enssib
École Nationale Supérieure des Sciences
de l'Information et des Bibliothèques

NOTE DE SYNTHÈSE
Sciences de l'Information et de la Communication

option :
Systèmes d'information documentaire

LES OUTILS DE RECHERCHE SUR INTERNET : TYPOLOGIE ET PRINCIPALES CARACTÉRISTIQUES

Cynthia Delisle

Directeur de recherche : Omar Larouk

Le 11 avril 2000

Université Lumière
Lyon 2

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

Université Jean Moulin
Lyon 3

Table des matières

« Sometimes searching is more of an art than a science. »
www.yahoo.com

1 Objectifs et méthodologie

Cette note de synthèse a été rédigée dans le cadre du *DEA lyonnais en sciences de l'information et de la communication*, diplôme dispensé conjointement par l'École Nationale Supérieure des Sciences de l'Information et des Bibliothèques (enssib) et les universités Lyon 2 et Lyon 3. Elle s'insère, de manière plus spécifique, à l'intérieur de l'option *Systèmes d'information documentaire*. Menée sous la direction de M. Omar Larouk, maître de conférences à l'Université de Bourgogne, elle constitue un travail préparatoire à notre mémoire de recherche, qui sera soutenu à l'automne 1999.

Dans la mesure où notre mémoire sera axé sur les outils linguistiques de recherche d'information sur Internet, il nous a semblé tout naturel – suivant en cela la proposition de notre directeur de recherche – de nous consacrer pour la note de synthèse à une version élargie de cette problématique, soit la question plus globale de la recherche d'information sur le Web. La note de synthèse demeurant un exercice de facture assez libre, nous avons choisi de réaliser, dans les pages qui suivent, moins une revue de littérature « traditionnelle » sur le sujet qu'une typologie des principaux genres d'outils disponibles sur Internet. Nous espérons ainsi clarifier certaines notions, opérer les distinctions nécessaires, avoir un aperçu de la diversité de l'offre actuelle en matière d'outils de recherche – bref, bâtir le cadre de référence où les instruments linguistiques pourront par la suite venir s'insérer. Pour chaque catégorie d'outil retenue, l'on trouvera ci-après une présentation des principales caractéristiques et quelques noms de produits. Nous avons également inclus, en annexe, des descriptions détaillées de quelques outils spécifiques.

Pour mener à bien notre tâche, nous avons identifié et dépouillé de nombreuses sources. La consultation de CD-ROM généralistes ou spécialisés (FRANCIS, LISA, Library Literature) nous a permis de repérer un certain nombre de monographies et d'articles de périodiques. Toutefois, la grande majorité des ressources consultées pour ce

travail proviennent directement d'Internet : sites des outils de recherche eux-mêmes (notamment les sections d'aide et de présentation), textes en ligne sur le sujet, barèmes comparatifs, etc. Ces documents nous ont souvent semblé plus à jour, plus pertinents et plus intéressants que les textes provenant de sources traditionnelles; c'est pourquoi nous les avons délibérément privilégiés. Enfin, nous avons également exploité les références présentes dans les ressources analysées pour découvrir de nouveaux documents.

La bibliographie qui conclut ce travail est tributaire de certains choix méthodologiques. Ainsi, pour des raisons de lisibilité et d'utilité pratique, nous n'y avons pas inclus de documents diffusés en d'autres langues que le français ou l'anglais. Nous avons, par ailleurs, privilégié les documents récents (1997-1999), de manière à tenir compte de l'évolution très rapide qui caractérise le domaine. Certaines références plus anciennes qui nous semblaient enrichissantes figurent, toutefois, dans la bibliographie, notamment en raison de leur « intérêt historique ». Enfin, puisqu'Internet prête le flanc à de nombreuses critiques en ce qui a trait à la validité des renseignements qu'on y trouve, nous avons tenté de limiter les risques liés à la désinformation en opérant une sélection rigoureuse parmi les très nombreuses ressources repérées initialement sur le Web (plus d'une centaine en excluant les sites d'outils), afin de ne conserver que celles qui offraient des garanties minimum d'authenticité et de sérieux : les sites spécialisés sur le sujet, les intervenants en provenance du monde académique ou professionnel, les documents faisant l'objet de nombreux liens et citations externes, etc.

Deux mises en garde découlent donc des commentaires précédents; le lecteur devrait idéalement les garder à l'esprit tout au long de la consultation de ce travail. La première concerne l'obsolescence qui sans doute sera bientôt son lot, ne serait-ce que de manière partielle; l'autre vise à souligner une nouvelle fois, et en dépit des précautions que nous avons prises, la circonspection avec laquelle il faut aborder des informations en provenance d'Internet, *a fortiori* quand elles émanent de textes à saveur commerciale ou publicitaire comme ce fut le cas pour les nombreux documents d'aide associés aux outils que nous avons inspectés.

Tous les liens hypertextuels figurant dans ce document ont été vérifiés dans la semaine du 2 mai 1999. Nous ne pouvons, bien sûr, garantir leur pérennité.

2 Introduction : De la nécessité des outils de recherche sur le Web

Internet semble destiné à jouer désormais un rôle de premier plan dans notre existence quotidienne, aussi bien au travail que dans la vie privée. Ce média relativement récent présente, certes, des avantages qui ne sont plus à démontrer : outre son utilité en tant qu'outil de communication (courriel, bavardage en direct, transactions financières sécurisées, etc.), il permet le noyautage du contenu de millions d'ordinateurs éparpillés aux quatre coins du globe. Une gigantesque base de données multidisciplinaire et multilingue s'élabore ainsi peu à peu, mêlant des documents aux formats variés où sont proposés pêle-mêle texte, images, photos, sons, vidéos, logiciels, etc. Les internautes ont désormais à leur disposition une énorme masse informationnelle dont la croissance – actuellement exponentielle – semble appelée à se maintenir au cours des prochaines années, et ce, d'autant plus que l'on assiste à l'explosion de nouveaux marchés qui menacent de concurrencer sérieusement les Nord-Américains, utilisateurs de la première heure : Europe (notamment Europe de l'Est), Asie, Amérique du Sud.

Le cyberespace apportera probablement – l'avenir le dira – une contribution appréciable à l'idéal démocratique d'une meilleure répartition de l'accès au savoir et à l'information, en permettant la consultation simultanée et à distance d'une même ressource par plusieurs personnes, et ce, pour un coût relativement minime. Toutefois, pour que ce médium devienne réellement un médium « de masse », il reste encore à résoudre un certain nombre de problèmes relatifs à sa convivialité et à son efficacité d'utilisation, dont le moindre n'est assurément pas celui du repérage des données.

En effet, dans la mesure où il n'existe aucun recensement centralisé de l'ensemble des ressources rendues disponibles par Internet (un idéal qui n'est pas en vue à court terme), un problème fondamental persiste actuellement à ce niveau. Les métaphores de l'aiguille dans la botte de foin et de la bibliothèque aux livres dispersés sur le sol ont été

employées abondamment déjà, mais demeurent efficaces pour exprimer la frustration qui est le lot de tout internaute – aussi bien chevronné que néophyte – lorsque vient le temps de mettre la main rapidement et au bon moment sur *le* document pertinent dont on subodore vaguement l'existence *quelque part* sur le réseau des réseaux. Outre son gigantisme, ce dernier pose également des problèmes de par sa mouvance intrinsèque : de nouvelles ressources s'ajoutent quotidiennement, des sites plus anciens sont modifiés, d'autres changent d'adresse ou disparaissent carrément... On fait donc face, comme le souligne P. Laublet, à un *paradoxe apparent* : « *L'information, sur Internet et le Web, est directement accessible mais difficile à trouver. On se retrouve rapidement perdu dans* " *l'hyperespace* " » [Laublet, s.d.].

Pour trouver l'information désirée sur le Web, une stratégie de base consiste tout bonnement à « surfer », c'est-à-dire à déambuler de lien en lien, au gré des pages. Souvent agréable et fructueuse, cette démarche exige toutefois habituellement un investissement considérable en temps; les résultats obtenus demeurent, par ailleurs, largement tributaires des pérégrinations de l'internaute et de l'inclusion subjective par les auteurs de liens dans leurs pages Web. C'est pourquoi il est plus courant, de nos jours, de recourir à ce que l'on appelle des *outils de recherche*, c'est-à-dire aux divers sites spécialisés dans le repérage de l'information sur Internet. Immensément populaires, ces outils sont également de plus en plus nombreux : il en existe désormais plusieurs centaines, et il n'est sans doute pas exagéré de prétendre qu'il en apparaît de nouveaux presque tous les jours.

La plupart des premiers outils de recherche ont été conçus pour donner accès à un seul type de ressource : les serveurs anonymes du protocole FTP, le Gopher, le Web. Le champ d'action, par la suite, s'est progressivement élargi : d'abord pour embrasser plusieurs protocoles à la fois, permettant ainsi d'offrir l'accès à un plus grand nombre de ressources[1]; ensuite via l'évolution contemporaine vers ce qu'il est désormais convenu d'appeler des *portails*, c'est-à-dire des sites cherchant à se positionner comme points d'entrée de l'internaute sur le Web. En addition à la fonction de recherche d'information

[1] Du moins en principe, car, dans les faits, il semble que nombre de ressources ne relevant pas du Web soient régulièrement absentes d'outils qui prétendent pourtant prendre en compte divers protocoles.

proprement dite (et parfois au détriment de la qualité de celle-ci...), les outils de recherche proposent désormais, en effet, une surenchère de services complémentaires : actualités, dépêches d'agence, météo, cours de la Bourse, résultats sportifs, horoscope, téléchargement de logiciels, petites annonces, dossiers spéciaux (ex. : crise du Kosovo, rapport Starr), Pages Jaunes et Blanches, babillards virtuels, bavardage en direct, adresse gratuite de courriel, personnalisation du site (c'est-à-dire la possibilité pour l'usager de configurer ses préférences d'interface), offres d'emploi, envoi de cartes « postales », calendriers et agendas en ligne, mise en place de « filtres familiaux » tels l'*AV Family Filter* d'ALTAVISTA afin de bloquer – en principe – l'accès aux pages à contenu disgracieux, dictionnaires, horaires des programmes de télévision... Les possibilités sont quasi infinies et la créativité des concepteurs semble sans bornes; c'est dire que les outils de recherche, déjà incontournables, sont appelés à devenir des acteurs de plus en plus importants sur la scène cybernétique.

Dans la suite de ce travail, après quelques considérations d'ordre général, nous nous proposons de présenter successivement les trois catégories principales d'outils de recherche : les répertoires, les moteurs et les métamoteurs. Nous évoquerons ensuite de manière plus succincte quelques autres outils qui en constituent des applications spécifiques ou qui s'avèrent de nature hybride, de même que certaines technologies d'avant-garde telles que les agents intelligents, les procédés de *push* et les outils de gestion des connaissances (*knowledge management*). Nous terminerons enfin cet exposé par l'énumération de critères pouvant servir à identifier un « bon » outil de recherche, ainsi que par quelques conseils visant à permettre des investigations plus efficaces et fructueuses sur Internet.

3 Quelques précisions préliminaires

Préalablement à la présentation par catégories, il nous paraît important d'énoncer quelques remarques et explications de nature plus générale sur les outils étudiés, afin de mettre en contexte et de clarifier la suite de notre propos.

Notre analyse portera ici essentiellement sur les outils qui procèdent par recherche et comparaison de chaînes de caractères[2], un mode opératoire qui ne tient aucun compte du sens des énoncés linguistiques. Ces outils appréhendent un texte comme une suite aléatoire de mots délimités entre eux par un signe de ponctuation, une espace typographique ou d'autres caractères tels $%&-/#_~.

Résumé grossièrement, leur fonctionnement est le suivant : chaque outil, ayant construit une base de données à partir des ressources recensées (où chaque enregistrement correspond à un site Web ou à une page spécifique, suivant les cas) confronte la requête de l'usager au contenu de la base. Les entrées qui apparaîtront sur la liste de résultats seront celles qui contiennent la ou les chaîne(s) recherchée(s), soit dans le texte même du document, soit dans d'autres champs de l'enregistrement (par exemple, les balises META du fichier HTML ou encore, s'il y a lieu, les rubriques de classification).

Ces outils fonctionnent donc via la saisie par l'usager de *mots clés*, c'est-à-dire d'un ou plusieurs terme(s) de recherche. Il est habituellement possible, pour une plus grande efficacité, d'agrémenter ces derniers de que l'on appelle des *opérateurs logiques*, lesquels reposent eux aussi essentiellement sur le principe de concordance de modèle (*pattern matching*)[3].

On distingue les opérateurs suivants :

1. les opérateurs booléens

- l'opérateur ET

  Il permet de rendre la présence d'un mot obligatoire. Il est également symbolisé par son équivalent anglais AND ou par le signe +.

  Exemple : *commerce ET électronique* repérera toutes les entrées où ces deux mots figurent.

[2] Les outils plus « évolués » – notamment ceux qui incorporent des technologies linguistiques – feront l'objet de notre mémoire de DEA.

[3] Il faut, bien sûr, ajouter à ce mode d'interrogation la navigation par catégories en ce qui concerne les répertoires (voir section 4). Par ailleurs, certains outils disposent de langages de recherche beaucoup plus sophistiqués, notamment HARVEST BROKER (installé à l'interne sur de nombreux sites Web, par exemple http://amelia.db.erau.edu/Harvest/brokers/LDP/) qui permet l'utilisation d'expressions régulières.

- l'opérateur OU

  Il permet de rendre la présence d'un mot optionnelle. Il est également symbolisé par son équivalent anglais OR ou par l'espace lorsqu'il est pris par défaut.

  Exemple : *commerce OU électronique* repérera toutes les entrées qui comprennent au minimum un de ces deux mots.

- l'opérateur SAUF

  Il permet d'exclure la présence d'un mot. Il est également symbolisé par ses équivalents anglais NOT, BUT NOT ou AND NOT, ou encore par le signe –.

  Exemple : *commerce SAUF électronique* repérera toutes les entrées où figure le mot *commerce* mais sans qu'y apparaisse le terme *électronique*.

- les parenthèses ( )

  Elles permettent de limiter la portée des opérateurs booléens et/ou d'introduire un ordre de priorité entre les différentes parties d'une requête.

  Exemple : *(commerce OU paiement) ET électronique* repérera les entrées qui contiennent à la fois *électronique* et soit *commerce* soit *paiement* soit ces deux termes.

- la troncature

  Elle consiste à recourir à l'emploi de masques (*jokers* ou *wild cards*). Généralement symbolisée par les signes *, ? ou $, la troncature permet d'effectuer des recherches sur des parties de mots. Elle est moins flexible dans le contexte de la recherche d'information sur le Web qu'en ce qui a trait aux logiciels documentaires traditionnels (impossibilité de l'appliquer en début de mot, nécessité fréquente de saisir un nombre minimum de lettres, etc.). Elle est toutefois intéressante en ce qu'elle permet de faire des recherches sur des mots de même famille et sur les variations de genre et de nombre.

  Exemples : *biblio** repérera *bibliothèque*, *bibliothèques*, *bibliothécaire*, *bibliophile*, etc. La troncature peut aussi s'utiliser à l'intérieur d'un mot, pour remplacer un ou plusieurs caractère(s) : *coll$sion* repérera *collision* et *collusion*.

- la recherche de locutions

  Elle fonctionne habituellement à l'aide des guillemets " " et permet la recherche exacte d'une séquence ordonnée de mots adjacents.

  Exemple : *"commerce électronique"* repérera toutes les entrées où ces deux mots figurent l'un à côté de l'autre et dans cet ordre.

2. l'opérateur de proximité

   Il permet de rechercher des entrées où les mots désirés apparaissent à l'intérieur d'une « fenêtre » de voisinage dont l'ampleur varie selon les outils (généralement entre 10

et 100 mots, parfois beaucoup plus). Les formulations les plus habituelles sont anglophones : NEAR ou FOLLOWED BY (dans ce dernier cas, on tient également compte de la linéarité, c'est-à-dire de l'ordre d'apparition des termes). Pour rechercher des termes côte à côte (un peu comme une recherche de locution), on emploie parfois également un opérateur de proximité spécifique, dit *opérateur d'adjacence*. Il est généralement symbolisé par ADJ.

Exemples : *commerce NEAR électronique* repérera les entrées où ces deux termes figurent près l'un de l'autre. *Commerce FOLLOWED BY électronique* exigera, de plus, que l'ordre de saisie des mots soit respecté. *Commerce ADJ électronique*, pour sa part, recherchera les entrées où ces deux termes apparaissent immédiatement l'un à côté de l'autre, peu importe l'ordre d'apparition.

Chaque outil de recherche n'offre pas nécessairement la possibilité d'utiliser l'ensemble de ces opérateurs. Les plus courants de ceux-ci, toutefois, font actuellement l'objet d'un mouvement officieux de normalisation au niveau des symboles employés (par exemple, les signes + et – pour forcer ou exclure la présence d'un mot), au contraire des syntaxes avancées d'interrogation (du genre title: ou host: sur ALTAVISTA) qui continuent, elles, de varier énormément d'un outil à l'autre.

Mentionnons, enfin, que les outils étudiés se comportent de manières diverses face aux signes diacritiques et à la distinction majuscule/minuscule : certains recherchent toujours les occurrences exactes, d'autres procèdent à un genre de « nivellement par le bas » (en ramenant tous les caractères à une « syntaxe pauvre » composée uniquement de minuscules et/ou de lettres désaccentuées), d'autres enfin optent pour des positions mitoyennes (par exemple, en recherchant exactement les occurrences lorsqu'elles sont saisies accentuées et les en recherchant à la fois avec et sans accents lorsqu'elles sont saisies désaccentuées).

Nous pouvons maintenant aborder les différents types d'outils de recherche d'information. Comme le souligne D. Jakob, « *la connaissance de l'outil à utiliser, c'est déjà la moitié de la bataille pour trouver les informations sur Internet* » [Jakob 1995, révisé 1997].

4 Premier type d'outils : les annuaires

4.1 Principales caractéristiques

Nous retiendrons comme premier type d'outils les *annuaires* – qu'on appelle également *guides, répertoires* ou *catalogues*. Les annuaires sont des guides par sujet des ressources d'Internet. Ils consistent en des classements arborescents où l'accès au thème souhaité s'effectue en parcourant une série de rubriques et de sous-rubriques de plus en plus pointues. Ils incorporent également, d'ordinaire, un moteur de recherche par mot clé qui permet d'arriver directement à la bonne rubrique. Ces listes thématiques de sites constituent en quelque sorte l'équivalent cybernétique (et moins élaboré) du plan de classification que l'on applique traditionnellement dans les bibliothèques et centres de documentation. Elles présentent également des similitudes avec les bibliographies thématiques, info-guides et autres listes imprimées de ressources que les bibliothécaires mettent à la disposition de leur clientèle, et avec ces pages Web personnelles qui proposent en compilation les « meilleures » ressources d'Internet ou, tout simplement, les sites préférés de leur auteur.

Le consultant Internet français Olivier Andrieu propose la définition suivante des annuaires :

> *Un annuaire est un outil de recherche qui recense un certain nombre de sites (et non de pages) Web au travers de fiches descriptives comprenant, en règle générale, le titre, l'adresse (l'URL) et un bref commentaire d'une longueur allant le plus souvent de 15 à 25 mots au maximum. Chaque site est inscrit dans une ou plusieurs catégorie(s) – on parle également de rubrique(s) –. Ces outils peuvent ainsi être considérés comme les pages jaunes du Web. Lorsqu'un mot clé est saisi dans le formulaire proposé, l'annuaire effectue une recherche sur les occurrences de ce terme dans ses fiches descriptives de site, et non pas dans le contenu des pages du site en question. Il s'agit là de la différence la plus notable avec les moteurs de recherche.* [www.abondance.com/]

On peut résumer ainsi les principales caractéristiques des annuaires :

- Ils recensent des *sites* et non des *pages* individuelles;
- Ils structurent leurs inventaires thématiques selon une classification en général propre à l'outil (certains ont recours aux classifications documentaires traditionnelles comme celle de Dewey ou de la Bibliothèque du Congrès de Washington, mais le cas demeure rare);
- Le repérage et la catégorisation des ressources s'effectuent souvent manuellement, au moins en partie. Les annuaires recourent, à cette fin, soit à des professionnels de la

documentation (bibliothécaires, documentalistes), soit à des spécialistes des diverses thématiques concernées (par exemple, des médecins pour la rubrique *Santé*), soit encore à des volontaires (rémunérés ou non).

Le principe des annuaires présente plusieurs avantages. Tout d'abord, ces outils permettent de guider l'utilisateur dans ses investigations; ils s'avèrent donc moins intimidants que la ligne vide des autres outils de recherche. Grâce à la catégorisation effectuée sur l'information, il s'avère aisé pour l'usager de « butiner » entre sites traitant d'un même sujet, un peu comme l'on bouquine devant les rayons d'une bibliothèque. La philosophie des annuaires permet également de limiter le taux de bruit[4], et s'accompagne d'une substantielle valeur ajoutée due à l'activité humaine de sélection, d'évaluation et de hiérarchisation des ressources. On note également, bien sûr, certains inconvénients : augmentation du taux de silence[5] (en supposant qu'un document soit classifié dans une seule catégorie), couverture relativement restreinte d'un bassin potentiel de millions de sites Web, mise à jour moins rapide que pour les autres outils, dépendance par rapport aux choix éditoriaux des réalisateurs (il n'y a souvent qu'un pas entre l'évaluation des ressources et la censure...). En outre, même si les requêtes de recherche sont possibles, elles offrent en général moins de souplesse et de précision que celles permises dans les outils de type moteur.

De manière globale, on peut donc dire que les annuaires, favorisant le repérage de sites généraux sur un sujet donné, s'avèrent surtout utiles pour des fouilles vastes et thématiques, ou encore pour débuter une recherche d'information encore mal définie. Comme, par ailleurs, leur convivialité en fait les outils de recherche les plus simples d'utilisation, ils sont également tout indiqués pour les débutants.

Il convient enfin de souligner que les annuaires disponibles en plusieurs versions linguistiques ne constituent pas autant de copies d'une même base de données simplement coiffées d'interfaces différentes. Il s'agit bien, dans les faits, de bases totalement dissociées; il importe donc de les interroger successivement et d'effectuer les

[4] En documentation, le *taux de bruit* (ou *taux de précision*) concerne le rapport entre le nombre de réponses pertinentes repérées et le nombre total de réponses repérées.
[5] Le *taux de silence* (ou *taux de rappel*) concerne le rapport entre le nombre de réponses pertinentes repérées et le nombre total de réponses pertinentes présentes dans la source interrogée.

requêtes dans la langue de l'interface (par exemple, en anglais dans YAHOO! INTERNATIONAL et en français dans YAHOO! FRANCE).

4.2 Procédures d'inscription

L'inscription dans les divers annuaires est gratuite[6]. Il est toutefois généralement nécessaire d'entreprendre une démarche délibérée à cette fin : le responsable du site à enregistrer doit se rendre sur le site de l'annuaire concerné, choisir la ou les catégorie(s) où il désire être répertorié, puis suivre un lien habituellement baptisé *submit* ou *suggérer un site* qui le dirigera vers un formulaire spécifique où il lui sera loisible de décrire son site : titre, URL, résumé, mots clés, courriel du webmestre, etc. Un assez long délai (quelques jours à quelques semaines) suit habituellement l'envoi de ces données au gestionnaire de l'annuaire, au cours duquel le site est visité, évalué et – si accepté – incorporé dans la base de données de l'annuaire. Il faut noter que la prise en compte est loin d'être assurée (la plupart des annuaires refuseraient plus de sites qu'ils n'en acceptent). Dans le cas des annuaires multilingues, il est nécessaire d'effectuer l'inscription dans la version linguistique correspondant à la langue du site : à titre d'exemple, un site dans une langue autre que l'anglais n'a pour ainsi dire aucune chance d'être accepté par YAHOO! INTERNATIONAL.

Pour favoriser la visibilité d'un site lors des recherches subséquentes des usagers dans les annuaires, il y a lieu de travailler soigneusement le descriptif proposé au moment de l'inscription : il faut éviter d'en faire une simple suite de mots clés sans intérêt (surtout si on en propose déjà séparément) et privilégier une description réelle, dynamique et « verbale » du contenu, laquelle devrait également être rédigée de façon à être encore pertinente un an ou deux plus tard. Mentionnons que les descriptions ainsi fournies sont

[6] Mentionnons, en passant, que la controverse fait rage quant à la « vénalité » réelle ou présumée des outils de recherche en général et des compagnies qui les exploitent. O. Andrieu [www.abondance.com] affirme, par exemple, qu'il est tout à fait faux de prétendre que l'on peut actuellement améliorer la visibilité d'un site contre rémunération, sauf en de très rares cas d'ailleurs clairement affichés, comme les moteurs GOTO et, désormais, ALTAVISTA. La position contraire est défendue par certains spécialistes, entre autres l'éditorialiste américain Jesse Berst dans « Search Sites' Shocking Secret », chronique publiée le 17 août 1998 dans ZDNN [www.zdnet.com/]. Voir les *Brèves du CEVEIL*, août 1998, pour un résumé du texte de Berst [« Les outils de recherche : petite montée de laid », www.ceveil.qc.ca/brevesaou98.html#a4].

susceptibles dans bon nombre d'annuaires de modifications par les gestionnaires – de même, d'ailleurs, que les choix de catégorisation.

4.3 Exemples d'annuaires internationaux

Nom	**URL**
ABOUT.COM	http://www.miningco.com/
GALAXY	http://galaxy.einet.net/
JASSAN	http://www.jassan.com
LOOKSMART	http://www.looksmart.com/
MAGELLAN	http://magellan.excite.com/
OPEN DIRECTORY PROJECT	http://dmoz.org/
SNAP	http://www.snap.com/
SUITE101.COM	http://www.suite101.com/
YAHOO! INTERNATIONAL [7]	http://www.yahoo.com/

4.4 Exemples d'annuaires francophones

Nom	**URL**
CARREFOUR	http://www.carrefour.net
CHALOOP	http://www.chaloop.com
CTROUVÉ	http://www.ctrouve.com/
FRANCITÉ	http://www.i3d.qc.ca/
NOMADE [8]	http://www.nomade.fr/
YAHOO! FRANCE	http://www.yahoo.fr/

[7] Voir portrait en annexe page 44.
[8] Voir portrait en annexe page 47.

5 Second type d'outils : les moteurs

5.1 Principales caractéristiques

Notre second type d'outils est constitué par ce que l'on appelle des *moteurs*. Si les annuaires évoquent le plan de classification des bibliothèques traditionnelles, les moteurs, pour leur part, ressemblent un peu à ces programmes qui produisent automatiquement des index primitifs en associant, à chaque mot d'un document, la ou les page(s) où il figure – du reste, on les appelle aussi parfois des *index*. Les moteurs permettent à l'usager de repérer l'information non suite à une navigation thématique, mais via l'interrogation à l'aide de mots clés et de commandes logiques d'une base de données indexée; leur fonctionnement rejoint ainsi celui des logiciels de gestion documentaire usuels. En général, deux modes de recherche sont permis : *simple* (proposé par défaut, avec plus ou moins de possibilités de recherche) et *complexe* (accessible en option et où des possibilités de recherche variées et approfondies, souvent paramétrables, sont disponibles).

Voici ce que dit O. Andrieu à propos des moteurs :

> *Lorsque l'internaute saisit un mot clé dans le formulaire proposé, le moteur va en rechercher les occurrences dans son index, c'est-à-dire dans le contenu (le texte) des pages Web sauvegardées au préalable. Une fois identifié le « lot » de pages contenant le terme demandé, le moteur classe les pages par ordre de pertinence, selon un ordre et un algorithme (basé sur certains critères de tri) qui lui est spécifique. Le moteur de recherche effectue donc ses recherches sur des pages Web, alors que l'annuaire, pour sa part, vous proposera des sites Web. Là est toute la différence qui explique qu'il est absolument impossible de comparer les résultats fournis par les deux types d'outils.* [www.abondance.com/]

Le fonctionnement des moteurs s'appuie sur la collecte de données par des *robots*, lesquelles sont ensuite indexées à l'aide des mots mêmes les constituant. De gigantesques bases de données sont ainsi élaborées; elles opèrent *grosso modo* sur le mode des « fichiers inverses » en établissant des correspondances entre des mots et des URL. Les utilisateurs sondent la base à l'aide d'un module d'interrogation qui recoure à un langage de requête plus ou moins standard; des interfaces conviviales sont généralement mises en place afin de faciliter l'interaction. L'activité des moteurs de recherche, contrairement à celle des annuaires, est entièrement automatisée.

Les robots – qui connaissent diverses autres appellations évocatrices, notamment *spider* (« araignée »), *ant* (« fourmi »), *worm* (« ver de terre » ou « se faufiler »), *wanderer* (« vagabond »), *crawler* (« nageur »), etc. – sont tout simplement des programmes informatiques qui tournent sur un ordinateur relié au réseau et qui explorent systématiquement celui-ci de manière à collecter l'information présente. Les robots procèdent en identifiant les liens hypertextuels d'un document pour ensuite aller visiter les pages sur lesquelles pointe ce dernier. Ils parcourent ainsi rapidement la totalité d'un site, puis d'autres sites qui lui sont liés, et ainsi de suite. Comme le fait remarquer J.-N. Plourde, « *c'est l'automatisation et la systématisation de ce que l'on fait de chez soi en se baladant dans le Web* » [Plourde 1996]. Il n'est pas rare que le même robot soit utilisé par plusieurs moteurs différents, avec seulement quelques différences de paramétrage.

La plupart des moteurs concentrent leur activité sur le Web, bien que certains polyvalents s'intéressent aussi à d'autres ressources d'Internet. Généralement, seuls les fichiers ASCII et HTML sont indexés (et non, par exemple, les fichiers compressés). La couverture de la base de données d'un moteur est également tributaire des sites utilisés comme points de départ, de la stratégie privilégiée pour la visite des liens (en largeur ou en profondeur), etc.

Mentionnons que, contrairement aux annuaires, les moteurs qui se déclinent en plusieurs versions linguistiques ne proposent, en général, que des versions localisées d'une même base de données (EXCITE FRANCE, LYCOS FRANCE). Beaucoup des grands moteurs internationaux ne se donnent, d'ailleurs, pas cette peine et se contentent de doter leur interface anglophone d'une option de recherche de restriction linguistique (ALTAVISTA, HOTBOT, INFOSEEK, NORTHERN LIGHT).

Un des avantages de la démarche de type moteur réside dans le fait que l'utilisateur n'a pas à connaître la catégorie (et la structure hiérarchique) dans laquelle pourrait se trouver l'information recherchée, puisque cette dernière n'est pas compartimentée de la sorte et que la recherche s'opère principalement par concordance avec un modèle (*pattern matching*). Par ailleurs, comme l'absence d'intervention humaine équivaut souvent à une

absence de déontologie, les moteurs sont en principe plus utiles que les annuaires pour repérer des documents à contenu sensible (violence, pornographie) ou carrément sujets à controverse (par exemple, des sites haineux, terroristes ou pédophiles), une caractéristique que l'on peut ou non applaudir mais qui est conforme à l'esprit libertaire et anarchiste du Net. Le taux de rappel obtenu par les moteurs est souvent bon, mais il s'accompagne malheureusement d'une grande quantité de bruit, c'est-à-dire d'une baisse du taux de précision : les moteurs suscitent des réponses très hétérogènes, où les doublons abondent parfois. La (non-)mise à jour des index constitue souvent également une source de problèmes. Autre inconvénient : contrairement aux annuaires, les moteurs abandonnent l'usager à lui-même (rien ne guide ni ne balise la recherche) et ne fonctionnent généralement pas sur le mode d'un ensemble de réponses qu'il est possible de restreindre et d'affiner successivement : la recherche se fait en un coup et un seul. Enfin, leur maniement demeure délicat et les recherches peuvent prendre beaucoup de temps.

Généralement plus appréciés des internautes aguerris que des débutants, les moteurs, en un certain sens, sont plus « puissants » que les annuaires. Ils sont donc tout indiqués pour des recherches sur des sujets fins et précis, mais risquent de générer des milliers de réponses d'intérêt inégal si la requête s'avère trop vague ou trop commune.

Comme on le voit, les moteurs se distinguent des annuaires à de nombreux points de vue. Toutefois, de plus en plus de sites de recherche proposent aux internautes l'accès aux deux genres d'outils, selon des formules qui privilégient l'un ou l'autre type : moteur agrémenté d'un répertoire (par exemple, VOILA) ou répertoire complété d'un moteur de recherche externe (par exemple, FRANCITÉ). Une autre tactique consiste à conclure des accords de partenariat avec des sociétés concurrentes : YAHOO!, par exemple, dirige l'internaute sur ALTAVISTA en cas de recherche infructueuse. INFOSEEK FRANCE et EXCITE FRANCE, pour leur part, affichent les catégories et les descriptions de sites de l'annuaire NOMADE.

5.2 Procédures d'inscription

Comme pour les annuaires, l'inscription aux moteurs est gratuite. Il y a deux manières de procéder. La première consiste tout simplement à attendre que le robot du moteur débusque le site concerné au détour d'un lien, le visite et en indexe les différentes pages[9]. Cette méthode demeure aléatoire et requiert habituellement l'écoulement d'un certain laps de temps. Il est donc nettement préférable d'opter pour la seconde tactique, soit la soumission manuelle des URL que l'on désire faire connaître. Pratiquement tous les moteurs offrent, en effet, une fonction de type *Add a site* ou *Add URL*.

Selon les outils, un délai de un jour à deux mois s'écoule d'ordinaire entre le moment où une page est signalée manuellement et celui où le robot vient la visiter. Lors de cette première inspection, seule la page indiquée est prise en compte. L'aspiration du reste du site aura lieu dans une seconde étape, après un nouveau délai de un jour à huit semaines. Le gestionnaire d'un site Web est donc libre d'indiquer une à une les différentes URL à référencer pour une prise en compte plus rapide, ou de se borner à informer le moteur de l'adresse de la page d'accueil puisque le robot viendra de toute façon déambuler de lien en lien et donc de page en page.

Une fois le site indexé, le robot reviendra régulièrement « capturer » une version plus récente des différentes pages. La fréquence des visites varie de quelques semaines à quelques mois. Il convient de noter que certains moteurs imposent des limites sur le nombre de pages prises en compte pour un même site (par exemple, ALTAVISTA et INFOSEEK, avec respectivement 400 et 600 pages environ).

Les webmestres soucieux de favoriser le classement de leurs pages dans les différents moteurs doivent garder à l'esprit que ceux-ci sont susceptibles d'inspecter les éléments suivants : balises HTML <TITLE>, <DESCRIPTION> et <KEYWORDS>; intitulé de l'URL; corps du texte; attributs ALT des balises <IMG>; images en

[9] Il n'est habituellement pas possible de retirer ou modifier manuellement les références ainsi incluses dans la base de données d'un moteur. Toutefois, on peut empêcher l'aspiration d'une page grâce à l'emploi de la balise HTML <ROBOTS> ou à l'insertion d'un fichier spécial (*robots.txt*) – du moins en théorie, car les moteurs de recherche ne tiennent pas toujours compte de la présence de ces éléments...

coordonnées (*imagemaps*); cadres (*frames*), etc.[10] Plusieurs moteurs tiennent compte également du nombre de liens qui pointent vers une page dans leur base de données – c'est-à-dire, en quelque sorte, de la popularité de celle-ci. Pour certains éléments (notamment le titre, les balises <DESCRIPTION> ou <KEYWORDS> et le texte visible), l'intégralité du contenu n'est pas toujours prise en compte : le moteur s'arrête, dans certains cas, après un nombre prédéfini de caractères (ex. : les 60 premiers caractères du titre). Les mots importants ont d'ailleurs intérêt à figurer dans la première partie de la page, puisque certains moteurs ne vont pas au-delà. Il semble, par ailleurs, que l'indexation des cadres soit problématique (la plupart des moteurs se limiteraient au traitement du fichier principal[11]). Soulignons que les moteurs ne prennent pas tous en compte les mêmes éléments; en outre, la pondération appliquée varie : le titre, le corps du texte ou la balise <KEYWORDS> peuvent être très importants pour un moteur et relativement insignifiants pour un autre.

Comme la fréquence d'apparition des termes de la requête figure parmi les critères les plus utilisés par les moteurs au moment du tri des réponses, il existe quelques procédés qui permettent de « tricher » afin de favoriser l'émergence d'une page dans les premières places des palmarès : impression de mots en ton sur ton sur le fond d'écran, répétition délibérée de mots clés dans les champs <TITLE>, <DESCRIPTION> et <KEYWORDS>, etc. Ces techniques de *spamdexing* (dixit Olivier Andrieu) sont désormais bien connues des moteurs de recherche et leur utilisation est souvent pénalisée, pouvant aller jusqu'à entraîner l'exclusion de la page fautive de la base de données de l'outil. Une autre tactique pour augmenter l'achalandage sur un site Web – celle-là plus difficile à combattre – consiste à saupoudrer les pages Web de mots clés très populaires mais sans rapport aucun avec le contenu, l'exemple canonique étant *sex* (toutefois, certains sites pornographiques jouent également du procédé inverse, attribuant par

[10] Les balises HTML <TITLE>, <DESCRIPTION> et <KEYWORDS>, toutes optionnelles, contiennent respectivement le titre de la page Web, un résumé de son contenu et des mots clés servant à la décrire. Les attributs ALT des balises <IMG>, optionnels eux aussi, contiennent de courtes descriptions textuelles associées aux images figurant dans une page Web, descriptions qui sont affichées par le fureteur de l'internaute lorsque le mode « chargement des images » est désactivé.

[11] Celui qui renferme la balise HTML <FRAMESET>, c'est-à-dire celui qui contient les liens vers les autres fichiers.

exemple à leurs pages des mots clés fréquemment employés par les internautes tels *sport*, *car*, etc.).

5.3 Exemples de moteurs internationaux

Nom	**URL**
ALTAVISTA [12]	http://www.altavista.com/ http://www.av.com/ http://altavista.digital.com/ etc.!
THE ELECTRIC MONK	http://www.electricmonk.com
EUROSEEK	http://www.euroseek.net/
EXCITE	http://www.excite.com/ version française : http://www.fr.excite.com
GOTO.COM	http://www.goto.com/
HOTBOT	http://www.hotbot.com/
INFOSEEK	http://infoseek.go.com/
LYCOS	http://www-english.lycos.com/ version française : http://www.lycos.fr/
NORTHERN LIGHT	http://www.northernlight.com/ http://www.nlsearch.com/
WEBCRAWLER	http://www.webcrawler.com/

5.4 Exemples de moteurs francophones

Nom	**URL**
ÉCILA	http://www.ecila.fr/
LOKACE	http://www.lokace.com/
VOILA [13]	http://www.voila.fr/ http://voila.fr/ version mondiale : http://www.voila.com/

[12] Voir portrait en annexe page 49.
[13] Voir portrait en annexe page 52.

6 Troisième type d'outils : les métamoteurs

6.1 Principales caractéristiques

Le troisième grand groupe d'outils de recherche est celui des *métamoteurs*. Fondamentalement, les métamoteurs visent à faciliter l'exécution d'une même requête sur plusieurs outils de recherche. Un premier type de métamoteur, assez primitif, est constitué par ce que l'on appelle les *CUSI (Configurable Unified Search Interface)* ou, plus prosaïquement, les *bibliothèques de moteurs* ou les *All in One*. Ce genre d'instrument recense habituellement un grand nombre d'outils de recherche, de manière à les rendre accessibles à l'interrogation à partir d'une même page. Utiles dans la mesure où ils fournissent un accès direct à beaucoup de services et disposent souvent d'une interface qui évite à l'usager d'avoir à retaper continuellement la requête de recherche, ces métamoteurs de première génération ne rendent toutefois que peu de services supplémentaires. Ils se chargent tout simplement de communiquer la requête concernée aux différents outils de recherche, généralement de façon séquentielle.

Quelques exemples de sites, parmi bien d'autres :

Nom	**URL**
ALL-IN-ONE SEARCH PAGE	http://www.allonesearch.com/
EASY SEARCH (japonais)	http://www.aist.go.jp/NIBH/~honda/EasySEARCH/index.cgi
GOLDENBRICK (francophone)	http://www.goldenbrick.fr/goldensearch/recherche.html
THE SEARCH PLACE	http://users.isaac.net/duane/search/

Ceci dit, le terme *métamoteur* est surtout attribué à une seconde catégorie d'outils, à valeur ajoutée ceux-là : les *SUSI (Simultaneous Unified Search Interface).* Ces métamoteurs fonctionnent en transmettant simultanément la requête de l'usager à plusieurs outils de recherche, principalement des moteurs. La quantité d'outils ainsi « interpellés » est très variable; elle se situe d'ordinaire entre 30 et 150. Les métamoteurs récupèrent par la suite les différentes listes de résultats et les façonnent en un document unique. Certains procèdent, en outre, à un classement de pertinence supplémentaire et à

l'élimination des doublons. Plusieurs d'entre eux permettent également de configurer la liste des sources à interroger.

Les principaux avantages de ce type de démarche ont trait au gain de temps (il n'est plus nécessaire de visiter les outils un à un) et au fait que les métamoteurs dispensent l'usager de l'obligation de s'initier aux modalités d'utilisation de chaque nouvel outil – une entreprise souvent laborieuse en ce qui concerne le mode de recherche expert. L'emploi des métamoteurs est, toutefois, confronté à certains problèmes pratiques. Tout d'abord, il s'avère impossible pour ces outils d'exploiter les fonctionnalités avancées des moteurs de recherche, précisément parce que la syntaxe en est très variable. Ensuite, comme le fait remarquer O. Andrieu :

> *[...] les métamoteurs font la synthèse de résultats fournis par plusieurs moteurs différents, classant chacun leurs résultats de façons différentes, sans utiliser les mêmes critères de pertinence. Une synthèse de documents classés de façons ainsi disparates est-elle si simple que cela à effectuer, et surtout, est-elle plus pertinente ? La question reste posée...* [www.abondance.com]

À un autre niveau, il souligne à juste titre les problèmes éthiques que suscite l'apparition de ce genre d'outils :

> *L'utilisation de ce type de métamoteurs engendre un autre problème de fond : quasiment tous les moteurs de recherche sur lesquels ils s'appuient se financent grâce aux bandeaux publicitaires qu'ils affichent. Or, les promoteurs de cette couche logicielle supplémentaire que sont les métamoteurs ne répercutent pas systématiquement (ou pas du tout) ces bandeaux, préférant même parfois proposer leurs propres annonces. Le recours à ces métamoteurs réduit donc de façon substantielle le nombre d'accès au moteur de recherche traditionnel, ce qui compromet ses recettes publicitaires et risque, à terme, de signer son arrêt de mort. D'autre part, se pose un problème d'éthique : est-il juste d'utiliser pour son propre compte les technologies et investissements mis en œuvre par d'autres sociétés, sans contrepartie financière ?* [www.abondance.com]

Ajoutons – ce qui n'étonnera personne – que les métamoteurs anglophones font généralement preuve de ce que l'on pourrait qualifier de « myopie anglo-saxonne » en ce qui concerne la liste des outils à sonder... Le concept de métamoteur, tout en étant intéressant en soi, demeure donc l'objet d'un certain nombre de réserves. Pris pour ce qu'il est, toutefois, et utilisé un peu à la manière d'un annuaire (pour des recherches larges et thématiques), ce type d'outil peut tout de même s'avérer d'une utilité non négligeable.

6.2 Exemples de métamoteurs internationaux

Nom	URL
BEELINE	http://www.transerve.com/beeline/
DOGPILE	http://www.dogpile.com/
INFERENCE FIND	http://www.infind.com/ version française : http://www.infind.com/infind_fr/
MAMMA	http://www.mamma.com
METACRAWLER	http://www.metacrawler.com
METAFIND	http://www.metafind.com/
SAVYSEARCH	http://www.savvysearch.com/
WEBSEEKER	http://www.bluesquirrel.com/products/seeker/webseeker.html

6.3 Exemples de métamoteurs francophones

Nom	URL
ARI@NE	http://www.espace2001.com/moteur
COPERNIC [14]	http://www.copernic.com/fr/
DEBRIEFING	http://www.debriefing.com/france/
TROUVEZ	http://www.trouvez.com

7 Outils particuliers

Les outils présentés dans cette section sont, pour la plupart, des annuaires ou des moteurs qui se distinguent par un trait spécifique : restriction volontaire de couverture (géographique, thématique, sur le type de ressource Internet, sur le genre de document visé), mode de fonctionnement inusité, etc. Certains outils combinent, du reste, plusieurs de ces caractéristiques – par exemple, une limite thématique et une limite géographique.

[14] Voir portrait en annexe page 55.

On peut prévoir que, dans l'avenir, ces divers outils seront de plus en plus souvent incorporés aux sites portails, puisque ces derniers aspirent à être à la fois généralistes et aptes à cibler certains domaines très finement.

Pour chaque rubrique, nous fournissons quelques exemples à titre indicatif.

7.1 Limite géographique

On y trouve des annuaires dont les sites sont classés par zones géographiques, ainsi que des outils dont la couverture est limitée à certaines régions.

EUROFERRET
http://www.euroferret.com/french/
Axé sur les ressources européennes.

MATILDA
http://www.aaa.com.au/images/logos/searches/world/
Sites d'intérêt (notamment outils de recherche) classés par pays.

POLARSEARCH
http://www.polarsearch.com/
Axé sur la Scandinavie : Suède, Danemark, Norvège, Finlande, Islande, Groenland.

LA TOILE DU QUÉBEC
http://www.toile.qc.ca
Axé sur le Québec.

WOYAA!
http://www.woyaa.com/indexFR.html
Moteur de recherche spécialisé sur l'Afrique.

7.2 Limite thématique

Ce sont des annuaires ou des moteurs dont la couverture se limite à un domaine précis. Ils comportent souvent une limitation additionnelle d'ordre géographique.

FOURCHETTE
http://www.fourchette.com
Recherche de restaurants sur le territoire québécois.

BOTTIN ENTREPRISES
http://www.bottin.fr
Recherche d'entreprises en France.

FÉDÉRATION FRANÇAISE DES SALONS SPÉCIALISÉS
http://salons.wcube.fr/
Recherche de salons en France et en Europe par date ou par thème.

KPL
Domaine de la philosophie.
http://www.chez.com/kphi/index_kpl.htm

LAWCRAWLER
http://lawcrawler.findlaw.com/
Domaine du droit.

D'une certaine manière, on pourrait également inclure dans cette catégorie ce que P. Nygren [http://perso.club-internet.fr/nygren/] appelle des *sites gateway*, c'est-à-dire des sites qui appliquent le principe du *Best of* et qui – comme leur nom l'indique – servent de porte d'entrée vers de nombreuses ressources pertinentes sur une thématique donnée.

Exemple :

SCICENTRAL
http://www.scicentral.com/index.html
Pour l'ensemble des sciences et techniques.

7.3 Limite sur le type de ressource Internet

Ces outils, dont plusieurs remontent aux premières années du réseau, se limitent à un type particulier de ressource : sites FTP ou Gopher, adresses de courriel, *newsgroups*, listes de diffusion, etc.

ARCHIEPLEX
http://www.lerc.nasa.gov/archieplex/
Recherche de sites FTP.

BIG FOOT
http://www.bigfoot.com
Version française : http://fr.bigfoot.com/
Recherche d'adresses de courriel (également recherche d'adresses postales). LOKACE , ALTAVISTA, HOTBOT, INFOSEEK, EXCITE et VOILA proposent également une option en ce sens.

DEJA
http://www.deja.com/
Recherche de *news* (forums Usenet). ALTAVISTA, HOTBOT, INFOSEEK, EXCITE, YAHOO!, VOILA (pour les *news* francophones) proposent également une option en ce sens.

FRANCOPHOLISTES
http://www.cru.fr/listes
Recherche de listes de diffusion francophones.

FTPSEARCH
http://ftpsearch.ntnu.no/
Recherche de sites FTP.

LISZT, THE MAILING LIST DIRECTORY
http://www.liszt.com/
Recherche de listes de diffusion.

PERSO-SEARCH !
http://www.perso-search.com/
Recherche de pages personnelles francophones.

VERONICA
gopher://mudhoney.micro.umn.edu:4326/7 (par exemple)
Recherche de sites Gopher.

7.4 Limite sur le genre de document

Outils dont les recherches sont restreintes à un type déterminé de documents.

INPI – RECHERCHE DE BREVETS
http://www.inpi.fr/espacenet
Recherche de brevets européens et internationaux.

LES JOURNAUX SUR LE WEB
http://www.webdo.ch/base/presse.web
Recherche de journaux classés par pays.

A WEB OF ON-LINE DICTIONARIES
http://www.facstaff.bucknell.edu/rbeard/diction.html
Répertoire de dictionnaires classés selon la langue.

SHAREWARE.COM
http://www.shareware.com/
Recherche de partagiciels.

TELEPHONE DIRECTORIES ON THE WEB
http://www.teldir.com/
Répertoire d'annuaires téléphoniques de différents pays.

7.5 Systèmes impliquant des intermédiaires humains

ASK JEEVES
http://www.askjeeves.com/
Cet outil fonctionne à l'aide d'une base de données – élaborée par des humains et non via des robots – où sont regroupées les questions les plus fréquemment posées par les internautes et les réponses précises qui y correspondent. Lorsqu'on pose une question déjà traitée par ASK JEEVES, on est donc certain d'obtenir une réponse pertinente (si la question ne figure pas dans la base, c'est une autre histoire... le logiciel tente alors de cerner le problème en posant des « questions connexes » ou renvoie tout bonnement la requête récalcitrante aux outils de recherche classiques).

KNOWPOST
http://www.humansearch.com
Système basé sur la participation des internautes : chaque réponse apportée à une question posée sur le site donne le droit de formuler une requête à son tour.

8 Systèmes innovants de recherche d'information[15]

Dans cette section, nous présenterons trois secteurs d'activité appelés vraisemblablement à connaître un fort développement dans les années à venir : les agents intelligents, les procédés de *Push* et les pratiques de gestion des connaissances. Mentionnons d'entrée de jeu, toutefois, que les frontières entre ces divers concepts ne sont pas toujours nettement établies : la gestion des connaissances, par exemple, peut être vue comme englobant le *Push* et les agents intelligents – lesquels sont eux-mêmes inextricablement liés entre eux.

Rappelons, en outre, que certains des aspects abordés ici feront l'objet d'une analyse plus détaillée dans notre mémoire de recherche.

8.1 Les agents intelligents

Le concept d'agent intelligent recouvre des réalités nombreuses et diverses. Au sens large, les agents intelligents peuvent être définis comme des outils « *permettant d'automatiser, périodiquement ou à la demande, des tâches de façon transparente pour l'utilisateur qui bénéficie des résultats* » [Philippe Courtot, CEO de Verity, cité dans Careil et de Frémont, s.d.]. Dans le contexte plus spécifique de la recherche d'information, ces logiciels sont généralement dotés, à des degrés divers, des caractéristiques suivantes :

- fonctionnement automatique et autonome;
- mobilité (aptitude à voyager sur les réseaux);
- capacité d'échange avec des interlocuteurs humains ou mécaniques;
- apprentissage dynamique (les agents évoluent au fil du temps et peuvent s'adapter aux circonstances, prendre des décisions et enrichir eux-mêmes leur propre comportement sur la base d'observations qu'ils effectuent).

J.-M. Careil et B. de Frémont [Careil et de Frémont, s.d.] proposent une typologie tripartite des agents intelligents :

[15] Nous empruntons cette expression à O. Andrieu [www.abondance.com].

- **Les *crawlers***
  À partir d'une URL de départ, le *crawler* suit les liens rencontrés et rapatrie le contenu des pages HTML. La puissance de ces agents réside essentiellement dans le traitement qui est ensuite fait de la masse d'information renvoyée : statistiques, indexation, nouveau départ du *crawler* à partir des premiers résultats... C'est ce genre d'agents qui est utilisé par les outils de recherche de type moteur pour élaborer leur base de données (d'ailleurs, le terme *robot* sert parfois aussi à désigner les agents intelligents).
- **Les agents intelligents personnalisés**
  Basés sur les techniques d'intelligence artificielle et de réseaux de neurones, ces agents s'adaptent à leur utilisateur, en analysant ses demandes successives, et/ou à leur environnement, en analysant en direct les informations. Ils sont dès lors capables d'ajuster leur recherche afin d'optimiser la pertinence des résultats. Certains vont jusqu'à accepter des requêtes en langage naturel. C'est ici l'adaptation dynamique du logiciel qui fait son efficacité.
- **Les agents privés**
  Contrairement aux deux types précédents, ces logiciels évoluent sur des bases de données privées. Leur usage implique que le prestataire d'information se soit conformé à une structure spécifique dans le stockage de ses données afin que l'agent puisse repérer l'information utile.

Pour sa part, P. Nygren [http://perso.club-internet.fr/nygren/] fait remarquer fort à propos que les agents intelligents sur Internet n'intègrent pas réellement de caractéristiques types. Il propose donc de les classer en fonction de leur mission, c'est-à-dire de leur capacité à exécuter des tâches spécifiques :

- **Agents de recherche d'information**
  - **Fédérateurs de recherche**
    Ces outils présentent de nombreuses fonctionnalités: recherche d'information simultanée sur plusieurs outils; rapatriement et indexation des pages en local; classement et gestion des informations; élimination des doublons; création de résumés; surveillance des modifications de sites selon une périodicité paramétrable, etc. Les métamoteurs évoqués précédemment s'inscrivent dans cette catégorie. L'on y retrouve également des produits comme WEBSEEKER, BEELINE, WEB FERRET et INFORIAN QUEST 98.
  - **Agents sectoriels**
    Ce sont des fédérateurs de recherche spécialisés dans un domaine précis, par exemple les sciences et techniques, la finance ou la littérature. Les agents sectoriels consultent des outils de recherche spécialisés dans les domaines concernés. WEBSEEKER, BEELINE et COPERNIC entrent aussi dans cette catégorie.

- **Agents pour la consultation hors ligne**
  Ces outils permettent d'aspirer un site Web (texte et images) pour le recopier sur un poste local, en respectant l'arborescence du site d'origine. Il est habituellement possible de spécifier le niveau de profondeur des pages à inclure. Exemples : WEBWHACKER, ECATCH.
- **Agents autonomes**
  Ces agents ont pour mission de dépister « toutes » les pages susceptibles de répondre à une requête donnée (ils peuvent éventuellement prendre l'initiative d'enrichir cette dernière). Ils filtrent et analysent les documents trouvés, ne rapatriant que ceux qui sont réellement pertinents. Ils permettent souvent l'emploi du langage naturel. Exemples : DIGOUT4U, UMAP, NET ATTACHÉ PRO.
- **Agents pour le commerce électronique**
  - **Assistants d'achat (*shopbots*)**
    Destinés aux consommateurs, ils enregistrent les préférences de ces derniers et visent à faciliter la sélection de boutiques virtuelles, de marques ou de produits. Ils peuvent ainsi parcourir les galeries marchandes du Web à la recherche d'un produit ou service particulier; comparer les prix; dresser un tableau récapitulatif des offres disponibles; recommander des produits ou procéder directement à l'achat. Ces assistants peuvent être généralistes (SHOPPING EXPLORER, ROBOSHOPPER) ou porter sur un domaine d'activité précis (PRICELINE pour les billets d'avion, chambres d'hôtel, etc., BARGAIN FINDER pour le prix des disques sur Internet).
  - **Agents d'analyse de la demande**
    Destinés aux commerçants, ils permettent de mieux connaître la demande et les consommateurs, pour une meilleure gestion des profils clients et la personnalisation de l'offre. Exemple : SELECTCAST.
- **« Autres agents »**

Certes, pour le moment, les agents intelligents ont quelque peu usurpé leur nom… Ils deviennent, toutefois, de plus en plus efficaces, et on commence à voir se réaliser les prédictions formulées à leur sujet par J. de Rosnay en 1995 :

> Les agents vont rapidement constituer une nouvelle population d'êtres virtuels. Comme des virus informatiques contrôlés, ils vont se reproduire, constituer des groupes, des « cultures ». Représentants de la vie artificielle, ils vont progressivement coloniser des continents entiers du cyberespace. Des agents travailleront en équipe. Munis de « permis » et « d'autorisations » (d'achat, de négociation), ils pourront se partager un travail et comparer des informations; leurs compétences s'accroissant au fur et à mesure de leurs travaux de recherche ou de préparation de dossiers. Circulant sur les réseaux, ces « intra-terrestres » d'un nouveau genre offriront leurs services. Grâce aux algorithmes génétiques des programmes d'agents pourront muter, s'autosélectionner, évoluer pour résoudre des problèmes de plus en plus complexes. Leur valeur augmentera à la bourse des emplois électroniques. Mais les agents représenteront aussi des dangers potentiels. Sachant tout sur les habitudes, préférences ou secrets de leurs patrons, ils pourront être kidnappés sur les réseaux et utilisés contre leurs employeurs. [de Rosnay 1995]

Les agents peuvent rencontrer les suffrages de nombreuses clientèles. Pour les particuliers, ils peuvent agir comme guides vers les informations recherchées sur le Web, comme assistants d'achat, ou encore pour la gestion documentaire personnelle (lorsque l'agent est configuré pour effectuer des recherches sur le poste même de l'utilisateur). Pour les entreprises, les agents intelligents s'avèrent, en outre, d'une utilité appréciable dans un contexte de veille concurrentielle et technologique sur Internet : « *L'agent intelligent est l'outil de prédilection du cyber-veilleur. De façon transparente ou active, il est obligé de passer par lui pour retrouver l'information pertinente au milieu de ce cyber-fatras* » [Careil et de Frémont, s.d.]. Les agents intelligents permettent ainsi aux veilleurs d'économiser du temps tout en effectuant une couverture exhaustive des sources d'information. De même, il devient possible pour les entreprises de mettre en place des pratiques de surveillance systématique de l'environnement en maintenant des agents en recherche permanente sur le site d'un concurrent : aucun des mouvements économique et stratégique de ce dernier n'échappera ainsi aux utilisateurs desdits agents. Les agents intelligents peuvent également, enfin, servir à élaborer des bases de données thématiques consultables hors ligne ou à analyser un serveur hors ligne.

Voici certains de ces produits :

Nom	**URL**
AURESYS	http://ms161u06.u-3mrs.fr/hom.html
BARGAIN FINDER	http://bf.cstar.ac.com/bf (description seulement)
BEELINE	http://www.transerve.com/beeline/
DIGOUT4U	http://www.arisem.com/index_fr.html
ECATCH	http://www.ecatch.com/accueil.htm
INFORIAN QUEST 98	http://www.inforian.com
MATA HARI	http://www.thewebtools.com
NEARSITE	http://www.nearsite.com
NET ATTACHÉ PRO	http://www.tympani.com/products/NAPro/NAPro.html
PRICELINE	http://www.priceline.com

ROBOSHOPPER	http://www.roboshopper.com/client.htm
SELECTCAST	http://www.aptex.com/products-selectcast.htm
SHOPPING EXPLORER	http://www.shoppingexplorer.com
UMAP	http://www.umap.com
WEB FERRET	http://www.ferretsoft.com/netferret/products.htm
WEBSEEKER	http://www.bluesquirrel.com/products/seeker/webseeker.html
WEBWHACKER	http://www.bluesquirrel.com/products/whacker/whacker.html
WEBZINGER	http://www.webzinger.com

8.2 Les technologies *Push*

Le *Push* – dans le contexte du Web, on parle également de *Webcasting* – réfère à des technologies et services spécialisés qui permettent la livraison automatique aux consommateurs des informations qui les intéressent, en continu et selon des critères personnalisés. Le *Push* s'oppose au *Pull*, méthode classique de recherche d'information où l'utilisateur va vers les données (et non le contraire). Selon P. Nygren, « *le push est donc un nouveau mode d'utilisation d'Internet* » [http://perso.club-internet.fr/nygren/].

Les technologies *Push* emploient diverses tactiques pour transmettre au consommateur les données ainsi récupérées : messagerie électronique, chaînes thématiques auxquelles on souscrit un abonnement, écrans de veille, pages HTML, applications Java, petites fenêtres virtuelles ou messages brefs à l'écran, bandeaux dans lesquels les informations défilent (c'est le système des *tickers*), etc.

P. Nygren distingue trois types de *Push* :

- *Push* pur

  Cette approche a été développée par AirMedia, seul fournisseur à proposer un système de ce type. Le serveur central émet en permanence des pages par radio; toute personne disposant d'un récepteur peut recevoir ces pages, les stocker et les lire sur son ordinateur personnel.

- *Push* sélectif

  Puisqu'un utilisateur particulier n'est pas concerné par toutes les informations émises, on n'émet vers un poste client que les informations susceptibles de l'intéresser. Cette solution est appliquée notamment par POINTCAST.

- *Push*/*Pull* distribué

  Cette solution est mise en œuvre, entre autres, par BACKWEB et EBUSINESS CENTER. Les demandes d'un groupe de clients connectés, se trouvant par exemple dans une même entreprise, sont regroupées au sein d'un serveur local et retransmises au serveur principal. Les informations sélectionnées sont ensuite envoyées sur le serveur local, à charge pour les clients de les consulter sur ce site. BACKWEB va même plus loin : les données sélectionnées sont directement chargées sur les disques durs des clients.

Les technologies *Push* se rencontrent également directement du côté client; elles constituent alors, en quelque sorte, un « *Pull* automatisé ». Un logiciel de *Push* peut ainsi, par exemple, être utilisé pour télécharger à intervalles réguliers certaines pages Web spécifiques sélectionnées par l'utilisateur.

Lancé en 1996, le *Push* a suscité un fort engouement initial, qui est retombé toutefois assez rapidement (suite aux ratés de la technologie ?). De nombreuses firmes spécialisées fondées dans les premiers temps du *Push* ont d'ailleurs fait faillite... À en croire le Yankee Group, cependant, le *Push* a de nouveau le vent dans les voiles puisque les revenus découlant de cette technologie devraient avoisiner les 5,7 milliards de dollars en l'an 2000. S'il s'est beaucoup amélioré, le *Push* demeure néanmoins un procédé qui souffre de nombreux handicaps. Tout d'abord, cette technique est actuellement très gourmande en bande passante, ce qui se traduit par un réel risque de saturation du réseau et requiert sur chaque poste de travail une fraction importante du disque dur. (Un utilisateur de POINTCAST, par exemple, consommerait en gros de l'ordre de 10 fois plus de bande passante qu'un internaute « traditionnel ».) L'absence d'une norme technique établie se fait également cruellement sentir. Enfin, à cause de la transmission d'information en continu, cette technologie est pour ainsi dire incompatible « de nature » avec l'emploi d'une connexion Internet par modem, ce qui contribue à lui fermer le marché des particuliers.

Les plus récentes versions de MICROSOFT INTERNET EXPLORER et de NETSCAPE COMMUNICATOR intègrent des fonctions de *Push*. Du reste, après la bataille des fureteurs, Netscape et Microsoft seront sans doute les deux principaux acteurs de la « guerre du *Push* » (qui débute à peine...). Mentionnons quelques autres produits :

Nom	**URL**
BACKWEB	http://www.backweb.com/
EBUSINESS CENTER	http://www.mediapps.com/web/ibc.nsf
MARIMBA	http://www.marimba.com/
POINTCAST	http://www.pointcast.com

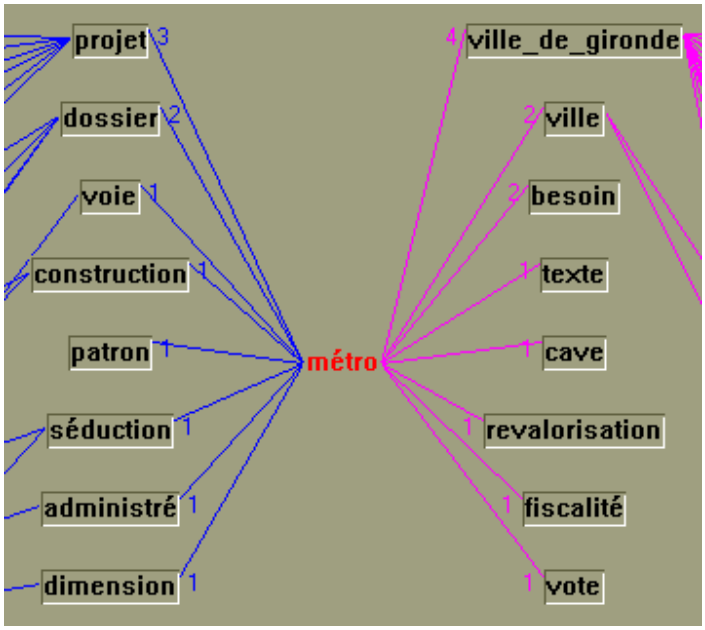
8.3 La gestion des connaissances

Très à la mode actuellement, la gestion des connaissances ou *knowledge management* (KM) recouvre un ensemble d'enjeux, de pratiques et de solutions orientés vers la gestion efficace de ce qui est désormais le capital le plus précieux de toute organisation : le savoir. Le domaine de la gestion des connaissances dépasse donc, en fait, largement le cadre d'Internet, et investit notamment celui des intranets et extranets.

Les outils de KM s'appuient d'ordinaire sur des bases de connaissances généralistes et/ou spécialisées, composées de règles structurelles et contextuelles et de lexiques. Cela leur insuffle une certaine « intelligence » qui, dans un contexte d'activités de traitement documentaire, leur permet d'appréhender un texte comme un ensemble de « molécules de sens » et non plus comme une suite de chaînes de caractères. Ces produits s'avèrent donc aptes, à des degrés divers, à prendre en charge automatiquement des tâches d'analyse textuelle et linguistique : repérage et extraction d'information; comparaison, tri, classement, archivage, indexation et condensation de documents; dépouillement de vocabulaire; correction grammaticale et orthographique; aide à la traduction ou à la rédaction; mise en place de filtres (ex. : pour la messagerie électronique), etc.

À titre d'exemple, TROPES est un logiciel qui permet (selon ses promoteurs) d'effectuer des analyses textuelles en profondeur, afin de comparer des documents entre

eux, de les classer ou de les indexer facilement, rapidement et automatiquement, sans même qu'il soit nécessaire de les lire. TROPES peut travailler sur un texte unique ou sur des corpus de plusieurs milliers de documents; il peut notamment être utilisé pour identifier les mots les plus significatifs, pour établir des listes de synonymes et termes liés, pour constituer des sous-corpus de documents semblables, et pour traduire graphiquement le contenu textuel sous forme de relations sémantiques, comme dans l'exemple suivant :



Source: http://www.acetic.fr/docum.htm

Voici quelques produits qui s'inscrivent dans cette tendance du KM :

Nom	**URL**
ARISEM	http://www.arisem.com/index_fr.html
AUTONOMY	http://www.autonomy.com/
EXCALIBUR	http://www.excalib.com/
NETOWL EXTRACTOR	http://www.isoquest.com/
NOMINO	http://www.ling.uqam.ca/nomino/
RIG VEDA	http://www.gespro.com/fr/produits/rig-veda.html
TEXTSMART	http://www.spss.com/software/textsmart/

9 En guise de conclusion

Nous aimerions clore cette brève présentation des différents types d'outils de recherche d'information existant actuellement sur Internet par quelques considérations relatives, d'une part, aux critères permettant d'évaluer un outil de recherche et, d'autre part, aux pratiques qui favorisent un repérage optimum des données à l'aide de ces outils.

9.1 Critères d'évaluation des outils de recherche

Outre les mesures de *rappel* et de *précision* (un bon outil devrait repérer les ressources pertinentes sans noyer l'usager sous celles qui ne le sont pas), les indicateurs suivants peuvent s'avérer utiles lorsque vient le moment de dénicher l'outil idéal :

- La crédibilité de l'organisation responsable
- La facilité d'utilisation et la convivialité
- La présence de procédures d'aide claires et détaillées. Ces dernières sont primordiales, comme le fait remarquer J.-N. Plourde :

  > La documentation pour les services de repérage aide les utilisateurs à atteindre deux objectifs. Le premier est d'évaluer la pertinence de la base, c'est-à-dire sa nature (objets répertoriés), ses objectifs, son autorité, etc. Le second est la maîtrise et l'utilisation efficace des services de repérage et la vérification du comportement de ces services (obtient-on les résultats escomptés ?). [Plourde 1996]

- L'ampleur de la couverture effectuée par la base de données; son objectivité; ses modalités d'ajout (notamment la possibilité pour les auteurs de sites Web de soumettre leur œuvre)
- La fréquence de mise à jour de la base de données
- La rapidité de fonctionnement
- L'originalité de l'outil
- Les fonctionnalités de recherche simple et avancée offertes, particulièrement en ce qui concerne :
  - ♦ Les domaines de recherche : recherche sur des champs spécifiques (titre, texte intégral, nom du serveur, nom de domaine, intitulé de l'URL, liens hypertextuels,

champ ALT des balises <IMG>); recherche sur des types de ressources déterminés (forums Usenet, adresses de courriel, cartes géographiques ou routières, pages personnelles, etc.); recherche sur des types précis de fichiers (texte, image, audio, vidéo, JavaScript, etc.); recherche sur des dates (date de début, date de fin, intervalle entre deux dates) ou sur un nom de personne.

- La syntaxe de saisie : recherches fondamentales et recherches booléennes (ET, OU, SAUF, locutions, troncature interne ou finale, parenthèses); distinction majuscules/minuscules; sensibilité aux caractères diacritiques; recherches de proximité et prise en compte de l'ordre des mots; possibilité pour l'usager d'appliquer une pondération aux différents termes de la recherche; possibilité d'emploi du langage naturel.

- L'affichage des résultats : facilité de consultation (par exemple, apparition en surbrillance des termes de la requête); configuration de la quantité de résultats à afficher (par page ou en tout) et du format d'affichage (minimal, standard, complet); possibilité de demander l'affichage du titre, du début du texte, de son résumé, de sa taille, de sa date de création et/ou d'indexation, de sa langue de rédaction, de son pays d'origine, d'un pourcentage de pertinence par rapport à la requête, etc., afin de pouvoir se faire rapidement une opinion sur l'intérêt d'un document; possibilité de tri des résultats selon de nombreux critères.

- L'accès à l'historique des recherches et la possibilité de pratiquer des requêtes récurrentes (c'est-à-dire d'effectuer une nouvelle recherche à l'intérieur des résultats d'une requête précédente).

- Les services complémentaires : par exemple, les différentes ressources inhérentes aux sites de type portail[16], la possibilité de traduire automatiquement les documents repérés (comme chez ALTAVISTA ou INFOSEEK), etc.

9.2 Pour une recherche efficace

Certes, les outils de recherche ne rendent pas compte de tout ce qui se trouve sur Internet. Plusieurs facteurs expliquent cette situation : immensité et métamorphoses du réseau; présence de *firewalls*[17]; sites non trouvés, non explorés en profondeur ou interdits aux robots; censure, etc. Une ressource qui n'a pas été recensée demeure évidemment fort difficile à découvrir, à moins de suivre un lien ou d'en connaître l'URL d'avance... Néanmoins, les outils de recherche restent actuellement la meilleure façon de mettre la

[16] Voir section 2.

[17] L'Office de la langue française du Québec définit le *firewall* ou *coupe-feu* comme un « dispositif informatique qui permet le passage sélectif des flux d'information entre un réseau interne et un réseau public, ainsi que la neutralisation des tentatives de pénétration en provenance du réseau public » [http://www.olf.gouv.qc.ca/].

main sur l'information disponible sur Internet. Les conseils suivants permettront d'en optimiser l'utilisation :

- Résumer son besoin d'information sous forme d'une phrase, puis identifier les principaux concepts ayant trait à la requête. Déterminer les termes les plus significatifs (plusieurs, de préférence). Les mots clés retenus devront, dans la mesure du possible, s'avérer « discriminants », c'est-à-dire être rares ou inhabituels. Les mots trop communs sont à éviter absolument, de même que les fautes d'orthographe et de frappe... Il faut également songer à trouver d'éventuels synonymes et traductions des mots clés.

- Ne jamais se limiter à un seul outil ou genre d'outil. Ainsi que nous l'avons déjà mentionné, aucun outil de recherche sur Internet n'offre une couverture parfaitement exhaustive. Les divers outils de même type doivent, du reste, être appréhendés comme étant complémentaires plutôt que concurrents, puisque les portions d'Internet prises en compte varient en dépit de certains recoupements. Inutile, cependant, de tomber dans la surenchère : deux ou trois outils de chaque type – annuaire, moteur, métamoteur – utilisés en parallèle suffisent généralement. Pour une requête simple, thématique ou générale, commencer par un annuaire ou un métamoteur. Pour une requête plus complexe ou pointue, utiliser directement un moteur adapté. Il est impératif de bien connaître le fonctionnement des outils employés en ce qui a trait au type d'indexation effectuée, aux domaines de recherche, à la formulation des requêtes (notamment les opérateurs booléens à utiliser et les questions de majuscules et de diacritiques), aux options d'affichage, etc. Pour une requête en français, il est souvent préférable de s'adresser en priorité à des outils disponibles en langue française, particulièrement en ce qui concerne les annuaires.

- En cas de doute quant au comportement de l'outil face aux majuscules et aux diacritiques, saisir la requête en lettres minuscules et désaccentuées. Il vaut souvent mieux affronter un peu de bruit que souffrir de trop de silence...

- Si les outils habituellement employés ne repèrent rien de satisfaisant pour une requête donnée :
  - recourir aux métamoteurs pour croiser les recherches;
  - se servir d'agents intelligents;
  - rechercher dans les *foires aux questions* (FAQ);
  - utiliser des *newsgroups* judicieusement choisis pour poser la question;
  - tenter une nouvelle requête à l'aide de mots clés plus génériques.

- Si, au contraire, une requête particulière s'avère trop fructueuse :
  - ajouter un ou des mot(s) clé(s) supplémentaire(s);
  - pour les outils qui le permettent, recourir aux opérateurs booléens, notamment à la recherche de locutions;
  - exploiter les possibilités de la recherche avancée;
  - tenter une nouvelle requête à l'aide de mots clés plus spécifiques.

- Penser à recourir aux outils de recherche spécialisés, dont on a souvent intérêt à se faire des signets.

- Pratiquer ce que P. Nygren [http://perso.club-internet.fr/nygren/] appelle la *Reverse Psychology* : cette technique consiste, à partir d'un site jugé pertinent, à rechercher systématiquement toutes les pages possédant un lien hypertextuel pointant vers son URL. Ce résultat peut être atteint soit par l'emploi des commandes avancées des moteurs de recherche soit par le recours aux annuaires, en retrouvant la ou les catégorie(s) où le site en question a été référencé.

10 Liste des sources consultées

Outre les sources mentionnées ci-après, nous avons eu recours, dans le cadre de ce travail, à l'information présente sur les sites mêmes des divers outils de recherche.

Nous avons tenté, pour les références de ressources électroniques, de fournir une description aussi exhaustive que possible. Certaines informations manquent cependant parfois (date, nom du responsable, etc.), car elles sont demeurées introuvables.

A) Ressources Internet spécialisées

Sites Web en français

Abondance : recherche d'information, référencement et promotion de sites Web
http://www.abondance.com/
Maintenu par Olivier Andrieu.

Les agents intelligents
http://ms161u06.u-3mrs.fr/
Maintenu par Bruno Mannina.

La Loupe : guide de recherche sur Internet
http://laloupe.magnit.com

Meta News
http://www.metanews.net/
Maintenu par la société La Mine.

Les moteurs de recherche francophones
http://www.idf.net/mdr/
Maintenu par la société IDF.net.

Les outils de recherche : pour enfin s'y retrouver
http://pages.infinit.net/popnet/recherche/
Maintenu par la société Services Pop.net.

Un outil de veille stratégique sur Internet
http://perso.club-internet.fr/nygren/
Maintenu par Pierre Nygren.

Sites Web en anglais

Search Engine Showdown
http://www.notess.com/search/
Maintenu par Gregg R. Notess.

Search Engine Watch
http://www.searchenginewatch.com/
Maintenu par Danny Sullivan.

Forum de discussion

alt.internet.search (anglophone)

Listes de discussion et listes de diffusion

Agents
Porte sur les agents intelligents.
Pour inscription : agents-subscribe@egroups.com

I-Search Digest (anglophone)
Porte sur les outils de recherche.
http://www.audettemedia.com/i-search/

Motrech
Porte sur les moteurs de recherche.
http://www.chez.com/jcharron/motrech/presentation.html
Pour inscription : motrech-subscribe@egroups.com

Référencement
Porte sur le référencement de sites Web.
Pour inscription : referencement-subscribe@egroups.com

B) Autres documents en ligne

Careil, J.M. et B. de Frémont. « Les agents intelligents ». Présentation interactive disponible pour consultation sur le site de Bruno Mannina [http://ms161u06.u-3mrs.fr/].

Conférence des recteurs et des principaux des universités du Québec (CREPUQ), Sous-comité des bibliothèques, Groupe de travail sur l'accès aux ressources documentaires, Sous-groupe de travail sur Internet. « GIRI – Guide d'initiation à la recherche dans Internet ». Édition du 1er juin 1998.
http://www.bibl.ulaval.ca/vitrine/giri/index.htm

Conférence des recteurs et des principaux des universités du Québec (CREPUQ), Sous-comité des bibliothèques, Groupe de travail sur l'accès aux ressources documentaires, Sous-groupe de travail sur Internet. « GIRI 2 – Guide des indispensables de la recherche dans Internet ». Édition du 1er mars 1999.
http://www.bibl.ulaval.ca/vitrine/giri/giri2/index.html

de Rosnay, Joël. « Les agents intelligents : robots logiciels ». 19 octobre 1995.
http://194.199.143.5/derosnay/agent.htm

Jakob, David. « Trouver des informations sur le Web ». *Flash Réseau*, no 15, Bibliothèque nationale du Canada, Services de technologie de l'information. 10 octobre 1995 (révisé le 29 juillet 1997). http://www.nlc-bnc.ca/pubs/netnotes/fnotes15.htm

Koster, Martijn. « Robots in the Web : threat or treat ? ». Avril 1995. http://info.webcrawler.com/mak/projects/robots/threat-or-treat.html

Laublet, Philippe. « Collecte d'information et recherche documentaire sur Internet ». CAMS, Université de Paris-Sorbonne. S.d. (postérieur à 1997). http://www.mpl.orstom.fr/CDROM/ch06/laublet/laublet.htm [Ce document est désormais inaccessible]

Plourde, Jean-Noël. « Définition et application de critères d'évaluation d'outils de recherche dans Internet ». *Cursus*, vol. 1, no 2. Printemps 1996. http://www.fas.umontreal.ca/EBSI/cursus/vol1no2/plourde.html

Vanhoolandt, Philippe. « Ask Jeeves et SearchMill : des recherches à coloration humaine ». Janvier 1998. http://www.vanho.com/ar980109.htm

C) Monographies

Andrieu, Olivier. *Créer du trafic sur son site Web*. Paris : Éditions Eyrolles, 1998. 500 pages.

Andrieu, Olivier. *Trouver l'info sur Internet*. Paris : Éditions Eyrolles, 1998. 460 pages.

Basch, Reva. *Researching Online for Dummies*. Foster City (Californie) : IDG Books Worldwide, 1998. 328 pages.

Centre d'expertise et de veille Inforoutes et Langues (CEVEIL). *Internet, intranet, extranet : comment en tirer profit*. Montréal : Les Éditions Transcontinental, 1998. 208 pages.

Lalonde, Louis-Gilles et André Vuillet. *Chercher et trouver dans Internet*. Montréal : Éditions Logiques, 1998. 139 pages.

Miles, Peggy. *Internet World Guide to Webcasting*. New York : John Wiley & Sons, 1998. 422 pages.

D) Articles de périodique

Balas, L. Janet. « Beyond Veronica and Yahoo! : more Internet search tools (6 lesser-known tools that pick out e-mail addresses, newsgroups, and more) ». *Computers in Libraries,* vol. 16, mars 1996, p. 34-38.

Balas, L. Janet. « Exploring some new search tools for librarians ». *Computers in Libraries*, vol. 19, mai 1999, p. 34-37.

Brandt, D. Scott. « What flavor is your Internet search engine ? ». *Computers in Libraries*, vol. 17, janvier 1997, p. 47-50.

Courtois, P. Martin; Baer, M. William et Marcella Stark. « Cool tools for searching the Web : a performance evaluation ». *Online*, novembre 1995, p. 15-32.

Courtois, P. Martin et Michael W. Berry. « Results ranking in Web search engines ». *Online*, mai/juin 1999, p. 39-46.

Dalloz, Xavier. « Les agents intelligents arrivent ». *L'Atelier*, no 46-47, septembre/décembre 1995, p. 24-27.

Dubin, David. « Search strategies for Internet resources ». *School Library Media Quarterly*, vol. 24, automne 1995, p. 53-54.

Garman, Nancy. « Meta search engines ». *Online*, mai/juin 1999, p. 74-78.

Grevet, Jean-Pascal. « Les moteurs de recherche sur Internet ». *Icônes*, no 55, février/mars 1996, p. 20-25.

Hock, Randolph. « Web search engines features and commands ». *Online*, mai/juin 1999, p. 24-28.

Lardy, Jean-Pierre. « Les outils de recherche d'information sur Internet ». *Documentaliste – Sciences de l'information*, vol. 33, no 1, 1996, p. 33-38.

O'Leary, Mick. « Portal wars ». *Online*, janvier/février 1999, p. 77-79.

Notess, R. Greg. « Searching the World-Wide Web : Lycos, WebCrawler and more (best known indexes to the Web and other Internet search tools) ». *Online*, juillet/août 1995, p. 48-53.

Notess, R. Greg. « Internet "Onesearch" with the mega search engines (SavvySearch and MetaCrawler) ». *Online*, novembre/décembre 1996, p. 36-39.

Notess, R. Greg. « Rising relevance in search engines ». *Online*, mai/juin 1999, p. 84-86.

Notess, R. Greg. « Search engines in the Internet age ». *Online*, mai/juin 1999, p. 20-22.

Randal, Neil. « Search engines : powering through the Internet ». *PC Computing*, septembre 1995, p. 165-168.

Repman, Judi et Randal D. Carlson. « Surviving the storm : using metasearch engines effectively ». *Computers in Libraries*, vol. 19, mai 1999, p. 50-55.

Scales, B. Jane et Elizabeth Caulfield Felt. « Diversity on the World Wide Web : using robots to search the Web ». *Library Software Review*, vol. 14, automne 1995, p. 132-136.

Sherman, Chris. « The future of Web search ». *Online*, mai/juin 1999, p. 54-61.

Sullivan, Danny. « Crawling under the hood : an update on search engine technology ». *Online*, mai/juin 1999, p. 30-38.

Thil, Jérôme. « Outils "intelligents" de recherche d'informations : mythe ou réalité ». *Technologies Internationales*, no 26, juillet/août 1996, p. 7-10.

Tomaiuolo, G. Nicholas et Joan G. Packer. « An analysis of Internet search engines : assessment of over 200 search queries ». *Computers in Libraries*, vol. 16, juin 1996, p. 58-62.

Vidmar, J. Dale. « Darwin on the Web : the evolution of search tools ». *Computers in Libraries*, vol. 19, mai 1999, p. 22-28.

11 Annexes

11.1 Portrait : YAHOO! INTERNATIONAL

http://www.yahoo.com/

Lancé en 1994.

Sous la responsabilité de la société Yahoo! Inc.

Taille de la base : plus de 500 000 sites distribués en plus de 25 000 catégories.

Sites régionaux : Canada, Amérique latine, France, Italie, Espagne, Irlande et Royaume-Uni, Allemagne, Danemark, Suède, Norvège, Australie et Nouvelle-Zélande, Hongkong, Japon, Corée, Taiwan, Chine, Singapour. Il existe également des Yahoo! spécifiques à certaines villes : Chicago, Los Angeles, etc.

YAHOO! est à la fois un pionnier d'Internet et un des principaux piliers du Réseau. Il compte parmi les sites les plus fréquentés du Web et constitue sans contredit l'outil de recherche le plus connu et le plus utilisé. Selon certaines études, il générerait à lui seul près de la moitié de tout le trafic dirigé vers un site Web par les moteurs de recherche. YAHOO! se présente également comme un immense portail où sont offerts une multitude de services : actualités, résultats sportifs, météo, cours de la Bourse, annonces classées, courriel gratuit, bavardage en direct, Pages Jaunes, horaire des programmes de télévision, cartes géographiques, calendrier, personnalisation individuelle du site – pour n'en nommer que quelques-uns.

Les sites jugés particulièrement intéressants par l'équipe éditoriale sont signalés dans le répertoire par une icône spécifique (représentant des verres fumés).

Modalités d'inscription

YAHOO! fonctionne principalement par le biais de la soumission volontaire de sites. La demande d'inscription comporte, entre autres, le titre du site, son URL et une brève description.

- On peut inscrire un même site dans deux catégories différentes. Celles-ci ne peuvent être choisies qu'à partir d'un certain degré de spécificité (postérieur au minimum aux deux premiers niveaux hiérarchiques). Les usagers sont libres, par ailleurs, de faire des suggestions de nouvelles catégories. Il est à noter que la structure hiérarchique de YAHOO! fonctionne selon un ingénieux système dit des « catégories liées » : en naviguant dans la hiérarchie catégorielle, l'usager croise bon nombre de rubriques dont l'intitulé est suivi du symbole @, qui sert à indiquer que la rubrique concernée est « officiellement » implantée, en fait, dans un autre endroit de l'arborescence de YAHOO!, mais accessible ponctuellement via un lien hypertextuel. Cette technique vise la multiplication des points d'accès en permettant de faire figurer une rubrique à tous les endroits où elle peut se révéler potentiellement intéressante, et augmente donc la visibilité des sites qui y sont répertoriés. C'est un peu comme si un même site était recensé dans des dizaines de rubriques différentes...

- La classification de YAHOO! opère une distinction entre les sites à contenu régional et les sites à contenu non régional : les premiers doivent figurer dans la rubrique *Regional*. YAHOO! différencie également les sites commerciaux et ceux qui ne le sont pas, en imposant l'inscription des sites

marchands à l'intérieur de la rubrique *Business and Economy*. Si un site présente un double caractère régional et commercial, il doit être recensé à la fois dans les rubriques commerciale (en priorité) et régionale adaptées. Les pages personnelles, pour leur part, doivent être soumises à une rubrique dédiée (*Society and culture/People/Personal Home Pages*). Enfin, les sites présentant un caractère limité dans le temps (foire, exposition, congrès, etc.) doivent faire l'objet de démarches spécifiques lors de l'inscription. Les événements « en direct » (par exemple, une séance de *chat* impliquant une célébrité), notamment, doivent être soumis à la catégorie *Yahoo! Net Events*.

- Les sites dans une langue autre que l'anglais doivent être répertoriés dans le site régional adapté (s'il existe) et non dans YAHOO! INTERNATIONAL. Cette restriction semble valoir aussi pour les pays anglo-saxons pour lesquels se propose un Yahoo! spécifique : Australie et Nouvelle-Zélande, Irlande et Royaume-Uni, Canada[18].

- L'équipe éditoriale se réserve le droit de remanier les données soumises par les webmestres, notamment en ce qui a trait au descriptif. Le choix des rubriques de classification, de même, est susceptible de modifications : les éditeurs peuvent remplacer une des rubriques choisies (ou les deux), inscrire le site dans une seule catégorie ou même en créer une nouvelle pour l'occasion...

La visite d'un membre de l'équipe éditoriale a lieu quelques jours – voire quelques semaines – après la soumission. Les milliers de demandes d'inscription reçues quotidiennement par YAHOO! font l'objet d'une épuration drastique : beaucoup plus de la moitié seraient rejetées. La procédure normale d'inscription prend en tout un ou deux mois; en cas d'échec, il est permis au webmestre de reformuler la candidature de son site autant de fois qu'il le désire.

Pour augmenter les chances d'acceptation d'un site :

- Miser sur le contenu : originalité, intérêt, forte valeur ajoutée.
- Assurer une navigation rapide et aisée à l'intérieur du site.
- Éviter les sections de type « serveur en chantier », « rubrique non disponible », de même que les liens hypertextuels brisés (« erreur 404 »).
- Ne pas imposer l'emploi de programmes informatiques supplémentaires (*plug-in*).
- Soigner le titre et la description fournis : l'intitulé du titre mérite une attention particulière puisque YAHOO! classe ses résultats par ordre alphabétique à l'intérieur d'une rubrique. Il doit être bref (pas plus de 5 mots ou 40 caractères) et être saisi en minuscules à l'exception, éventuellement, de la lettre initiale. Si le site présente une entreprise, il est recommandé d'employer le nom de celle-ci comme titre. Quant à la description, également brève (elle ne doit pas dépasser 25 mots), il lui faut être précise et efficace. Plutôt qu'une liste de mots clés séparés par des virgules, il vaut mieux proposer une phrase grammaticalement correcte, incluant un ou plusieurs verbes. Ne pas répéter le titre ni le nom de la catégorie dans le commentaire afin de ne pas gaspiller inutilement un espace précieux. Ne pas non plus saisir en majuscules le texte de la description – ni même les initiales de chaque mot – ni y inclure de chiffres publicitaires (du type *Offering over 200 products*) ou de noms propres propriété d'autrui (*We sell Rolex, Timex and Tag Heuer watches*). Enfin, il est impératif de ne jamais recourir à des affirmations égocentriques comme *The best site on the Internet* ou *We're the number one dealer*...

YAHOO! ne semble pas vérifier périodiquement la validité des URL des sites recensés.

[18] Sauf les sites québécois, qui sont répertoriés dans YAHOO! FRANCE.

Syntaxe de saisie

- **Casse**
  Elle n'est pas prise en compte.

- **Opérateurs booléens**
  ET (+)
  OU (par défaut, donc espace)
  SAUF (–)
  Locution (" ")
  Troncature (*) Elle s'applique par défaut à droite pour tout mot de plus de trois lettres (par exemple, *moon* retrouvera *moons*, *moonlight*, *moonlighting*, etc.). Elle est donc automatique sur le pluriel des mots.

  Il faut employer les symboles algébriques + – et l'espace, non les conjonctions AND, OR, NOT.

- **Recherches spécifiques**
  Sur le titre du document (t:)
  Sur l'intitulé d'une URL (u:)

- **Options additionnelles de recherche avancée**
  Sur la date (ajouts remontant à 1 jour, 3 jours, 1 semaine, 1 mois, 3 mois, 6 mois, 3 ans)
  Sur le type de ressource (sites recensés ou forums Usenet)
  Sur diverses sections de la base de données (catégories YAHOO! ou sites recensés)
  Nombre de résultats à afficher par page (10, 20, 50, 100)

Les opérateurs suivants peuvent être combinés à l'intérieur d'une même requête, à la condition expresse de respecter l'ordre dans lequel ils sont énumérés : + – t: u: " " *.

Une requête infructueuse est automatiquement transférée à la base de données du robot INKTOMI.

Affichage des résultats

Les résultats sont affichés dans l'ordre suivant :

- Catégories YAHOO! dont l'intitulé contient les termes de la requête.
- Sites correspondant à la requête (y compris la ou les catégorie(s) où ils sont recensés).
- Le cas échéant, pages repérées dans la base de données du robot INKTOMI.

YAHOO! se sert des éléments suivants pour le classement des sites suite à une requête, par ordre décroissant d'importance :

- Nom des catégories.
- Titres.
- Descriptifs.
- URL.

Sont également pris en compte le nombre de mots clés de la requête trouvés et la présence de correspondances exactes plutôt qu'approximatives.

11.2 Portrait : NOMADE

http://www.nomade.fr/

Sous la responsabilité de la société Objectif Net.

NOMADE est un répertoire francophone qui recense les sites Internet en français, indépendamment de la nationalité de leur webmestre ou de l'emplacement géographique des serveurs d'hébergement. Outre la navigation par catégorie, NOMADE permet une recherche à l'aide de mots clés dans sa base de données. Le site offre également des services de type portail : météo, finance, actualité, etc.

L'équipe rédactionnelle de NOMADE propose chaque semaine une sélection de sites particulièrement dignes d'intérêt.

Modalités d'inscription

NOMADE étant un guide par soumission, les webmestres fournissent la description de leur site ainsi que la ou les catégorie(s) de classification. L'adéquation de ces informations, toutefois, est contrôlée par l'équipe éditoriale avant la mise en ligne. Cette équipe prend également l'initiative d'ajouter certains sites dans le guide, rédigeant alors une description détaillée des services concernés.

- Un site ne peut être soumis que dans le dernier niveau de catégorie de l'arborescence (les catégories supérieures sont uniquement destinées à orienter l'utilisateur). Le webmestre peut inscrire un même site dans deux catégories différentes; l'équipe éditoriale se réserve le droit d'ajouter une troisième catégorie.

- Plusieurs sous-ensembles de pages d'un même site peuvent faire l'objet d'inscriptions indépendantes, à condition que le site soit constitué de rubriques développées et dont le contenu est réellement varié. Il est nécessaire, dans ce cas, de rédiger une description particulière pour chaque sous-ensemble de pages soumis.

- Les sites édités par une entreprise commerciale ou contenant une offre de produit ou service sont répertoriés en priorité dans la rubrique générale *Entreprises et services*. (Cependant, si un site commercial contient de l'information générale à caractère non promotionnel, ce sous-ensemble de pages peut être inscrit dans des catégories supplémentaires.)

  Tout site édité par un particulier et qui présente cette personne et ses activités doit être inscrit dans la catégorie *Pages personnelles* (située dans la rubrique *Loisirs et tourisme*) et ses région et ville correspondantes. Si un site édité par une personne physique contient des informations détaillées sur un sujet d'intérêt général (art, sport, région touristique, etc.), toutefois, il est possible d'inscrire ce site dans la ou les catégorie(s) correspondante(s).

  Les autres sites sont classés dans les différentes catégories de NOMADE en fonction des sujets qu'ils traitent.

- Le navigateur du webmestre doit accepter les *cookies* pour permettre l'inscription en ligne d'un site (autrement, il demeure possible de procéder à une inscription par courriel).

- L'inscription peut être refusée, notamment dans les cas suivants :
  - ♦ Descriptifs et choix de catégories ne correspondant pas au contenu du site.

- Site inscrit plus d'une fois sous des URL différentes.
- Sites proposant un contenu trop succinct ou insuffisamment cohérent.
- Sites portant « atteinte à l'ordre public ou aux bonnes mœurs ».
- Annonces immobilières de particuliers.

Le contenu de la base de données de NOMADE s'enrichit quotidiennement (un lien permet de visualiser les dernières nouveautés). Un délai d'une semaine environ s'écoule entre l'inscription d'un site et son apparition dans l'annuaire (s'il est accepté). Un robot tourne « régulièrement » sur toutes les adresses répertoriées afin de repérer celles qui sont inactives. Par ailleurs, les webmestres et utilisateurs sont invités à collaborer à l'entretien du répertoire : suggestion de nouvelles catégories; mise à jour des URL et descriptifs; signalement des liens inactifs, des changements d'adresse, des sites non répertoriés, des erreurs de classification, des descriptions incorrectes, etc.

Syntaxe de saisie

- **Casse, caractères diacritiques**
  Ils ne sont pas pris en compte.

- **Opérateurs booléens**
  ET (par défaut)
  OU
  Locution (" ")
  Troncature (appliquée par défaut à la racine des mots : *voiture* retrouvera *voiturette*, etc.).

  Il faut sélectionner les choix du formulaire de recherche et non se servir de conjonctions de coordination (ET, OU) ou de symboles algébriques (+ –).

- **Recherches spécifiques**
  Sur les différentes sections de la base (intégralité, dépêches AFP, sites récents, « sélections »)
  Sur le domaine (pays/région francophone ou États-Unis, type d'organisme)
  Sur le type de public (adultes, enfants, professionnels, etc.)
  Par défaut : sur le pluriel et les formes conjuguées d'un mot clé, y compris les formes irrégulières (ex. : alternances du type *cheval* vs *chevaux*).

La recherche par mot clé s'effectue sur les descriptions des sites – à savoir le titre, le résumé, les mots clés complémentaires et les informations concernant le webmestre. Les mots non significatifs de la requête (articles, conjonctions, prépositions, etc.) ne sont pas pris en compte, sauf si on utilise des guillemets. En cas de recherche infructueuse, la requête est automatiquement redirigée vers ALTAVISTA.

Affichage des résultats

Les informations suivantes y figurent :

- Catégorie(s) du site.
- Nom du webmestre.
- Ville, région, pays.
- Nature (association, entreprise, personnel, public, éducation).
- Public visé.
- Résumé.

11.3 Portrait : ALTAVISTA

http://www.altavista.com/
http://www.av.com/
http://altavista.digital.com/
Et plusieurs autres...

Lancé en décembre 1995.

Sous la responsabilité de la société Compaq Computer Corporation.

Taille de l'index : 140 millions de pages.

Délai de rafraîchissement de l'index[19] : six semaines.

Nom du robot utilisé : *Scooter*. Ce robot indexait 10 millions de pages par jour en 1998, en plus de surveiller de façon continue 2 000 sites majeurs (parmi les plus populaires du Web) afin qu'ALTAVISTA en reflète le contenu le plus fidèlement possible.

Six miroirs régionaux : Canada, Europe du Sud, Europe du Nord (inactif), Asie, Australie et Amérique latine. Les bases de données de ces sites sont des recopies du site original californien.

Beaucoup d'experts considèrent ALTAVISTA comme l'un des meilleurs outils de recherche en texte intégral sur le Web. Il figure assurément, en tout cas, parmi les plus connus et les plus populaires...

Le temps de réponse sur une requête est d'environ une demi-seconde.

Modalités d'indexation

ALTAVISTA tient compte des éléments suivants :

- Contenu des champs <TITLE>, <DESCRIPTION> et <KEYWORDS> du fichier HTML (limite de 100 caractères dans le premier cas et de 1 024 caractères dans les deux autres).
- Attributs ALT des balises <IMG>.
- Intitulé de l'URL.
- Images en coordonnées.
- Fichier principal des cadres, et parfois également le contenu de chaque cadre.
- Corps du texte jusqu'à concurrence de 100 Ko (entre 100 Ko et 4 Mo, seuls les liens hypertextuels sont indexés; aucune indexation n'est effectuée au-delà de 4 Mo).

L'évaluation de pertinence s'appuie surtout sur les éléments suivants, par ordre décroissant d'importance : titre, corps du texte, balise <KEYWORDS>.

[19] C'est-à-dire le délai qui s'écoule entre deux renouvellements complets du contenu de la base de données de l'outil.

Syntaxe de saisie

- **Casse**
  Requête en lettres minuscules : toutes les occurrences sont recherchées.
  Requête comportant des lettres majuscules : seule l'occurrence exacte est recherchée.

  ALTAVISTA observe un respect absolu de la casse employée lors de la requête, sauf dans le cas d'une requête entièrement en minuscules – laquelle repêchera toutes les occurrences du motif recherché, majuscules et minuscules confondues. Par exemple, *montréal* repêchera *montréal*, *Montréal*, *MONTRÉAL*, etc., alors que *Montréal* ne repêchera que *Montréal*.

- **Caractères diacritiques**
  Requête sans accents : toutes les occurrences sont recherchées.
  Requête avec accents : seule l'occurrence exacte est recherchée.

  AltaVista recherche les occurrences exactes de la requête en ce qui a trait à l'accentuation, à l'exception des requêtes entièrement sans accents qui repêcheront toutes les occurrences du motif concerné, accentuées ou non. Par exemple, *cote* repérera *cote*, *coté*, *côte*, *côté*, etc. Cette caractéristique vaut également pour les majuscules accentuées, d'où certains problèmes dus aux différences d'usage à ce niveau entre les diverses régions de la Francophonie (une requête avec *École* ne repérera pas *Ecole*...).

- **Ordre des mots**
  Il est pris en considération.

- **Opérateurs booléens**
  OU (par défaut, donc espace)
  ET (+)
  SAUF (-)
  Locution (" ")
  Troncature (*) Elle s'applique pour 0 à 5 caractères apparaissant après une séquence de 3 lettres au moins, et n'est valide ni pour les chiffres ni pour les majuscules. La requête sera annulée s'il y a trop de réponses... Par ailleurs, la troncature du pluriel en *s* est automatique (par exemple, *bibliothèque* retrouvera *bibliothèques* et vice-versa).

- **Recherches spécifiques**
  Sur le titre du document (title:)
  Sur le nom de domaine (domain:)
  Sur le nom du serveur (host:)
  Sur l'intitulé d'une URL (url:)
  Sur le nom d'une applet Java (applet:)
  Sur le nom d'une image (image:)
  Sur les adresses des liens hypertextuels (link:)
  Sur les intitulés des liens hypertextuels, c'est-à-dire les portions de texte (généralement soulignées et de couleur contrastée) qui servent à indiquer ces liens (anchor:)
  Sur le texte visible (text:)
  Limitation linguistique (25 langues)

- **Options additionnelles de recherche avancée**
  ET, OU, SAUF deviennent respectivement : AND (&), OR (|), AND NOT (ou OR NOT)
  Opérateur de proximité (« fenêtre » de 10 mots maximum) : NEAR (~)
  Parenthèses
  Recherche sur la date des documents (from: et to: pour indiquer une fourchette de dates)

Affichage des résultats

Les résultats sont limités à 200 par requête. Chaque page de résultats propose 15 entrées. Plusieurs choix s'offrent à l'usager au niveau du format d'affichage des réponses : *standard*, *détaillé*, *condensé*. Les informations suivantes peuvent y figurer :

- Contenu de la balise <TITLE> (80 premiers caractères environ) ou, à défaut, la mention *No Title*.
- En guise de résumé : contenu de la balise <DESCRIPTION> (150 premiers caractères environ), ou, à défaut, les premières lignes du texte.
- Taille du document.
- Date de la dernière modification (si le document a été indexé par *Scooter* lors d'une visite « classique » depuis un lien externe) ou date de l'entrée dans l'index (si la page a été proposée manuellement par l'intermédiaire de la fonction *Add URL*).
- URL.
- Langue de la page.

Le format standard comporte le titre, le résumé, l'URL et la taille du fichier. Le format condensé se limite au titre et aux premiers mots du résumé, le tout étant tronqué pour tenir en une demi-ligne environ. Le format détaillé comprend tous les éléments ci-haut.

Mentionnons que, depuis le printemps 1999, ALTAVISTA vend au plus offrant les deux premières positions de ses listes de résultats, du moins ce qui concerne les requêtes effectuées à l'aide des mots clés les plus fréquemment employés (du style *woman* ou *picture*...).

11.4 Portrait : VOILA

http://www.voila.fr/
http://voila.fr/
http://www.voila.com/ (version mondiale)

Lancé en juillet 1998.

Sous la responsabilité de France Télécom, en collaboration avec la société Echo.

Taille de l'index : plus de 100 millions de pages, dont 6 millions de pages francophones[20].

Délai de rafraîchissement de l'index[21] : une semaine.

Nom du robot utilisé : *Echo*.

Miroirs régionaux disponibles pour les pays suivants : Canada (en collaboration avec CARREFOUR), États-Unis, Pays-Bas, Danemark, Espagne, Portugal et Italie.

VOILA inclut et remplace les services PAGESWEB et QUIQUOIOÙ [http://www.wanadoo.fr/bin/frame.cgi?service=quiquoiou]. Ce moteur incorpore un annuaire composé de 13 « chaînes » (actualité & média, administration & politique, arts & culture, économie & finance, enseignement & emploi, informatique & Internet, santé, sciences & recherche, shopping, société & religion, sports, tourisme & voyages, vie pratique) de même que divers services de type « portail » : actualités, courriel gratuit, bavardage en direct, Pages Jaunes et Blanches, plans et itinéraires, agenda culturel, etc. VOILA est à la fois un moteur francophone et un moteur international (que ses promoteurs espèrent sous peu positionner avantageusement sur le marché américain, entre autres). Il est par ailleurs possible d'installer une version personnelle du moteur sur un site Web, afin de s'en servir comme robot interne.

Modalités d'indexation

VOILA tient compte des éléments suivants :

- Contenu des champs <TITLE>, <DESCRIPTION> et <KEYWORDS> du fichier HTML (limite d'environ environ 100 caractères dans le premier cas et de 400 caractères pour les deux autres).
- Intitulé de l'URL.
- Corps du texte.
- Fichier principal des cadres, et parfois également le contenu de chaque cadre.

[20] 150 millions prévus pour la mi-1999.
[21] C'est-à-dire le délai qui s'écoule entre deux renouvellements complets du contenu de la base de données de l'outil.

Les attributs ALT des balises <IMG> et les images en coordonnées sont ignorées. L'évaluation de pertinence s'appuie surtout sur le titre (très important), ainsi que sur l'URL et le positionnement de mots dans des balises spécifiques (<H1> ou <EM>, par exemple). La balise <KEYWORDS> et le corps du texte sont considérés comme moins importants.

Syntaxe de saisie

- **Casse, caractères diacritiques et ordre des mots**
  Ils ne sont pas pris en compte.

- **Opérateurs booléens**
  Pas de ET, OU, SAUF ni de recherche de locutions en recherche simple (ET est appliqué par défaut).

- **Recherches spécifiques**
  Sur les diverses sections de la base de données (Web mondial, Web francophone, *newsgroups*, dépêches AFP)
  Sur le type de fichier (sons, images, vidéos)
  Sur le domaine (pays francophone ou type d'organisme)
  Par thèmes : VOILA offre une possibilité intéressante et novatrice, celle de pratiquer des recherches thématiques sur 26 sujets différents, grâce à une technologie algorithmique qui permet, lors de la constitution de la base de données suite aux investigations du robot, de classer automatiquement les pages recueillies à l'intérieur d'une arborescence de thèmes (la base de données du moteur évoque donc un peu la structure d'un annuaire). La restriction thématique d'une recherche peut s'effectuer en amont (en optant pour un thème spécifique dans le formulaire de recherche) ou en aval (suite à une requête, le moteur propose à l'usager un liste de thèmes susceptibles de correspondre à la thématique de recherche, ce qui permet de filtrer les réponses obtenues dans un premier temps).

- **Options additionnelles de recherche avancée**
  Opérateurs booléens
  Recherche sur le féminin, le pluriel et d'autres mots proches (orthographe voisine)
  Choix supplémentaires pour les restrictions sur le domaine ou le type de fichier (ex. : pays non francophones, fichiers de type Windows/Mac/Unix)
  Nombre et format d'affichage des résultats

Affichage des résultats

En recherche avancée, chaque page de résultats propose 10, 50 ou 100 entrées, selon les préférences de l'utilisateur. Les informations suivantes y figurent :

- Contenu de la balise <TITLE> (environ 100 caractères maximum) ou, à défaut, la mention *Pas de titre*.
- En guise de résumé : contenu de la balise <DESCRIPTION> (environ 400 caractères maximum) ou, à défaut, les premières lignes du texte. L'affichage du résumé demeure optionnel.
- Taille du document.
- Date de la dernière modification du fichier.
- URL.

Puisque VOILA combine moteur et annuaire, les sites indexés manuellement apparaissent (s'il y en a) en tête de liste pour améliorer la pertinence des réponses. L'option *Réponses récentes* permet d'obtenir un classement des résultats par ordre chronologique décroissant. VOILA offre également la possibilité de visualiser l'emplacement du mot recherché pour chaque résultat proposé. L'affichage, par ailleurs, est limité à une seule page par site (technique du *clustering*).

11.5 Portrait : COPERNIC

http://www.copernic.com/fr/ (pour téléchargement ou achat)

Sous la responsabilité de la société Agents Technologies Corp.

Le métamoteur COPERNIC se décline en une version personnelle ou standard gratuite (COPERNIC 98) et en une version professionnelle qui se détaille 29,95$ US (COPERNIC 98PLUS). COPERNIC 98PLUS est également disponible en version d'essai de 30 jours.

La version gratuite interroge près de 30 sources d'information généralistes (les plus importants moteurs du Web) réparties en quatre domaines de recherche : Web, Web francophone, groupes de discussion et adresses de courriel. La version professionnelle incorpore, en outre, près de 125 moteurs et répertoires spécialisés, pour un total de plus de 140 sources d'information distribuées en 20 domaines : Web, Web francophone, groupes de discussion, adresses de courriel, emplois, livres, nouvelles, archives de presse, société, affaires, finances, connaissances, jeux, enfants, cinéma, musique, logiciels, sports, technologies, voyages. Il est également possible, dans le cas de la version professionnelle, de configurer la liste des sources consultées via la désactivation ou l'ajout d'un nombre illimité de moteurs.

Les résultats retournés par les moteurs sont emmagasinés sur le disque dur de l'utilisateur pour une meilleure gestion et une consultation plus rapide. COPERNIC nécessite au minimum un processeur 486-33 MGz, 8 Mo de mémoire vive et 5 Mo d'espace disque. Il fonctionne sous Windows 95, 98, NT 4.0 ou plus récent. Il requiert également MICROSOFT INTERNET EXPLORER 3.0 ou plus récent (fureteur avec lequel il présente des liens privilégiés, notamment à partir de la version 4.0) ou NETSCAPE NAVIGATOR 3.0 ou plus récent.

Principales caractéristiques de fonctionnement

- Interroge de manière simultanée jusqu'à 32 sources.
- Permet l'emploi d'opérateurs booléens lors de la formulation de la requête (ET, OU, locutions).
- Permet quatre types de requête : rapide, normal, détaillé, personnalisé. Ces types se distinguent entre eux par le nombre maximum de résultats permis par moteur et par requête (fixé par défaut respectivement à 300 et à 1000).
- Détruit les doublons et les liens désuets.
- Affiche, pour chaque document rapporté :
  - un score de pertinence
  - le titre
  - une description
  - l'URL
  - la date
  - l'outil de repérage initial
  - l'état (accessible, inaccessible, nouveau, ignoré, visité, téléchargé, raffiné).

- Les résultats peuvent être :
  - ♦ consultés hors ligne
  - ♦ raffinés à l'aide d'opérateurs booléens (ET, OU, SAUF, locutions)
  - ♦ triés selon diverses clés (titre, date, pertinence, adresse)
  - ♦ sauvegardés en différents formats (texte, HTML, XML, dBASE, etc.).

- Possibilité de réactiver une requête pour en mettre à jour les résultats (les nouveaux documents repérés sont alors mis en évidence).

- Peut être utilisé pour « aspirer » des sites afin de permettre une consultation hors ligne (là aussi, jusqu'à 32 sites en même temps).

- Historique détaillé des recherches.

- Mise à jour hebdomadaire automatique du logiciel et des différents moteurs.