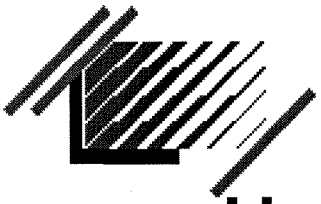
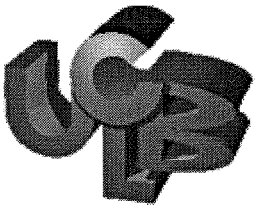


enssib

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



Université
Claude Bernard
Lyon 1

DESS en INGENIERIE DOCUMENTAIRE

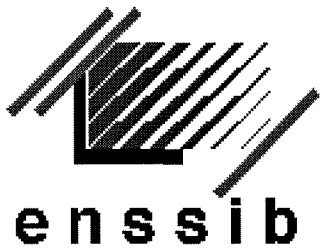
Rapport de Stage

Etude comparative de moteurs d'indexation et de recherche

ABDEL ADIM Leïla

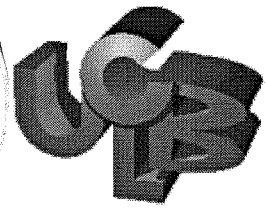
Sous la direction de
Jacques Kouloumdjian – professeur au Laboratoire d'Ingénierie des Systèmes d'Information, INSA Lyon.

Année 2000



enssib

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



Université
Claude Bernard
Lyon 1

DESS en INGENIERIE DOCUMENTAIRE

Rapport de Stage

Etude comparative de moteurs d'indexation et de recherche

ABDEL ADIM Leïla

Sous la direction de
Jacques Kouloumdjian – professeur au Laboratoire d'Ingénierie des Systèmes d'Information, INSA Lyon.

Année 2000

Etude comparative de moteurs d'indexation et de recherche

Résumé :

Le sujet de ce stage est l'étude de la mise en place d'un moteur d'indexation et de recherche sur les sites Web des laboratoires de l'INSA. Deux des objectifs visés par ce projet sont d'offrir d'autres moyens d'accès à l'information des laboratoires que ceux prédéfinis par les hyperliens et de mettre à disposition des documents HTML ou non. Le moteur devra donc permettre l'indexation de multiples formats de fichiers et être capable de rendre des résultats « contextuels ».

Mots clés :

Outils d'indexation et de recherche, étude comparative, indexation en texte intégral, recherche contextuelle, indexation multi-formats.

Comparative study of search engines

Summary :

The subject of this training course is the study of the installation of a search engine on the Web sites of the laboratories of the INSA. Two of the objectives pursued by this project are to offer other means of access to the information of the laboratories that those predefined by the hyperlinks and to place at the disposal of documents HTML or not. The engine will have to thus allow the indexing of multiple formats of files and to be able to make "contextual" results.

Keywords :

Tools of indexing and search, comparative study, indexing in plaintext, contextual search, indexing multi-formats.

Sommaire

Sommaire

Introduction

1. Description du projet servant de support au stage

1.1. Contexte

L'INSA est actuellement en train de définir une structure de base de données pour les laboratoires de recherche. Cette base de données doit être au cœur du système d'information des laboratoires. Elle doit permettre en particulier l'alimentation des pages Web des laboratoires et garantir des informations à jour en permanence.

Cependant cette approche de la diffusion de l'information des laboratoires reposant sur l'interface Web classique est insatisfaisante à plusieurs égards :

- elle est limitée à la diffusion de documents au format HTML,
- elle contraint l'utilisateur à employer les chemins d'accès définis dans les pages,
- elle ne permet pas de feed-back sur les attentes de ceux qui utilisent le site,
- elle ne favorise pas la promotion automatique des sites auprès des moteurs de recherche qui indexent Internet.

Le projet vise à répondre à ces insuffisances.

1.2. Projet

Le projet se situe dans le thème fédérateur **Documentique**. Les techniques à mettre en œuvre et les problématiques sont celles qui sont sous-jacentes aux bases de documents répartis, à l'indexation des documents et à leurs accès.

Il s'agit d'offrir de nouvelles possibilités d'accès en consultation aux informations mises sur le réseau par les laboratoires tout en permettant une exploitation de ces accès grâce à une surveillance sur le nombre et la nature des interrogations faites. Actuellement, les pages Web des laboratoires offrent des chemins d'accès prédéfinis aux personnes désireuses de s'informer. Les possibilités de recherche d'information

sont donc celles prévues par les liens attachés au document HTML et ne correspondent pas nécessairement à la vision de celui qui recherche une information précise. En outre, ces chemins ne permettent qu'un accès aux pages Web et non aux autres documents qui pourraient être mis en ligne (rapports d'activité, rapports de recherche et publications du laboratoire, thèses ou documents techniques en texte intégral...). Enfin, aucune trace des interrogations n'est conservée pour une exploitation a posteriori des accès.

Les objectifs techniques du projet sont les suivants :

- ***étudier l'intégration d'un moteur de recherche*** aux sites des laboratoires de façon à permettre une interrogation libre par mots clés sur les différents documents en ligne. Pour cela, l'idée est d'utiliser en local les possibilités d'indexation totale et automatique d'un moteur de recherche (comme Altavista). L'utilisateur peut alors se servir des mots clés qu'il souhaite pour interroger le système d'information du laboratoire en plus des accès prévus dans les pages Web.
- ***élargir le champ des documents offerts en accès au public :*** il ne sera plus nécessaire que les documents soient au format HTML spécifique à l'affichage des pages Web. Un document pourra être sous les nombreux formats reconnus par le navigateur (Word, PDF, PostScript,...). Il s'agira d'étudier les contraintes associées à la rédaction de ces documents pour qu'ils puissent être disponibles en ligne et exploitables pour affichage. A terme, on étudiera si des formats d'échange de type XML peuvent servir de format commun d'affichage.
- ***surveiller les accès faits sur les informations du laboratoire*** mises sur le réseau : au-delà du simple comptage du nombre d'accès réalisés sur les pages Web, il est particulièrement intéressant de connaître les mots clés qui ont été utilisés dans les recherches d'information pour savoir ce que recherche ceux qui viennent consulter les pages des laboratoires ainsi que les adresses électroniques des auteurs. Il s'agit en quelque sorte d'une veille à partir des accès externes.
- ***étudier comment promouvoir de façon automatique les pages des laboratoires*** en signalant aux moteurs de recherche du Web les nouvelles

pages qui sont créées (amorce d'une technologie « push » de diffusion de l'information).

- D'une manière plus générale, étudier de quelle façon ***l'utilisation conjointe de bases de données*** servant à l'indexation et à la recherche de données factuelles et de ***robots d'indexation et de recherche*** pour de l'information textuelle peut améliorer le fonctionnement de site Web en terme de diffusion et recueil d'information.

Calendrier d'étude :

La proposition est faite de façon conjointe entre le LISI (Laboratoire d'Ingénierie des Systèmes d'Information) qui est le pilote du projet, qui mettra en place les moyens de l'étude et les outils et servira de premier terrain d'expérimentation, DocINSA, qui travaille sur l'écriture d'un cahier des charges d'un Système Intégré de Gestion de Bibliothèque, le LAI (Laboratoire d'Automatique Industrielle) et l'URGC (Unité de Recherche en Génie Civil) qui serviront de terrains d'expérimentation complémentaires. Un ou deux membres de chaque structure s'occupent du projet : en ce qui concerne le LISI, M. Kouloumdjian bien sûr et Anne Tchounikine; pour DocINSA, Monique Joly et Jean-Michel Mermet; pour le LAI, M. Wilfrid Marquis-Favre et pour l'URGC Mme Bernadette Escalier. Ces six personnes forment le groupe de travail du projet.

Ce projet concerne l'ensemble des laboratoires de l'INSA, la liste des laboratoires participant au projet pourra être élargie au besoin. Sa durée est de trois ans.

Année 1 :

- Etude bibliographique sur les techniques avancées de gestion de sites Web.
- Etude des moteurs d'indexation disponibles et choix d'un moteur.
- Etude des documents des laboratoires exploitables sur réseau et leur nature (INSA et extérieur).
- Aspects juridiques liés à la mise en consultation des documents.
- Implantation du moteur choisi pour indexation des documents des laboratoires.

- Maquette d'un site intégrant un moteur de recherche aux informations d'un laboratoire.

Année 2 :

- Etude de l'impact de la forme des documents pouvant être mis en ligne en vue de leur exploitation par le moteur d'indexation et de recherche, en liaison avec les formats d'échange de documents émergents de type XML.
- Réalisation d'un prototype opérationnel intégrant la mise en ligne de documents de nature différente.
- Définition des paramètres des accès à surveiller en liaison avec les caractéristiques des moteurs d'indexation.

Année 3 :

- Mise en place d'outils de surveillance des accès au site du laboratoire.
- Expérimentation et publication des résultats.

2. Objectif du stage

Les chercheurs des laboratoires de l'INSA produisent toute une littérature scientifique qui n'est actuellement pas mise en valeur. Un utilisateur ne peut la consulter par Internet car elle n'est pas accessible de manière simple et conviviale. Le même problème se pose pour un utilisateur interne à l'INSA.
Il y a donc un réel besoin de la part des laboratoires de valorisation des travaux de recherche. Ce besoin se retrouve à DocINSA. Toutefois, cette dernière a déjà commencé la mise à disposition des documents scientifiques en mettant en ligne et accessibles à tous les thèses soutenues à l'INSA.

Dans un contexte plus général, les besoins sont similaires à l'ENSSIB et à l'UCBL. La valorisation de la production des chercheurs par la mise en ligne et l'accès de leurs travaux à tout public est devenue importante pour les organismes d'enseignement supérieur.

Mon stage consiste donc en une première étude du projet, un débroussaillage des objectifs techniques. Mais le sujet étant trop large, nous avons dû le restreindre et donc

se concentrer uniquement sur l'étude de l'intégration d'un moteur de recherche et l'élargissement du champ des documents offerts en accès au public, soit les deux premiers objectifs techniques du projet.

Les objectifs généraux du stage se placent dans ce cadre là et sont les suivants :

- étude de l'installation d'un moteur d'indexation contextuel et de recherche dans les sites de l'INSA : prise en compte des besoins, étude de l'offre des moteurs, recherche d'étude similaires, écriture d'un cahier des charges pour l'acquisition d'un moteur, réalisation d'une maquette. Le but à long terme est l'accès facile et convivial pour tous aux documents des laboratoires mis en ligne.
- (éventuellement) recherche d'outils pour l'analyse des accès faits sur les sites Web de l'INSA et définition d'une méthodologie pour l'installation d'outils de surveillance d'accès.

Bien que ces deux aspects soient liés, je ne me suis attachée qu'au premier objectif, par manque de temps.

3. Méthodologie de travail

J'ai procédé de la façon suivante :

- étude de l'offre du marché en matière de moteurs d'indexation et de recherche à partir d'une première évaluation des besoins, rédaction d'un premier comparatif,
- choix de certains logiciels parmi la dizaine de logiciels susceptibles d'être intéressants,
- recherche de versions d'évaluation et demande de démonstration,
- Installation et manipulation des moteurs, tests à partir d'un protocole d'expérimentation défini au préalable, rédaction d'un comparatif détaillé.

Le stage s'est donc divisé en 3 grandes périodes :

- 1ère période : étude de l'offre, recherche d'information (durée : un mois et demi).

Durant cette période, j'ai eu des réunions dans un premier temps avec M. Kouloumdjian puis dans un deuxième temps avec le groupe de travail. Au cours des deux réunions (l'une le 21 juin et l'autre le 16 juillet), j'ai montré au groupe les différents moteurs qui avaient été sélectionnés, nous avons discuté de leurs avantages et de leurs inconvénients. Les différents membres du groupe m'ont donné leur avis sur les différents critères utilisés, ce qu'ils n'arrivaient pas à bien comprendre et donc qui devait être explicités, ce qu'il faudrait approfondir ou améliorer. De plus, les objectifs à court terme apparaissaient, comme la nécessité de tester les logiciels pour faire une étude plus complète.

- 2ème période : installation et tests des versions d'évaluation, mise en place d'un protocole d'expérimentation, affinement des critères, rédaction du comparatif final (durée : un mois et demi).

- 3ème période : rédaction du rapport (durée :un mois).

Etude de l'offre logicielle

1. Esquisses des besoins

A la suite d'entretiens avec M. Kouloumdjian d'une part et Mme Joly et M. Mermet d'autre part, il est apparu qu'à première vue, le moteur devra remplir certaines fonctionnalités :

- indexer d'une part des pages HTML et d'autre part des documents qui seront dans la machine où se trouvera la base de donnée du laboratoire. De plus, il y a différents formats de documents à indexer (HTML, PDF, PS, Word pour les principaux). Une évolution à long terme vers le format XML étant prévu, ce format devra lui aussi être indexé par le moteur. Donc le moteur devra parfaitement bien supporter un nombre important de formats de fichiers,
- donner le contexte de la réponse. Dans un document-réponse, les termes recherchés lors de la requête par l'utilisateur devront apparaître impérativement en surbrillance. Cette fonctionnalité est importante, notamment pour DocINSA qui, dans le cadre du projet Cither, indexe les thèses soutenues à l'INSA. Ces dernières faisant plusieurs centaines de pages, la surbrillance des termes recherchés faciliterait grandement la lecture de l'utilisateur,
- ouvrir le document trouvé et sélectionné. Une fois la recherche effectuée, l'utilisateur doit pouvoir accéder au document contenant les termes recherchés. Pour cela, le moteur doit localiser où se trouvent les différents documents indexés, quelque soit l'endroit, et doit permettre à l'utilisateur de les ouvrir pour que ce dernier puisse les lire,
- indexer tout le texte d'un document. Le moteur doit pouvoir effectuer une indexation en texte intégral ou full text (voir lexique en annexe),
- éventuellement, possibilités de recherche avancée. Outre la recherche simple avec les différents opérateurs et des mots clés, le moteur peut proposer une recherche par champs ou une recherche de phrases, etc.

- éventuellement, possibilité d'indexer des images.

Mme Joly et M. Mermet ont insisté sur l'importance de la réponse contextuelle, la reconnaissance des méta-tags et des méta-données (donc des balises HTML), l'indexation du XML (car évolution à long terme vers ce format) .

2. Choix des critères d'évaluation et de sélection

Pour avoir une définition précise des termes utilisés dans ce paragraphe, se référer au lexique situé en annexe.
Au vu des fonctionnalités recherchées pour les moteurs, il m'a semblé pertinent de les étudier suivant les critères suivants :

- Formats de documents indexés : HTML, XML, Doc, PDF, Postscript,
- Type d'indexation : indexation automatique, indexation en texte intégral ou full text, indexation incrémentale,
- Type de recherche proposé : recherche booléenne, recherche floue, recherche en langage naturel, recherche par proximité, recherche par similarité (Query By Example), recherche pondérée...
- Présentation des réponses : catégorisation des réponses, portail, tri des réponses, surbrillance des termes trouvés...

Cela a permis de définir un tableau de comparaison des différents moteurs avec ces critères.

3. Sources d'informations utilisées pour l'étude

Les informations contenues dans ces différentes sources citées dans la bibliographie se recoupaient. Les deux premières sources étant très détaillées, la consultation des autres sources n'a pas réellement apporté d'éléments nouveaux. Cela a surtout permis de vérifier que la recherche était exhaustive et que l'essentiel des logiciels pertinents avait été repéré. Toutefois, le livre de Catherine Leloup datant un peu l'étude d'autres documents plus récents était indispensable. Deux autres sources

d'informations se sont rajoutées : un document donné par Anne Tchounikine d'après une étude sur les moteurs de recherche faite par des étudiants et une étude de cas informatique d'étudiants conservateurs de bibliothèques de l'ENSSIB faite en mai 2000 intitulée « Robots d'indexation et classement automatique ».

4. Panorama de l'offre logicielle

4.1. Première étude : premier recensement des moteurs

Lors la première semaine, l'étude des différents documents constituant les sources a permis d'avoir une vue d'ensemble sur les moteurs d'indexation et de recherche disponibles sur le marché.

- Etude de l'étude de cas informatique des étudiants conservateurs de bibliothèques. Quatre logiciels sont étudiés :
  - Altavista de COMPAQ,
  - Sampler de CS SI,
  - Ultraseek Server+ CCE d' INFOSEEK,
  - MoreSence4U+InformationMiner4U+Class4U d'ARISEM.

  J'ai décidé de n'en retenir que trois ou plutôt deux sachant que Altavista est déjà utilisé sur les sites de l'ENSSIB, de l'UCBL et sur l'Intranet de l'INSA. Sampler étant un outil très tourné vers la veille, j'ai considéré qu'il valait mieux l'éliminer.

- Consultation du document « logiciels texte intégral » [3]. Trois logiciels semblent intéressants parmi les quatorze proposés :
  - ZyImage Web Serveur de ZYLAB,
  - Search'97 Information Server de VERITY,
  - RetrievalWare d'EXCALIBUR TECHNOLOGIES.

- Etude du guide d'achat 2000 [1]. Dans la partie « logiciels de texte intégral », outre certains logiciels précédemment cités (comme ZyImage, RetrievalWare, Search'97 Information Server et

MoreSence4U+InformationMiner4U+Class4U), d'autres semblent intéressants :

- Searchway de BULL,
- LexiGuide Web de LEXIQUEST,
- Askam Pro 3.0 d'OBLIVEA,
- SearchServer de FULCRUM.

- Etude du document donné par Anne Tchounikine. Cinq moteurs sont sélectionnés :
  - Harvest-ng et Harvest Indexer de HARVEST,
  - BDDBot de TIM MACINTA,
  - Webglimpse de l'Université d'Arizona,
  - Webinator,
  - Ht-dig,
  - i4m de SINEQUA.

- Etude du livre de Catherine Leloup [2] pour approfondir les informations concernant les moteurs suivants : Altavista Search Intranet, RetrievalWare, Search'97 de Verity, SearchServer de Fulcrum, Spirit, ZyIndex.

Soit au total 16 moteurs de recherche et d'indexation.

Après une réunion avec M. Kouloumdjian le 9 juin, il a été décidé que:

- quatre logiciels sont éliminés définitivement car leurs fonctionnalités sont trop limitées : ce sont BDDBot, Webglimpse, Ht-dig et i4m,
- six logiciels sont sélectionnés car ils remplissent la plupart si ce n'est toutes les conditions : MoreSence4U+InformationMiner4U+Class4U, UltraSeek Server + CCE, Altavista Search Intranet, Verity Information Server , Fulcrum Search Server et Spirit V2.
- les autres logiciels demandent des recherches complémentaires et plus approfondies pour décider s'ils sont réellement intéressants.

4.2. Deuxième étude : élargissement de l'étude

La rencontre lors de la deuxième semaine avec Monique Joly et Jean-Michel Mermet de DocINSA a permis de centrer les priorités sur :

- l'importance de la réponse contextuelle,
- la reconnaissance des méta-tag et des méta-données (donc des balises HTML),
- l'indexation du XML car évolution à long terme vers ce format de document,
- l'analyse des requêtes effectuées par les utilisateurs.

Je me suis donc concentrée sur les formats des documents indexables (notamment le XML) et sur la réponse contextuelle (critère décisif).

- Parmi d'autres moteurs suggérés par M. Kouloumdjian (Alkaline, Web Crawler et Web Clim), il s'est révélé que seul Alkaline semblait intéressant.

- Etude du livre d'Henry Samier et Victor Sandoval [5]. Le chapitre sur les outils et les méthodes de recherche automatique contient un tableau comparatif de 48 produits avec comme critères la recherche, l'indexation, le filtrage, la présentation, la distribution, l'aide à la décision suivant que ces fonctions sont principales, remplies, optionnelles ou pas remplies par l'outil.

  Dix produits semblaient intéressants. J'ai trouvé de la document sur sept produits. Je n'en ai retenu que deux : WorldScanning de CREATEAM et Callable Personnal Librarian de PLS.

- Etude du document « Search Enabling Web Site » [4]. Deux logiciels sont retenus :DeepSearch 3.0 de NANO et HomePage Search Engine.

Soit au total dix-huit moteurs de recherche et d'indexation.

4.3. Classification des moteurs étudiés

J'ai décidé de répartir les logiciels en trois catégories :

- les logiciels les plus intéressants : ce sont ceux qui correspondent à tous les critères de sélection.
- les logiciels susceptibles d'être intéressants : ils répondent à certains des critères demandés et ont d'autres fonctionnalités qui a priori ne sont pas demandées à l'origine mais qui peuvent être intéressantes. Ils peuvent être sélectionnés après réflexion.
- les logiciels à approfondir : il manque des informations pour trancher et savoir s'ils peuvent ou non être intéressants. Pour cela, il faut des informations supplémentaires.

4.3.1. Logiciels intéressants

Ils ont au nombre de six : leurs caractéristiques sont testées ci-dessous en fonction des premiers critères retenus.

Nom de la société	Nom du produit	Formats des documents indexés	Type d'indexation	Type de recherche proposée	Présentation de réponse	remarques
ARISEM www.arisem.com	**MoreSense4U+L4U + Class4U** (portail automatique+analyseur sémantique +catégoriseur)	HTML, XML, PDF, TXT, XML	Full text	-recherche par catégorie, -accès simultané à de multiples sources d'info selon une vue structurée, hiérarchisée et orientée selon ses besoins stratégiques	-Catégorisation des réponses, -portail documentaire -accès à l'info selon une arborescence, sous forme d'un plan de classement composé de thèmes et de sous-thèmes	-livré avec une base de connaissance -module d'étude statistique pour connaître la fréquentation des catégories, des documents trouvés -350 000 F HT
INFOSEEK http://www.ultraseek.com/fr/ultrahop_fr.htm *Site des clients :* http://www.ultraseek.com/demos/customer_sites.htm	**Ultraseek server + CCE 3.0** (Moteur de recherche et d'indexation + moteur de classification de contenu)	HTML, XML, PDF, MS Office, Texte ASCII, RTF, PostScript, PDF/ PostScript ... +100 autres formats	-Full text (organisation en catégorie avec CCE) -indexation automatique -création d'index (mise à jour automatique quand suppression ou ajout de document) -reconnaît les balises XML -thésaurus personnalisable	-Mot clé -pointage et cliquage (développement arborescence thème) avec CCE -recherche par date, +sieurs langues -recherche en langage naturel -recherche par champs (titre, URL, lien ou champ précis dans un document)	-Catégorisation des réponses (arborescence de catégorie à la Yahoo) avec CCE -tri par degré de pertinence ou par date -résultats avec titre ou titre et résumé (+ URL, date et taille du document) -possibilité d'avoir des « sites similaires » -sur brillance des termes trouvés dans	-visite plus souvent les pages modifiées que celles qui restent inchangées -choix des fréquences max et min et du nombre d'indexation à un moment donné -indique la fréquence de modification d'1 documents et modifie automatiquement le taux de nouvelles visites

			-reconnaît les méta-tags		les résumés (voir site	-Sun Solaris, NT, Linux -Ultraseek 4995$ (version gratuite 30 jours) +CCE 4995$
COMPAQ	**Altavista Search Intranet** (Moteur de recherche et d'indexation)	HTML, TXT, PDF, WORD, ZIP, EXCEL PostScript 150 formats différents	-indexation de chaque mot et de chaque n° -création d'un index -Prise en charge améliorée des méta tags HTML pour personnaliser les recherches par champ	-Recherche simple ou avancée -troncature , opérateurs booléens, guillemets -prend en charge la recherche par mot clé. Les mots clés de type hypertexte, applet, domaine, hôte, image, lien, texte, titre et url sont ainsi supportés de façon immédiate.	-Tri automatique des résultats en fonction de la pertinence avec sélection de la longueur du résumé -affichage standard (titre, 1ère ligne, @, date de dernière modification, taille du document) -affichage détaillé (+ rang de pertinence) -affichage compact pour la recherche élaborée (nombre de réponses) -Les recherches effectuées par le biais d'une interface de requête avancée sont triées en fonctions de termes spécifiques que l'utilisateur a choisi.	
VERITY http://www.verity.com/international/fraweb/datas_index.html *exemple d 'application* http://www.verity.com/international/fraweb/in_action.html	**Verity informati on server** (indexation et recherche en texte intégral à travers un navigateur Web standard)	-indexe plus de 200 formats de documents tel que HTML, texte, SGML, XML, Office 95, 97 et 2000, PDF	-spider capable d'indexer automatiquement le contenu des réponses et les sites Web -indexation incrémentale (seuls les nouveaux documents ou ceux qui sont modifiés sont ré-indexés) -reconnaît les méta-données	- requête simple (mots et phrases séparés par des virgules), texte libre (longue phrase) et Internet (basé sur les services de recherche Web courants) -recherche Booléenne , de proximité, recherche par zone, densité / fréquence, par champs, par concepts, avec pondération et typographique (recherche floue)	-sur brillance des termes trouvés -classement par ordre de pertinence (4 critères principaux : -catégorisation -résumé automatique -requête par l'exemple(rechercher les documents similaires à un documents trouvé)	-navigation terme à terme -1 serveur pour 50 postes désignés : 59500 F HT
FULCRUM http://www.hummingbird.com/products/dkm/french/index.html http://www.hummingbird.com/	**SearchServer**	Support de plus de 150 formats de documents (y compris PDF, HTML, SGML, XML Excel, ...) grâce à FulView, le visualisateur de document	-thésaurus international - création d'un index global comprenant tous les termes - dans un catalogue, les infos sont référencées et structurées dans des tables (à	-recherche à partir d'un mot, d'une phrase, de la racine d'un mot, expression du langage naturel - sélection des termes à partir d'1 liste de mots -parcours des mots d'un	-classement par degré de pertinence puis par titre (possibilité de structurer autrement. Ex : taille du document, auteur, date...) -mise en sur brillance des termes recherchés dans le texte -navigation d'une	-recherche intuitive permettant de trouver des documents similaires à la sélection opérée par l'utilisateur -Unix, NT -pour NT, 50 postes : 100 à 150 KF

products/dkm/pdf/SearchServerfr.pdf			chaque ligne d'une table correspond un documents ou un objet. On peut définir des colonnes contenant des infos supplémentaires par ex auteur, titre ou mot clé.) -mise à jour de l'index en mode différé ou en mode immédiat, quand la table est modifiée	documents par ordre alpha -taper 1 ou +sieurs lettres avant de sélectionner les mots pertinents -opérateurs booléens -recherche par proximité -recherche par champs (auteur, sujet, texte, date)	occurrence à l'autre à l'aide des icônes « Next Term » et « Previous Term »	
Tgid http://www.t-gid.com/produits.html	**Spirit V2+ Collector web ou collector sgbd**	Info textuelles dont HTML +utilisation de convertisseur pour 200 formats (Word, Pdf, Excel, wp)	-Indexation en texte intégral	-recherche en langage naturel sur champs textuels -recherche au travers d'un navigateur ou en client/serveur standard -recherche mixte sur champs factuels et champs textuels -opérateurs booléens sur champs factuels -proposition de re-formulation (synonymie, mot de la même famille, concept associé)	-mise en sur brillance des termes recherchés -ordre de pertinence (priorité des documents contenant les concepts identiques à ceux de la question) -info complémentaires (nature du document, date) avant visualisation -un bouton permet de naviguer dans les pages informationnelles par ordre décroissant de pertinence	-choix de stocker ou non les documents dans la base -toute partie de documents peut devenir une nouvelle question. Cela permet de rechercher des infos complémentaires sur une partie du texte (hypertexte dynamique) -monoposte : 30KF 10 postes : à partir de 60

4.3.2. Logiciels susceptibles d'être intéressants

Au départ au nombre de sept, trois ont été sélectionné, les quatre autres étant éliminés car au moins une de leurs caractéristiques importantes ne convenaient pas.

Nom de la société et site	Nom du produit	Format des documents indexés	Type d'indexation	Type de recherche proposé	Présentation des réponses	Remarques
EXCALIBUR TECHNOLOGIES http://www.forum-geide.com/Excalibur/FR-	**Retrieval Ware Retrieval Ware web server** (logiciel permettant la recherche documenta	Word, Excel, Pdf, Tiff, PostScript HTML, SGML, XML	-Indexation textuelle intégrale et automatique -indexation par le contenu et le texte associé pour les images -reconnaît les méta-données	-Langage naturel -recherche multimédia pour les images -recherche contextuelle (recherche sur un résultat de recherche) --recherche par	-sur brillance des termes recherchés (différentes couleurs selon qu'1,2 ou plusieurs termes ont été trouvé) -possibilité de QBE (Query By Example) -possibilité de cliquer sur « Go to the best	-Retrieval-Ware text serveur +visual retrievalware+screening room (video) Pour 10 postes : 20 000$ HT -sous Unix et Windows NT

RW.htm et http://www.excalib.com/products/rw/index.shtml	ire par mots clés, texte intégral et langage naturel)			séquence binaire -recherche par « concept », « pattern », « boolean » -« power search » : pour chaque mot, on choisit dans 1 liste le sens du mot (s'il existe plusieurs sens). Possibilité de pondération si la recherche comprend plusieurs mots -possibilité de recherche par champs	hit » : pour aller à l'expression qui correspond le mieux à la recherche - classement des résultats par ordre de pertinence	
ZYLAB http://www.zylab.fr/	**ZyImage 98 et ZyImage Webserver 98**	ASCII, Word, Excel, Pdf..(plus de 200 formats de fichiers) TIFF	-module d'indexation pour les images numérises et les documents déjà numérisés. -moteur d'indexation construit et maintient un fichier inversé d'index de documents	-recherche booléenne, troncature -recherche de proximité -possibilité de recherche floue (i.e donne des réponses pour les mots approchants, en plus de mots exacts) -recherche dans vocabulaire, thésaurus, concept champs, référence	-mot recherché en sur brillance -résultats peuvent être triés par densité de réponse, par date ou taille de document -possibilité d'aller d'1 occurrence à l'autre grâce à « hit »	-Windows NT - 50 postes : 90 000 F HT
THUNDERSTONE www.thunderstone.com/webinator Démo sur le Web : http://www.thunderstone.com/jump/Demonstrations.html	**Webinator**	HTML+ Adobe Acrobat / PDF file support (versions commercial et full Texis) XML +de 100 formats (pour la version commerciale et full Texis)	-Index des sites multiples dans un index commun -mise à jour de l'index pendant l'utilisation de la base -permet des bases de données multiples sur un site -permet aux sites lointains d'être recopié sur le système local -offre une interface de requête SQL pour la maintenance	-langage naturel -opérateurs booléens -opérateurs de proximité -recherche floue	-recherche de documents similaires -liens avec les documents -sur brillance des termes trouvés -classé par ordre de pertinence	-Windows NT -commercial (Intranet able, SQL SELECT, SQL DELETE, multiple database, 700 $) -Full Texis (Intranet able, SQL SELECT, SQL DELETE, SQL INSERT, SQL CREATE,,multiple database, API Interface, Unlimited Schemas, Full Texis Web script) -The Adobe Acrobat PDF and Word Processor plug-in option is $600

4.3.3. Logiciels à approfondir

Au nombre de cinq au départ, ils ne sont plus que trois à être sélectionnés. Après une étude plus approfondie, il est apparu qu'aucun ne correspondait réellement à ce qui était recherché. Aucun moteur de cette catégorie n'a donc été retenu.

Lors d'une réunion en juillet avec le groupe de travail, il a été décidé que sept logiciels seraient retenus : les six logiciels les plus intéressants plus un des logiciels susceptibles d'être intéressant. Pour choisir parmi ces sept moteurs d'indexation et de recherche, il était indispensable de les tester. La deuxième phase du stage allait commencer.

Recherche, installation et manipulation des versions d'évaluation

1. Les différentes politiques des entreprises

Selon les entreprises, trois politiques différentes de versions d'évaluation sont apparues :

- certaines entreprises, parmi les plus récentes, proposent des versions gratuites pour un temps limité téléchargeables d'Internet. C'est le cas d'Infoseek qui permet le téléchargement d'Ultraseek pour 1 mois et de Compaq qui autorise le téléchargement d'Altavista SearchEngine 3.0 pour 45 jours.
- d'autres entreprises envoient des versions d'évaluation de leur logiciels sur CD-ROM. Généralement, ils ne sont pas très enthousiastes pour le faire car ils considèrent que pour réellement maîtriser leurs logiciels il faut que le demandeur vienne faire une formation dans leur entreprise ou qu'un de leur ingénieur vienne dans l'entreprise demandeuse. Quand je leur ai expliqué que dans mon cas ce n'était pas possible, ils m'ont quand même envoyé les versions en me conseillant d'avoir un informaticien à mes côtés pour m'aider lors de l'installation. De fait, un informaticien s'est révélé indispensable pour mener à bien l'installation. Les entreprises qui ont cette politique sont Verity pour Information Server, Fulcrum pour Search Server et Excalibur Technologies pour RetrievalWare. Leurs produits datent d'un certain nombre d'années – une dizaine d'années voire plus – ce qui explique le maniement relativement difficile des versions d'évaluation contrairement aux versions directement téléchargeables d'Internet.
- la dernière catégorie d'entreprise refuse tout net d'envoyer des versions d'évaluation. En fait, elles n'en proposent pas, les logiciels qu'elles commercialisent étant trop lourd. Trop complexes, trop volumineux, dépassant de loin les simples fonctionnalités d'un moteur de recherche et

d'indexation, ces logiciels ne peuvent être testés simplement et nécessitent dans ce cas une intervention de l'entreprise. Ainsi, Arisem propose bien de tester son produit WebPortal4U mais cela coûte 30 KF pour un test limité d'un mois et nécessite qu'un ingénieur de l'entreprise s'en occupe . Il en est de même pour T-GID avec son logiciel Spirit.

2. Contact avec les entreprises

2.1. Demande de versions d'évaluation

J'ai donc réussi à obtenir, malgré quelques réticences, des versions d'évaluation pour les trois logiciels suivants :

- SearchServer de FULCRUM,
- Information Server de VERITY,
- RetrievalWare de EXCALIBUR TECHNOLOGIES.

Bien sûr, la durée du prêt était limitée (contrôle effectué par une clé de licence qui m'a été donnée quand j'ai reçu les versions d'évaluation).

2.2. Demande de démonstration

Puisqu'il est impossible d'avoir des versions d'évaluation, je me suis renseignée pour savoir s'il était possible d'avoir une démonstration. Dans le cas WebPortal4U d'ARISEM, le groupe de travail a jugé qu'il serait intéressant d'avoir une présentation et une démonstration du produit et comme cela était possible, elle a été fixé au 5 septembre.

En ce qui concerne Spirit de T-GID, le représentant m'a proposé à partir de l'URL du site du laboratoire de faire une démonstration. Après un premier contact en juillet, j'ai repris contact avec lui fin août. Depuis, n'ayant pas de nouvelles et vu qu'une décision a été prise, je n'ai pas jugé bon de le recontacter.

3. Installation des logiciels

3.1. Téléchargement à partir d'Internet

L'installation d'Altavista SearchEngine 3.0 de COMPAQ et d' Ultraseek Server d'INFOSEEK ne prend pas beaucoup de temps et peut se faire quand on le désire. Il suffit d'envoyer un mail – à n'importe quel moment – et une clé de licence est envoyée quasiment immédiatement. La seule chose dont il faut s'assurer, c'est qu'on possède la configuration informatique adéquate. En effet, Altavista SearchEngine 3.0 nécessite une configuration assez puissante qui ne correspond pas forcément à une configuration standard (Intel Pentium 300 MHz avec Windows NT 4.0, 256 MB RAM, 5 GB d'espace disque libre...). Ultraseek Server ne nécessite que Windows NT 4.0 avec 64 Mo de RAM uniquement. L'installation ne présente donc pas de difficulté.

3.2. Installation des versions d'évaluation

L'installation des versions d'évaluation fut plus complexe.

Pour Verity Information Server, il a fallu changer la configuration de l'ordinateur. En effet, ce logiciel nécessite Windows NT Server avec Service Pack 6 et Microsoft Internet Information Server (MIIS). N'ayant ni les compétences ni les logiciels pour faire cette installation moi-même, j'ai sollicité l'intervention d'un informaticien pour m'aider.

Pour Fulcrum SearchServer, a priori Windows NT 4.0 est la seule configuration nécessaire. En réalité, le logiciel n'a marché que lorsque MIIS a fonctionné – alors qu'il n'est spécifié nulle part dans le document que ce serveur Web est nécessaire – et lorsque qu'on trouvé le mot de passe administrateur – qui lui aussi n'est indiqué nulle part et qui se trouve être « password ».

En ce qui concerne RetrievalWare, Windows NT Server est aussi indispensable ainsi que 128 MB de RAM. Avec 6 CD-ROM et une installation qui, d'après la documentation, semblait très compliquée, j'ai décidé de ne pas tester ce logiciel, me contentant uniquement de la documentation fournie – qui est très détaillée.

4. Manipulation des logiciels

4.1. Temps de prise en main

Une fois la partie installation terminée, la manipulation a pu commencer pour les quatre logiciels installés. Pour Ultraseek Server, la prise en main fut facile. Pour Altavista SearchEngine, elle fut un peu plus longue car le logiciel est plus complexe. Il nécessite une certaine habitude de l'environnement informatique car il faut intervenir dans certaines fonctionnalités comme l'activation des javascripts. Verity Information Server, du fait qu'une clé de licence pour Verity Spider – qui indexe les sites distants – ne m'a pas été donnée, a nécessité l'import du site test en local. A part cela, le prise en main s'est passée sans problème. Quant à Fulcrum SearchServer, le manque de convivialité et de simplicité du logiciel au niveau administrateur a rendu très difficile voire impossible la prise en main de ce logiciel.

4.2. Complexité de la manipulation

Ultraseek Server est assez simple d'utilisation. Il suffit de rentrer l'URL du site que l'on veut indexer, de changer les paramètres des requêtes si besoin est – le logiciel les définit par défaut –, de préciser le format des documents que l'on veut indexer...Tous les paramètres sont soit définis lors de l'installation soit définis par défaut par le logiciel. Dans tous les cas, ils peuvent être modifiés par l'administrateur.

Altavista SearchEngine marche à peu près sur le même modèle. Son robot indexeur appelé Mercator indexe à partir de l'URL du site – distant ou non. Contrairement aux autres logiciels, ce moteur indexe aussi les bases de données et les systèmes de fichiers.

Verity Information Server guide l'administrateur avec son module Quick Start qui permet de vérifier la configuration de notre système, d'indexer notre site Web, de chercher sur le site indexé. On peut soumettre une nouvelle indexation, voir la liste des collections et la situation du serveur...Il est aussi possible d'indexer des répertoires locaux. Dans l'ensemble, l'administrateur est bien guidé et l'utilisation de ce logiciel est facile et conviviale.

Quant à Fulcrum SearchServer, ce n'est pas vraiment la même chose. L'administration est loin d'être facile. Cela provient du fonctionnement du logiciel : Fulcrum SearchServer est basé exclusivement sur le modèle des tables SQL, un

enregistrement correspondant à un document. Avant l'indexation, il faut donc créer la table qui correspond aux documents que l'on veut indexer, définir les colonnes de la table qui correspondent à certains attributs d'un document comme le titre, l'auteur, le sujet, les mots clés et le texte du document. Après il faut préciser les options de la table (notamment le type d'indexation) puis donner le chemin d'accès aux documents que l'on veut indexer, etc. Cette gestion est très lourde et n'est pas évidente à appliquer pour l'administrateur. Pour vérifier que l'indexation s'est bien passée, il a la possibilité de faire des requêtes en SQL – ce qui n'est pas convivial – . Le module WordSense permet aux utilisateurs de faire des requêtes en langage naturel, dans la langue de leur choix. N'ayant pas compris son fonctionnement, je n'ai pas pu le faire marcher. Je dirais que ce moteur est le plus difficile, le plus complexe à utiliser des quatre moteurs. De plus, il n'est pas convivial. Je crois qu'un tel outil nécessite impérativement qu'un représentant de la société Fulcrum explique le fonctionnement à la personne chargée d'administrer ce moteur.

4.3. Complexité de la documentation

On peut noter deux politiques différentes :

- d'un côté, les moteurs d'indexation et de recherche directement téléchargeables d'Internet qui ont une documentation très limitée. Généralement, c'est une aide en ligne qui se limite à une aide à la recherche. Il n'y a aucune aide sur la manière d'utiliser le logiciel. Pour ce qui est de l'installation, tout est expliqué sur le site d'où on peut télécharger le logiciel. Généralement, se trouve aussi des FAQ (Frequently Asked Questions) qui permettent éventuellement de trouver une solution lors d'un problème de manipulation de logiciel. Dans tous les cas, une notice expliquant le fonctionnement général, les avantages et les spécificités du moteur est disponible. Cette politique ne pose pas de problème car la manipulation et l'utilisation de ces outils sont faciles et conviviales donc ne nécessitent pas de documentation développée.
- à l'opposé se trouvent les versions d'évaluation disponibles sur CD-ROM. La politique est totalement différente et s'expliquent par le fait que ce sont des logiciels complexes, aussi bien au niveau de l'installation que de la

manipulation. Il y a des notices explicatives – dont le nombre de pages est assez conséquent – pour tout : l'installation, l'intégration, l'administration, le guide de l'utilisateur... De plus, ces logiciels permettent généralement de faire de la programmation informatique : une partie non négligeable de la documentation est consacrée à cela. Dans tous les cas, la documentation est très complète. Mais elle n'est pas forcément très accessible. C'est le cas de Fulcrum dont la documentation est aride et complexe – à l'image du logiciel. Il est à noter qu'une partie voire toute la documentation est installée avec le logiciel. Elle est donc accessible en ligne mais ce n'est pas forcément plus facile pour chercher des informations , sauf quand il y a un index.

5. Visite de sites utilisant des moteurs

Visite des sites suivants :

- site de l'Adit (www.adit.fr) qui utilise Verity information Server,
- site du W3C (http://www.w3.org/) qui utilise Altavista SearchEngine 3.0,
- site de la DIST du CEA (www-dist.cea.fr) qui utilise Spirit,
- site de l'Education Nationale (http://educlic.education.fr/Arisem23/iClass4U/) qui utilise WebPortal4U.

5.1. Facilité d'utilisation du moteur utilisé sur le site

En ce qui concerne le site du CEA, les fonctionnalités proposées par le logiciel Spirit sont intéressantes. L'interrogation se fait en langage naturel soit dans la partie thème – où sont pris en compte les mots du titre et les mots-clés – soit dans le titre – où sont pris en compte les mots du titre des livres et des congrès uniquement –.

Pour la recherche simple, le traitement linguistique permet de rechercher le vocable donné, et ses variantes singulier/pluriel pour les noms, singulier/pluriel/masculin/féminin pour les adjectifs, ou de temps pour les verbes.
Pour la recherche élargie, le traitement linguistique permet de plus d'étendre la recherche aux synonymes et termes équivalents du vocable donné.

Bien entendu pour chaque terme jugés équivalent toutes ses variantes de type genre/nombre/masculin/féminin ou temps des verbes sont aussi cherchées comme dans la recherche simple.

Il est aussi possible de chercher par auteur, par année de publication, par date de congrès et par numéro de rapport.
L'utilisateur peut changer le nombre maximum de documents à afficher.

A première vue, l'utilisateur peut faire une recherche assez complète d'une manière simple et conviviale. Cependant, il ne faut pas oublier que c'est le CEA qui s'est occupé du développement informatique de Spirit pour T-GID mais maintenant ne le fait plus. La version qu'il y a sur le site du CEA peut donc être différente de la version proposée par T-GID. Mais cela donne quand même une idée de ce que propose le logiciel.

Pour ce qui est du site de l'ADIT, une recherche simple par mot-clé est actuellement proposée pour l'interrogation de la base de donnée, la recherche avancée par concepts associés étant en cours de réactualisation. Il est aussi possible de rechercher par zone comme le titre, l'auteur, le sujet, les mots-clés du document... Certains des opérateurs proposés par le moteur de Verity sont utilisables ainsi que quelques caractères génériques. En ce qui concerne la présentation des réponses, le score – entre 0 et 1 –, le titre du document, le résumé et la prévisualisation. Le résultat de la recherche est présenté uniquement sous le format PDF. Bien que toutes les fonctionnalités du moteur Verity ne soient pas exploitées, ce site est intéressant et montre combien ce moteur est convivial et facile d'utilisation.

WebPortal4U d'Arisem est un produit assez particulier. Le CNDP l'a utilisé pour créer un site portail pour les professionnels de l'éducation. L'utilisateur peut donc faire une recherche par catégories et sous-catégories ou une recherche directe sur les documents des 200 serveurs éducatifs. Cette dernière recherche n'est possible que parce que le moteur Verity a été intégré à la solution d'Arisem. Les utilisateurs ont la possibilité de personnaliser leur portail. L'écran d'accueil de ce site est convivial et agréable contrairement au site www.atmedica.com qui utilise aussi le produit d'Arisem. Il montre ce que peut faire en apparence WebPortal4U, sachant que derrière se cache

tout un traitement sémantique des documents, une base de connaissance ainsi qu'un plan de classement automatique.

En apparence, l'interrogation sur le site du World Wide Web Consortium avec Altavista Search Engine n'est pas différente de l'interrogation que l'on peut faire sur le site d'Altavista. Toutefois, il existe des différences. On peut par exemple choisir pour la présentation des réponses la forme standard, la forme compacte et la forme détaillée. Les deux modes de recherche simple et avancée sont toujours proposées. L'utilisateur peut limiter sa recherche à une partie du site.

5.2. Contact avec certains webmasters

J'ai envoyé des mails – par deux fois – aux webmasters du site du CEA, de l'ADIT et du CNDP. Je leur demandais des renseignements sur l'installation, l'administration et la gestion de l'outil de recherche qu'ils utilisaient sur leur site. Aucun d'entre eux m'ayant répondu, je n'ai pu avoir de témoignages de personnes utilisant ces outils. C'est dommage car cela aurait été très instructif.

Tests des logiciels

1. Approfondissement des critères d'évaluation

1.1. Site du CEA [6]

La Direction de l'Information Scientifique et Technique (la DIST) du CEA a effectué une comparaison assez poussée en 1998 de treize différents moteurs de recherche accessibles sur le Web (dont Altavista et Infoseek). Ils se sont attachés particulièrement à la collecte des documents, l'indexation des documents, la recherche des documents et la présentation des résultats. Seul un certain nombre de critères et de renseignements étaient intéressants pour moi parmi la quarantaine proposée. Le tableau comparatif qui en ressort est très complet.

Les principaux critères que j'ai sélectionnés et que j'ai un peu modifiés dans certains cas sont les suivants :

- en ce qui concerne l'indexation des documents, je me suis attachée à plusieurs choses :
  - la méthode d'indexation : on distingue ici deux méthodes d'indexation, l'indexation automatique et l'indexation manuelle. A cela j'ai rajouté l'indexation incrémentale.
  - les données indexées : les critères d'indexation peuvent être multiples et variés. Nous avons retenu :
    - Le titre du document,
    - ses différents sous-titres (balises <H1>...<Hn>),
    - son en-tête (le <META> tag),
    - sa date de création et/ou modification,
    - sa taille,
    - les URLs qu'il cite,
    - le texte des URLs qu'il cite,
    - d'autres balises éventuelles,
    - un résumé du document,

- un extrait du document,
- et enfin le texte intégral du document.

- pour ce qui est de la recherche de document :
  - le type de recherche : on distingue trois types de recherche : booléenne, avec pondération et floue (ou typographique)
  - le type de question : trois modes d'interrogation peuvent être proposés : la question en langage naturel (formulation d'une question en langage libre), interrogation par mot clé et la requête avec des expressions booléennes,
  - les opérateurs booléens proposés :

    Nous distinguons ici,
    - tout d'abord les opérateurs booléens classiques (ET, OU, NON ou AND, OR, NOT),
    - la combinaison des opérateurs, qui doit permettre dans une même requête booléenne de mêler les différents opérateurs disponibles afin d'effectuer des recherches plus élaborées que celle ne pouvant contenir qu'un type d'opérateur,
    - le parenthésage des expressions donnant alors la possibilité de créer des requêtes très complexes comprenant plusieurs niveaux de parenthèses et plusieurs types d'opérateurs,
    - la possibilité de spécifier une certaine distance en terme de mots entre certains termes de la question (opérateur de proximité),
    - enfin, le support de l'opérateur d'adjacence qui est en fait un cas particulier de l'opérateur de proximité. Il permet de spécifier que certains des mots de la recherche ne doivent être séparés par aucun autre terme. Nous pouvons trouver dans les systèmes de recherche actuels deux formes d'opérateurs d'adjacence. Tout d'abord sous la forme d'un opérateur du type <mot1> ADJ <mot2> qui signifie que l'on recherche des documents où <mot1> et <mot2> sont adjacents. Mais on trouve plus couramment ce que les moteurs de recherche appellent la recherche de phrase qui revient au même (mais qui est

cependant moins souple dans le cas d'expressions parenthésées complexes). La syntaxe généralement utilisée est "<mot1> <mot2>" (phrase entre guillemets) pour recherche les documents dans lesquels les mots <mot1> et <mot2> sont adjacents.

- o la troncature manuelle : on distingue la troncature gauche (par exemple, pour la chaîne *matique, le moteur de recherche va récupérer les documents contenant les termes informatique, télématique, mathématique, ...,), la troncature droite (pour la chaîne inform*, le moteur de recherche doit retourner les documents contenant les mots informatique, informaticien, information, informationnel, informe, informel, informulé,...) et la troncature interne (qui permet de spécifier le début et la fin d'un mot, en laissant une partie libre. Par exemple, poi*on donnera poison, poisson, poivron, ...).
- o les champs de recherche : cette rubrique énumère les divers champs dans lesquels le système effectue la recherche. Nous avons dégagé plusieurs champs : URL, titre, sous-titre, auteur, sujet, date, mots clés, type de document, taille, résumé, texte intégral, méta-tag, URL cité.

1.2. Site Abondance [7]

Ce site est très développé et constitue une référence en ce qui concerne les outils de recherche. Il propose un tableau comparatif des principaux moteurs de recherche mondiaux (dont Altavista et Infoseek). Pour chaque moteur, sont développés :

- la syntaxe de saisie donc les possibilités proposées pour la recherche simple (minuscules/majuscules, lettres accentuées, ordre des mots, opérateurs booléens, troncature, recherche sur différents champs...) et pour la recherche avancée (opérateur de proximité, recherche sur d'autres champs..),
- les champs indexés (titre, différentes balises META, importance relative des différents critères...),
- les champs affichés dans la pages de résultats (titre, résumé, taille du document, date, URL, autres...).

Ce tableau comparatif, bien que prenant en compte beaucoup de critères, n'est pas aussi complet que le tableau proposé par la DIST du CEA. Toutefois, il est intéressant car il permet de mettre à jour certaines informations concernant Altavista et Infoseek.

1.3. Site du Service de recherche documentaire DSI [8]

C'est un autre site intéressant proposant un comparatif des différents moteurs disponibles sur le Web (dont font toujours partie Altavista et Infoseek). Le comparatif est divisé en plusieurs tableaux :

- tableau des opérateurs logiques,
- tableau des champs de recherche (texte intégral, casse, accent, vérification orthographique, langue, méta-mot, méta-titre, adresse URL, domaine, pays, hyperlien, serveur), affichage des réponses...

Dans ce cas non plus, ce tableau ne m'a pas apporté d'éléments nouveaux. Il m'a permis de vérifier que les informations concernant Altavista et Infoseek produites par le CEA étaient toujours bonnes.

2. Définition d'un protocole d'expérimentation

2.1. Création d'un site test

J'ai demandé à Anne Tchounikine, webmaster au LISI, de faire une copie du site. J'ai jugé plus prudent de ne pas faire les tests directement sur le site du LISI au cas où il y aurait un problème lors des tests. Un site test a donc été créé auquel j'avais accès en tant qu'administrateur, pouvant ainsi rajouter, modifier ou supprimer des documents comme je le souhaitais.

2.1.1. Choix des documents initiaux

A l'origine, le site comportait presque exclusivement des documents HTML, quelques documents ZIP, des images GIF, un ou deux documents Word et PostScript. Ce panel de documents était insuffisant pour que les tests sur le site soient complets et détaillés.

2.1.2. Rajout et amélioration des documents en vue des tests

Les discussions lors de la présentation du premier tableau comparatif détaillé ont montré les faiblesses de ce dernier. Il est apparu qu'il y avait un certain nombre de critères à développer. Pour cela, il fallait effectuer certains tests sur le site test et donc, pour se donner les possibilités de le faire, modifier et améliorer le site test. Cela s'est traduit par :

- des rajouts de méta-tags description, keywords, author dans certains documents,
- des rajouts de documents au format PDF (13 au total dont certains ont les champs informations générales remplis), au format XML (6 en tout), ZIP (3 fichiers), Postscript (4 fichiers), au format DOC.

  Il est à noter que le rajout des documents PS et XML a nécessité respectivement le téléchargement d'un viewer Postscript et de la version β de Netscape 6 pour pouvoir les visualiser,
- le rajout de deux documents HTML avec dans l'un, des mots en minuscule, des mots au singulier et des mots sans accent et dans l'autre, les mêmes mots en majuscule, au pluriel et avec accent.

2.2. Conduite de l'expérimentation

2.2.1. Données indexées par rapport aux formats de documents

Suivant les conseils du groupe de travail, j'ai décidé de développer les données indexées par rapport au format des fichiers indexés :

- pour le format HTML, les critères d'indexation sont : le titre, les sous-titres (<Hn>), les méta-tags author, description, keywords, l'attribut ALT des balises IMG, l'intitulé de l'URL, l'hôte, le domaine, l'adresse des liens (link) , l'intitulé des liens hypertextes (anchor), l'applet, frame, imagesmaps, corps du texte.
- pour le format XML, les critères d'indexation sont : title, description, keywords, body, date et corps du texte.

- pour le format PostScript, les critères d'indexation sont : titre, type, date, corps du texte.
- pour le format PDF, les critères d'indexation sont : auteur, titre, sujet, date de modification, mot clé, corps du texte.

2.2.2. Gestion des minuscules/majuscules

En ce qui concerne les minuscules et les majuscules, j'ai cherché à savoir si :

- quand on écrit un mot en minuscule, il cherche le mot en minuscule ET le mot en majuscule c'est-à-dire : mot → mot et Mot ?
- quand on écrit un mot en majuscule, il cherche le mot en majuscule ET le mot en minuscule c'est-à-dire : Mot → mot et Mot ?
- quand on écrit un mot tout en lettres capital, il cherche aussi les mots en minuscule ET en majuscule c'est-à-dire : MOT→ mot et Mot ?

2.2.3. Gestion des accents

J'ai procédé de la même façon pour vérifier la gestion des accents :

- quand on écrit un mot avec accent, cherche-t-il le mot avec accent ET le mot sans accent c'est-à-dire : si mot accentué → mot accentué et mot non accentué ?
- quand on écrit un mot sans accent, cherche-t-il le mot avec accent ET le mot sans accent c'est-à-dire : si mot non accentué → mot accentué et mot non accentué ?
- si on cherche un mot accentué (comme événement), cherche-t-il le même mot accentué différemment (comme évènement) ?

2.2.4. Gestion du singulier/pluriel

De même pour le singulier et le pluriel.

- si le mot recherché est au singulier, cherche-t-il le pluriel aussi ?
- si le mot recherché est au pluriel, cherche-t-il le singulier aussi ?

2.2.5. Gestion des mots vides

Je n'ai considéré que le cas où la gestion des mots est faite par une liste (non pas par l'élimination des mots dépassant un certain nombre d'occurrence dans la base). J'ai essayé de savoir si la gestion était automatique (i.e. la liste est directement définie par le moteur sans intervention de l'administrateur) ou manuelle (c'est l'administrateur qui doit définir lui-même sa propre liste de mots vides).

2.2.6. Etude des possibilités d'amélioration de la recherche

- Thésaurus

  J'ai essayé de déterminer deux choses :

  - d'une part, au niveau de la gestion c'est-à-dire au niveau de l'administrateur : est-elle automatique (i.e. le thésaurus est-il directement constitué par le moteur de recherche et d'indexation sans intervention de l'administrateur ?) ou manuelle (l'administrateur peut-il intervenir dans la constitution de ce thésaurus ?),
  - d'autre part, au niveau de l'utilisation c'est-à-dire au niveau de l'utilisateur courant : est-elle automatique (le moteur utilise-t-il directement la fonction thésaurus lors d'une recherche de l'utilisateur sans que celui-ci le demande ?) ou manuelle (l'utilisateur doit-il préciser d'une façon quelconque au système qu'il veut utiliser le thésaurus pour sa recherche ?).

- Lemmatisation

  Le raisonnement pour la lemmatisation est similaire :

  - d'un côté, au niveau de la gestion c'est-à-dire au niveau de l'administrateur : est-elle automatique (i.e. la lemmatisation est-elle directement constituée par le moteur de recherche et d'indexation sans intervention de l'administrateur ?) ou manuelle (l'administrateur peut-il intervenir dans la constitution de la lemmatisation?),
  - d'un autre côté, au niveau de l'utilisation, c'est-à-dire au niveau de l'utilisateur : est-elle automatique (le moteur utilise-t-il directement la fonction lemmatisation lors d'une recherche de l'utilisateur sans que

celui-ci le demande ?) ou manuelle (l'utilisateur doit-il préciser d'une façon quelconque au système qu'il veut utiliser la lemmatisation pour sa recherche ?).

- correction orthographique :

  Dans ce cas-là, j'ai restreint à l'utilisation à savoir si un utilisateur quelconque devait préciser qu'il voulait la fonction correction orthographique (utilisation manuelle) ou le moteur l'appliquait systématiquement (utilisation automatique).

2.2.7. Etude des méthodes de recherche en fonction des données

Vu que j'ai développé les données indexées en fonction des formats de documents, il m'a semblé pertinent de développer les possibilités de recherche en fonction des données indexées et donc en fonction du format de documents :

- pour le format HTML :
  - est-il possible de rechercher par méta-tags, c'est-à-dire est-il possible de préciser que l'on ne recherche que sur les champs author, description ou keywords ?
  - est-il possible de rechercher sur d'autres champs par exemple uniquement sur le titre du document ?
- pour le format PDF, est-il possible de rechercher uniquement sur les champs auteur, titre, sujet, mot-clé ?
- pour le format PostScript, est-il possible de rechercher sur les champs type, title, date ?
- pour le format XML, peut-on faire une recherche sur les champs title, description, keywords, body, date ?

Résultats

1. Données qualitatives, synthèse

Dans l'ensemble, je dirai d'une part que ces logiciels ne nécessite pas de compétences informatiques particulièrement développées. En ce qui concerne Ultraseek et Altavista, aucun problème. Pour les autres, il ne faut pas oublier que les entreprises proposent une formation à leurs logiciels. Je ne pense pas qu'un informaticien soit indispensable pour la gestion des moteurs d'indexation et de recherche, une personne ayant des connaissances et des compétences en informatique sans être spécialisée dans le domaine conviendrait. D'autre part, au niveau des possibilités d'évolution de ces logiciels, vu la lourdeur et la complexité de Fulcrum SearchServer, Verity Information Server et d'Excalibur RetrievalWare, je pense que les nouvelles versions sont assez rares et les mises à jour d'une ancienne version ne sont guère aisées. Pour les logiciels téléchargeables d'Internet, c'est plus facile vu que les mises à jour peuvent être directement téléchargées. Mais là aussi je ne suis pas sûre que les nouvelles versions soient si fréquentes que ça.

2. Données quantitatives

	indexation des	automatique	incrémentale	manuelle	formats des fichiers indexés	HTML	données indexées	titre	sous-titre (<Hn>)	méta-tag: author	méta-tag:description	méta-tag:keywords	attribut ALT des balises IMG	intitulé d'URL	hôte	domaine	lien	anchor	applet	frame	imagesmaps	coprs du texte
AltaVista SearchEngine 3.0		1	0	1		1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ultraseek Server d'Infoseek		1	1	1		1		1	0	1	1	1	1	1	0	0	0	0	0	1	1	1
Veriry Information Server		1	1	1		1		1	1	1	1	1	0	0	0	0	0	0	0	0	0	1
Fulcrum SearchServer		1	?	1		1		1	0	1	1	1	0	0	0	0	0	0	0	0	0	1
RetrievalWare d'Excalibur		1	?	1		1		?	?	?	?	?	?	?	?	?	?	?	?	?	?	1
Spirit V2 de T-GID		1	?	1		1		?	?	?	?	?	?	?	?	?	?	?	?	?	?	1

	PostScript	**données indexées**	titre	type	date	corps du texte		WORD	**données indexées**	corps du texte
AltaVista SearchEngine 3.0	1		?	?	?	?		1		1
Ultraseek Server d'Infoseek	1		?	?	?	?		1		1
Veriry Information Server	1		?	?	?	?		1		1
Fulcrum SearchServer	1		?	?	?	?		1		1
RetrievalWare d'Excalibur	1		?	?	?	?	?	1		1
Spirit V2 de T-GID	?		?	?	?	?		1		1

	PDF	**données indexées**	auteur	titre	sujet	date de modification	mot-clé	corps du texte	XML	**données indexées**	titre	description	keywords	body	date	corps du texte
AltaVista SearchEngine 3.0	1		?	?	?	?	?	1	?							
Ultraseek Server d'Infoseek	1		?	?	?	?	?	1	1		1	1	1	1	1	1
Veriry Information Server	1		1	1	1	1	1	1	1		?	?	?	?	?	1
Fulcrum SearchServer	1		1	1	1	0	1	1	1		?	?	?	?	?	1
RetrievalWare d'Excalibur	1		?	?	?	?	?	?	1		?	?	?	?	?	?
Spirit V2 de T-GID	1		?	?	?	?	?	?	?							

Recherche des	Type de recherche				Types de question				Opérateurs booléens							Troncature manuelle			
	booléenne	avec pondération	floue(typographique)		langage naturel	mots-clés	expressions booléennes		AND	OR	NOT	combinaison des opérateurs	opérateur de proximité	opérateur d'adjacence	parenthèsage		ganche	droite	interne
AltaVista SearchEngine 3.0	1	0	0		0	1	1		1	1	1	1	1	1	1		0	1	1
Ultraseek Server d'Infoseek	1	0	0		1	1	1		1	1	1	1	1	1	1		0	0	0
Veriry Information Server	1	1	1		1	1	1		1	1	1	1	1	1	1		0	1	0
Fulcrum SearchServer	1	1	1		1	1	1		1	1	1	1	1	?	?		1	1	1
RetrievalWare d'Excalibur	1	1			1	1	1		1	1	1	1	1	1	1		?	1	1
Spirit V2 de T-GID	1	0	1		1		1		1	1	1	1	?	?	1		?	1	?

Champs de recherche	HTML																XML							
AltaVista SearchEngine 3.0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	?							
Ultraseek Server d'Infoseek	1	1	1	1	0	1	1	0	0	1	0		1	0	1	1	1	1	1	1	1	1	1	1
Veriry Information Server	1	0	1	0	0	0	0	0	0	0	0		1	0	1	1	1	1	?	?	?	?	?	
Fulcrum SearchServer	1	0	1	0	0	0	0	0	0	0	0		0	0	1	1	1	1	?	?	?	?	?	
RetrievalWare d'Excalibur	1	?	?	?	?	?	?	?	?	?	?		?	?	?	?	?	1	?	?	?	?	?	?
Spirit V2 de T-GID	1	?	?	?	?	?	?	?	?	?	?		?	?	?	?	?	?						

	PDF		title	author	subject	keywords	corps du texte		Gestion des mots vides			Gestion des MAJ/min			Gestion de l'accentuation			gestion du singulier/pluriel	
									gestion manuelle	gestion automatique		minuscule -> MAJUSCULE et MAJUSCULE -> minuscule et MAJUSCULE			accentué -> non accentué et accentué non accentué -> accentué et non			singulier-> pluriel et singulier pluriel->singulier et pluriel	
AltaVista SearchEngine 3.0	1		?	?	?	?	1		?	?		1	0		0	1		0	0
Ultraseek Server d'Infoseek	1		?	?	?	?	1		?	?		1	0		0	1		0	1
Veriry Information Server	1		1	1	1	1	1		1	0		1	0		0	0		1	1
Fulcrum SearchServer	1		1	1	1	1	1		1	1		1	0		?	1		1	?
RetrievalWare d'Excalibur	1		?	?	?	?	1		?	?		?	?		?	?		?	?
Spirit V2 de T-GID	1		?	?	?	?	1		?	1		1	1		1	1		1	1

	Amélioration de la recherche																	
	thésaurus				lemmatisation					correction orthographique			recherche par concept	recherche documents similaires(QBE)				
	gestion manuelle	gestion automatique	utilisation manuelle	utilisation automatique	gestion manuelle	gestion automatique	utilisation manuelle	utilisation automatique		utilisation manuelle	utilisation automatique							
AltaVista SearchEngine 3.0	1	1	1	0	1	1	1	0		0	1		0	0				
Ultraseek Server d'Infoseek	1	1	?	?	1	1	0	1		?	?		0	1				
Veriry Information Server	1	1	1	0	1	1	1	0		0	1		1	1				
Fulcrum SearchServer	1	1	1	0	0	1	1	0		0	1		0	1				
RetrievalWare d'Excalibur	?	?	?	?	?	?	?	?		?	?		0	1				
Spirit V2 de T-GID	0	0	0	0	0	1	0	1		0	1		0	?				

	Présentation des résultats																							
	Informations générales				organisation des documents-réponses						Informations concernant les documents													
	1	0			0	0	1	0	0	1	1	1	0	1		1	1	0	0	0	1	0	0	0
AltaVista SearchEngine 3.0	nombre de documents-réponses	navigation terme à terme			tri par date	tri par titre	par mot	tri par URL	tri par auteur	tri par pertinence	URL du document	titre du document	mots clés	résumé (contenu de la balise META Description ou 1ères lignes du doc)		taille du document	date de dernière modification	date de dernière visite	score (mesure de pertinence)	mise en évidence de mots de la question				
Ultraseek Server d'Infoseek	1	0			1	1	0	0	0	1	1	1	0	1		1	1	0	1	1	*			
Veriry Information Server	1	1			0	1	0	1	0	1	1	1	0	1		1	1	0	1	1				
Fulcrum SearchServer	1	1			1	0	0	0	1	1	0	1	1	0		0	0	0	1	1				
RetrievalWare d'Excalibur	1	1			0	1	0	0	0	1	0	1	1	1		1	1	0	1	1				
Spirit V2 de T-GID	1	1			?	?	0	0	0	1	0	1	0	0		0	1	0	0	1				

*uniquement dans le titre et le résumé

3. Affinement du cahier des charges, conclusion

La démonstration faite par le représentant d'ARISEM sur son produit WebPortal4U fut très instructive. Il est primordial de comprendre que ce produit n'est pas un moteur de recherche et d'indexation. Il crée des sites portails en utilisant une analyse sémantique des documents à partir d'une base de connaissance intégrée au logiciel et développée ensuite suivant le secteur d'activité de l'entreprise. Il offre des portails personnalisés pour une communauté d'utilisateurs et intègre des possibilités de veille sur profil. Ces fonctionnalités dépassent donc de loin les besoins de l'INSA définis au début de mon stage. Mais c'est un produit intéressant et très riche. Il ne faut toutefois pas oublier que si l'on veut que l'utilisateur fasse une recherche par mot – et non pas uniquement par catégorie et sous-catégorie –, il faut intégrer un moteur de recherche et d'indexation du genre ceux que proposent Verity et Sinequa.

Une réunion avec Mme Joly et M. Mermet de DocINSA et M. Kouloumdjian en vue de l'affinement du cahier des charges a permis de faire un bilan à partir du comparatif que j'ai effectué. Il est apparu qu'Altavista Search Engine 3.0 et Infoseek offraient le meilleur rapport qualité/prix et répondaient le mieux aux besoins définis lors du projet. De plus, Ultraseek Server permet de créer des sites portails grâce à son module CCE (Content Classification Engine) et Altavista Search Engine 3.0 le permet aussi (voir le site d'Amazon.com). D'après PCWEEK LABS qui a testé le logiciel (consulter l'article http://www.zdnet.com/eweek/stories/general/0,11011,2553306,00.html), les coûts d'acquisition, de mise en œuvre et de formation, de maintenance sont faibles et le temps d'attente des bénéfices est court.

Bibliographie

Format des références : en ce qui concerne les pages Internet, du fait qu'il n'est pas toujours possible d'avoir une référence complète, j'ai donné à chaque fois :

- le nom de l'auteur –quand il est indiqué- sinon le nom de l'organisme produisant le document
- l'organisme auquel appartient l'auteur
- le nom de la page
- la date de dernière mise à jour - ou la date de dernière visite
- l'URL

En ce qui concerne les autres références, elles ont présentées en appliquant les deux normes suivantes :

- Z44-005 "Documentation, références bibliographiques : contenu, forme et structure" de décembre 1987 qui elle-même reproduit intégralement la norme internationale ISO 690-1987.
- ISO/DIS 690-2 -1995 "Information et documentation – Références bibliographiques – Documents électroniques ou parties de ceux-ci".

[1] *Guide d'achat 2000 : les systèmes d'information documentaire.* Paris : Archimag, 2000.

[2] **LELOUP** Catherine. *Moteur d'indexation et de recherche : environnement client-serveur, Internet et Intranet.* Paris : Eyrolles, 1998. 285 p.

[3] **APROGED** (**A**ssociation des **PRO**fessionnels de la **GED**). Le site du Forum Geide : logiciels texte intégral.
Date de dernière visite : juin 2000.
<URL : http://www.forum-geide.com/>

[4] **SOFTSEEK (Your source for Shareware, Freeware and Evaluation Software)**. Search Enabling Web Site.
Date de dernière visite : juin 2000.
<URL : http://www.softseek.com/Internet/Web_Publishing_Tools/Search_Enabling_Web_Sites/>

[5] **SAMIER Henry, SANDOVAL Victor**. *La recherche intelligente sur l'internet : outils et méthodes*. Paris :Hermès, 1998. 155 p.

[6] **DIST/CEA.** Les principaux problèmes la recherche d'informations sur Internet, tableaux comparatifs, l'évaluation des moteurs de recherche.
Date de dernière mise à jour : 1998
<URL : http://www-dist.cea.fr/ext/neuf/moteur/ >

[7] **ANDRIEU Olivier (Site Abondance, le site des outils de recherche)**. Comparatifs des fonctionnalités des principaux moteurs mondiaux.
Date de dernière mise à jour : mai 2000
<URL : http://www.abondance.com/outils/comparatif.html>

[8] **DUVAL Marc (Service de recherche documentaire DSI)**. Les moteurs de recherche.
Date de dernière mise à jour : septembre 2000.
<URL : http://pages.infinit.net/duvalm/dossiers/moteurs_de_recherche.html>

ANNEXE

1. LEXIQUE

Analyse syntaxique : technique destinée à attribuer aux termes d'un texte leurs attributs grammaticaux : nom, verbe, adjectif,...

Analyse sémantique : technique d'analyse du sens des mots, des phrases, des textes et des discours.

Applet : petit programme Java, intégré tel quel dans une page HTML et qui accomplit une fonction. Ce programme tourne quel que soit la plate-forme qui le supporte.

Base de connaissance : partie d'un système expert contenant l'ensemble des règles, d'hypothèses, d'opinions, de convictions et de faits qui constituent le domaine de compétences du système.

Domaine : dans l'Internet et les autres réseaux, subdivision la plus haute d'un nom de domaine dans une adresse de réseau, qui identifie le type d'entité possédant cette adresse. Il peut représenter un pays (par exemple : .fr), une activité (par exemple : .edu pour éducation, .com pour commercial), une institution (par ex : .org pour organisation), etc.

Filtre : outil qui récupère le texte d'un document bureautique - pour l'indexer - , sous la forme d'un flot de texte, dans le jeu de caractère exploitable par le moteur et l'outil de visualisation. Le cas échéant, si le document est structuré, on conservera les balises des champs que l'on souhaite identifier pour l'indexation.

Hits : nombre d'occurrences trouvées d'un terme sur les documents d'un lot résultat.

Hôte (Host) : ordinateur distant qui reçoit les appels d'autres machines (connexions sur un site Web, par exemple). Cet ordinateur est relié à un réseau et offre des services aux autres ordinateurs.

HTTP (HyperText Transfer Protocol) : norme pour transférer des données entre des serveurs Web et des clients. C'est une méthode utilisée pour transporter des pages HTLM du WWW sur le réseau.

Index : liste ordonné alphabétiquement des éléments ou données contenus dans un document ou un fichier, qui permet de les localiser.

Indexation automatique : technique d'indexation passant par une analyse de texte conduisant à extraire les principaux concepts, puis à représenter ceux-ci par les termes du langage contrôlé choisi (typiquement un thésaurus).

Indexation en texte intégral ou « full text »: technique d'indexation et de recherche permettant d'accéder à des textes par des requêtes portant sur des mots, expressions et conditions de présence dans tout ou une partie du texte. Les fichiers d'index (ou fichiers d'index positionnels) construits lors de l'indexation comprennent davantage d'informations que les fichiers d'index normaux. Pour chaque document, l'information gérée dans le fichier d'index texte intégral comprend outre la clé – ou numéro d'enregistrement – , la position du mot dans le texte.

Indexation incrémentale : cette méthode évite la ré-indexation complète du site, en prenant en compte uniquement les documents modifiés.

Indexation multi-formats : indexation de multiples formats de fichiers (PDF, WORD, PostScript, XML...).

Lemmatisation ou « Stemming » : fonction que possèdent certains moteurs de recherche et répertoires permettant d'obtenir des résultat sur les mots qui ont la même base que le mot clé saisi. Par exemple, lorsque vous sélectionnez cette recherche élargie et que vous voulez avoir des informations sur la danse, vous pouvez saisir dans* comme mot clé et vous aurez dans les résultats danse, danseur, danseuse et dansant.

Matches : nombre de réponses à une requête (nombre de documents dans un lot résultat). .

Métadonnées : données au sujet d'éléments de données, y compris leurs descriptions de données, ou données au sujet de la propriété des données, des chemins d'accès, des droits d'accès et de la volatilité des données.

Méta Tag ou balise Méta : une construction placée dans l'entête HTML de votre page Web, fournissant des informations qui ne sont pas visibles par les navigateurs. Les méta-tags les plus courants (et les plus utiles pour les Moteurs de Recherche) sont KEYWORDS (mots-clés) et DESCRIPTION.

- Le méta-tag KEYWORD permet à l'auteur de souligner l'importance de certains mots et phrases utilisés dans sa page. Certains Moteurs de Recherche tiendront compte de cette information - d'autres l'ignoreront.
- Le méta-tag DESCRIPTION permet à l'auteur de contrôler le texte affiché quand la page paraît au niveau des résultats d'une recherche. Certains Moteurs de Recherche peuvent ignorer cette information.
- Le méta-tag HTTP-EQUIV est employée pour émettre des commandes HTTP et est fréquemment employée avec le tag REFRESH pour remettre à jour le contenu de page après un nombre donné de secondes. Les pages passerelle emploient parfois cette technique pour forcer les navigateurs à aller vers une page ou un site différent. La plupart des Moteurs de Recherche en sont conscients et classeront la page à la fin

et/ou réduiront le placement du site. Infoseek est contre cette technique et pénalise le site ou même l'interdit.

- D'autres méta-tags sont GENERATOR (pour ceux utilisant un logiciel d'assisté :-) à la création de pages) et AUTHOR (utilisé pour créditer l'auteur de la page qui contient souvent son adresse e-mail, l'URL de son site et toute autre information utile).

Occurrence : fréquence d'apparition d'un terme de recherche dans un lot résultat

Recherche booléenne : mode de recherche fondée sur la logique ensembliste et utilisant des opérateurs ET, OU, SAUF (AND, OR, NOT). Exemple : la vie de Mozart : (life* OR vie OR vies OR biograph*) AND mozart

Recherche floue ou « fuzzy search » : mode de recherche autorisant la déformation des termes de la question et/ou des documents, tolérante aux erreurs d'orthographe, par exemple.

Recherche en langage naturel : l'utilisateur a la possibilité de formuler sa requête sous la forme d'une phrase en langage parlé. Sa question est ensuite analysée par un outil linguistique qui va construire la requête à poser au moteur d'indexation et de recherche.

Recherche par proximité : mode de recherche permettant de définir les textes proches composant une recherche textuelle , en précisant éventuellement les structures du texte (phrase, paragraphe).

Recherche par similarité ou « Query By Example » : mode de recherche consistant à fournir les textes proches d'un texte servant de modèle (ce texte peut être un texte trouvé dans une recherche précédente). En général, cette recherche est fondée sur une recherche booléenne pondérée, voire floue.

Recherche pondérée : lorsqu'on compose une équation de recherche, et afin de mettre en évidence les termes les plus importants, certains outils permettent de leur affecter un poids – généralement coté de 0 à 100% – de façon à privilégier les documents qui contiennent certains de ces termes.

SGML (Standard Generalized Markup Language) : langage de description universel de la structure d'un document. Il permet de baliser et de signaler la structure d'un document, de manière à pouvoir échanger des fichiers entre partenaires et fournisseurs.

Spider : programme automatisé qui recherche sur l'Internet de nouveaux documents Web, et qui indique leurs adresses et les informations qu'ils contiennent dans une base de données à laquelle on accède par un moteur de recherche.

Site Portail : terme générique pour désigner un site qui sert de point d'entrée sur Internet pour un nombre significatif d'utilisateurs.
Les exemples de sites portail sont les Moteurs de recherche, les annuaires, la page par défaut d'un navigateur, la page de base du site d'un fournisseur d'accès, les sites offrant de l'hébergement gratuitement ou des services de messageries.

Thésaurus : dictionnaire de termes structurés avec des relations hiérarchiques (relation père-fils), associatives (relation liant deux descripteurs indiquant une proximité de sens) et d'équivalence (relation liant un descripteur à un non-descripteur, indiquant un sens voisin ou assimilé). Les termes représentent des concepts.

Visualiseur : outil logiciel permettant d'afficher des documents électroniques de formats d'origine diverse.

XML (eXtensible Markup Language) : langage à balises personnalisables. Version simplifiée de SGML destinée aux applications Internet, il permet aux développeurs de créer leurs propres marqueurs, pour fournir des fonctionnalités qui ne sont pas disponibles en HTML. XML ne fonctionne pas tout seul et à besoin d'être exploité par un analyseur. XML donne la description du contenu et la structure du document, la forme étant décrite par XLS.

XSL (eXtensible Scripting Language) : un langage XML de feuilles de style décrivant la mise en forme pour l'affichage d'un document. Il est supporté par les derniers navigateurs Internet Explorer 5 et Netscape Communicator 6.