Centre
NTE
Nouvelles Technologies
et Enseignement

UNIVERSITÉ DE FRIBOURG - SUISSE
UNIVERSITÄT FREIBURG - SCHWEIZ

« Trouver de l'information sur le web »

Gérald Collaud & Jacques Monnard

Centre NTE, Université de Fribourg

Table des matières

1. Introduction

1.1 Internet, le web et les navigateurs

Internet est une structure, un support qui permet d'acheminer des flux d'information d'un endroit de la planète à l'autre en respectant certaines règles (par exemple le protocole TCP/IP). Pour communiquer, les utilisateurs d'internet ont également besoin de logiciels qui utilisent différents protocoles de communication. On parle alors de services. Les principaux services de l'internet sont le courrier électronique (*email*), les forums de discussion (*news*), l'accès aux machines (*telnet*), le transfert de fichier (*ftp*), le dialogue interactif à deux ou à plusieurs (*chat*, *irc*).

Le World Wide Web (*WWW*) sert de fédérateur pour les autres services. Le World Wide Web (ou simplement web) est un immense réseau hypertexte contenant des ressources multimédia accessible au niveau planétaire à travers internet.

Inventé initialement au CERN à Genève, il s'est développé de manière fulgurante à partir de 1992; il comprend aujourd'hui des milliards de documents et le volume d'information disponible double environ tous les 6 mois.

Pour naviguer sur internet il existe de nombreux logiciels, dont par exemple, *Netscape* (Navigator ou Communicator) (http://www.netscape.com), *Internet Explorer* (Microsoft) (http://www.microsoft.com/ie), *Opera* (http://www.operasoftware.com/), *Mosaic* (http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/NCSAMosaicHome.html), ou *Lynx* (ftp://ftp2.cc.ukans.edu/pub/WWW/lynx/).

Voici quelques dates qui résument le développement d'internet :

- 1969: création d'ARPAnet, par le DoD, un réseau qui permet de partager les ressources des super-ordinateurs; utilisation de Telnet et de FTP;
- 1977: création du courrier électronique, par l'université du Wisconsin;
- 1985: NFSnet, réseau dissident pour les universités américaines (1000 ordinateurs connectés) et internationalisation;
- 1990: fusion entre NFSnet et ARPAnet et création d'i*nternet* (150'000 ordinateurs connectés);
- 1992: naissance du *World Wide Web*;
- début 1994: plus de 2'000'000 d'ordinateurs connectés; explosion du nombre d'internautes; début de l'utilisation commerciale.

En 2000, l'internet est un gigantesque réseau de réseaux informatiques, qu'on décrit souvent en utilisant la métaphore de la toile d'araignée. Cette interconnexion mondiale permet à des personnes, éventuellement éloignées géographiquement les unes des autres, d'échanger des données (textes, images, sons), ceci très rapidement et à un coût réduit.

1.1.1 Notion d'URL

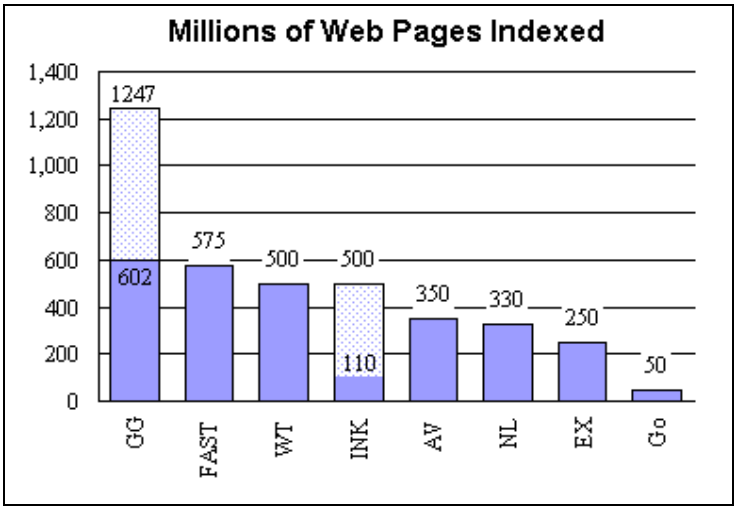
L'URL (*Universal Ressource Locator*) est un schéma universel pour la localisation des ressources. L'URL permet en principe de localiser n'importe quelle ressource sur le web. Une URL a le format suivant :

http://	iiufpc01.unifr.ch	/nte/guides/	default.html	#ancre
protocole utilisé	adresse de la machine	répertoire	nom de la ressource	position dans le document

1.1.2 Quelques chiffres

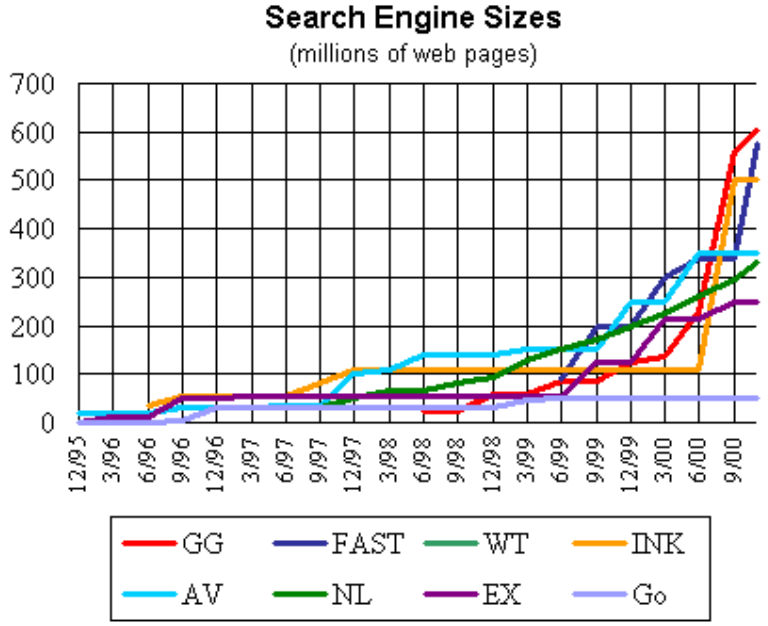
Il est intéressant de donner quelques chiffres actuels relatifs à l'internet, tout en signalant leur croissance exponentielle :

- plus de 300 mio. d'utilisateurs
- plus d'un milliard de pages
- environ 7 mio. de serveurs web
- 2.5 % des pages sont en français, et 85% en anglais
- 55 % des pages web se trouvent dans des domaines *.com* (commercial)
- Les moteurs de recherche ne peuvent indexer toutes les pages du web. Leur taille estimée est la suivante :



GG=Google,
FAST=FAST,
WT=WebTop.com,
INK=Inktomi,
AV=AltaVista,
NL=Northern Light,
EX=Excite,
Go=Go (Infoseek).

- L'évolution des moteurs de recherche est rapide comme le montre le graphique suivant :



1.1.3 Types d'information

La majorité des informations présentes sur le web se présentent sous la forme de textes et d'images. Le texte se trouve généralement dans des fichiers avec l'extension *.html* ou *.htm*.

HTML (*HyperText Markup Language*) est un langage de balisage (ou langage de marquage) pour la publication de documents sur le web. Grâce à HTML, les auteurs peuvent publier des textes formatés incluant des tableaux des images, etc., relier différents documents à l'aide de liens hypertextes, concevoir des formulaires interactifs pour mener des transactions avec des services à distance, ou encore inclure des éléments multimédia (son, vidéo, etc.).

Un document HTML est un texte dont le contenu est enrichi par des balises de la forme <*balise*>; ces marques sont des directives que le navigateur doit interpréter de manière appropriée. Certaines balises délimitent une portion de texte; dans ce cas elles sont formées d'une balise ouvrante <*portion*> et d'une balise fermante </*portion*>. Certaines balises peuvent être munies d'un ou de plusieurs attributs; elles prennent alors la forme <*balise attribut=valeur ...*>.

Les images sur le web se présentent généralement dans des fichiers avec les extensions *.gif* ou *.jpg*

Le GIF (*Graphic Interchange Format*) est certainement le format le plus utilisé sur le Web, il est limité à 256 couleurs donc il ne convient pas dans le cas de photographies très colorées ou avec beaucoup de nuances, par contre pour insérer un logo, icône ou même une banderole, il est imbattable rapport qualité / taille. Ce format propose deux caractéristiques intéressantes; il offre la possibilité de définir une couleur dite de transparence et l'affichage progressif (ou encore entrelacé). Le cas du GIF entrelacé se rencontre lorsqu'une image s'affiche progressivement d'abord floue puis de plus en plus nette. Il est utilisé pour charger un gif assez important (plus de 30 ko) sans pour autant pénaliser le lecteur qui explore vos pages pendant que les images se transfèrent petit à petit, sa taille est alors légèrement supérieure.

Le JPEG (*Joint Photographic Experts Group*) est un format particulièrement utilisé pour les photographies scannées riches de milliers de couleurs. En fait le JPEG réduit la taille d'une image en jouant sur sa qualité. il ne gère pas l'effet de transparence comme le GIF mais il dépasse la limite des 256 couleurs. Le taux de compression d'une image peut être déterminé entre 1 et 99 %. Plus le taux de compression est augmenté moins la qualité sera bonne. Le meilleur taux se situe certainement dans la fourchette 10-30%. Le JPEG gère également l'affichage progressif.

Format	Description	Utilisation
GIF	Pas de pertes de détails 256 couleurs maxi Affichage progressif possible (entrelacé) Gère la transparence Reconnu par tous les navigateur	Logo Icônes Banderoles
JPEG	Pertes de détails 16,7 millions de couleurs Affichage progressif Ne gère pas la transparence Reconnu par tous les navigateur	Banderoles Photographies

Comparaison des formats GIF et JPEG

Il existe un troisième format en développement qui devrait à terme remplacer GIF et JPEG, il s'agit du format PNG. Le PNG (*Portable Network Graphics*) est un format développé par le consortium W3C afin de mettre à disposition des internautes un format graphique qui ne soit pas propriétaire. La compression est meilleure qu'avec un GIF et ce format permet d'avoir plus de 256 couleurs (24 voire même 32 bits). A noter qu'il offre aussi un affichage progressif. De plus, ayant été développé spécifiquement pour le web, il permet de mieux tenir compte des différences de plateformes du point de vue du contrôle de la luminosité par exemple.

Il existe encore de nombreuses autres extensions comme par exemple *.asp* ou *.php* pour les pages web générées dynamiquement ou *.pdf* pour des documents pour lesquels le format de présentation doit être préservé.

2. Recherche d'informations générales sur le web

2.1 Stratégie de recherche

1. Définir le type d'information que l'on recherche
2. Sélectionner un outil de recherche :
   - pour une recherche très générale: moteur de recherche, méta-moteur
   - pour une recherche dans un domaine spécifique: index thématique, site spécialisé
3. Le cas échéant, choisir des mots-clés appropriés, définissant assez bien ce qu'on recherche, et lancer la recherche.
4. Si les résultats obtenus ne conviennent pas, affiner le choix des mots-clés, et relancer la recherche

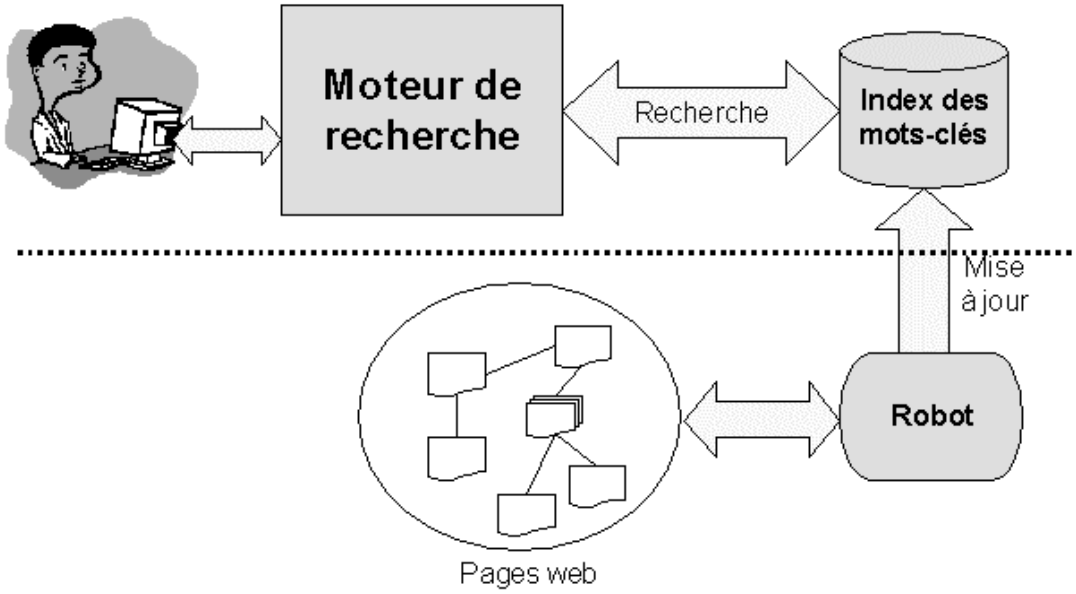
Si, au cours d'une recherche, vous trouvez des sites intéressants (en particulier des centres de ressources spécifiques à un domaine qui vous intéresse), ajoutez-les à votre liste de signets. Vous pourrez plus tard y revenir directement pour rechercher des informations.

Pour plus d'informations sur les moteurs de recherche et leur utilisation :

- *Guide d'initiation à la recherche dans internet* (http://www.bibl.ulaval.ca/vitrine/giri/)
- Vous pouvez consulter aussi le guide des indispensables de la recherche dans internet (http://www.bibl.ulaval.ca/vitrine/giri/giri2/tableau.htm), qui, selon le type d'informations que vous recherchez (adresse, bibliothèque, journal, etc.), vous suggère les répertoires et outils de recherche appropriés.
- *Finding Information on the Internet: a tutorial* (http://lib.berkeley.edu/TeachingLib/Guides/Internet/FindInfo.html)

2.2 Moteurs de recherche

Avec ces outils, la recherche s'effectue par mots-clés dans un index généré par un *robot* (un programme, appelé parfois *crawler* ou *spider*) qui parcourt plus ou moins régulièrement les pages du web, et en extrait tous les mots.



Fonctionnement d'un moteur de recherche

- Avantages :
  - énorme quantité de pages indexées,
  - recherche rapide et puissante.
- Inconvénients :
  - On n'a aucune garantie sur la qualité des résultats retournés, qui sont très hétérogènes: le contenu du site du journal *Le Monde* est indexé de la même façon qu'un site personnel créé par un amateur (ce qui permet cependant d'avoir parfois de bonnes surprises!).

- La couverture du web par les différents moteurs varie, et le contenu n'est pas toujours à jour. Selon une étude de 1999 (*Accessibility and Distribution of Information on the Web*, http://www.wwwmetrics.com/), le taux de couverture cumulé des principaux moteurs de recherche ne dépasse pas 40% du web, et la couverture individuelle varie entre 2 et 16%. La partie du web "ignorée" par les moteurs est donc énorme. De plus, les catalogues locaux (bibliothèques) et autres pages web générées dynamiquement par un site (par exemple à partir d'une base de données) ne sont souvent pas indexés.
- La mise à jour est souvent lente, et peut parfois prendre plusieurs mois. Ainsi, les nouvelles pages tardent parfois à apparaître dans un moteur, alors que des pages obsolètes subsistent encore.
- Les moteurs de recherche n'indexent pas tous les formats de document (Word, PDF, etc.)

- Voici une liste non exhaustive des moteurs de recherche francophones :

Nom	Adresse W3	Origine
AltaVista France	http://www.altavista.fr	France
Carrefour.net	http://www.carrefour.net	Québec
Ecila	http://www.ecila.nomade.fr/	France
Francité	http://www.i3d.qc.ca	Québec
La Toile du Québec	http://www.toile.qc.ca	France
Lokace	http://www.lokace.com	France
Nomade	http://www.nomade.fr	France
*Voilà**	http://www.voila.fr	France
Yahoo! France	http://www.yahoo.fr	France

- Moteurs suisses:
  - *search.ch* (http://www.search.ch/index.html.fr),
  - *the Blue Window Search* (http://search.bluewindow.ch/search?pg=home&lang=fr)
- Autres moteurs:
  - *Fast* (http://www.alltheweb.com/): le plus grand nombre de pages indexées
  - *Google* (http://www.google.com): résultats tenant compte de la popularité des pages
  - *Go.com* (http://www.go.com/): classe les résultats par site, et permet de continuer la recherche en partant des résultats obtenus (option *Search within results*)
  - *HotBot* (http://www.hotbot.com/)

On retrouve en général les mêmes fonctionnalités de base, sous une forme ou une autre, dans la plupart des moteurs de recherche. Cependant, chacun a ses particularités, et il vaut la peine de jeter un coup d'oeil à l'aide en ligne pour les connaître. Vous trouverez ci-dessous un tableau de comparaison de différentes fonctionnalités de quelques moteurs de recherche :

Outils et adresses Internet	Pages Web	Usenet	Images	Sons	Accents	Booléens
AltaVista http://www.altavista.com	oui	oui	oui	non	oui	oui
Excite http://www.excite.com	oui	oui	non	non	non	oui
Francité http://www.i3d.qc.ca	oui	non	non	non	oui	non
HotBot http://www.hotbot.com	oui	oui	oui	oui	oui	oui
Go http://www.go.com	oui	non	non	non	oui	oui
Lycos http://www.lycos.com	oui	oui	oui	oui	oui	non

2.2.1 Recherche avec Altavista

Le moteur de recherche AltaVista

a) Recherche simple

Dans le cas le plus simple, on introduit deux ou trois mots-clés, et on lance la recherche (sélectionnez d'abord l'option *Tout le web* si vous ne voulez pas vous limiter aux pages françaises). Si les termes ne sont pas très spécifiques, on obtient souvent des milliers (voire millions) de résultats. Différentes possibilités existent pour affiner la recherche:

- Chaque terme qui doit obligatoirement apparaître dans les pages trouvées doit être précédé d'un signe + dans le champ de recherche:
  - pour rechercher des pages contenant à la fois les mots *géographie* et *Suisse*, introduisez *+géographie +Suisse* (et non pas seulement *géographie +Suisse*).
- A l'inverse, pour exclure un terme de votre recherche, précédez-le du signe –:
  - pour exclure les pages les pages se rapportant à Fribourg, introduisez *–Fribourg*
- Pour chercher une expression composée de plusieurs mots, mettez-la entre guillemets dans le champ de recherche du moteur:

- pour trouver des pages concernant la formation continue, introduisez *"formation continue"* (avec les guillemets)

- Si l'on introduit les mots-clés en minuscule, le système retourne des résultats aussi bien en minuscules qu'en majuscules. Une recherche sur *paris* vous retournera ainsi les pages contenant *paris*, *Paris*, *PARIS*, etc. Par contre, dès qu'on utilise des majuscules, le système ne recherche que la forme correspondant exactement à ce qui a été introduit. Ainsi, une recherche sur *Paris* ne retournera que les pages contenant exactement *Paris* ainsi écrit.
- Vous pouvez utiliser l'astérisque (*) pour substituer les lettres manquantes dans un mot:
  - si vous entrez *cheva**, vous obtiendrez (entre autres) les pages contenant les mots *cheval*, *chevaux*, *chevals* (!), etc.

  L'astérisque ne peut être utilisé qu'après trois caractères au minimum, et peut remplacer de zéro à cinq lettres.

  Il faut noter aussi que certains moteurs gèrent automatiquement les terminaisons (verbes, pluriels), de sorte que toutes les formes possibles soient retournées sans effort supplémentaire (cela est souvent le cas pour les moteurs anglophones).
- L'option *langages* vous permet de choisir la langue des pages recherchées. Attention: le moteur de recherche essaye de déterminer automatiquement la langue des pages, ce qui peut parfois poser des problèmes (p. ex. pour des pages plurilingues).
- Vous pouvez utiliser plusieurs mots-clés spéciaux pour rechercher des pages présentant certaines caractéristiques particulières. Les plus intéressants sont:
  - *host* pour trouver des pages sur un site spécifique, p. ex. *host:unifr.ch* pour trouver des pages sur le site de l'Université de Fribourg,
  - *domain* pour trouver des pages dans un domaine spécifique, p. ex. *domain:ch* pour trouver des pages en Suisse.
  - *link* pour trouver des pages contenant des liens vers un site, p. ex. *link:www.monsite.ch.* En combinant *link* avec *host*, on peut trouver les pages pointant sur son propre site, à l'exclusion des pages du site lui-même, p. ex. *–host:www.unifr.ch +link:www.unifr.ch/nte* pour trouver des pages hors de l'Université de Fribourg pointant sur le site du Centre NTE.

b) Recherche avancée

La recherche avancée permet de définir plus précisément les critères de recherche:

- On peut définir le critère de recherche avec une expression logique (booléenne), en combinant les termes avec des opérateurs:
  - *AND* (et): recherche de documents comportant tous les termes ou expressions saisis. Exemple: *bourse AND Suisse.*
  - *OR* (ou)*:* recherche de documents comportant au moins l'un des termes ou expressions saisis (y compris les pages contenant les deux termes). Exemple: *pomme OR poire.*
  - *AND NOT* (et non) : rejette les documents comportant le terme ou l'expression saisis. Exemple: *enseignement AND NOT primaire* pour trouver des pages consacrées à l'enseignement, à l'exclusion du primaire.

- *NEAR* (proche de): recherche de documents comportant tous les termes ou expressions saisis, sous réserve que 10 mots au plus les séparent. Exemple: *gravité NEAR lune.*
- On peut créer des expressions complexes en utilisant des parenthèses, par exemple: *(gâteau OR tarte) AND (pomme OR poire)*.

- On peut aussi effectuer une recherche sur la date de modification des pages, le pays, etc.
- On peut demander de ne retourner qu'une page par site (*Afficher un résultat par site*), pour limiter le nombre de résultats.

Pour aller au-delà de 200 résultats (seulement avec *Altavista.com*): effectuer une recherche (simple ou avancée), cliquer sur la page no. 20, qui contient les résultats 190 à 200, et remplacer le chiffre *190* qui apparaît dans *stq=190* à la fin de l'adresse de la page par le chiffre souhaité (par exemple *200* pour obtenir les résultats 201 à 210). On peut continuer ainsi aussi loin qu'on le souhaite.

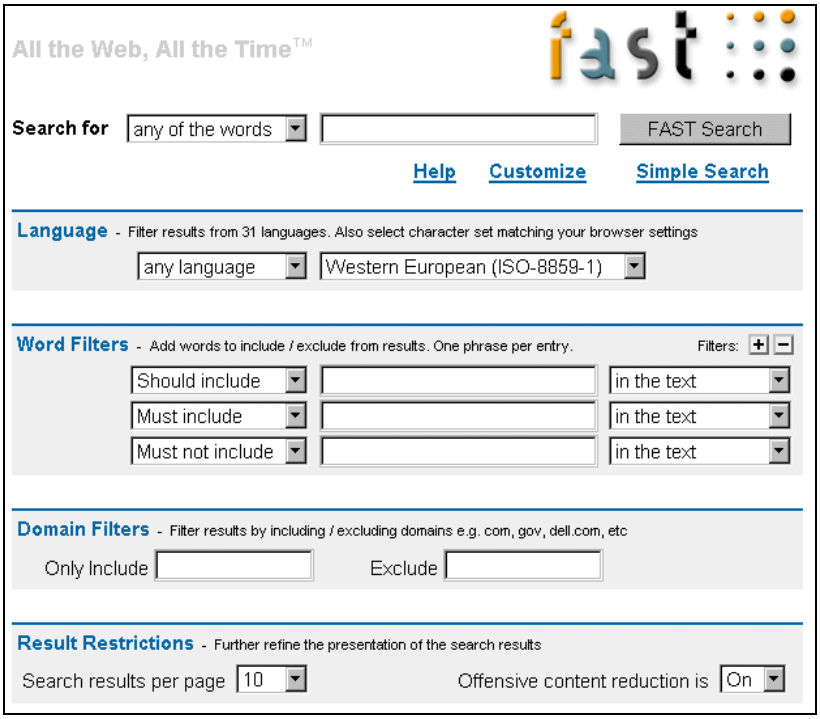
c) Recherche multimédia

Pour rechercher des images, sons, vidés, cliquez sur un des onglets *Images*, *MP3/Audio* ou *Vidéo*.

2.2.2 Autres moteurs de recherche

a) Fast

Le moteur de recherche *Fast* (http://www.alltheweb.com), qui présente une bonne couverture du web, offre des fonctions similaires à celles d'*Altavista*, et permet aussi de définir les options de recherche avancée dans un formulaire:



Le formulaire de recherche avancée de Fast

b) Go

Le moteur de recherche *Go* (http://www.go.com, anciennement *Infoseek*) permet de faire une recherche par étapes, en affinant la recherche à partir des résultats déjà obtenus (option *Within Results*).

2.3 Méta-moteurs de recherche

Les "méta-moteurs" de recherche permettent d'effectuer une recherche simultanée dans plusieurs moteurs. Après que vous avez introduit les termes à rechercher, ces outils transmettent votre demande simultanément à différents moteurs (p. ex. *Altavista*, *Go* et *Lycos*), et vous présentent les résultats obtenus sous une forme agrégée:

- *MetaCrawler* (http://www.metacrawler.com/): l'option *Power Search* permet de choisir les moteurs de recherche dans lesquels la recherche sera effectuée.
- *Search.com* (http://www.search.com/): propose des catégories dans lesquelles sont classés différents moteurs de recherche spécifiques, par exemple pour la santé, pour trouver des personnes, etc.
- *Dogpile* (http://www.dogpile.com/)

Il faut noter que les méta-moteurs sont bien adaptés pour des recherches assez simples (deux ou trois termes). Avec une recherche plus complexe, les résultats risquent de ne plus être complets et/ou cohérents. En effet, tous les moteurs de recherche ne fonctionnent pas de la même manière, et certains d'entre eux ne comprendront pas forcément votre requête telle que vous l'avez formulée dans le méta-moteur.

2.4 Index thématiques (annuaires web)

Dans ces sites, appelés aussi répertoires, les adresses sont classées par catégories et sous-catégories (arborescence). Ce travail étant effectué manuellement par les gestionnaires du site, ce type d'outils ne recense qu'un nombre limité de sites. Pour chaque site, on trouve généralement une brève description du contenu.

- Avantages :
  - Tous les sites sur un même thème sont regroupés dans une catégorie.
  - La qualité des adresses est meilleure que dans un moteur de recherche (limitation du "bruit").
- Inconvénients :
  - Par rapport aux moteurs de recherche, le taux de couverture des index thématiques est beaucoup plus faible, et la mise à jour souvent moins rapide.
  - L'organisation des catégories du site ne correspond pas toujours à l'idée qu'en a l'utilisateur.

Quelques exemples :

- *Yahoo France* (http://www.yahoo.fr), qui contient aussi une catégorie pour la Suisse, *Yahoo USA et monde* (http://www.yahoo.com), *Yahoo Allemagne* (http://www.yahoo.de, contient également une catégorie pour la Suisse)

L'index thématique Yahoo France

- *Francité* (http://www.i3d.qc.ca/)
- *Nomade* (http://www.nomade.fr)
- *Voila* (http://www.voila.fr)
- *Romandie.com* (http://www.romandie.com/): Suisse romande
- *dmoz* (http://www.dmoz.org/World/Français/): partie francophone de l'*Open Directory Project*, qui vise à créer le répertoire Web le plus complet possible grâce à une équipe importante d'éditeurs volontaires.

De nombreux sites ont maintenant tendance à combiner les deux fonctionnalités moteur de recherche et index thématique. C'est le cas par exemple pour *Altavista* et *Yahoo*.

2.4.1 Recherche avec Yahoo

On peut procéder de différentes manières pour effectuer une recherche avec *Yahoo*:

- On peut naviguer dans l'arborescence des catégories jusqu'à trouver celle qu'on recherche, par exemple *Sciences humaines*, puis *Philosophie*, puis *Philosophes* pour trouver des sites consacrés à divers philosophes.
- On peut effectuer une recherche avec des mots-clés. Yahoo retourne alors aussi bien des catégories que des sites correspondant à la recherche. En entrant par exemple *Services web* dans le champ de recherche, on trouvera des catégories *Yahoo* et des sites proposant ce genre de services.
- On peut combiner les deux types de recherche précédente pour cibler la recherche. Ainsi, si l'on cherche des entreprises proposant des services web en Suisse, on naviguera d'abord dans les catégories *Exploration_geographique/Pays/Suisse/*. On entre ensuite *Services web* dans le champ de recherche, et on sélectionne l'option *Uniquement cette catégorie* avant de lancer la recherche.

Faire apparaître un site dans un moteur de recherche ou un index

Pour qu'un site soit indexé par un moteur de recherche, il est préférable de l'annoncer soi-même, plutôt que d'attendre que le moteur l'indexe "par hasard". Cette fonction s'appelle parfois *Annoncer un site*, ou bien *Ajouter une URL*. Pour ce qui est des index thématiques, l'annonce est nécessaire. Dans les deux cas, on peut en général se limiter à une dizaine des principaux moteurs et index, qui regroupent la grande majorité des recherches. Il faut noter aussi qu'une assez longue période peut s'écouler après l'annonce jusqu'à ce que le site apparaisse dans le moteur de recherche ou l'index.

2.5 Sites spécialisés

Les sites spécialisés sont des centres de ressources qui rassemblent des informations sur un domaine particulier. Il peut s'agir d'informations créées par les auteurs du site ou d'adresses externes sur le web. Ils sont gérés par un spécialiste du domaine considéré. Leur contenu est plus limité que les annuaires commerciaux, mais aussi souvent de meilleure qualité et mieux organisé.

Pour trouver le site spécialisé qui vous intéresse, vous pouvez utiliser un index thématique comme Yahoo. Il existe aussi des « meta-sites » spécialisés :

- *About* (http://about.com) regroupe toute une série de sites spécialisés gérés par des spécialistes de chaque domaine
- *WWW Virtual Library* (http://cuisung.unige.ch/vl/Home.html): regroupement de pages thématiques maintenues par des experts de chaque domaine.

et des annuaires de sites spécialisés :

- *Les Signets de la Bibliothèque Nationale de France* (http://www.bnf.fr/web-bnf/liens/index.htm)
- *The Argus Clearinghouse* (http://www.clearinghouse.net/)
- *TheBigHub* (http://www.thebighub.com/): répertoire de plus de 3'000 base de données; à partir de ce site, on peut également effectuer une recherche directe dans un des sites trouvés.

On trouve des sites spécialisés dans de nombreux domaines. Il existe par exemple de nombreux sites fournissant des ressources utiles pour la création et la gestion de sites web :

- *AaZ Webmasters* (http://aaz-webmasters.platomic.com/)
- *Annuaire du webmaster* (http://www.annuaire-du-webmaster.com/pages/index.html)
- *WebReference* (http://www.webreference.com/index2.html)
- *CNET Builder.com* (http://builder.cnet.com/)
- *Developer.com* (http://www.developer.com/)

De même, on trouve des centres de ressources consacrés aux nouvelles technologies dans l'enseignement et la formation:

- *Educa* (http://www.educa.ch/f/index.html), site du Centre suisse des technologies de l'information dans l'enseignement

- *Edutech* (http://www.edutech.ch/edutech/index_f.asp), site suisse consacré aux nouvelles technologies dans l'enseignement supérieur
- *EducaSource* (http://www.educasource.education.fr/): répertoire français de ressources électroniques sélectionnées et commentées pour les enseignants

3. Recherche d'autres types d'informations

3.1 Personnes – adresses électroniques

Plusieurs sites web permettent de rechercher l'adresse électronique d'une personne, voire son adresse (surtout aux USA). Les plus connus et utilisés sont :

Nom	Adresse W3
Four11	http://www.four11.com
WhoWhere	http://www.whowhere.com
Excite People Finder	http://www.whowhere.com/ wwphone/excite_world.html
MESA	http://mesa.rrzn.uni-hannover.de/

Comme pour la recherche de pages web, il est possible d'utiliser un meta moteur, comme par exemple :

MESA : http://mesa.rrzn.uni-hannover.de/

Remarque : si on cherche les coordonnées d'une personne travaillant pour une organisation internationale, un service gouvernemental, une institution d'enseignement ou une entreprise possédant sa propre vitrine internet, il est souvent plus simple de se rendre directement sur le site Web de l'employeur. On a alors de bonnes chances de trouver la liste du personnel et les coordonnées professionnelles à partir de la page d'accueil.

3.2 Logiciels (*software*)

De nombreux logiciels sont disponibles sur le web soit gratuitement (*freeware*), soit moyennant le paiement d'une légère contribution (*shareware*). Les sites suivants permettent de rechercher ces logiciels :

Nom	Adresse W3
Logithèque Yahoo	http://fr.shareware.yahoo.com/
C|Net Shareware.com	http://www.shareware.com
Download.com	http://www.download.com
ZDNet	http://www.hotfiles.com/index.html
Tucows	http://tucows.wau.nl/

3.3 Forums de discussion (*news, newsgroups*)

Les groupes de discussion sont globalement organisés par grands thèmes. Les grandes catégories[1] sont :

préfixe	Contenu
alt	discussions "alternatives" abordant les sujets les plus variés
misc	Le mot misc désigne *miscellaneous*, qui veut dire divers. On trouve dans cette catégorie les inclassables
bionet	recherche en sciences de la vie, biologie
comp	Pour les sujets qui intéressent les professionnels et les passionnés d'informatique, de logiciels et d'informations sur les matériels
sci	Regroupe les discussions relatives à la recherche et à leurs applications pour les sciences exactes
talk	Les gens qui parlent pour parler
news	Pour les administrateurs de News, les logiciels de lecture de News, les annonces de News
soc	Regroupe les discussions relatives aux problèmes de société et/ou relatives aux différentes cultures du monde

Les sites qui permettent de recherche des news sont les suivants :

Nom	Adresse W3
Voila (francophone)	http://www.news.voila.fr/
DejaNews	http://www.dejanews.com
Reference.com	http://reference.com
AltaVista	http://altavista.com
HotBot	http://www.hotbot.com/usenet.html

3.4 Listes de distribution (*mailinglists*)

Il existe des listes de distribution dans d'innombrables domaines. Les sites suivants vous en donneront une liste :

[1] Il faut noter qu'il existe aussi des hiérarchies par pays : fr.test, ch.market, etc.

Nom	Adresse W3
Francopholistes	http://www.cru.fr/listes
Liszt	http://www.liszt.com
Tile.net	http://tile.net/lists

3.5 Images et icones

Si vous voulez trouver des images et icones pour décorer des pages web, de nombreux sites vous proposent des séries d'images groupées par exemple par thème.

Nom	Adresse W3
Abed's Icon Collection	http://darkwing.uoregon.edu/~alquds/icons.html
Uni Karlsruhe Icons	http://www.rz.uni-karlsruhe.de/Icons/
Leo's Icon Archive	http://www.silverpoint.com/leo/lia/
Anthony's Icon Library	http://www.cit.gu.edu.au/~anthony/icons/index.html
CERN Icons	http://www.w3.org/hypertext/WWW/Icons
WWW Icons and Images	http://www.lirmm.fr/bib-icons/AIcons/
Banque d'icones	http://www.iconbazaar.com/
Graphic Element Samples	http://www.lirmm.fr/bib-icons/Stanford/
Lines - Graphic Element Samples	http://www.lirmm.fr/bib-icons/Stanford/lines.html
Graphic Element Samples	http://www.lirmm.fr/ bib-icons/Stanford/othericons.html

3.6 *Ask an expert*

En désespoir de cause, vous pouvez aussi tenter de vous adresser à un "expert". Certains sites offrent la possibilité de poser une question à laquelle un être humain plus ou moins spécialisé dans le domaine de votre question tentera de répondre. Il s'agit par exemple de :

http://www.allexperts.com/

http://experts.yahoo.com/

http://www.askanexpert.com/

http://www.inquiry.com/ (technologies de l'information)

Quant à *AskJeeves* (http://www.askjeeves.com), il propose une réponse automatique à une question posée en langue naturelle (en anglais). Malheureusement, les résultats ne correspondent pas toujours à ce qu'on cherche.

3.7 Le web invisible (*invisible web* ou *deep web*)

Le web invisible est constitué d'informations mémorisées dans des bases de données disponibles sur le web. Ces bases de données concernent généralement un domaine spécifique ou un aspect particulier d'un domaine, mais peuvent aussi contenir des sites web entiers. Les moteurs de recherche traditionnels ne les inventorient généralement pas dans leurs archives. De plus, les bases de données sont souvent payantes sur le web. Toutefois, le site http://urfist.univ-lyon1.fr/gratuits/index.html présente une liste de bases de données gratuites.

Certains sites sont spécialisés dans la collecte d'information provenant des bases de données. Un de ces sites est justement appelé *The InvisibleWeb* (http://www.invisibleweb.com/) et contient des références vers environ 10'000 bases de données accessibles par le web. Il est aussi possible de trouver ces informations sur d'autres sites :

Nom	Adresse W3	Description
The BigHub.com	http://www.thebighub.com/	Propose des modèles de recherche organisés par sujet
CompletePlanet	http://www.completeplanet.com/	Propose un accès à des milliers de bases de données avec un résumé pour chacun des sites
Direct Search	http://gwis2.circ.gwu.edu/~gprice/direct.htm	Une large compilation de liens vers des moteurs de recherche (compilé par Gary Price de l'Université George Washington)
Lycos Directory: Searchable Databases	http://dir.lycos.com/Reference/Searchable_Databases/	Une grande collection de bases de données organisé par sujet; presque identique au InvisbleWeb
Search.Com	http://www.search.com/	Des dizaines de bases de données organisées par sujet (proposées par Cnet)
Search Engines and News	http://www.internets.com/	Une grande collection de moteurs de recherche par sujets, et de nouvelles.
Subject Directory of Search Engines	http://www.searchiq.com/subjects/	Une liste de bases de données organisée par sujets, venant de SearchIQ
WebData.com	http://www.webdata.com/webdata.htm	Une grande liste de bases de données organisée par sujet
PresseWeb	http://www.presseweb.ch	Base de données de journaux dans le monde

4. Trucs et astuces

a) Utilisez votre bon sens

La plupart des entreprises choisissent un nom de site web que les clients potentiels peuvent retenir facilement. Ainsi, si vous cherchez l'entreprise *abcd* vous avez de fortes chances de trouver son site web en tapant l'URL :

http://www.abcd.com

De même, connaître le pays d'origine de l'entreprise pourra vous aider. Par exemple :

http://www.ubs.ch

b) Effacer des parties de l'URL

Lorsque vous vous trouvez sur une page web quelconque, vous pouvez presque toujours en apprendre plus sur le site en supprimant des éléments de l'URL. Ainsi par exemple, en supprimant le chemin d'accès d'une page web, vous obtiendrez dans la plupart des cas, la page d'accueil.

Par exemple, en supprimant la partie */fr/nutrition/default.asp* de l'adresse *http://www.nestle.ch/fr/nutrition/default.asp*, vous trouverez la page de départ du site suisse de l'entreprise Nestlé.

c) Utiliser l'aide en ligne du moteur de recherche

S'il est possible de faire une recherche simple de manière plus ou moins similaire avec la plupart des moteurs de recherche, certains offrent pourtant des options supplémentaires. Par exemple, certains moteurs offrent la possibilité de faire une recherche en posant une question en "langue naturelle", alors que d'autres nécessitent l'usage de mots clés. De même, Google offre la possibilité de visualiser une page web telle qu'elle a été mémorisée dans leur index, ce qui peut être très utile lorsque le site est momentanément non disponible.

C'est en lisant l'aide en ligne que vous découvrirez ce que permet un moteur. N'hésitez donc pas à passer un peu de temps à lire l'aide en ligne de quelques moteurs de recherche.

d) N'oubliez pas les signets (ou favoris)

Les signets sont souvent la meilleure source d'information. A chaque fois que vous trouvez un site intéressant, n'oubliez pas de l'ajouter à vos signets. Vous verrez à l'usage qu'une série de signets bien gérés est souvent aussi utile qu'une recherche sur l'un ou l'autre des moteurs.

e) Qui a fait un lien vers votre site ?

Avec Altavista, il est possible de découvrir les pages web qui ont établi un lien vers l'une ou l'autre de vos pages web. Il suffit d'utiliser le mot clés link dans votre requête. Par exemple :

link:www.unifr.ch/nte (la page web du Centre NTE) retourne 57 entrées.

f) Etes-vous un voyeur ?

Enfin, si vous êtes curieux, vous pouvez voir ce que d'autres internautes sont en train de chercher sur le web. Tapez l'URL :

http://www.excite.com/search/voyeur/

et amusez-vous !

5. L'avenir de la recherche sur le web

Différents logiciels installés localement permettent parfois de simplifier et d'accélérer la recherche, par exemple :

- *Sherlock* (http://www.apple.com/sherlock/) sur Mac
- *Copernic* (http://www.copernic.com) sous Windows

Une tendance se dessine aussi vers des systèmes de recherche décentralisés, comme *gnutella* (http://gnutella.wego.com/, principalement utilisé pour la musique).

Annexe A : Glossaire des termes liés à la recherche d'information

Achat de mots-clés
: L'achat de mots-clés dans les moteurs de recherche est uniquement possible à travers les bandeaux. Tous les principaux moteurs de recherches (sauf *EuroSeek* et *GoTo*) insistent sur le fait que cet achat de mots-clés n'est lié qu'à l'apparition de bannières et n'influence en rien le résultat des requêtes. Un service *Bannerstake* proposé par Thomson et Thomson à l'adresse suivante : http://www.namestake.com permet de comprendre la logique d'affichage de bandeaux en fonction de la requête faite.

Adjacency
: Cf Juxtaposition

Adresse électronique
: Tout utilisateur de messagerie doit avoir une adresse qui l'identifie personnellement. Sur l'internet elle est de la forme : *nom@nom_organisation*

Algorithme de Pertinence
: C'est la méthode qu'utilise un Moteur de Recherche ou un Répertoire pour relier les mots-clés d'une requête avec le contenu de chaque page, de telle sorte que les pages Web trouvées correspondent bien au sujet de la requête. Chaque Moteur de Recherche ou Répertoire est susceptible d'utiliser un algorithme différent et de le changer ou de l'améliorer de temps en temps.

Altavista
: C'est un moteur de recherche très populaire avec une des plus grandes bases de données sur internet. Son URL principale est http://www.altavista.com. Jusqu'en 1998, ce moteur était utilisé par Yahoo pour la recherche d'informations. Altavista indexe tous les mots d'une page et les nouvelles pages sont rajoutées dans la base de données très rapidement, généralement, dans les deux jours ouvrables. Il vous est demandé de soumettre juste la première page de votre site, le robot d'Altavista explorera votre site et indexera vos pages. Quelques problèmes de *spamming* ont été notés. L'utilisation des mots-clés dans les *meta-tags* est pénalisée. Altavista propose différentes options alternatives avant les résultats de sa recherche tel que des suggestions de questions (en utilisant les services de *Ask Jeeves*) et *RealNames*. Les premières places achetées commencent à apparaître dans les pages de résultats.

Anonymous FTP
: Permet de se connecter à un serveur de fichiers comme utilisateur *anonyme*

Araignée
: C'est la partie du Moteur de Recherche qui surfe sur la toile, enregistre les URLs, classe les mots-clés et le texte de chaque page qu'il trouve. En français, le terme plus souvent employé est robot. Vous pouvez trouver plus d'information sur chaque araignée au niveau du *Search Engine Watch.*

ArchitextSpider
: C'est le petit nom de l'araignée du Moteur de Recherche d'*Excite.*

Ask Jeeves
: Un méta moteur de recherche a qui il est possible de poser des questions en anglais. Ce service est utilisé par *Altavista* et trouvable à http://www.askjeeves.com.

Coup
: Dans le contexte de visiteurs d'une page Web, un coup est une demande simple d'accès à

un fichier texte ou à un graphique sur le serveur. Si, par exemple, votre page contient dix boutons (10 images séparées), la simple visite d'une personne utilisant un navigateur web avec l'option graphique mise en place implique onze coups sur le serveur. (Souvent, il arrive que les accès ne vont pas aussi loin que le serveur du site que vous visitez parce que la page se trouve dans la mémoire cache de votre fournisseur d'accès local).
Dans le contexte d'un Moteur de Recherche, un coup est la mesure du nombre de sites apparaissant lors de la réponse à une requête d'un visiteur.

Crawler
ou Chenille, cf Araignée

Domaine
Un sous-ensemble des adresses internet. La partie la plus significative de l'adresse se trouve à la fin. Les domaines généralistes sont *com*, *net*, *org*, *edu*, *gov*, *mil* qui correspondent à des domaines spécifiques d'utilisation; *com* pour commercial, *net* pour network, etc. Il y a également des domaines correspondant à chaque pays, par exemple *ar* (Argentine), *ca* (Canada), *fr* (France), *us* (Etats-Unis), etc.
La logique des Moteurs de Recherche est telle que les sites qui ont leur propre Nom de Domaine (par exemple http://www.nativetongues.com/) auront souvent un meilleur positionnement que les sites qui sont des sous-répertoires d'une autre organisation, société. (par exemple, http://ourworld.compuserve.com/homepages/tijana/).

Domaine Virtuel
Un domaine qui est hébergé sur un Serveur Virtuel.

Enregistrement
C'est l'action d'informer un Moteur de Recherche ou un Répertoire qu'une nouvelle page ou un nouveau site doit être indexé.

En-Tête
Plusieurs moteurs de recherche donnent plus d'importance et de poids au texte trouvé entre les commandes d'en-tête au niveau du html (*heading tags*). Il est généralement conseillé d'utiliser ces commandes d'en-tête dans une page web et d'y mettre des mots clés dedans. (<*h1*> à <*h6*>)

Euroseek
Un moteur de recherche qui se concentre sur les informations ayant rapport avec l'Europe. L'adresse est http://www.euroseek.com.

Excite
Il est regardé comme un des meilleurs moteurs de recherche avec sa base de données de 250 millions de pages. Il peut être lent à indexer de nouveaux sites. Son adresse est http://www.excite.com et pour la version française, c'est http://www.excite.fr. Les sites utilisant des cadres doivent avoir l'option *noframes* pour se retrouver indexé. Excite a la possibilité d'effectuer des recherches sur de l'audio et de la vidéo qui est une partie du *RealNetworks' RealPlayer G2*.

FTP (*File Transfer Protocol*)
Protocole de base de transfert de fichier

GIF (*Graphics Interchange Format*)
Technique de compression d'image

Heading
Cf En-Tête

Hit

Cf Coup

Hotbot
C'est un des plus grands moteurs de recherche. Il utilise la base de données, la puissance de *Inktomi*. Les nouvelles inscriptions sont prise en compte sous deux semaines voire plus. Son adresse est http://www.hotbot.com.

HTML
HyperText Markup Language - le (principal) langage utilisé pour écrire des pages Web.

HTTP
HyperText Transfer Protocol - le (principal) protocole de communication entre les serveurs web et les navigateurs (clients).

Image Cliquable
Cf Image Map

Image Map
C'est une série de liens hypertextes attachés à une image. Ils sont définis dans la page ou à travers un fichier externe.
Si l'image cliquable est définie comme un fichier externe, les Moteurs de recherche peuvent avoir des problèmes pour indexer vos autres pages, à moins que vous ayez défini aussi des liens hypertextes plus conventionnels.
Si l'image cliquable est définie dans votre page, les Moteurs de recherche n'auront pas de problèmes pour suivre les liens. Il est quand même conseillé de fournir aussi sur votre page des liens au format texte pour aider ceux qui ont des problèmes de vision ainsi que ceux qui accèdent au site sans les graphismes ou en utilisant un navigateur texte.

Index
Cf Répertoire - Fait aussi référence à la base de données qui contient les pages web d'un Moteur de Recherche et/ou d'un Répertoire.

Inktomi
Cette base de données est utilisée par certains des plus gros moteurs de recherche, dont HotBot. Inktomi est aussi utilisé par Yahoo quand une requête n'est pas trouvée dans la base de données de Yahoo.

JPEG (*Joint Photographic Expert Group*)
Technique de compression avec perte d'image

Keyword
Cf Mot Clé

Lien à l'arrivée
Un lien hypertexte vers une page particulière venant de quelque part et apportant du trafic à cette page. Les liens à l'arrivée sont souvent un instrument de mesure pour connaître la popularité d'une page. La recherche des liens à l'arrivée est faisable sur *Altavista*, *Go* et *Hotbot*.

Lien Mort
Un lien qui ne mène plus à une page ou à un site, probablement parce que le serveur est en panne ou que la page a été déplacée ou alors n'existe plus. La plupart des Moteurs de Recherche ont des techniques pour retirer de telles pages de leurs listes automatiquement. internet continuant à augmenter en taille quotidiennement, il devient de plus en plus difficile pour un Moteur de Recherche de contrôler régulièrement toutes ces pages. Reporter des liens morts aide à maintenir les Moteurs propres et précis. On peut le faire en soumettant le lien mort au Moteur de Recherche.

Looksmart
: Un répertoire de taille moyenne, son adresse URL est http://www.looksmart.com/.

Lycos
: Un des moteurs de recherche le plus importants. Lycos semble se transformer peu à peu en répertoire en utilisant le projet *Open Directory* pour résultat de sa recherche. Il peut être assez lent à indexer votre page. Le robot de Lycos ignore les commandes méta dans les pages des sites. Son adresse URL est http://www.lycos.com/ et pour la France, http://www.lycos.fr/.

Mauvais Coup
: La page de résultats affichée par le Moteur de Recherche ou le Répertoire ne correspond pas à la requête effectuée. Plusieurs raisons peuvent l'expliquer:
  - La page contient bien les mots-clés, mais ils sont utilisés dans un mauvais contexte, ou alors avec une signification différente ou une corrélation différente que celle que vous avez prévue.
  - La page est une tentative de *spamdexing*.
  - Le Moteur de Recherche a un problème dans sa base de données ou une anomalie dans son programme de requête.

Metacrawler
: Un méta-moteur de recherche qu'il est possible de trouver à l'adresse suivante: http://www.metacrawler.com/. Le résultat d'une requête dans différents moteurs est résumé sur une page facile à lire.

Metafind
: Un méta-moteur de recherche qui est trouvable à http://www.metafind.com/.

Méta-Moteur de Recherche
: Un serveur qui passe des requêtes à plusieurs moteurs de recherche et/ou répertoires et résume les résultats. *Ask Jeeves*, *Debriefing*, *Dogpile*, *Infind*, *Metacrawler*, *Metafind* et *Metasearch* sont des exemples de méta-moteurs de recherche.

Méta Recherche
: La recherche des recherches. Une requête est soumise à plus d'un Moteur de Recherche ou Répertoire. Les résultats de tous les moteurs sont affichés après élimination des doubles et un triage.

Metasearch
: Un méta-moteur de recherche qui est trouvable à http://www.metasearch.com/.

Méta Tag
: 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. N'utilisez pas des guillemets autour des mots ou phrases clés.

  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 la balise *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).

Mining Company
: Un grand répertoire présent sur plusieurs adresses URL. L'adresse principale est http://www.miningco.com.

Mot Clé
: Un mot qui forme (une partie de) la requête dans un Moteur de Recherche.

Moteur de Recherche
: Un serveur ou un groupe de serveurs qui se consacre au référencement des pages internet. Lors de requêtes particulières, ces Moteurs renvoient des listes de liens correspondants à la demande. L'enregistrement dans ces moteurs se fait par les robots, la plupart du temps. Les principaux Moteurs de Recherche sont Altavista, Excite, Hotbot, Lycos, Infoseek, Northern Light et *Webcrawler*. Notez que Yahoo n'est pas un Moteur de Recherche mais un Répertoire. Le terme *Moteur de Recherche* est bien souvent employé pour décrire les deux, Répertoire et Moteurs de recherche.

Multicrawl
: Multicrawl est un moteur de recherche qui offre à ceux qui le désirent leur propre version personalisée du moteur. http://www.multicrawl.com/

Netfind
: Le moteur de recherche par défaut pour les usagers du FAI (Fournisseur d'Accès à internet) AOL. C'est un site qui est très "occupé". Son adresse URL est http://www.net-find.com. Netfind utilise le même moteur de recherche qu'Excite.

Northern Light
: Un moteur de recherche avec la possibilité d'accéder de manière payante à une collection spéciale d'articles sur les affaires, la santé et la consommation. Le premier moteur de recherche à bannir les méta moteur de recherche de sa base de données. L'adresse URL est http://www.northernlight.com.

Page de Garde
: Equivalent à une Page Passerelle mais affiche un texte avant de transporter le visiteur vers la page principale. C'est extrêmement ennuyant.

Page d'Entrée
: Cf Page Passerelle

Page Vue
: En anglais, c'est *page view*. Le nombre de pages vues est souvent utilisé dans les logiciels de statistiques et mis en avant par rapport aux *hits* du serveur. Bien souvent, plusieurs *hits* sont nécessaires pour mesurer l'affichage d'une seule page causant plusieurs entrées dans les fichiers *logs*. La plupart du temps, les logiciels d'analyse peuvent déterminer les *hits* appartenant au même visiteur et les regrouper pour fournir des informations plus pertinentes et ainsi compter plus facilement le nombre de personnes visitant un site.
Voir aussi Hit et Visiteur Unique.

Phrase Clé
: Une phrase qui forme (en partie) la requête dans un Moteur de Recherche.

Placement
: Cf Positionnement

Popularité d'une Page, des Liens
: Mesure le nombre et la qualité des liens pointant vers une page particulière (des liens à l'arrivée). Plusieurs moteurs de recherche utilisent de plus en plus ce procédé dans le processus de positionnement. Le nombre et la qualité des liens à l'arrivée se révèle être de plus en plus important comme l'optimisation du contenu de la page. Un service gratuit qui mesure la popularité de votre page peut être trouvé à http://www.linkpopularity.com.

Portail
: Cf. la page passerelle mais peut aussi désigner le site portail.

Positionnement
: C'est le processus de classement des sites, des pages web dans un Moteur de Recherche ou un Répertoire de façon à ce que les sites les plus appropriés apparaissent en premier sur la page résultat lors d'une requête spécifique. Des logiciels tel que *AgentWebRanking Freeware* (Logiciel gratuit permettant de suivre la position de son site ou d'un site concurrent dans les principaux moteurs de recherche et index thématiques en fonction de mots clés. Ce graticiel permet aussi d'améliorer sa position dans les moteurs, d'inscrire votre site, de vérifier les liens et vos pages web.) *PositionAgent*, *Rank This* et *Webposition* peuvent être utile. Ils vous aident à définir votre position dans la page résultat d'un Moteur de Recherche en utilisant pour la recherche une phrase ou un groupe de mots particulier. Le site *GoHip Search* vous permet d'avoir des informations sur votre positionnement dans les principaux moteurs de recherche et l'ensemble affiché sur une même page.

Protocole
: Ensemble de règles définissant le dialogue entre systèmes informatisés.

Query
: Cf Requête

Rang
: ou Ranking, Cf Positionnement

RealNames
: Un système d'adresse alternatif de sites webs qui est en fonction sur Altavista. Les marques déposées utilisées dans les requêtes sont directement redirigées vers le site web approprié, généralement parce que la société qui possèdent la marque à payer un loyer à *RealNames*. http://www.realnames.com.

Recensement
: c'est le processus de se promener sur le web, d'emmagasiner des URLs, d'indexer des mots clés, des liens et du texte.
: Même les plus gros Moteurs de recherche ne peuvent pas recenser toutes les pages du réseau. Les raisons en sont la quantité énorme d'informations disponible, la vitesse d'apparition de nouvelles données, la pratique d'une certaine courtoisie et une certaine limite dans le nombre page visitable en une fois. Les Moteurs de recherche ont trouvé des compromis dans leur méthode d'indexation. Par exemple, certains Moteurs indexent uniquement la page de garde des sites, d'autres visitent uniquement les sites pour lequel ils ont eu une demande, d'autres jugent de l'importance du site en fonction du nombre de liens externes avant d'indexer plus profondément.

Recherche booléenne
: Une recherche qui permet d'inclure ou d'exclure certains mots par l'usage d'opérateurs tels que AND, NOT et OR.

Recherche élargie
: *Stemming* correspond, en gros, à une recherche élargie. C'est une 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.

Referer
: Le referer, c'est l'URL d'où vient votre visiteur. Le fichier de *referer-log* de votre serveur vous indique cette information. Si un visiteur vient directement d'un résultat de Moteur de Recherche, la requête utilisée pour trouver la page sera encodée dans le *referer-log*, rendant plus facile la connaissance des mots-clés qui amènent des visiteurs. L'information referer peut également être consultée à partir du document.referer dans *JavaScript* ou par l'intermédiaire de la variable d'environnement *http_referer*.

Registration
: Cf Enregistrement

Regroupement
: Le regroupement consiste en l'affichage d'une seule page et donc adresse pour chaque site web sur la page des résultats après une requête auprès des moteurs de recherche ou des répertoires. Cette méthode permet d'éviter qu'un petit nombre de sites occupe toutes les premières positions de résultats et par la même occasion, cela rend la liste affichée par les moteurs plus claire et surtout beaucoup plus pratique pour l'utilisateur.

Relevancy Algorithm
: Cf Algorithme de Pertinence

Répertoire
: Un serveur ou un groupe de serveurs dédiés à l'indexation des pages du web. Ces répertoires retournent une liste de pages de liens selon les requêtes particulières faites par le visiteur. Les répertoires (aussi connu comme des index thématiques ou annuaires) sont généralement mis à jour manuellement, le plus souvent suite à la demande de l'utilisateur (comme à *Whatsnew.com*) et la plupart mettent en place un procédé éditorial de sélection et/ou de catégorisation (comme *Yahoo* et *Looksmart*).

Requête
: ou bien *query* en anglais. Un mot, une expression ou un groupe de mots employés pour passer des instructions à un Moteur de Recherche ou à un Répertoire afin de localiser des pages sur le sujet recherché.
  Pour des détails sur quelles requêtes sont utilisées, visitez le site en anglais *GoTo.com Search Inventory*. Un résumé de ce que recherchent avant tout les gens peut être trouvé à cette adresse http://www.synergy-marketing.com/search.html. Un programme gratuit s'appelant *Word Market* vous récupère les mots utilisés dans les moteurs de recherche et est disponible à http://www.softwaresolutions.net/free.htm. Le *Réseau Canadien de Courrier Electronique* (*CEBN*) propose un moteur de recherche cherchant sur les commandes méta mots-clés à http://www.cebn.com/metatags.htm et qui permet la recherche à travers des centaines de données existantes (attention, très, immensément long à la détente!).

Robot
: Tous les programmes de navigation qui suivent les liens hypertexte des pages de Web mais qui ne sont pas directement sous contrôle humain. Les exemples sont les araignées des Moteurs de Recherche, les programmes (*harvester*) qui extraient les adresses *E-mail* à partir des pages Web ou groupes de News ainsi que différents programmes de recherche intelligents. Une base de donnée des robots est maintenue par *Webcrawler*.

robots.txt
: C'est un fichier texte déposé dans le répertoire principal de votre site pour interdire l'accès aux robots de certains pages ou sous-répertoires du site. Seuls les robots qui sont conformes à la norme d'exclusion de robots (*Robots Exclusion Standard*) liront et obéiront aux commandes dans le fichier. Les robots liront ce fichier à chaque visite, de sorte que des pages ou les zones des sites puissent être rendues publiques ou privées à tout moment en changeant la teneur du fichier robots.txt. L'exemple simple ci-dessous permet d'empêcher tous les robots de visiter le répertoire /secret.
  User-agent: *
  Disallow: /secret
  Pour plus d'information, visitez la page d'Altavista sur le robots.txt.

Scooter
: C'est le petit nom de l'araignée du Moteur de Recherche d'Altavista.

Search Engine
: Cf Moteur de Recherche

Searchking
: Un petit moteur de recherche qui autorise les visiteurs à voter sur la pertinence des pages renvoyées par leurs requêtes afin de mieux classer les sites en fonction de l'opinion des visiteurs. http://www.searchking.com.

Serveur
: Un ordinateur, un programme ou un processus qui répond aux demandes d'informations d'un client. Sur l'internet, toutes les pages web sont stockées sur des serveurs y compris les Moteurs et Répertoires de recherche qui sont accessibles de l'internet.

Serveur Virtuel
: Un compte ouvert sur le serveur d'une société d'hébergement généralement lié vers son propre domaine. Ca permet de posséder son propre site web avec son propre nom de domaine à un coût moindre. Cette manière permet de posséder son propre site comme une grande société sans avoir à investir dans une machine complète et son entretien.

Sidewinder
: C'est le petit nom de l'araignée du Moteur de Recherche d'*Infoseek.*

Site Hit
: Cf Hit

Slurp
: C'est le petit nom de l'araignée utilisée par *Inktomi.*

Spider, Spyder
: Cf Araignée

Spidering
: Cf Recensement

Stop Word

Un mot qui est ignoré lors d'une requête auprès d'un moteur de recherche. Le mot est trop souvent utilisé que son utilisation n'améliore en rien la pertinence des résultats. Comme exemples, en anglais, les mots liés au net comme computer, web et des mots plus généralistes comme *get*, *I*, *me*, *the*, *you*, etc.

URL
: Universal Ressource Locator - Une adresse qui peut indiquer n'importe quelle ressource spécifique à internet. Le début de l'adresse indique le type de ressource - par exemple *http*: pour des pages Web, *ftp*: pour des transferts de fichier, *mailto*: pour des adresses *E-mail*, etc

Web
: ou Toile. Le web désigne plus largement l'ensemble du réseau de sites où vous pouvez naviguer avec votre navigateur.

Webcrawler
: Un moteurs de recherche important dont l'adresse URL est http://www.webcrawler.com/.

Yahoo
: Yahoo est équivalent à un moteur de recherche mais avec une base de données gérée à la main. C'est l'outil de recherche le plus utilisé mondialement. L'adresse URL est http://www.yahoo.com et pour la France, c'est http://www.yahoo.fr. Il est très difficile de se faire enregistrer sur yahoo.com et quand c'est fait, il est encore plus difficile de faire modifier des données voire même de les supprimer ! Pour améliorer vos chances d'être indexé, suivez ces conseils :

  - Sélectionnez correctement les trois catégories où vous désirez être affiché. Tenez compte aussi des catégories régionales. Vérifiez bien que les catégories correspondent au contenu de votre site.
  - Soumettez votre site à l'une de leurs filiales dans votre pays ou ville.
  - Soyez sûr que votre site est bien fait et facilement navigable.
  - Soyez sûr que votre site n'a pas de liens morts.
  - Soyez sûr que vos pages se chargent rapidement.
  - Fournissez de bonnes informations pour rentrer en contact avec vous sur votre site.

  Si vous arrivez à faire rajouter votre site dans Yahoo, gardez l'email que Yahoo vous envoi. Vous pouvez écrire à cette même personne si vous avez des modifications à apporter à votre référencement, par la suite.

Annexe B : Quelques lectures

B.1 Livres

Abramatic, J.F. *Développement Technique de l'Internet*, Rapport de Mission, 1999
[visité le 13/11/2000] http://mission-dti.inria.fr/Rapport/rapport.html

Gavrilut, G., Letranchant, M., St-Jacques, N., Tellier, S. *Internet : Les aides à la recherche*, Les Éditions du Trécarré, 1996

Huitéma, C. *Le routage dans Internet*, Eyrolles, 1994

Huitéma, C. *Et Dieu créa l'Internet*, Eyrolles, 1995

Lardy, J.P. *Les accès électroniques à l'information*, ADBS, 1993

Maire, G. *UNGI97 - Un Nouveau Guide Internet*, Editions First, 1997
[visité le 13/11/2000] http://www.imaginet.fr/~gmaire/toc.htm

B.2 Articles

16 outils de recherche pour bien surfer sur le Net Netsources, novembre 1998
[visité le 13/11/2000] http://www.fla-consultants.fr

A brief history of the Internet, article collectif de plusieurs pionniers
[visité le 13/11/2000] http://www.isoc.org/internet-history/brief.html

n/e/tsurf Predictions Voici ce que l'Internet pourrait nous réserver l'année prochaine selon plusieurs spécialistes de la presse francophone
[visité le 13/11/2000] http://www.netsurf.ch/quoideneuf/predictions99.html

Sapristi un ensemble de documents très pédagogiques
[visité le 13/11/2000] http://csidoc.insa-lyon.fr/sapristi/digest.html

Web search results still have human touch, News.com, 27 Décembre 1999
[visité le 13/11/2000] http://news.cnet.com/news/0-1005-200-1507039.html

Berst, J. *Smarter Searches : Why Search Engines Are *Again* the Web's Next Big Thing*, ZDnet, 23 Décembre 1998
[visité le 13/11/2000] http://www.zdnet.com/anchordesk/story/story_2913.html

Koster, M. *The Web robots FAQ*
[visité le 13/11/2000] http://info.webcrawler.com/mak/projects/robots/faq.html

Tillman, H.N. *Evaluating Quality on the Net*
[visité le 13/11/2000] http://www.hopetillman.com/findqual.html

B.3 Sites web

L'Urfist de Strasbourg diffuse une série de pages thématiques
[visité le 13/11/2000] http://www-scd-ulp.u-strasbg.fr/urfist/home.htm

Andrieu, O. *Site Abondance*
[visité le 13/11/2000] http://www.abondance.com

How do search engines work ?
[visité le 13/11/2000] http://www.cnet.com/Content/Features/Dlife/Search/

Notess, Greg R. *Search Engine Showdown*
[visité le 13/11/99] http://www.notess.com/search/