurer un certain nombre de références dès le début des investigations. Ces références constituent une assise pour construire une recherche plus avancée. Pour aller dans le même sens, on peut dire que les premiers documents identifiés fournissent une ensemble de descripteurs auxquels on n'avait pas forcément pensé au tout départ.

II-1-2 Critères de sélection des références trouvées

Les références trouvées par les différents axes de recherche ont été conservées ou éliminées selon les critères de pertinence par rapport au sujet de la recherche.

Ont été conservées :

- Les références générales sur les moteurs de recherche notamment les monographies qui consacrent le plus souvent un chapitre aux méthodes d'indexation des moteurs.
- Les références présentant un état des lieux du fonctionnement, de la couverture, de la pertinence des moteurs de recherche.
- Les références présentant les projets de recherche des principaux acteurs de l'informatique mondiale sur les futurs moteurs de recherche.
- Les références faisant mention des conséquences possibles des changements actuels sur le comportement des utilisateurs des moteurs de recherche.

Ont été éliminées :

- Les références qui font état des résultats de pertinence à des requêtes standardisées dans le but de mettre en place une hiérarchie des moteurs de recherche.

Les références conservées se veulent les plus cohérentes possibles et sont rassemblées dans le chapitre « Bibliographie ».

II-1-3 Sélection et localisation des bases de données

Localisation des bases de données

Des CD-ROM interrogeables à Lyon grâce à la liste des CD-ROM disponibles en Rhône-Alpes édité par l'URFIST de Lyon.

Des bases de données interrogeables sous le serveur DIALOG.

Le critère de choix des bases a été le thème général de ces bases qui devait concerner soit les sciences de l'information, soit l'informatique.

Bases de données bibliographiques

INSPEC

Base de données en ligne accessible via le serveur DIALOG (fichier N°2) ou un site Internet
Période couverte : depuis 1969
Contenu : sciences de l'ingénieur (électronique, informatique, physique...)
Editeur : IEE (Institution of Electrical Engineers)

LISA : Library Information Science Abstract

Base de données bibliographique sur CD-ROM spécialisée dans les sciences de l'information
LISA PLUS en est la version disponible sur CD-ROM
La base de données en ligne est accessible via le serveur DIALOG (fichier N° 61)
Période couverte : depuis 1969 pour la version en ligne et 1981 pour la version CD-ROM
Editeur : Bowker-Saur

PASCAL

Base de données bibliographique sur CD-ROM et en ligne via le serveur DIALOG (fichier N°24)
Période couverte : depuis 1992 pour les CD-ROM
Contenu : sciences et techniques
Editeur : INIST (Institut National de l'Information Scientifique et Technique)

Information Science Abstracts

Base de données bibliographique accessible en ligne via le serveur DIALOG (fichier N°202)
Période couverte : depuis 1966
Couverture : environ 300 périodiques
Contenu : sciences sociales
Utilisation d'un vocabulaire contrôlé

Bases de données spécialisées en Informatique

Computer News Fulltext

Base de données en texte intégral accessible en ligne via le serveur DIALOG (fichier N°674)
Mise à jour hebdomadaire
Contenu : Informatique (technologies, industrie, marchés, réseaux internationaux et locaux, télécommunications…)
Intégralité des publications : Computerworld et Network World
Editeur : IDG Communication

Gale Group Computer Database

Base de données bibliographique accessible en ligne via le serveur DIALOG (fichier N°275)
Contenu : informations sur les industries de l'informatique, de l'électronique et de télécommunications (hardware, software, réseaux,périphériques)
Couverture : environ 20000 enregistrements

Microcomputer Abstracts

Base de données bibliographique accessible en ligne via le serveur DIALOG (fichier N°233)
Contenu : utilisation des micro-ordinateurs en économie, industrie, éducation, bibliothèques et hardware, software avec systèmes d'exploitation, systèmes en ligne, réseaux…
Intégralité des publications de Microcomputer Abstracts
Editeur : Information Today, Inc.

JICST-Japan Science Abstracts

Base de données bibliographique couvrant les publications au Japon concernant les sciences et techniques et la médecine (fichier N°94)
Couverture : plus de 6000 périodiques
Editeur : Japan Science and Technology Corporation, Information Center for Science and Technology (JICST)

Il faut distinguer deux stratégies de recherche distinctes. Une mise en place pour les bases de données en texte intégral telles que Computer News Fulltext et l'autre pour les bases de données bibliographiques.

Interrogation d'une bases de données bibliographique

Avant d'effectuer les interrogations, il était nécessaire de consulter les thésaurus respectifs de chacune des bases et de s'adapter, lors des requêtes, aux termes sélectionnés.
Du fait de la spécificité de la plupart des termes employés pour effectuer les requêtes, les problèmes de synonymie n'ont été que très peu nombreux. Seuls les termes « collecte » et « indexation » pouvaient parfois porter à confusion mais rapportés au contexte de l'Internet, le problème était résolu.

Interrogation d'une base de données en texte intégral

Pour la recherche dans la base de données Computer News Fulltext, il a été nécessaire de restreindre les requêtes à certains champs sous peine de retrouver un trop grand nombre de

références. Un autre problème s'est posé spécifiquement à cette base de données dû au fait que c'est une base à vocation commerciale. De nombreuses références concernant des produits commerciaux apparaissaient, si bien qu'il a fallu éliminer les articles portant sur ces produits commerciaux.
La recherche en texte intégral a pour conséquences l'obtention de résultats beaucoup moins précis dans un premier temps du fait de l'absence de champ « mot-clé » comme c'est le cas dans les bases de données bibliographiques.

II-2 Consultation de fonds spécialisés à Lyon

II-2-I Bibliothèque et centre de documentation de l'ENSSIB

Plusieurs outils ont été interrogés ou consultés :

- le catalogue de la bibliothèque
- le fonds de revues
- les produits proposés tels que le bulletin des sommaires
- les mémoires ou rapports de synthèse bibliographique dont le thème est voisin de celui-ci

II-2-2 Bibliothèque Universitaire de Sciences de Lyon 1

De la même manière qu'à la bibliothèque de l'ENSSIB étaient disponibles :

- le catalogue de la bibliothèque
- le fonds de revues
- certaines bases de données sous forme de CD-ROM

II-3 Recherche sur Internet

La recherche sur Internet a représenté une part très importante du travail. Le sujet s'y prêtant très favorablement, on pouvait s'attendre à retrouver une très grande somme d'informations, qui plus est, de factures très récentes. Ce dernier argument est celui qui plaide le plus en faveur de la consultation d'Internet pour traiter ce type de sujet.

II-3-1 Utilisation des moteurs de recherche

Celle-ci s'est effectuée de deux façons bien distinctes.

Tout d'abord, l'utilisation traditionnelle des moteurs en leur soumettant des requêtes précises, courtes et surtout bien construites. La difficulté a résidé dans l'élaboration de ces requêtes. Celles-ci ont été différentes d'un moteur à l'autre suivant les règles de syntaxe de ces différents moteurs. La règle a été d'utiliser des moteurs généralistes ainsi que des annuaires pour une recherche thématique.

Face à un sujet qui tourne en rond (un moteur de recherche qui cherche des pages sur le fonctionnement des moteurs de recherche), les résultats sont bien moindres que ceux escomptés. En effet, de nombreux sites se proposent d'effectuer des comparaisons entre moteurs au niveau des résultats fournis et ainsi d'évaluer leur pertinence. Mais ils sont beaucoup moins nombreux à se proposer d'expliquer leurs fonctionnements et leurs modes d'indexation. Malgré tout, il existe quelques sites très spécialisés.

Ensuite, il a été fait appel aux pages d'aide des moteurs de recherche. Le plus souvent, ces pages d'aide se trouvent en lien sur la page d'accueil du moteur. Les menus qui sont proposés aux utilisateurs laissent à penser qu'ils auront la possibilité de tout savoir sur leur fonctionnement. Mais là aussi, la déception est au rendez-vous, puisque les pages qui nous sont proposées évoquent seulement des généralités qui sont loin de répondre aux questions sur la collecte et l'indexation des pages Web.

Les moteurs de recherche utilisés ont été :

➢ AltaVista ➢ HotBot ➢ Infoseek ➢ WebCrawler	➢ NorthernLight ➢ Excite ➢ Lycos ➢ Et le annuaires Voilà et Yahoo!

II-3-2 Listes de discussion

De gros serveurs de listes permettent un accès facilité par thèmes aux différentes listes. Cette source d'informations peut apparaître comme secondaire mais il ne faut surtout pas la sous-estimer. Ce système est basé sur l'entraide et la bonne volonté des personnes qui vous répondent et cela donne parfois des résultats surprenants.

Les listes de discussion utilisées mises à contribution ont été :

netplus@sorengo.com Aide à la recherche d'information sur Internet. webonetnews@poplist.net Les dernieres nouvelles du net. webrevue@cec.fr Revue de Web hebdomadaire.	webtheme@coollist.com Conseils, Astuces et Echanges autour du Web. adbs-info@cru.fr L'Association des Professionnels de l'Information et de la Documentation. info@cybion.fr Veille, recherche d'informations, agents intelligents et prospective.

La stratégie de recherche sur Internet a été constituée de deux temps. Tout, d'abord il a fallu s'attacher à consulter les moteurs de recherche et annuaires de façon traditionnelle. Sous la forme de requêtes avancées utilisant les opérateurs booléens mis à la disposition des utilisateurs par les moteurs, on a pu recueillir un certain nombre de réponses pertinentes sous la forme de sites entiers consacrés aux moteurs ou de documents de la presse spécialisée disponibles en ligne.

Dans un second temps, il a fallu explorer le contenu des pages internes des moteurs de recherche. Celles-ci renferment des renseignements sur leur fonctionnement propre. Ces informations sont mises à disposition des personnes voulant des informations dans le but de faire référencer leur site.

II-4 Recherche dans les bases de données et analyse des résultats

II-4-1 Recherche dans les bases de données CD-ROM

Un premier temps a été consacré à l'apprentissage des CD-ROM pour repérer les éléments suivants qui nous permettent d'optimiser les interrogations :

- structure de la base
- consultation des thésaurus, des index, des listes de vedettes matières...
- sélection des descripteurs appartenant au langage contrôlé des différentes bases
- repérage de la syntaxe à utiliser pour les différents opérateurs boléens, les opérateurs d'adjacence, les troncatures
- sélection des champs permettant d'éviter de faire une interrogation en texte libre, source de bruit

Les équations de recherche ont été conservées et seront adaptées au langage d'interrogation de DIALOG pour les bases de données qui existent sous les versions CD-ROM et en ligne.

Les résultats obtenus ont d'abord été visualisés en format court pour éliminer les réponses non pertinentes puis en format long et enfin conservés dans des fichiers ou imprimés. Un travail de dédoublonnage assez important a dû être effectué

Un exemple d'interrogation ayant eu des résultats particulièrement pertinents. Lors de cette interrogation effectuée à l'ENSSIB, le CD-ROM PASCAL 99 a été utilisé. Ce CD-ROM contient pour information 203641 notices.

Requête : DXA= « search engine » ET DEA= « indexing »

A cette requête, le CD-ROM a trouvé 6 documents en retour. Lors des interrogations sur Pascal, on s'aperçoit que le descripteur « moteur de recherche » ou « search engine » est toujours employé au singulier ce qui permet d'éviter d'effectuer une troncature. Les opérateurs DEA et DXA ont des significations bien précises. En effet, DEA s'utilise devant un mot-clé en anglais (équivalent de DEF en français). Quant à DXA, il devance une expression en anglais (DXF en français).

II-4-2 Consultation des bases de données en ligne

Elle a été réalisée lors des séances de travaux pratiques de recherche documentaire informatisée et lors de séances autonomes à partir du serveur DIALOG.

Les bases de données en ligne ont été utilisées pour la mise à jour des références obtenues dans la recherche sur CD-ROM. La recherche avec les descripteurs retenus ou termes associés a donc été réduite par l'année de publication.

Sous le serveur DIALOG, il a été possible d'utiliser l'option "One Search" qui a permis d'interroger plusieurs bases (INSPEC, Information Science Abstracts et LISA) de façon groupée. Ceci permet d'effectuer une opération de dédoublonnage directe.

On peut prendre un exemple de recherche dans la base LISA (spécialisée dans les sciences de l'information) sous connexion DIALOG.

S1 « search engine ? »
S2 indexing
S3 collect ?
S4 web
S5 (S2 OR S3) AND S1 AND S4

Cette interrogation permet de mettre en évidence les principaux types d'outils qui sont mis à la disposition des utilisateurs par DIALOG. L'expression « moteur de recherche » se situe entre parenthèses. On utilise l'opérateur « ? » pour indiquer une troncature : dans ce cas, on effectue une recherche sur le singulier et le pluriel de « search engine ».Dans la requête S5, on récapitule l'ensemble des requêtes précédentes en fonction de leurs résultats propres. Ici, on demande de croiser les résultats pour obtenir des documents où les termes « search engine » et « web » sont présents simultanément et où sont présents ou l'un ou l'autre des termes « indexing » et « collect ? ».

II-4-3 Tableaux des résultats

Dans les tableaux récapitulatifs, on ne peut pas distinguer des documents traitant exclusivement de la collecte, de l'indexation des pages Web ou des conséquences sur le comportement des utilisateurs. En effet, les documents retrouvés mêlent les trois thèmes le plus souvent.

De la même manière, un certain nombre de documents traitant des résultats obtenus par les moteurs en terme de pertinence à des requêtes, abordent les thèmes recherchés et sont inclus dans les résultats.

La difficulté de la recherche a, en fait, résidé dans le choix des documents qui ne devait pas seulement se faire à partir du titre mais le plus souvent après avoir au moins examiné le résumé en détail. Ainsi, les références bibliographiques ne seront pas classées par thèmes puisqu'elles englobent tout ou partie du sujet de départ.

Tableaux de résultats

Bases de données en ligne	Domaines couverts	Nombre de références pertinentes
✓ INSPEC	Electronique, informatique, physique	30
✓ Computer News Fulltext	Informatique : technologie, industrie, telecom…	15
✓ Gale Group Computer Database	Industrie de l'informatique et de l'électronique	9
✓ Microcomputer Abstracts	Actualité des micro-ordinateurs	3
✓ JICST	Sciences et techniques	10
✓ Information Science Abstract	Sciences de l'information	22

Bases de données sur CD-ROM	Domaines couverts	Nombre de références pertinentes
✓ PASCAL	Domaines scientifiques	35
✓ LISA	Sciences de l'information	26

D'après les résultats obtenus, on s'aperçoit tout d'abord que les documents les plus anciens ne sont pas antérieurs à 1993. Le sujet est d'apparition très récente.

II-4-4 Estimation des coûts

Coûts financiers

- Utilisation de DIALOG : *dialindex :* 148 Fr.
  Interrogation : 450 Fr.
- Photocopies/achats/poste : 150 Fr.

Coûts horaires

- Utilisation DIALOG : 4 h
- Internet : 15 h
- Interrogations CD-ROM : 8 h
- Analyse des résultats : 40 h
- Saisie et mise en forme du document : 20 h

Conclusion sur l'interrogation des bases

Les deux bases les plus pertinentes au niveau des documents fournis ont été LISA et Information Science Abstracts. Mais des articles très intéressants ont été trouvés dans d'autres bases même s'ils n'étaient pas aussi nombreux.

Le principal avantage d'utiliser DIALOG pour trouver des références est la « fraîcheur » des informations. Les interrogations sont rapides même sur DIALOG, mais l'affichage des formats « résumé » ou « texte intégral » est beaucoup plus long. L'ensembles de données a été téléchargé sur une disquette lors de l'interrogation puis traité après déconnexion afin de limiter le plus possible le coût de la connexion.

Par contre le coût de téléchargement des notices est très élevé. On constate que les recherches sur DIALOG coûtent très cher, même si on utilise le format « résumé » pour effectuer une présélection. Cependant, dans un cadre professionnel, ces interrogations sous DIALOG présentent un avantage : celui de pouvoir obtenir immédiatement le texte intégral de la plupart des références retenues, d'où un gain de temps non négligeable.

II-5 Enseignements tirés de la méthodologie

Dans un premier temps, on a pu s'apercevoir de la difficulté de trouver les descripteurs exacts pour chacune des bases consultées. Dans un domaine aussi évolutif que l'informatique et les sciences de l'information (puisque le sujet se situe à cheval sur les deux domaines), la terminologie ne cesse de changer et cela pose des problèmes d'indexation pour ces bases qui ont quelques fois du mal à s'adapter.

Ensuite, le second point à souligner est l'extrême nécessité dans un domaine comme celui-ci, d'avoir des informations les plus récentes possibles. L'existence des moteurs de recherche ne remonte pas très loin dans le temps et les documents trouvés en réponse aux requêtes posées sont de facture très récentes.

Enfin, pour introduire la seconde partie (Synthèse), on peut insister sur le fait que les documents utilisés pour rédiger celle-ci, sont vieux d'au plus, un an. Ceci permet d'avoir une vision actuelle de la situation de collecte des pages Web par les moteurs ainsi que des

problèmes que vont avoir à affronter ces mêmes moteurs dans un avenir proche pour rester l'outil de référence en matière de recherche sur le Web.

On s'attachera particulièrement dans cette étude aux moteurs les plus utilisés et qui sont aussi les plus étudiés ou décrits.

Deuxième Partie

SYNTHESE BIBLIOGRAPHIQUE

III - SYNTHESE BIBLIOGRAPHIQUE

III-1 Les moteurs de recherche sur le Web

Ils font partie des outils de recherche du Web, au même titre que les annuaires ou les meta-moteurs **[23]**.

III-1-2 Définition

Les moteurs de recherche sont des programmes accessibles aux utilisateurs du réseau qui consultent une gigantesque mémoire, où est enregistrée pour chaque mot-clé présent dans les pages d'Internet, une liste de pages contenant ce mot. Un groupe de telles listes est nommé « index »**[1]**.

La création et la maintenance de ces index sont des tâches très lourdes, et c'est toujours un problème redoutable que de déterminer quelle information retourner en réponse à la question d'un utilisateur**[2]**.

III-1-3 Caractéristiques des moteurs de recherche

Les moteurs de recherche se composent de deux parties distinctes **[24]** :

- un logiciel de recherche : c'est un robot propre à chaque moteur qui traverse automatiquement la structure hypertextuelle du Web, collecte les informations nécessaires et suit les liens de proche en proche. Les informations collectées peuvent varier d'un moteur à l'autre mais généralement elles recouvrent l'URL, le titre, les premiers paragraphes, des mots-clés dans le texte ou texte intégral et/ou meta-informations. On appelle ces robots des « crawlers » ou « spiders ». Certains ont même un nom comme le robot d'AltaVista : « Scooter ». Ces informations sont ensuite transmises à une autre station en charge de la sauvegarde et de la gestion de ces données.

- la base de données qui reçoit les informations collectées par le robot. Elle est mise à jour à intervalles réguliers dépendants de la rapidité du robot, ceci pour éliminer les liens morts. Les fonctions principales de la base de données sont :

    - la recherche ou requête
    - l'insertion des données
    - la mise à jour des données
    - la suppression des données

Les réponses trouvées dépendront étroitement de l'indexation faite par le robot et stockée dans la base. C'est la raison pour laquelle, il appartient aux professionnels de l'information et de la documentation de bien connaître les différentes approches documentaires des différents outils de recherche.

La plupart des moteurs proposent une possibilité de soumission manuelle du site à indexer : les robots se basent alors sur l'URL pour visiter ces nouveaux sites sans attendre d'y accéder par un hyperlien.

On peut ajouter à cela que les annuaires, avec pour meilleur exemple Yahoo !, requièrent une intervention humaine pour l'étape d'indexation des pages. Il y a d'abord

soumission des pages à indexer de la part de leur auteur avec proposition de catégorie. Le choix final d'indexation dans tel ou tel thème se fait après visite du site par un spécialiste.

De leur côté, les meta-moteurs n'ont pas de technique d'indexation spécifique puisqu'ils permettent d'effectuer une recherche simultanée sur plusieurs autres outils (MetaCrawler).

III-2 Caractéristiques des données à indexer

Lorsqu'on parle d'indexation **[24]**, il appartient de préciser quelles sont les données à indexer. Sur le Web, l'indexation peut porter sur :

- des textes
- des images
- du son

III-2-1 Les textes

Les données textuelles peuvent prendre plusieurs formes sur Internet selon les formats employés **[2]** :

- les fichiers ASCII
- le format SGML
- le format HTML
- les formats de traitements de texte
- le format RTF
- le format PDF
- le format PostScript
- le format XML…

III-2-2 Les images

Les fichiers images peuvent aussi avoir plusieurs formats :

- GIF
- JPEG
- TIFF…

III-2-3 Les sons

Les fichiers sons de la même manière peuvent avoir pour extensions :

- WAV
- MID
- MP3…

III-3 Les stratégies de collecte et d'indexation

Les principes d'indexation et de présentation des résultats diffèrents entre un annuaire et un moteur de recherche (respectivement utilisation d'un robot et indexation manuelle) **[6]**. De plus, l'indexation porte sur des sites Web pour les annuaires et sur des pages Web pour un moteur. Une comparaison permet de mettre en évidence les propriétés de chacune des méthodes.

III-3-1 Couverture du Web par les moteurs de recherche

On estime aujourd'hui la taille du Web entre 800 millions et 1 milliard de pages. Les textes retrouvés vont de quelques mots à plusieurs centaines de pages pour les plus gros sites. Aucun moteur de recherche n'est exhaustif. La plupart d'entre eux ne dépasse pas 30% d'indexation du Web

Selon les moteurs et leur politique d'indexation, le nombre de documents indexés varie énormément **[33]**.

Tailles des index rapportées par chaque moteur au 3 février 2000

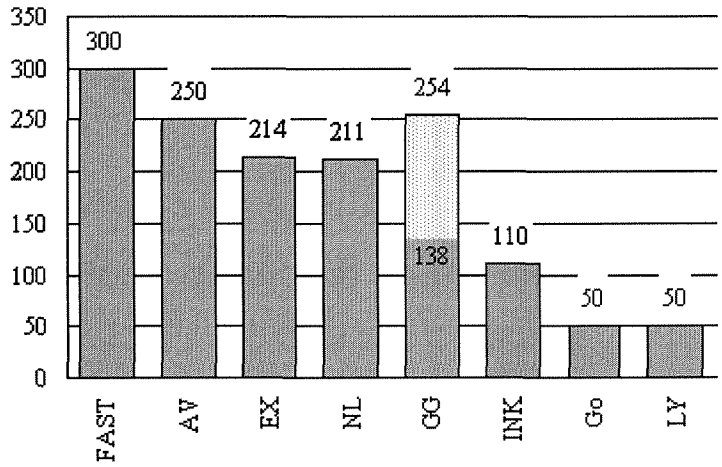


Fig.1

Abréviations : FAST=FAST, AV=AltaVista, EX=Excite, NL=Northern Light, GG=Google, INK=Inktomi, Go=Go (Infoseek), LY=Lycos.

Les utilisateurs qui ont une information très précise et pointue à trouver ont intérêt à sélectionner un moteur avec un gros index. Mais pour des recherches plus générales, un index large ne donne pas de meilleurs résultats **[58]**.
AltaVista annonce un index à 270 millions de pages, Excite : 250 millions, Google : 138 millions de pages enregistrées.

La structuration des données

Sur le Web, les documents sont par nature fortement structurés. En effet la structure logique du langage HTML permet d'indexer de façon méthodique la grande masse d'informations présente**[33]**.

III-3-2 Comparaison indexation automatique/indexation humaine

Processus d'indexation automatique : exemple *AltaVista*

Les moteurs de recherche de type « crawler » effectuent une indexation automatique du Web. Ils le parcourent automatiquement et indexent toutes les pages qu'ils rencontrent. L'indexation peut également s'effectuer par le biais d'un formulaire dans lequel le responsable du site indique au moteur l'existence de son service. On effectue ici une comparaison des deux méthodes.

Processus d'indexation automatique

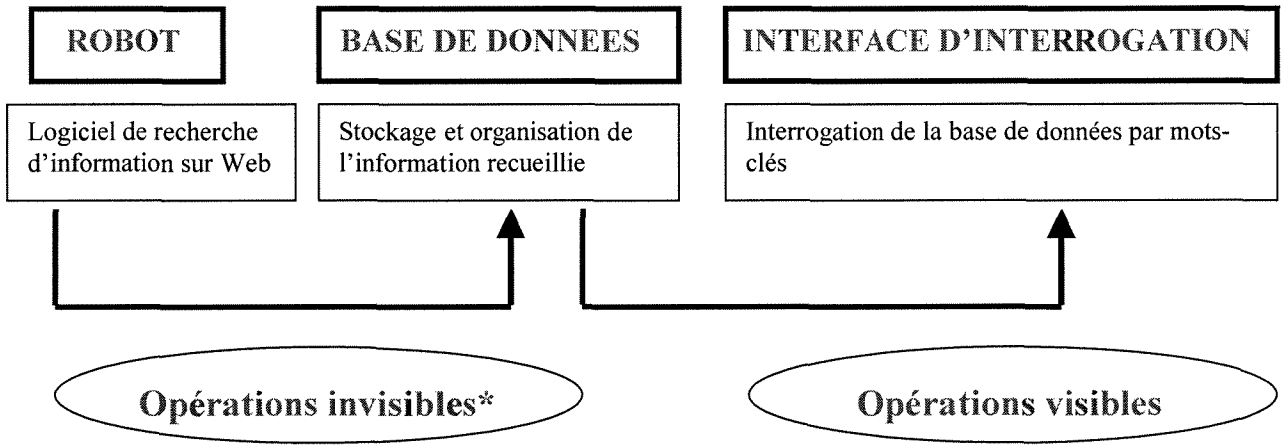


Fig.2

Processus d'indexation humaine : exemple *Nomade*

Dans les annuaires, la recherche est faite sur la description associée aux sites référencés et non sur le contenu des pages.

Processus d'indexation humaine

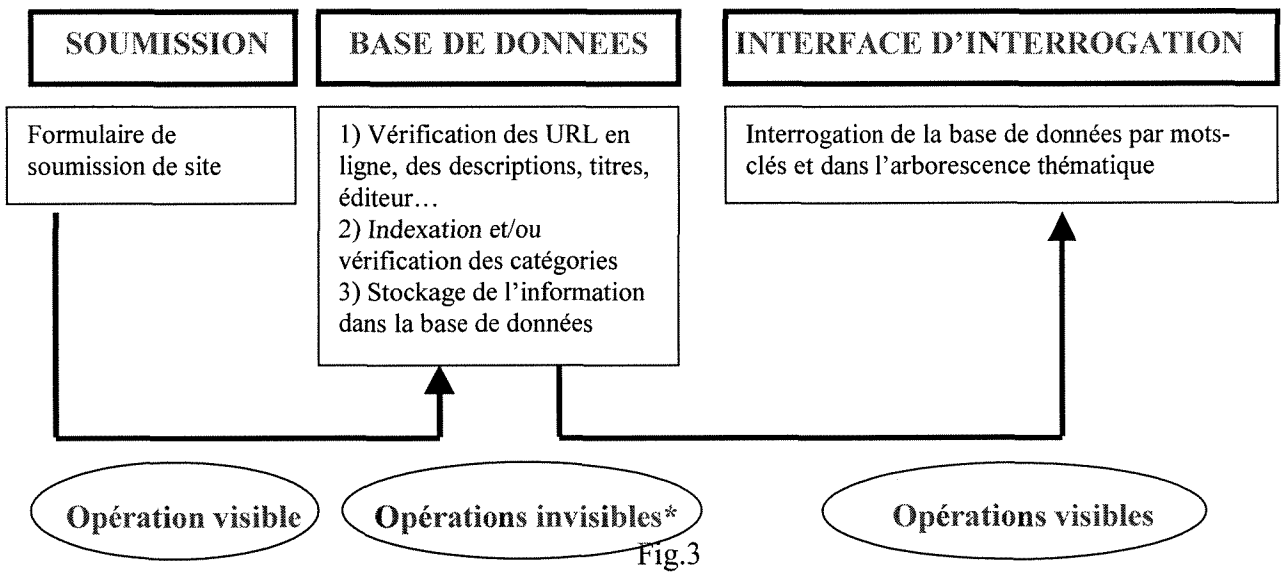


Fig.3

*Les opérations invisibles s'effectuent à l'insu de l'utilisateur et sont contrôlées par l'organisme qui gère l'outil de recherche.

III-3-3 Les champs d'indexation

Le tableau ci-dessous présente les champs pris en compte lors de l'indexation par les robots sur la base de ce que fournissent les moteurs eux-mêmes. Les moteurs étudiés sont : *AltaVista, HotBot, Excite, NorthernLight, Voilà, Lycos, WebCrawler, Infoseek.* Ces informations ont tirées du site abondance.com **[74]**.

Champs pris en compte lors de l'indexation

Champs/Moteur	AltaVista	Excite	HotBot	Lycos	Infoseek	WebCrawler	Northern Light	Voilà
Titre	**Oui**	**Oui**	***Oui***	**Oui**	**Oui**	**Oui**	**Oui**	**Oui**
Balise META description	**Oui**	**Non**	**Oui**	**Non**	**Oui**	**Non**	**Non**	**Oui**
Balise META keywords	**Oui**	**Non**	**Oui**	**Non**	**Oui**	**Non**	**Non**	**Oui**
Commentaires <!-- -->	**Non**	**Non**	**Non**	**Non**	**Non**	**Non**	**Non**	**Non**
Attributs ALT des balises IMG	**Oui**	**Non**	**Non**	**Oui**	**Non**	**Non**	**Non**	**Non**
Intitulé de l'URL	**Oui**	**Oui**	Uniqt le **nom du serveur**	**Non**	Uniqt le **nom du serveur**	**Non**	**Oui**	**Oui**
Frames	Le fichier principal est indexé.	Seul le fichier principal est pris en compte	Fichier principal parfois indexé.	Seul le fichier principal	Le fichier principal est indexé.	fichier ppal	Oui	fichier principal
Imagemaps	**Oui**	**Non**	**Non**	**Non**	**Oui**	**Oui**	**Oui**	**Non**
Corps du texte	**Oui** (tout le texte de la page est indexé jusqu'à 100 Ko).	**Oui.** Tout le texte de la page est indexé.	**Oui** Tout le texte de la page est indexé.	**Oui.** Le début du texte est très important.	**Oui** Tout le texte de la page est indexé.	**Oui.** La première partie Importa-te	**Oui** Tout le texte de la page est indexé.	**Oui**

Fig.4

On peut noter, lorsqu'on étudie ce tableau, des disparités importantes au niveau du traitement des informations présentes dans les pages à indexer. Les robots « spiders » de chaque moteur ont des particularités fortes et s'attachent à des portions de pages qu'ils pensent essentielles.

Tout d'abord, on peut noter que la seule information indexée par tous les moteurs est le titre et ceci, sans exception. Mais dès que l'on aborde les meta-informations, les disparités resurgissent. *AltaVista, HotBot, Infoseek, Voilà* traitent ces informations qui apparaissent en code source et qui sont remplies par le concepteur du site. Chez ces moteurs, la meta-description apparaît en résumé lors de la fourniture de résultats de recherche **[74]**.

Ensuite, on remarque une certaine unité dans la non prise en compte des commentaires en code source ainsi que dans la difficulté d'indexer les « frames » qui apparaissent pourtant

de plus en plus fréquemment dans les sites sous la forme de bandeaux qui permettent à l'utilisateur de naviguer plus aisément dans le site grâce à un menu.

Autre disparité entre moteurs : l'indexation des « images maps ». Ce sont des images ayant différentes zones réactives qui sont des liens hypertextes. De nombreux sites utilisent cette méthode pour leur « home page ».

Enfin, tous ces moteurs indexent la totalité du texte présent dans les pages **[74]**.

III-3-4 Les priorités dans les champs indexés

Priorités dans les champs indexés

Champs/Moteur	AltaVista	Excite	HotBot	WebCrawler	Infoseek	Lycos	Northern Light	Voila
Titre	***	***	*	***	**	***	***	***
Balise META keywords	*	**Non** (pas pris en compte)	***	**Non** (pas pris en compte)	***	**Non** (pas pris en compte)	-	*
Indice de popularité de la page	**	***	*	***	**	**	**	**Pas pris en compte** (à l'étude)
Corps du texte	**	**	**	**	*	***	**	*
Modalités d'exclusion des pages								
Fichier robots.txt	**Oui** : pris en compte	**Oui** : pris en compte	**Oui** : pris en compte	**Oui** : pris en compte	**Oui** : pris en compte	**Oui** : pris en compte	**Oui** : pris en compte	**Oui** : pris en compte
Balise META robots	**Oui** : prise en compte	**Non** : pas de prise en compte	**Oui** : prise en compte	**Oui** : prise en compte	**Oui** : prise en compte	**Oui** : prise en compte	**Oui** : prise en compte	**Oui** : prise en compte

Fig.5

De l'indexation va dépendre l'apparition plus ou moins précoce sur les pages réponses à des requêtes d'utilisateurs. Les constructeurs de sites doivent tenir compte de priorités qui sont accordées à tel ou tel champ pour accéder à un classement optimal chez ces différents moteurs **[49]**. Mais on va montrer que tout est histoire de compromis puisque les prises en compte ne sont pas les mêmes. Les différents champs indexés par les robots ne revêtent pas la même importance les uns par rapport aux autres. Chez certains moteurs comme *Altavista ou WebCrawler, Lycos, NorthernLight*, sont privilégiés le titre et l'indice de popularité des pages **[74]**.

Par contre, des moteurs comme *HotBot* et *Infoseek* font confiance aux meta-informations pour les classements. On s'aperçoit avec le tableau précédent que des pages peuvent être exclues de l'indexation par l'utilisation de noms particuliers pour nommer les fichiers.

III-3-5 Les modalités de référencement

Modalités de référencement

Réf/Moteur	AltaVista	Excite	Hotbot	Infoseek	Lycos	WebCrawler	Northern Light	Voila
Informations demandées	URL uniquement	URL, E-mail, langue, pays, thème du site.	URL, E-mail	URL, E-mail.	URL	URL, E-mail, langue, pays, thème	URL, nom, E-mail.	URL et adresse E-mail
Vérification du moteur en temps réel	Oui	Non	Oui	Oui	Oui	Non	Non	Non (à l'étude)
Délai de prise en compte de la page lors d'une soumission manuelle	1 ou 2 jours	1 à 3 semaines	1 jour à 2 semaines (2 jours en général)	2 à 4 semaines et plus	1 ou 2 jours	1 à 8 semaines et plus	1 ou 2 semaines	1 ou 2 semaines
Nombre maximal de pages soumises dans une journée	1 ou 2.	25	50 au maximum	Pas de limite	50	25	Pas de limite	Pas de limite
Délai avant indexation "naturelle" des autres pages par le spider	1 jour à 1 mois	3 semaines, parfois plus	2 semaines environ	4 semaines	1 ou 2 mois	1 sem à 2 mois	1 ou 2 sem	1 ou 2 sem
Délai de rafraîchissement de l'index	6 semaines	6 semaines	4 semaines	2 à 3 semaines	2 à 3 semaines	1 semaine	2 à 4 semaines	2 semaines
Limites sur le nombre de pages indexées pour un même site	400 pages environ.	Pas de limites connues.	Pas de limites connues.	Pas de limite	600 environ.	Pas de limite	Pas de limite	Pas de limite

Fig.6

Ces informations sont particulièrement utiles aux concepteurs de sites lors de la soumission de nouvelles pages à un moteur. Elles indiquent les délais de prise en compte des pages soumises ainsi que le temps mis par le robot pour accéder au reste du site **[49]**.

Le délai de rafraîchissement de l'index permet aux personnes à la recherche d'informations d'avoir un œil critique sur les pages proposées en réponse à leurs requêtes. Ce délai correspond en fait au temps mis par le robot pour parcourir l'ensemble des pages indexées et revenir à son point de départ **[74]**.

Toutes ces informations proviennent des responsables des moteurs eux-mêmes et doivent parfois être modérées. Lorsqu'on lit « pas de limite au nombre de pages indexées pour un même site », il faut en fait lire « pas de limite connue ». En ayant connaissance de toutes ces données, on se fait une meilleure idée sur les capacités et les limites de chacun des moteurs de recherche et les utilisateurs assidus du Web devraient, à partir de là, avoir un changement de comportement, devenant plus critiques vis-à-vis des performances de chacun d'eux **[49]**.

III-4 Les limites de l'indexation

III-4-1 Le rafraîchissement de l'index

On entend par rafraîchissement de l'index, le moment où le « spider » revient sur les pages.

Ce rafraîchissement ne s'effectue qu'au mieux tous les quinze jours, voire un temps beaucoup plus long (plusieurs mois) chez certains moteurs **[67]**. Si le contenu des pages est modifié quotidiennement, seules certaines versions de ces pages sont présentes dans l'index. Elles sont alors obsolètes à l'échelle d'Internet.

Une solution à ce problème est l'indexation manuelle par l'intermédiaire des formulaires de soumission de pages.

III-4-2 L'indexation partielles des sites

Même si tous les moteurs indexent la totalité d'une page, ils n'indexent pas la totalité du site.

Ceci a pour conséquence que l'on ne retrouve pas dans l'index 100% des pages d'un site **[67]**. Il n'est pas possible de savoir quel sont les algorithmes utilisés par les spiders pour définir le pourcentage des pages prises en compte et les choix qui prévalent à un rejet.

III-4-3 Le Web « invisible »

On entend par Web « invisible » **[67]**, les pages qui demeurent invisibles aux moteurs de recherche. On peut en extraire trois catégories :

- Les « frames » : de nombreux moteurs ne savent pas indexer les sites basés sur des « frames » (terme utilisé pour désigner les cadres qui permettent d'avoir constamment à l'écran le menu initial). Il est courant de voir des sites de 100 pages seulement référencés par leur « Home Page ». *Northern Light* et *AltaVista* comprennent les « frames » mais elles ne sont pas dans leur contexte (contenu indexé mais navigation impossible). Pour la plupart des moteurs, résoudre le problème des « frames » n'est pas une priorité **[67]**.

- Les pages dynamiques : ce sont les pages typiquement liées à des bases de données, il yen a de plus en plus sur le Web. Quand l'utilisateur fait sa demande, la base de données assemble les différentes pièces et délivre une page Web comme réponse. La marque de fabrique d'une page dynamique est la présence d'un « ? » dans l'URL. La plupart des moteurs ne vont pas plus loin que le « ? », qu'ils comprennent comme une erreur et qui empêche la page d'être indexée. *Google*, grâce à sa particularité d'indexer les mots proches des liens, permet de retrouver ces pages par leurs liens **[67]**.

- Le XML : ce format nouveau ouvre des perspectives intéressantes et est le langage d'avenir sur le Web. Tous les moteurs de recherche disent qu'ils supporteront le XML mais avec des « si ». Si un standard émerge, si les « tags » sont utilisés de façon uniforme. Les documents pourraient être indexés par auteur, par éditeur, par dates. L'indexation du XML n'est pas encore mise en place **[67]**

III-5 L'avenir des moteurs de recherche sur le Web

III-5-1 Le projet *CLEVER* de la société *IBM*

Le projet se base sur le fait que l'indexation ou la récupération de pages perd beaucoup de temps à ne se fonder et qu'il faut tirer partie du milliard de liens qui existent entre les pages.

En effet, de nombreux problèmes sont posés par les modes d'indexation usuels. On peut prendre l'exemple des concepteurs de sites d'entreprises commerciales qui cherchent plutôt à transmettre une image plutôt que décrire une activité réelle au sein de leurs pages (liste de discussion).

Pour faire face à ces problèmes d'indexation, les chercheurs ont mis au point un système automatique qui trouve des sites qui font autorité sur des sujets larges. L'approche est fondée sur l'hypothèse que chaque lien matérialise une parenté sémantique entre la page initiale et la page vers laquelle il pointe. Il existe des contre-exemples à cette hypothèse comme les « Retour au menu principal » ou les liens pour dénigrer un site mais les liens sont le plus souvent une marque d'adhésion. Ceci a permis de définir les sites « référence ».

Le robot d'indexation utilise un algorithme itératif qui définit aussi des sites « pivots » qui ont eux la particularité d'avoir des liens vers les sites « référence ». Des notes sont attribuées à chacun des sites visités en tant que « référence » ou « pivot ». Des classements en découlent en fonction des requêtes des utilisateurs.

CLEVER est un moteur prototype. Après traitement matriciel, l'algorithme de *CLEVER* a l'avantage de regrouper tous les sites en catégories. De par son principe, il ressemble à une analyse de citations. Le très connu moteur de recherche *Google*, mis au point à Stanford, utilise un classement fondé sur les poids d'influence. Il diffère de *CLEVER* par son mode de classement qui n'est pas adaptatif aux requêtes des utilisateurs ce qui lui permet d'être plus rapide.

Les améliorations qui devront être apportées à *CLEVER* porteront sur l'analyse du texte proche des liens qui permettra de pondérer les notes accordées aux sites.

III-5-2 Nouvelles stratégies (les exemples d'*AltaVista* et de *Lycos*)

Ces nouvelles stratégies se basent le plus souvent sur la volonté de fraîcheur de l'indexation. Ils ont tous la volonté d'éviter de fournir des documents obsolètes.
Lycos a introduit de nouvelles méthodes dans le monde des moteurs de recherche. Jusqu'ici, toutes les pages étaient considérées de la même façon et se trouvaient sur un même pied d'égalité **[75]**. Ce n'est plus le cas pour ce moteur qui fait désormais visiter par son robot les pages les plus prisées par les utilisateurs toutes les semaines. Les autres pages le sont toutes les 2 ou 3 semaines. Le « Wise Wire System » permet aux visiteurs de voter pour les pages qu'ils aiment. Ce vote est la base des URL visitées par le robot chaque jour. *Infoseek* a une stratégie se rapprochant de celle de *Lycos* : son robot visite plus souvent les sites de News ou sites conventionnels.

Quant à *AltaVista* **[75]**, les concepteurs se basent sur « l'intelligence du robot pour rafraîchir l'index ». Pour l'instant, le robot n'est pas intelligent car il compare la page avec sa copie déjà présente dans l'index pour savoir si elle a changé. *AltaVista* veut changer cela.

Dans un nouveau système, *AltaVista* garderait dans son index des pages en synchronisation avec le Web. Un robot serait chargé de retrouver les pages modifiées et un autre ferait le changement. Cela permet de concentrer l'attention du moteur uniquement sur ce qui en vaut la peine, c'est à dire les pages modifiées.

IV - CONCLUSION

La quête de documents sur un sujet aussi évolutif et récent que l'indexation des pages Web dans les moteurs de recherche trouve sur sa route de nombreux écueils. En effet, les sources de renseignement sont peu nombreuses et peu loquaces : ces méthodes sont d'une extrême importance dans ce milieu très concurrentiel. Il faut même parfois savoir remettre en question certaines affirmations en tenant compte de leurs origines.

Ce milieu, bien qu'indispensable aux utilisateurs est en pleine mutation, les différences entre moteurs et annuaires s'estompent pour laisser la place à des « portails » d'accès à l'Internet. Cette diversification en sites d'information continue, de services, d'échanges, de « chat », de vente à distance parfois qui s'accompagne de rapprochements, d'achats, de regroupements parmi ces acteurs, ne doit pas détourner les moteurs de recherche de leur fonction première : faciliter l'accès des utilisateurs à l'information sur le Web.

Ce réseau Internet est aujourd'hui très différent de ce qu'il était il y a seulement cinq ans. Personne ne peut dire ce qu'il sera dans cinq ans. L'indexation du réseau deviendra-t-elle impossible ? Et comment la notion de recherche changera-t-elle alors ? Les utilisateurs seront-ils encore plus mis à contribution ? Aujourd'hui, la seule chose dont on soit certain, est que la croissance du réseau continuera à lancer des défis informatiques à ceux qui veulent naviguer dans cet océan d'informations.

Troisième Partie
BIBLIOGRAPHIE

V - BIBLIOGRAPHIE

V-1 Monographies

[1] **ANDRIEU Olivier.** *Méthodes et outils de recherche sur Internet.* Eyrolles, 1997. 235 p.

[2] **ANDRIEU Olivier.** *Trouver l'info sur* Internet.Eyrolles, 1998. 426 p.

[3] **DILLON Martin.** *Assessing information on the Internet : toward providing library services for computer-mediated communication : results of an OCLC research project.* OCLC Online Computer Library Center, Inc., Office of Research, 1993.

[4] **GAYRILUT Gabriela, LETRANCHANT Maryline , SAINT-JACQUES Nathalie, TELLIER Sylvie.** *Internet : les aides à la recherche.* Editions du Trécarré, 1996. 183 p.

[5] **GLOSSBRENNER Alfred, GLOSSBRENNER Emily.** *Moteurs de recherche : le guide Quick Start.* Editions First Interactive, 1999. 303 p.

[6] **GOURBIN Géraldine.** *Repérage de l'information et indexation sur leWorld Wide Web : applications sur les annuaires et moteurs de recherche.* Nanterre : [s.n.], 1997.

[7] **LARDY Jean-Pierre.** *Recherche d'informations sur Internet : outils et méthodes.* ADBS Editions, 1999.

[8] **LARDY Jean-Pierre.** *Recherche d'informations sur Internet : outils et méthodes.* ADBS Editions, 1996.

[9] **LARDY Jean-Pierre.** *Recherche d'informations sur Internet : outils et méthodes.* ADBS Editions, 1997.

[10] **LELOUP Catherine.** *Moteurs d'indexation et de recherche : environnements client-serveur, internet et intranet.* Eyrolles, 1998. 285 p.

[11] **LIU Cricket.** *Systèmes d'information sur Internet : configuration et mise en œuvre.* O'Reilly International Thomson,1996. 727 p.

[12] **NEWBY Gregory B.** *Directory of directories on the Internet : a guide to information sources.* Meckler, 1994.

[13] **NICHOLSON Dennis.** *Cataloguing the Internet : CATRIONA feasibility study.* British Library Research and Development Department, 1995.

[14] **REVELLI Carlo, DE ROSNAY Joël.** *Intelligence stratégique sur internet : comment développer efficacement des activités de veille et de recherche sur les réseaux : moteurs de recherche, réseaux d'experts, agents intelligents.* Dunod, 1998. 212p.

V-2 Thèses

[23] **DELISLE Cynthia.** *Les outils de recherche sur Internet : typologies et principales caractéristiques.* Thèse sci. inf. et com. : Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques. Editions ENSSIB, 1999.

V-3 Comptes-rendus de congrès ou de colloques

[15] **BRUNELLE B S.** Online Information 96. *Smart systems, smart searches. Proceedings of the 20th international Online information meeting, London, 3-5 December.* Londres : Online, 1996, p.387-390.

[16] **CAROLL D J, LELE P.** IAMSLIC. *Human intervention in the networked environment : metadata alternatives. Proceedings of the 23th annual conference of the international association of aquatic and marine science libraries and information centers, Charleston, South Carolina, 5-9 October 1997.* Fort Pierce, Florida : IAMSLIC, 1998, p.59-71.

[17] **CHU H, ROSENTHAL M.** American Society for Information Science. *Search engines for the World Wide Web : a comparative study and evaluation methodology. Proceedings of the 59th annual meeting of the American society for Information Science, Baltimore, Maryland, 21-24 October 1996.* Medford, New Jersey : Information Today, 1996.

[18] **DING W, MARCHIONINI G.** American Society for Information Science. *A comparative study of Web search service performance. Proceedings of the 59th annual meeting of the American society for Information Science, Baltimore, Maryland, 21-24 October 1996.* Medford, New Jersey : Information Today, 1996, p.136-142.

[19] **Ecole nationale supérieure des sciences de l'information et des bibliothèques.** International society for knowledge organization. *L'indexation à l'ère d'Internet : [actes] du 2ème colloque du chapitre français de l'ISKO, Lyon, 21 et 22 octobre 1999.* Lyon : [s.n.], 1999.

[20] **HENZINGER M R, HEYDON A, MITZENMACHER M, NAJORK M, MENDELZON A.** Compaq Computer Corporation Systems Research Center. *Measuring index quality using random walks on the Web. Proceedings of the eighth international World Wide Web Conference, Toronto, Canada, 11-14 Mai 1999.* Toronto, Computer Networks, 1999.

[21] **KEILY L.** Learned Information. *Improving resource discovery on the Internet : the user perspective, Online information 97 : London, 9-11 December 1997.* Newcastle : [s.n.], 1999.

[22] **LE MOAL Jean-Claude, HIDOINE Bernard.** Association des professionnels de l'information et de la documentation (France) , 30 septembre - 4 octobre 1996, Trégastel (Côtes d'Armor). *La recherche d'information sur les réseaux : Internet, pour en savoir plus : cours INRIA..* ADBS Editions, 1996. 253 p.

V-4 Articles de périodiques

[24] **ANDRIEU Olivier.** Les surprises du référencement. *Technologies internationales,* 1999, n°51, p 19-22.

[25] **ANSTEAD Mark.** How to come top in search rankings. *Internet Magazine,* 1999, N°103.

[26] **ARDO Anders.** Indexing the nordic countries. *Electronic Library,* 1998, vol 16, N°2, p.117-118.

[27] **BELBENOIT-AVICH P M.** Des phares dans la nuit : la recherché documentaire sur Internet. *Bulletin des bibliothèques de France,* 1996, vol 41, N°4, p.52-57.

[28] **BONISTEEL Steven.** Compaq's new search engine indexes Net Audio. *Newsbytes News network PM,* 1999.

[29] **BRANDT DS.** What flavour is your Internet search engine ? *Computers in libraries,* 1997, vol 17, N°1, p.47-50.

[30] **BRIN S., PAGE L..** The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems conference,* 1998, vol 30, N°1-7, p. 107-117.

[31] **BYERS D..** Full-text indexing of non-textual ressources. *Computer Networks and ISDN Systems,* 1998, vol 30, N°1-7, p. 141-148.

[32] **CUNNINGHAM J.** Getting the most from AltaVista. *Behavioral and Social Sciences Librarian,* 1996, N°15, p.53-56.

[33] **DONG X.** Search engines on the World Wide Web and information retrieval from the Internet : a review and evaluation. *Online and CD-ROM review,* 1997, N°21, p.67-82.

[34] **DUCHEMIN P Y.** La recherche d'informations sur l'Internet : répertoires et moteurs de recherche. *Bulletin d'information de l'association des bibliothécaires de France,* 1997, N°174, p.91-96.

[35] **ELLIS D., VASCONCELOS A.** Ranganathan and the Net : using facet analysis to search and organize the World Wide Web. *Aslib perspectives : New Information perspectives,* 1999, vol 51, N°1, p.3-10.

[36] **ESLER S L, NELSON M L.** NASA indexing benchmarks: evaluating text search engines. *Journal of Network and Computer Applications,* vol 20, N°4, p.339-353.

[37] **FELDMAN Susan.** New study of Web search engine coverage published. *Information Today,* 1999, vol 16, N°8, p.29.

[38] **HATTERY M.** Online World : the bumby ride of the Web engine. *Information Retrieval & Library Automation,* 1997, N°33, p.1-2.

[39] **HIRATA Kyoji.** Multimedia Web retrieval system AMORE. Multimedia retrieval content on Internet. *Gazo Rabo,* 1999, vol 10, N°2, p.21-26.

[40] **HUDSPETH Lee.** Search engine secrets. *PC Computing,* 1999, vol 12, p.174-184.

[41] **LARDY Jean-Pierre.** Les outils de recherche d'informations sur Internet : guides, listes thématiques et index. *Documentaliste,* 1996, vol 33, N°1-2, p.33-39.

[42] **LUH James.** Sifting through Web videos—A new breed of search technology is catching on. *Internet World,* 1999, vol 5, N°28, p.40.

[43] **JACSO Peter.** More search engine hype and reality : how well do Google's indexing and page rank features measure up ?. *Information Today,* 1999, vol 16, N°4, p.30-31.

[44] KIENTZLE T.. A Java applet search engine. *Dr Dobb's Journal,* 1999, vol 24, N°2, p. 32, 36, 38-40.

[45] LAURSEN JV. Somebody wants to get in touch with you : search engine persuasion. *Database,* 1998, vol 21, N°1, p.42-46.

[46] LAWRENCE Steve, GILES C. Lee. Accessibility of information on the Web. *Nature,* 1999, vol 400, p 107-109.

[47] LAWRENCE Steve, GILES C. Lee. Searching the World Wide Web. *Science,* 1998, vol 280, p 98-100.

[48] LE GUELVOUIT Arnaud. Les outils de recherche du Web. *Documentaliste-Sciences de l'Information,* 1998, vol 35, n°6, p 315-320.

[49] LE GUELVOUIT Arnaud. Les outils de recherche du Web : limites et aléas du référencement. *Documentaliste,* 1999, vol 35, N°6, p.315-320.

[50] Mc MURDO G. How the Internet was indexed. *Journal of Information Science,* 1995,vol 21 N°6, p.479-489.

[51] MESERVE Jason. Searching for XML. *Network World,* 1999.

[52] MIZOBUCHI Sachi. An analysis of the effect of indexes on WWW information selection process. *Joho Shori Gakkai Hokoku,* 1998, vol 98, N°116, p.31-36.

[53] MOODY G. Searching the Web for gigabucks. *New Scientist,* 1997, N°150, p.36-40.

[54] MUKHERJEA S.. Towards a multimedia World-Wide Web information retrieval engine. *Computer Networks and ISDN Systems,* 1997, vol 29, N°8-13, p. 1181-1191.

[55] NANFITO M.. The indexed Web : engineering tools for cataloging, storing and delivering Web-based documents. *Information Outlook,* 1999, vol 3, N°2, p. 18-23.

[56] NEELAMEGHAN A. Indexing and search engines for the Web. *Information Studies,* 1999, vol 5, N°1, p.1-6.

[57] NICHOLSON S.. Indexing and abstracting on the World Wide Web. An examination of six Web Databases. *Information technology and libraries,* 1997, vol 16, N°2, p.73-81.

[58] NOTESS G.R.. Measuring the size of Internet Databases. *Database,* 1997, vol 20, N°5, p. 69-70, 72.

[59] NOTESS G.R.. Comparing Net Directories. *Database,* 1997, vol 20, p.61-64.

[60] POYNDER R. Web research engines ? *Information World Review,* 1996, N°120, p.47-48.

[61] RODRIGUEZ G., NAVARRO L. Aleph Web : a search engine based on the federated structure. *RIST,* 1997, vol 7, N°1, p.87-98.

[62] SCALES B J, FELT E C. Diversity on the World Wide Web : using robots to search the Web. *Library Software Review,* vol 14, N°3, p.132-136.

[63] **SMITH J.R., SHI-FU Chang.** An image and video search engine for the World Wide Web. *Proceedings of the SPIE – The International Society for optical engineering conference,* 1997, vol 3022, p. 84-95.

[64] **SNYDER Herbert.** How public is the Web ? : robots, access, scholary communication. *Proceedings of the ASIS annual meeting,* 1998, vol 35, N°1998, p.453-462.

[65] **SONGINI Marc.** Search engines skimming the surface of Web – and may be missing yours. *Network World,* 1999.

[66] **STANLEY T.** Searching the World Wide Web with Lycos and Infoseek. *New Review of Information Networking,* 1995, vol 1, p.191-202.

[67] **SULLIVAN D.** Crawling under the hood : an update on search engine technology. *Online,* 1999, vol 23, N°3, p.30-38.

[68] **TUNENDER H, ERVIN J.** How to succed in promoting your Web site : the impact of search engine registration on retrieval of a World Wide Web site. *Information technology and libraries,* vol 17, N°3, p.173-179.

[69] **Van der WALT M S.** Browsing guides on the Web : the use of classification for the organisation of information sources on the Internet. *South African Journal of Library and Information Sciences,* 1998, N°66, p.56-66.

[70]**WALKER D.** AusSI Web indexing prize 1998. *Indexer,* 1998, vol 21, N°3, p.108-110.

[71] **WALSTER D.** Search engines on the World Wide Web. *Emergency Librarian,* 1997, N°24, p. 21-23.

[72] **YAMADA Hideaki.** A 3D image search engine based on VRML Logical Structure. *Eizo Media Gakkai Hokoku,* 1998, vol 22, N°23, p.31-36.

V-5 Sites Internet

[73] **Abeille**. Les carnets de l'intelleigence compétitive. <http://www.abeille.org>, 7mars 2000.

[74] **Abondance**. Abondance : les moteurs de recherche. <http://www.abondance.com/outils/moteurs.html>, 18 février 2000.

[75] **AltaVista**. About AltaVista. La technologie Cow 9 sur AltaVista. <http://altavista.digital.com/av/content/about_our_technology_cow9.htm>, 8 juillet 1999.

[76] **BOURDONCLE François**. Ecole Nationale Supérieure des Mines de Paris. Panorama et perspectives des outils de recherche d'Information textuelle sur Internet. <http://www.cma.ensmp.fr/Francois.Bourdoncle/idt99.html>, 25 février 2000.

[77] **Écila Off-Shore**. Moteur de recherche déporté pour rechercher par mots clés dans le texte d'un site. <http://offshore.ecila.com/home-french.html>, 7 mars 2000.

[78] **EGGHE Leo, ROUSSEAU Ronald**. Introduction to Informetrics, in Elsevier Science Publishers.
<http://www.amazon.com/exec/obidos/ASIN/>, 1997.

[79] **FELLBAUM Christiane**. WordNet: An Electronic Lexical Database par Christiane Fellbaum.
<http://www.amazon.com/exec/obidos/ASIN/>, 20 février 2000.

[80] **IDF** : les moteurs de rechercehe francophones. Vocabulaire des moteurs de recherche - Un glossaire sur la technologie de recherche Internet.
<http://www.idf.net/mdr/>, 2 janvier 2000.

[81] **IsMap**. Moteur de recherche cartographique par mots clés, permet de localiser des sites sur une carte géographique (restaurants, hôtels, etc.).
<http://www.ismap.com/>, 3 mars 2000.

[82] **KLEINBERG Jon M.** Authoritative Sources in a Hyperlinked Environment in Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms.
<http://simon.cs.cornell.edu/home/kleinber/>, 22 janvier 1999.

[83] **LE COADIC Yves**. Recherche d'information sur l'Internet. Explication des concepts de base des moteurs et annuaires, méthodes de recherche et glossaire.
<http://www.adbs.fr/adbs/prodserv/edit/001/html/10000021.htm>, 28 janvier 2000.

[84] **MEGECAZE Stéphane**. Recherche sur Internet. Conseils pratiques, annuaires versus moteurs, les bons mots-clés et le référencement.
<http://www.no-bug.com/techno/recherche.htm>, 4 janvier 2000.

[85] **NGUON Hao Ching**. ISI Information Science Institute. Database publisher with a focus on Web-based products.
<http://www.isinet.com/>, 7 février 2000.

[86] **Présence Web**. Moteur de recherche des noms de domaines.
<http://www.presenceweb.com/>, 7 mars 2000.

[87] **RSACi North America Server**. Search Engine Watch: Tips About Internet Search Engines & Search Engine Submission.
<http://www.searchenginewatch.com/>

[88] **Science Citation Index Database: ISI**. L'Index des citations scientifiques (Science Citation Index).
<http://www.isinet.com/prodserv/citation/citsci.html>, 29 octobre 1999.

[89] **3 Dup.com**. Moteur de recherche 3D, 2D, multimédia, conception graphique et industrie audiovisuelle.
<http://3dup.com/search_french.shtml>, 7 mars 2000.

VI - ANNEXE

Adresses des moteurs de recherche cités :

http://www.altavista.com

http://www.excite.com

http://www.hotbot.com

http://www.infoseek.com

http://www-english.lycos.com

http://www.northernlight.com

http://www.voila.fr

http://www.webcrawler.com