

Typologie et mode de fonctionnement des outils de recherche d'information sur internet en biologie/médecine

Christophe Boudry

> L'objectif de cet article est d'exposer le mode de fonctionnement des outils de recherche rencontrés sur internet (moteurs, méta-moteurs et annuaires) et de proposer des pistes pour mieux maîtriser leur usage. Les difficultés rencontrées lors de l'utilisation de ces outils de recherche d'information y sont exposées et l'intérêt potentiel des outils spécifiquement dédiés à la recherche d'information en biologie/médecine discuté. <



Unité Régionale de Formation à l'Information Scientifique et Technique de Paris / École nationale des Chartes, 17, rue des Bernardins, 75005 Paris, France.
e-mail: boudry@ccr.jussieu.fr

> Les outils nécessaires au repérage des informations pertinentes dans le domaine de la biologie ainsi que leur mode de fonctionnement sont souvent trop mal connus des utilisateurs du réseau internet. L'objet de cet article est de détailler ce fonctionnement, afin de permettre aux biologistes de mieux les utiliser et ainsi d'augmenter leur potentialité à localiser des informations pertinentes sur internet.

De façon assez schématique et en se fondant sur leur mode de fonctionnement, il est possible de distinguer trois catégories d'outils de recherche d'informations sur internet : les moteurs ou robots de recherche, les méta-moteurs et les annuaires (Tableau 1).

Moteurs de recherche (ou robots)

Il existe à l'heure actuelle plusieurs milliers de moteurs de recherche sur internet [1]. La principale caractéristique de ces outils réside dans leur mode de fonctionnement totalement automatisé. Ils sont principalement constitués de trois parties: le robot (appelé aussi *spider* ou *crawler*), l'index ou base de données et le logiciel/interface d'interrogation. L'objectif du robot est

de repérer les pages web par suivi récursif des liens présents dans les pages (Figure 1). Il assure ainsi la lecture des données des pages HTML (voir glossaire) et le repérage des liens vers d'autres pages afin de constituer l'index. L'index est en effet le lieu de stockage et d'indexation des pages web visitées et repérées par le robot. L'entrée de pages web dans l'index est également rendue possible par la visite du robot dans des pages, soumises volontairement par leurs créateurs, au moteur de recherche. Ce référencement peut être gratuit ou payant avec, dans le second cas, l'assurance que la page web apparaisse dans les premières réponses à la suite de la saisie d'un mot clé donné et/ou que sa fréquence de mise à jour dans l'index soit élevée. En tout état de cause, l'économie marchande semble être amenée dans l'avenir à prendre une place de plus en plus importante dans ces processus de référencement [2]. L'indexation du code source HTML des pages web est réalisée de façon totalement automatisée (Figure 1), de manière non descriptive, quasiment en texte intégral (seules sont exclues certaines parties comme par exemple les commentaires figurant dans le code source HTML). Certaines données, comme l'adresse, le titre de la page, ou les « méta-données » (description

et mots clés associés par leurs auteurs aux pages web) sont parfois indexées dans des champs spécifiques. Afin d'actualiser le contenu de ces pages web dans l'index, le robot se rend à intervalle défini sur les pages web déjà indexées, pour mettre à jour leur contenu dans l'index. De par son caractère automatisé, l'indexation réalisée est très imparfaite et présente de nombreuses limites. Il n'y a en effet pas d'indexation du contenu associé aux pages web (images et autres fichiers), ni des pages orphelines (voir *glossaire*) ou en accès réservé ayant un contenu dynamique provenant d'une base de données (à quelques exceptions près). Il en résulte que la quantité de données accessibles via les moteurs de recherche est faible comparée à la totalité des données présentes sur internet (d'où la création de l'expression «web invisible», voir *glossaire*). Si les moteurs généralistes rivalisent actuellement pour inclure le plus grand nombre de pages possibles dans leur index (plus de 1 milliard pour *google* selon [3]), certains moteurs, pour proposer un contenu plus spécifique, conditionnent l'entrée des pages web dans leur index, à leur appartenance à un champ d'activité donnée (c'est le cas de *www.scirus.com* et *www.search4engine.com* pour les sciences, de *www.bioview.com* pour la biologie).

Le logiciel et l'interface de recherche permettent à l'utili-

sateur de saisir sa requête en utilisant un ou plusieurs termes, qu'il pense être représentatifs de son besoin d'information, tout en respectant une certaine syntaxe (voir *glossaire*). Les termes de la requête sont alors recherchés dans l'index et une liste de pages web est proposée à l'utilisateur, selon un ordre de pertinence donné, pages sur lesquelles il peut ensuite se rendre via un lien hypertexte (voir *glossaire*) (Figure 1). En ce qui concerne la requête, certains moteurs, tels que *www.search4science.com* proposent aux utilisateurs de restreindre ou d'étendre leurs recherches, soit en saisissant directement des termes complémentaires à ceux de la requête, soit en choisissant des termes qui se rapportent à la requête et sont suggérés par l'interface. D'autres moteurs offrent l'opportunité de restreindre la recherche à des champs spécifiques. Par exemple, sur *www.altavista.com*, l'utilisateur peut restreindre sa recherche aux pages dont l'adresse contient un terme précis (la saisie de «host:univ» restreindra la recherche aux pages dont l'adresse contient le terme «univ» et qui par conséquent, sont des pages web universitaires) ou aux titres des pages web (par la saisie de «title:terme_recherché»). Un des

problèmes importants des moteurs de recherche (et aussi des annuaires) est l'extrême variabilité de leurs syntaxes d'interrogation (Tableau II), qui oblige les usagers à un effort d'adaptation important dès que l'utilisation d'un nouveau moteur s'avère nécessaire [4]. Les résultats de la requête sont classés selon un ordre de pertinence. Celui-ci peut reposer par exemple sur le calcul de la fréquence d'apparition des termes de la requête dans une page web et/ou sur la popularité de la page appréciée par le nombre de pages ayant un lien vers elle. L'ordre de pertinence des réponses pourrait bien, à l'avenir, être de plus en plus influencé par la présentation systématique, dans les premiers résultats,

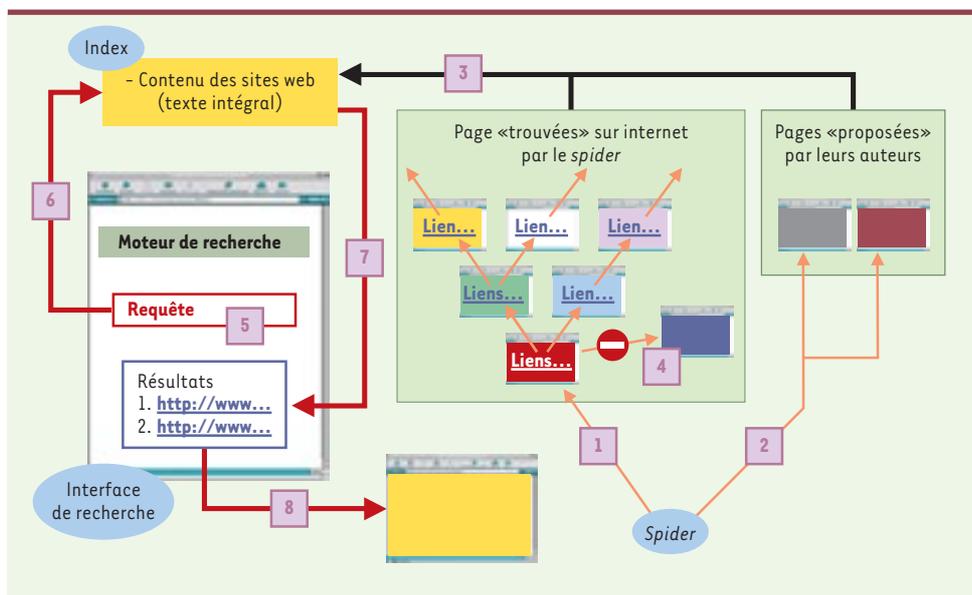


Figure 1. Mode de fonctionnement des moteurs de recherche. Le moteur de recherche contient trois éléments, le robot ou *spider*, l'index et l'interface de recherche. Le robot repère les pages web par suivi récursif des liens présents dans les pages présentes sur internet (1) ou proposées par les auteurs (2). Les pages repérées sont stockées dans l'index (3) et indexées de façon automatique en texte intégral. Les pages qu'il ne peut pas repérer (pages orphelines ou d'accès réservé...) ne sont pas indexées (4). L'interface de recherche permet à l'utilisateur de faire une requête (5) dont les termes seront recherchés dans l'index (6). Les résultats sont donnés selon un ordre de pertinence dépendant du type de moteur de recherche utilisé (7) et l'utilisateur peut se rendre sur une page donnée grâce à un lien hypertexte (8).

Figure 2. Mode de fonctionnement des méta-moteurs. Les méta-moteurs permettent une interrogation simultanée de plusieurs index de moteurs de recherche différents. La requête saisie à l'aide d'une interface unique (1) est transmise à ces différents index (2). Les résultats sont ensuite fournis (3) après élimination des doublons et l'utilisateur peut ensuite se rendre sur les pages grâce à un lien hypertexte (4).

de pages provenant d'un référencement payant. Certains moteurs de recherche proposent également un regroupement des réponses par répertoires thématiques (www.northernligh.com) ou par mots clés qui se rapportent aux termes de la requête (www.exalead.com) ou parfois une représentation sous forme de carte graphique (www.kartoo.com).

Méta-moteurs

Le principe des méta-moteurs est de permettre l'interrogation simultanée de plusieurs index de moteurs de recherche différents. La saisie de la requête (Figure 2) s'effectue à travers une interface unique qui peut être accessible *via* un site web (www.metacrawler.com) ou *via* un logiciel qu'il est nécessaire d'installer sur un poste client (c'est le cas de Copernic qu'il est possible de télécharger à l'adresse www.copernic.com). La requête est alors soumise aux différents moteurs de recherche interrogés (Figure 2) dont le nombre varie de quelques-uns à plusieurs dizaines selon les méta-moteurs considérés. Les réponses provenant de ces différents moteurs subissent le plus souvent une élimination des doublons et sont présentées par ordre de pertinence à l'utilisateur, qui peut se rendre sur chaque page proposée, *via* un lien hypertexte (Figure 2). Le principal intérêt de ces outils est d'augmenter la taille de l'index interrogé. Cependant, même si la plupart des méta-moteurs assurent une traduction de la requête pour l'adapter à la syntaxe de chacun des moteurs interrogés, l'utilisation de requêtes complexes entraîne le plus souvent des réponses très bruitées.

Annuaire

L'objectif principal de ce type d'outils n'est pas de chercher à indexer un maximum de pages web, mais de privilégier la qualité des pages présentes dans leur index. L'entrée dans l'index est en effet supervisée

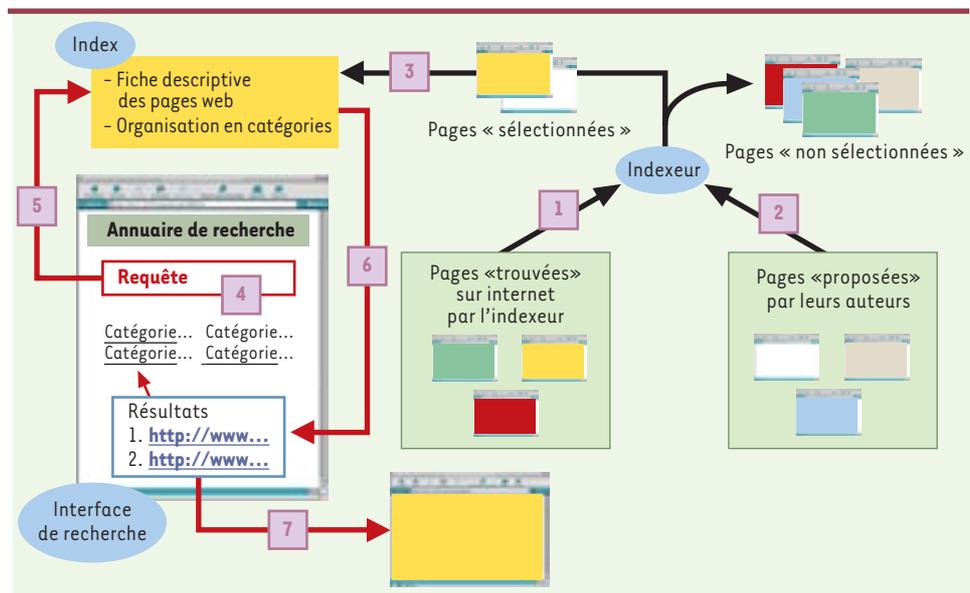
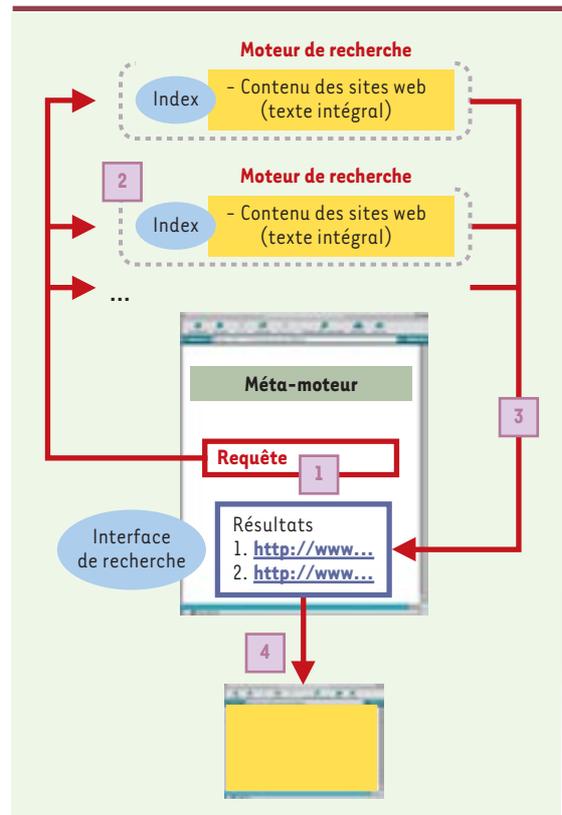


Figure 3. Mode de fonctionnement des annuaires. La différence majeure entre les annuaires et les moteurs de recherche ou les méta-moteurs est l'intervention humaine et la sélection des pages. En effet, un indexeur examine les pages soit sur internet (1), soit proposées par les auteurs (2). Seules les pages sélectionnées par l'indexeur (3) seront stockées dans l'index, accompagnées d'une fiche descriptive et classées en fonction de leur contenu dans des catégories définies. En ce qui concerne la requête (4), l'utilisateur peut soit saisir un ou plusieurs termes qui sont recherchés dans les fiches descriptives de l'index (5) et consulter la liste des sites correspondant à sa requête (6), soit naviguer lui-même dans les catégories grâce à des liens hypertextes. Dans les deux cas, l'accès au contenu des pages web s'effectue *via* un lien hypertexte (7).

par des «indexeurs» qui sélectionnent des pages web à partir de pages trouvées sur internet ou soumises par leurs créateurs (Figure 3), selon des critères tels que la qualité et l'intérêt de leur contenu et/ou leur appartenance à une discipline ou à une communauté donnée. La sélection de certaines pages web, notamment des pages commerciales, est parfois payante. L'indexeur crée alors une fiche décrivant le contenu de chaque page sélectionnée, lui affecte une catégorie d'appartenance et la stocke dans l'index. Les pages web sont ensuite classées en fonction de leur contenu dans des catégories et des sous-catégories pré-définies. L'archétype de ce type de classification est la classification Dewey, principalement utilisée dans les bibliothèques, mais également par certains annuaires de recherche sur internet (publ.ac.uk/link). Le nombre de pages web indexées varie d'un annuaire à l'autre, de quelques milliers pour les annuaires spécifiques à plusieurs millions pour certains annuaires généralistes. Pour effectuer sa recherche, l'utilisateur a deux possibilités (Figure 3). La première est de saisir une requête en utilisant un ou plusieurs termes qui sont alors recherchés dans les fiches descriptives présentes dans l'index, la liste des fiches descriptives correspondant à sa requête lui est alors proposée. La seconde est de se

déplacer en «furetant» dans les catégories (*browsing* pour les Anglo-Saxons) *via* des liens hypertextes à la recherche de pages pouvant répondre à son questionnement. Dans les deux cas, l'accès au contenu des pages web est obtenu *via* un lien hypertexte (Figure 3).

Problématique de la recherche d'informations en biologie/médecine

La qualité d'une recherche d'information peut s'exprimer en termes de silence et de bruit. Le bruit correspond aux résultats non pertinents trouvés par l'utilisateur tandis que le silence correspond aux résultats pertinents non trouvés. Les utilisateurs des différents outils de recherche disponibles sur internet se trouvent confrontés au quotidien à ces deux paramètres. Qui ne s'est jamais retrouvé face à un nombre faramineux de réponses (évoquant un bruit important) ou un trop faible nombre de réponses à la suite de sa requête (signe d'un silence important) ? Cette surabondance ou l'absence de réponses dépendent principalement de deux facteurs.

- La qualité de la requête formulée par l'utilisateur. En effet, l'utilisation d'un terme non approprié dans une requête peut aboutir à la présence de bruit si le terme utilisé est trop commun, ou à du silence si le concept

	Moteurs	Méta-moteurs	Annuaire
Généralistes	www.google.com www.altavista.com www.northernlight.com www.exalead.com www.kartoo.com ...	www.copernic.com www.ixquick.com www.metacrawler.com www.dogpile.com	www.yahoo.com www.dmoz.org www-english.lycos.com www.about.com vlib.org sunsite.berkeley.edu/InternetIndex/
Spécifiques du domaine scientifique	www.scirus.com www.search4science.com		infomine.ucr.edu/Main.html www.scicentral.com www.sciseek.com
Spécifiques du domaine de la biologie	www.bioview.com		www.pasteur.fr/recherche/BNB/bnb-fr.html publ.ac.uk/link bioresearch.ac.uk www.biofinder.org/index_e.html www.library.ucsf.edu/biosites/biology.miningco.com/education/biology www.sciencekomm.at www.123genomics.com/

Tableau 1. Principaux outils de recherche disponibles sur internet pour la recherche d'information en biologie. Les outils sont dits généralistes s'ils couvrent tous les domaines de connaissance, spécifiques s'ils couvrent uniquement le champ scientifique ou une discipline donnée. Des panoramas plus complets des outils généralistes et spécifiques au domaine «scientifique» et au domaine de la biologie peuvent être consultés respectivement aux adresses suivantes : (www.abondance.com) et (www.ccr.jussieu.fr/urfist/biolo/bioguide2/frame.htm, http://www.bioexplorer.net/Search_Engines).

recherché peut être décrit par plusieurs synonymes et qu'est utilisé dans la requête un synonyme peu usité. Les opérateurs booléens (*voir glossaire*) permettent, sous certaines conditions, de lutter efficacement à la fois contre le bruit (opérateurs ET et SAUF) et le silence (opérateur OU). L'utilisation de vocabulaire contrôlé de type Thésaurus (*voir glossaire*), hélas très rarement disponible sur les outils de recherche sur internet, contri-

bue également à augmenter la qualité des requêtes.

• La qualité de l'index interrogé (nombre de documents, qualité intrinsèque et spécificité des documents présents dans cet index). Si l'on suppose que la requête est parfaitement formulée, le nombre de réponses dépend du nombre de documents présents dans l'index de l'outil de recherche interrogé, et donc directement de sa nature (moteur, méta-moteur ou annuaire). Ainsi, dans le cas des moteurs

	<i>Bioview</i>	<i>Scirus</i>	<i>Search4 science</i>	<i>Altavista</i>	<i>Google</i>	<i>Copernic</i>	<i>Infomine</i>	<i>Open Directory Project</i>
Opérateur par défaut ⁽¹⁾	ET	ET	ET	OU	ET	ET	ET	ET
Opérateur ET ⁽²⁾	NON	OUI (choisir «All the words»)	NON	OUI (AND)	NON	OUI (choisir «Recherche de tous les mots» ou saisir AND)	OUI (AND)	OUI (AND)
Opérateur OU ⁽³⁾	NON	OUI (choisir «Any of the words»)	NON	OUI (OR)	OUI (OR)	OUI (choisir «Recherche de n'importe quel mot» ou saisir OR)	OUI (NOT)	OUI (OR)
Opérateur SAUF ⁽⁴⁾	NON	OUI (choisir AND NOT)	NON	OUI (AND NOT)	NON	OUI (NOT)	NON	OUI (AND NOT)
Opérateur arithmétique + ⁽⁵⁾	NON	NON	OUI (+)	OUI (+)	OUI (+)	NON	NON	OUI(+)
Opérateur arithmétique - ⁽⁶⁾	NON	NON	NON	OUI (-)	OUI (-)	NON	NON	OUI(-)
Recherche de phrase ⁽⁷⁾	NON	OUI (choisir «Exact phrase»)	NON	OUI (« »)	OUI (« »)	OUI (choisir «Recherche de l'expression exacte»)	OUI (« »)	OUI (« »)
Troncature ⁽⁸⁾	NON	NON	NON	OUI (*)	NON	OUI (*)	OUI	OUI (*)

Tableau II. Principaux opérateurs de recherche et syntaxe d'interrogation (entre parenthèses). Sont représentés les moteurs de recherche Bioview, Scirus, Search4science, Altavista et Google, le méta-moteur Copernic et les annuaires Open Directory Project et Infomine.

(1) L'opérateur par défaut est l'opérateur qui est utilisé «par défaut» lorsque l'utilisateur saisit deux termes à la suite sans aucune autre précision (ex: terme1 terme2). Les opérateurs booléens ET, OU et SAUF permettent de créer des requêtes complexes. (2) La saisie de «terme1 ET terme2» permet de localiser des pages web où apparaissent simultanément les deux termes recherchés. (3) La saisie de «terme1 OU terme2» permet de localiser des pages web où apparaissent au moins l'un des deux termes recherchés. (4) La saisie de «terme1 SAUF terme2» permet de localiser des pages web où apparaît le mot «terme1» sans qu'apparaisse le mot «terme2». (5, 6) Les opérateurs arithmétiques (*voir glossaire*) + et - permettent de localiser des pages web où apparaît un terme (syntaxe : +terme) ou d'exclure les pages web où apparaît un terme (syntaxe : -terme). Si ces opérateurs arithmétiques permettent d'obtenir des résultats équivalents à l'utilisation de l'opérateur booléen ET («terme1 ET terme2» équivaut à «+terme1 + terme2») ou de l'opérateur booléen SAUF («terme1 SAUF terme2» équivaut à «+terme1 - terme2»), il n'est en aucune façon possible d'obtenir l'équivalent de l'opérateur booléen OU. (7) La recherche de phrases permet de localiser des pages web contenant une expression donnée. La syntaxe la plus courante consiste à inclure l'expression entre guillemets. (8) La troncature, dont la syntaxe la plus courante est «*» permet la recherche de variants orthographiques. Par exemple, la saisie de apoptos* permet de localiser des pages où figure *apoptosis* ou apoptose. Pour connaître la syntaxe exacte d'autres outils, il est possible de consulter www.abondance.com ou www.calvin.edu/library/searreso/internet/searengi.stm pour les outils de recherche généralistes, www.ccr.jussieu.fr/urfist/biolo/frame.htm pour les outils de recherche dans les domaines scientifiques et biologiques.

et des méta-moteurs généralistes, le nombre de réponses à une requête peu spécialisée est généralement extrêmement élevé (parfois jusqu'à plusieurs millions !), avec habituellement un bruit important sur l'ensemble des réponses. En revanche, l'avantage de ces outils est de quasiment toujours fournir des réponses, même si le sujet de la requête s'avère extrêmement spécialisé. Dans le cas des annuaires généralistes, le nombre de réponses est toujours plus faible par rapport aux moteurs et aux méta-moteurs du même type. Si cela s'avère être un avantage indéniable dans le cas de requêtes peu spécialisées (car le nombre de réponses proposé est non seulement plus faible, mais aussi limité aux seuls sites répondant à certains critères de qualité pour lesquels ils ont été retenus), cela s'avère être un inconvénient pour des requêtes spécialisées qui produisent souvent très peu ou pas de réponses.

Les outils de recherche spécifiques au domaine scientifique ou au domaine plus restreint de la biologie/médecine ont été développés, avec pour objectif de pallier les principaux inconvénients des moteurs et des annuaires généralistes décrits plus haut. Pour les moteurs spécifiques, la sélection par le robot de sites appartenant uniquement à un domaine particulier permet en théorie de limiter grandement le bruit. Pour les annuaires spécifiques, la restriction au domaine scientifique ou à celui de la biologie/médecine permet de sélectionner dans l'index des sites plus spécifiques, minimisant ainsi le silence lié aux requêtes très spécialisées dans ces domaines. La difficulté pour ces outils spécifiques est de trouver un équilibre dans la sélection des sites présents dans leurs index afin de satisfaire le plus grand nombre d'utilisateurs d'un domaine donné. Pour améliorer son efficacité, l'utilisateur doit toujours adapter sa stratégie de recherche en choisissant, à chaque fois, le type d'outil adéquat (moteur, méta-moteur, annuaire) ainsi que son degré de spécialisation (généraliste, scientifique, biologie/médecine). C'est en effet l'unique possibilité de limiter au maximum le bruit et le silence, et par conséquent d'améliorer sensiblement la qualité de ses recherches. Cela nécessite bien entendu de connaître le mode de fonctionnement de chacun de ces outils et d'avoir toujours à sa disposition un panel d'outils de recherche d'informations généralistes et spécifiques valides.

Un secteur en pleine évolution...

Si les différences persistent entre moteurs et annuaires quant à leur mode de fonctionnement, ces deux types d'outils peuvent désormais être interrogés fréquemment à partir de la même interface. Par exemple, l'annuaire Yahoo! (www.yahoo.com) utilise les résultats du moteur de

recherche *Google* (www.google.com) en cas de recherche infructueuse dans son index et, à l'inverse, *Google* donne la possibilité à ses utilisateurs d'effectuer une recherche dans l'annuaire *Open Directory Project* (www.dmoz.org).

Si le monde des annuaires et des méta-moteurs semble maintenant relativement stable du point de vue des technologies déployées, celui des moteurs est en revanche encore en pleine ébullition. Il ne se passe en effet pas un mois sans que de nouveaux moteurs de recherche, proposant des technologies et des modes de fonctionnement innovants, apparaissent. Souhaitons simplement que ces évolutions aillent dans le sens d'une plus grande facilité d'utilisation pour les usagers, de l'augmentation des performances de ces outils afin de localiser des informations toujours plus pertinentes dans le domaine de la biologie sans que l'économie marchande ne vienne (trop) fausser la donne. ♦

REMERCIEMENTS

L'auteur remercie L. Salamatian pour sa participation à l'étude comparative des syntaxes d'interrogation des différents outils de recherche présentés et P. Herlin pour ses conseils et la relecture du manuscrit.

SUMMARY

Operating mode of internet search engines : applications in biology and medicine

The aim of this article is to present the operating mode of internet search engines (search engines, meta-search engines, and directories) and to propose advices for better controlling their use. The difficulties encountered during the use of these tools are exposed and the interest of tools specifically dedicated to the search of information in biology/medicine discussed. ♦

RÉFÉRENCES

1. <http://www.searchengine-guide.com/>. Page consultée le 5 octobre 2001.
2. Andrieu O. L'avenir du référencement sur les outils de recherche. *Bases/Netsources* 2001 ; 33 : 10 - 2.
3. <http://www.searchengine-watch.com/reports/sizes.html>. Page consultée le 4 octobre 2001.
4. Dong X, Su L. Search engines on the world wide web and information retrieval on the internet: Fa review and evaluation. *Online and CD ROM Review* 1997 ; 21 : 67 - 81.

TIRÉS À PART

C. Boudry

GLOSSAIRE

HTML (*hypertext markup langage*)

Langage informatique avec lequel sont conçues les pages web. Il permet de décrire la présentation des pages web et les liens hypertexte qu'elles renferment vers d'autres documents situés sur internet. Le code source d'une page web correspond au langage HTML décrivant cette page.

Lien hypertexte

Élément (mot, phrase ou image) contenu dans une page web à partir duquel il est possible d'accéder à un autre document en cliquant dessus.

Opérateurs arithmétiques

Opérateurs dérivés des mathématiques permettant, dans le cas des outils de recherche d'information, de construire des requêtes d'interrogation complexes. Les opérateurs arithmétiques mis à disposition des utilisateurs par les outils de recherche d'information sont le «+» et le «-».

Opérateurs booléens

Opérateurs émanant des travaux de George Boole permettant, dans le cas des outils de recherche d'information, de construire des requêtes d'interrogation complexes. Les opérateurs booléens mis à disposition des utilisateurs par les outils de recherche d'information sont «ET», «OU» et «SAUF».

Page orpheline

Page web n'ayant aucun lien hypertexte pointant vers elle. Autrement dit, il s'agit d'une page web sur laquelle il n'est pas possible d'accéder par le biais d'un lien hypertexte à partir d'une autre.

Syntaxe d'interrogation (ou de saisie)

Désigne les opérateurs disponibles sur l'interface d'interrogation d'un outil de recherche d'information permettant la construction de requêtes d'interrogation complexes. La disponibilité ainsi que l'orthographe des opérateurs peut varier d'un outil de recherche à l'autre: on parle de variabilité de syntaxe d'interrogation.

Thésaurus (vocabulaire contrôlé)

Ensemble de termes normalisés, utilisés pour interroger un outil de recherche d'information, permettant une description univoque des concepts présents dans les documents indexés.

Web invisible

Pages web dont le contenu n'est pas indexé par les moteurs de recherche. Il s'agit par exemple des pages web orphelines ou des pages en accès réservé.