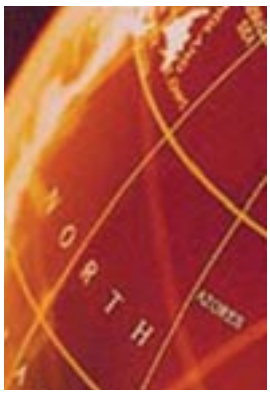
MEDECINE/SCIENCES 2002 ; 18 : 1107-12

Évaluation des performances des outils de recherche d'informations sur internet en biologie

Christophe Boudry

> Le réseau internet révolutionne peu à peu les pratiques des chercheurs tant du point de vue de leurs recherches documentaires que du point de vue de la diffusion de leurs travaux. Si l'opportunité d'accéder à une masse d'informations colossale « d'un simple clic » est une perspective particulièrement séduisante pour tout chercheur en biologie, la découverte d'informations pertinentes dans cet « océan » s'avère en réalité relativement difficile, et cela malgré l'existence d'un nombre croissant d'outils de recherche mis à la disposition des internautes. <



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

(➔) m/s 2002, n°5, p. 616

Le nombre de pages web accessibles à partir des différents outils de recherche d'informations est en constante progression (320 millions de pages disponibles en décembre 1997 [1]; 800 millions en juillet 1999 [2]). Cette profusion d'informations disponibles sur le web est bien entendu bénéfique aux utilisateurs, car l'augmentation de la taille de la base de données interrogée accroît la probabilité de trouver des informations pertinentes sur un sujet donné. Le corollaire à cette situation est que la localisation d'informations pertinentes dans cette masse d'informations s'avère de plus en plus délicate. C'est en quelque sorte l'illustration parfaite de la recherche de l'aiguille dans la botte de foin...

Dans ce contexte, si un nombre important d'outils de recherche d'informations sur internet, spécifiquement dédiés ou non à la biologie, sont à la disposition des internautes du domaine, la qualité de leurs performances et leur facilité d'utilisation s'avèrent primordial. Un certain nombre d'études ont été menées afin d'évaluer la pertinence des outils de recherche destinés au repérage d'informations de type « généraliste » [3-7] ou de type médical sur internet [8, 9]. En revanche, aucun travail de ce type ne semble avoir été mené à l'heure actuelle dans le domaine de la biologie.

Dans un article précédent (➔), nous avions présenté le mode de fonctionnement des outils de recherche sur internet. Ceux-ci peuvent être classés en trois catégories:

- les moteurs de recherche (ou robots) dont l'objectif est d'indexer en « texte intégral » les pages web, sans intervention humaine et sans critères de qualité associés, dans une base de données (on parle également d'index) dans laquelle les usagers peuvent effectuer des recherches par mots clés, *via* une interface spécifique.
- les méta-moteurs, dont le principe est d'interroger simultanément, *via* une interface unique (site web ou logiciel installé sur un poste client), un nombre plus ou moins élevé de moteurs de recherche différents.
- les annuaires (ou répertoires), dont les principales caractéristiques sont leur mode d'organisation hiérarchique et la possibilité de recherche en « furetant » dans différentes catégories. Contrairement aux pages web présentes dans l'index des moteurs de recherche, les pages répertoriées par les annuaires sont sélectionnées sur des critères de qualité et ne sont pas indexées en texte intégral; ce sont des notices des-

criptives créées par des éditeurs qui sont indexées.

En outre, les domaines de connaissances couverts par les moteurs et annuaires de recherche sont variables. On distingue ainsi les outils dits généralistes, qui concernent tous les domaines de connaissance, des outils sélectifs ou spécifiques qui couvrent un champ scientifique ou disciplinaire donné.

L'objectif de cette étude était de tester l'éventail des possibilités de recherche d'informations sur internet offertes aux biologistes, tant du point de vue du type d'outils (moteur, méta-moteur, annuaire) que du point de vue du domaine de connaissance couvert (généraliste ou spécifique).

Outils de recherche étudiés

Le nombre d'outils de recherche d'informations disponibles sur internet étant très élevé, pour des raisons de faisabilité, seul un nombre limité de 8 outils a été inclus dans cette étude, 5 moteurs, 1 méta-moteur et 2 annuaires.

Moteurs de recherche

Altavista (http://www.altavista.com) est, probablement en raison de son ancienneté (1995), un des moteurs généralistes les plus populaires [6] et les plus fréquemment testés [5].

Google (http://www.google.com), apparu assez récemment (1998), est le moteur généraliste qui possède à l'heure actuelle l'index le plus important en terme de nombre de pages indexées (plus de 1,5 milliards) [10].

Bioview (http://www.bioview.com) est un moteur spécifiquement dédié aux biologistes avec un index composé de pages web uniquement sélectionnées dans le domaine de la biologie.

Scirus (http://www.scirus.com) permet d'interroger l'index généraliste de *Fast* (http://www.fast.com) limité aux seules pages web ayant un contenu scientifique. Il permet également d'interroger 4 bases de données d'Elsevier dont l'accès est subordonné à la souscription d'un abonnement. Cette dernière possibilité n'a pas été exploitée pour cette étude car l'information proposée n'est pas en libre accès.

Search4science (http://www.search4science.com) est un moteur spécifiquement dédié à la recherche d'informations en sciences. Il offre la possibilité de rechercher dans deux index différents: (1) celui de *Northern Light* lorsque l'utilisateur opte pour l'option de recherche *Dynamic search*. Cette interface propose pour chaque terme de recherche saisi par l'utilisateur une série de synonymes, permettant d'élargir ou de restreindre la recherche en cours. Il présente la spécificité de regrouper les résultats dans des répertoires pour faciliter leur exploitation; (2) celui de *Google* lorsque l'utilisateur opte pour l'option *Direct search*. Cette dernière option étant très proche d'une interrogation directe de *Google via* sa propre interface, seule l'option *Dynamic search* a été évaluée.

Méta-moteur

Copernic (http://www.copernic.com) est un méta-moteur « client » au sens où son utilisation nécessite l'installation préalable d'un logiciel sur l'ordinateur de l'utilisateur. Deux versions sont disponibles: une version payante qui offre des possibilités avancées de recherche par domaine ou type d'informations recherchés et est plus spécialement dédiée au public des professionnels de l'information, et une version gratuite permettant d'interroger 10 moteurs de recherche généralistes simultanément. Une grande majorité d'utilisateurs en biologie n'ayant pas accès à la version payante, seule la version gratuite a été testée dans cette étude.

Annuaires

Infomine (http://infomine.ucr.edu/Main.html) est un annuaire « spécifique », développé par un réseau de bibliothèques universitaires californiennes. Il répertorie environ 20000 sites qui font autorité en sciences pour un public d'universitaires et de chercheurs.

Open Directory Project (ODP) (http://www.dmoz.org) est un annuaire généraliste, dont la sélection et l'indexation des pages web sont réalisées par des éditeurs volontaires, spécialistes chacun dans leur domaine d'activité. Il s'agit d'un des annuaires généralistes dont la taille d'index est la plus importante.

Mots	Phrases	Expressions contenant des opérateurs booléens
apoptosis	*cell morphometry*	*single cell gel electrophoresis* OU *comet assay*
fibronectin	*G-protein coupled*	*circadian rythm* ET *melatonin*
cytotoxicity	*receptor*	*DNA ploidy* ET image SAUF *flow cytometry*
immunostaining		
ghrelin		

Tableau I. ***Requêtes testées dans les différents outils de recherche.*** Les requêtes testées appartiennent toutes au domaine de la biologie et ont été élaborées afin de couvrir différents niveaux de complexité et de possibilités syntaxiques de recherche: mots, phrases, expressions contenant les opérateurs booléens ET, OU, SAUF.

Méthodologie d'évaluation des performances

L'objectif de cette étude était de tester les performances des 8 outils retenus, en utilisant des requêtes du domaine disciplinaire de la biologie, formulées sous forme d'équations de recherche *via* l'interface de recherche de chacun de ces outils. Ce type d'étude nécessite de définir le nombre et le type de requêtes utilisées ainsi que les paramètres spécifiques permettant d'évaluer la qualité de l'information obtenue.

Le nombre de requêtes testées dans des études similaires est très variable (de 3 à 30 [3, 4, 11-13]) et dépend grandement du nombre d'outils de recherche testé. Vu le nombre important d'outils testés dans la présente étude, pour des raisons de faisabilité, seules 10 requêtes différentes ont été utilisées, ce qui représente un nombre théorique de pages web dont le contenu à expertiser est égal à 1400.

Comme c'est le cas dans la majorité des études similaires [5], ces requêtes ont été spécialement construites, afin de couvrir différents niveaux de complexité et de possibilités syntaxiques de recherche: mots, phrases, expressions contenant les opérateurs booléens ET, OU, SAUF *(Tableau I)*. En outre, la variabilité de syntaxe des différents outils étudiés impose de traduire chaque requête afin de l'adapter au mieux à la syntaxe de chaque outil testé.

Un autre aspect important est d'expérimenter chaque requête, dans un délai le plus court possible afin de ne pas favoriser ou défavoriser certains outils par rapport à d'autres [14]. En effet, les outils testés en dernier sont théoriquement favorisés car ayant plus de chance d'avoir de nouvelles pages web indexées dans leur base de données.

Enfin, le nombre généralement très élevé de réponses proposées par les outils de recherche contraint à limiter le nombre de réponses considérées lors d'études de ce type [4, 9, 11, 13]. Ainsi, seules les 20 premières réponses ont été prises en compte. Ce choix a tenu compte du fait qu'un utilisateur qui ne trouve pas de réponses satisfaisantes dans les 20 premières réponses, a instinctivement tendance à reformuler sa requête plutôt que de consulter la totalité des réponses proposées.

Les paramètres étudiés ont été les suivants.

- La précision est un paramètre très largement utilisé [5, 14, 15]. Il correspond au nombre de réponses pertinentes proposées par un outil de recherche divisé par le nombre de réponses étudiées (20 dans notre cas). Comme la mise en œuvre de tout paramètre découlant d'une décision humaine, l'appréciation de la pertinence revêt un caractère subjectif incontournable [16]. Afin de minimiser au maximum ce caractère subjectif, ce paramètre a été évalué comme suit [17]: une réponse pertinente a été définie comme une page web présentant la propriété d'être informative par rapport à la requête formulée. Une réponse a été considérée comme non pertinente lorsque la page web proposée par l'outil de recherche n'était pas informative, ou n'était pas accessible *via* le lien hypertexte proposé par l'outil de recherche (page introuvable à l'adresse indiquée, page à accès réservé ou soumis à un abonnement préalable). De plus, la pertinence de chaque page proposée par chaque outil de recherche a été évaluée, non pas à partir de la notice ou du résumé proposé par l'outil de recherche comme pour [4] mais en visitant et consultant le contenu de chaque page proposée.

  La précision moyenne correspond au nombre total de page web pertinentes obtenues pour les 10 requêtes utilisées divisé par le nombre total de réponses étudiées.

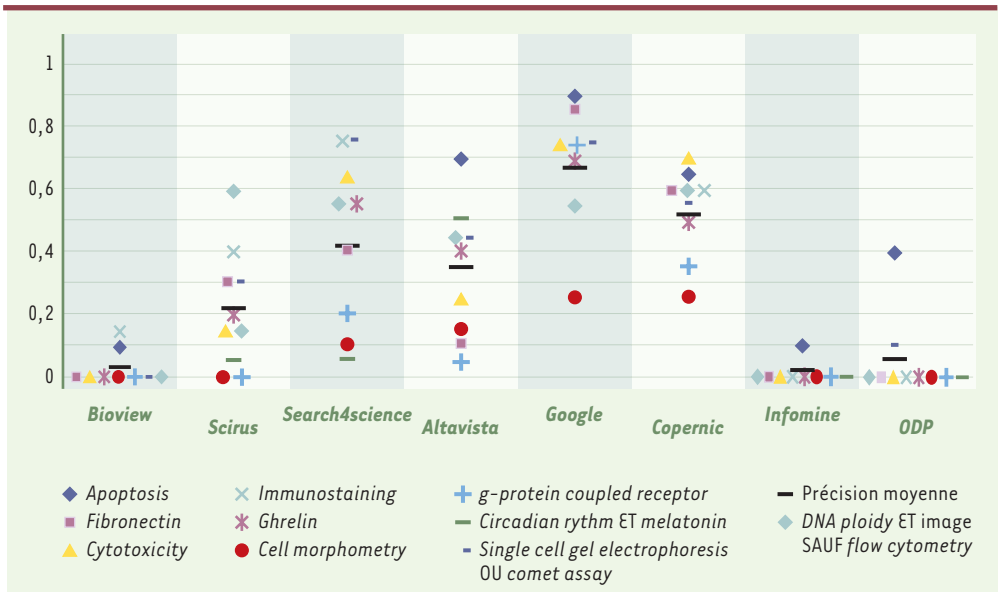


Figure 1. ***Précision des outils de recherche étudiés pour chacune des 10 requêtes testées et précision moyenne.*** La précision correspond à la fraction de réponse pertinentes proposées par un outil de recherche en réponse à une requête donnée divisé par le nombre de réponses étudiées (20 dans notre cas).

- La couverture relative correspond au nombre de page pertinentes trouvées par un outil de recherche, divisé par le nombre total de pages pertinentes trouvées par la totalité des outils de recherche étudiés, pour les 10 requêtes utilisées [13, 15]. Il permet de connaître la proportion de réponses pertinentes proposées par un outil de recherche donné par rapport à celles proposées par tous les autres outils étudiés.
- Le pourcentage de lien en erreur qui n'aboutissent pas à la page liée [2, 5] correspond, pour un outil de recherche donné, au nombre total de liens en erreur obtenus, divisé par le nombre total de réponses proposée par chaque outils de recherche pour les 10 requêtes utilisées. Il fournit une indication sur la fréquence de mise à jour des index des outils de recherche et, indirectement, sur la qualité de ces index (plus le nombre de liens en erreur est élevé, moins l'index est mis à jour fréquemment).

Performances des outils de recherche étudiés

La précision des outils de recherche étudiés pour les différentes requêtes testées est présentée sur la *Figure 1*. Il faut d'abord souligner la dispersion importante, pour un outil donné, des valeurs qui peuvent varier pratiquement du simple au triple en fonction de la requête considérée (0,25 à 0,67 pour *Google* par exemple). Ceci semble indiquer que les performances des outils présentés dépendent grandement de la requête formulée. Seulement deux outils permettent d'obtenir plus de 1 réponse pertinente sur 2 proposées, en réponse à l'ensemble des requêtes (*Google* et *Copernic,* avec une précision moyenne respective de 0,67 et 0,51), tandis que trois d'entre eux fournissent une précision moyenne inférieure à 0,1, ce qui correspond à moins de 1 réponse pertinente sur 10 proposées (*Bioview*, *ODP* et *Infomine*).

Même si les résultats des études visant à évaluer les performances des outils de recherche sont à comparer avec beaucoup de prudence à cause d'un manque de standardisation des méthodologies employées [17,18], les données disponibles concernent principalement *Altavista* et font état de valeurs un peu plus élevées (0,46 pour [13]; 0,48 pour [9] et 0,78 pour [4]) que celle que nous observons pour cette outil (0,34). Cette différence pourrait être due à la moins grande spécificité des requêtes testées dans les autres études.

Concernant la couverture relative des outils étudiés *(Figure 2)*, *Google* fournit à lui seul plus d'un tiers de la totalité des réponses pertinentes recueillies dans cette étude (37,9 %). Trois des 8 outils étudiés fournissent quant à eux une très faible proportion des réponses pertinentes sur l'ensemble des requêtes effectuées (*Bioview*, *Infomine* et *ODP*). On peut également souligner que l'utilisation des 2 outils les plus performants (*Google* et *Copernic*) permet d'obtenir près de 60 % des réponses pertinentes proposées par l'ensemble des 8 outils étudiés (214 réponses pertinentes fournies par ces 2 outils sur 356 fournies par l'ensemble des 8 outils testés).

Enfin, si la majorité des outils étudiés présente des pourcentages de liens en erreur voisins de ceux classiquement rencontrés dans les autres études [1, 2, 9], ce qui reflète une mise à jour assez régulière de leur index, deux outils (*Bioview* et *Infomine*) semblent avoir un index dont la fréquence de mise à jour n'est pas satisfaisante, à la lumière des requêtes testées *(Tableau II)*.

Conclusions

Les résultats présentés dans cette étude illustrent la difficulté à localiser des informations pertinentes dans le domaine de la biologie (précision et couverture relative faibles de certains outils, nombre de liens en erreurs parfois relativement important). Ils doivent être cependant interprétés avec certaines réserves: leur période de validité est en effet limitée dans le

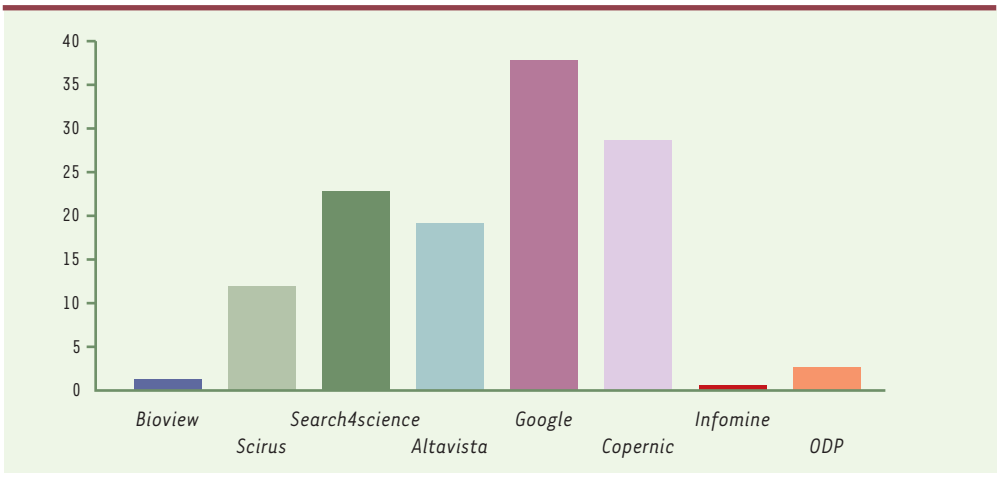


Figure 2. ***Couverture relative moyenne des huit outils de recherche étudiés.*** Ce paramètre correspond à la proportion de réponses pertinentes fournies par chaque outil de recherche pour l'ensemble des requêtes étudiées.

temps à cause du caractère « mouvant » des outils testés et ils ne sont valables que pour le type et le domaine de connaissance des requêtes testées.

La difficulté à repérer des informations intéressantes sur internet est principalement due à la structure même du web, qui provoque une dilution de l'information pertinente (les pages web au contenu « biologique ») dans un océan de pages aux contenus divers et variés.

Afin de tenter de résoudre ce problème, les concepteurs d'outils de recherche ont élaboré trois stratégies.

- La première consiste à sélectionner de façon interactive les pages web présentes dans l'index interrogé, en se basant sur la qualité de leur contenu. La conséquence principale est la taille généralement faible de l'index ainsi créé (cas des annuaires *ODP* et *Infomine*). Il semble qu'une sélection de ce type, très draconienne, ne retienne que des pages relativement généralistes et aboutisse à la formation d'index peu ou mal adaptés à la recherche d'informations spécifiques.
- Une autre stratégie consiste à construire automatiquement un index contenant le plus grand nombre possible de pages web, sans critères de tri préalable. L'élimination des pages « indésirables » est réalisée dans un deuxième temps, après la saisie de la requête de l'utilisateur, et grâce à des algorithmes de tri des résultats permettant leur présentation par ordre de pertinence. Cette stratégie, utilisée par les outils généralistes comme *Altavista*, *Google* et *Copernic*, semble la plus efficace : dans cette étude, 2 de ces 3 outils de recherche offrent globalement les meilleures performances. Les bons résultats de *Google* sont probablement dus à la grande taille de la base de données interrogée, couplée à un algorithme efficace de tri des résultats basé notamment sur un calcul de la « popularité » des pages web qui prend en compte le nombre de pages ayant un lien vers chaque page.
- Une troisième stratégie, hybride, consiste en une sélection automatisée des pages web, basée sur leur appartenance à un champ disciplinaire donné, sans qu'aucun critère de qualité n'intervienne, couplée a un tri des résultats par ordre de pertinence. Cette stratégie est utilisée par *Bioview*, *Scirus* et *Search4science*. Dans le cas de *Bioview*, les critères de sélection des pages web présentes dans l'index s'avèrent inadaptés au type de requêtes testées dans notre étude. En effet, l'index, dont la fréquence de mise à jour est insuffisante, possède une proportion de pages à caractère commercial très importante, ce qui nuit visiblement aux performances de cet outil. La sélection des pages web présentes dans un index déjà existant, réalisée par *Scirus* et *Search4science*, combinée à l'utilisation de technologies de tri des résultats ayant fait leurs preuves dans le cadre d'outils de recherche généralistes, semble beaucoup plus adaptée et prometteuse.

En tout état de cause, malgré l'apparition récente d'outils plus ou moins spécifiques dont l'objectif est de faciliter la localisation d'informations sur internet en sciences et/ou en biologie, il semble que les performances obtenues soient très éloignées des espoirs suscités [19] et qu'il faille encore privilégier les outils de type généralistes pour obtenir les meilleures performances.

Une augmentation significative des performances de tous ces outils sera possible uniquement si leurs concepteurs ont pour objectifs de généraliser l'utilisation de vocabulaires contrôlés associés, de prendre en compte le caractère polysémique de certains termes ou bien encore d'augmenter la fraction des pages indexées par rapport à la totalité des pages web existantes. La généralisation de l'utilisation de méta-données associées aux pages web par leurs créateurs [2, 20] (données fournissant une description du contenu des pages web ayant pour objectif de faciliter en particulier leur indexation), semble également totalement indispensable pour espérer obtenir une augmentation significative de ces performances à plus ou moins brève échéance. ◊

REMERCIEMENTS

L'auteur remercie P. Herlin pour ses conseils et pour la relecture du manuscrit.

	Bioview	Scirus	Search4-science	Altavista	Google	Copernic	Infomine	Open Directory Project
Pourcentage de liens en erreur	33	6,8	3,3	8,5	4	9	33,3	0

Tableau II. ***Pourcentage de liens en erreur n'aboutissant pas à la page liée pour l'ensemble des requêtes testées et des réponses proposées par chaque outil testé.***

SUMMARY

Evaluation of web search engine performances in the field of biology

The internet network revolutionizes little by little the practices of the researchers as well from the point of view of their information retrievals as from the point of view of the diffusion of their work. If the opportunity to reach a colossal mass of information very simply is a particularly attractive perspective for every researcher in biology, this study puts in evidence that the discovery of relevant information in the field of biology in this « ocean » turns out relatively difficult, and this in spite of the existence of an increasing number of search tools at the disposal of the Internet users. Furthermore, the results presented suggest that the use of non-specialized search engines and meta-engines seems preferable with that of specific search engines in the field of the biology and with that of non-specialized or specific directories in biology. ◊

RÉFÉRENCES

1. Lawrence S, Giles CL. Searching the world wide Web. *Science* 1998; 280: 98-100.
2. Lawrence S, Giles CL. Accessibility of information on the web. *Nature* 1999; 400: 107-9.
3. Winship IR. World wide web searching tools - an evaluation. *Vine* 1995; 99: 49-54.
4. Chu HT, Rosenthal M. Search engines for the world wide web: a comparative study and evaluation methodology. *Proc ASIS Annu Meet* 1996; 33: 127-35.
5. Dong X, Su L. Search engines on the world wide web and information retrieval on the internet: a review and evaluation. *Online CD ROM Rev* 1997; 21: 67-81.
6. Xie M, Wang H, Goh TN. Quality dimensions of internet search engines. *J Inf Sci* 1998; 24: 365-72.
7. Wang H, Xie M, Goh TN. Service quality of internet search engines. *J Inf Sci* 1999; 25: 499-507.
8. Akaho E, Ahmad SR. A comparative study of internet search engines by applying 'cost effective treatment for myocardial infarction' as a search topic. *Drug Inf J* 1998; 32: 921-32.
9. Wu G, Jie L. Comparing web search engine performance in searching consumer health information: evaluation and recommendations. *Bull Med Libr Ass* 1999; 87: 456-61.
10. http://searchenginewatch.com/reports/sizes.html. Page consultée le 12 septembre 2002.
11. Ding A, Marchionini G. Comparative study of web search service performance. *Proc ASIS Annu Meet* 1996; 33: 136-42.
12. Venditto G. Search engine showdown. *Internet World* 1996; 7: 79-86.
13. Clarke SJ, Willett P. Estimating the recall performance of web search engines. *Aslib Proc* 1997; 49: 184-9.
14. Oppenheim C, Morris A, McKnight C. The evaluation of WWW search engines. *J Doc* 2000; 56: 190-211.
15. Landoni M, Bell S. Information retrieval techniques for evaluating search engines: a critical overview. *Aslib Proc* 2000; 52: 124-9.
16. Chignell MH, Gwizdka J, Bodner RC. Discriminating meta-search: a framework for evaluation. *Inf Process Manag* 1999; 35: 337-62.
17. Green R. Topical relevance relationships. Why topic matching fails. *J Am Soc Inf Sci* 1995; 6: 646-53.
18. Hawking D, Craswell N, Thistlewaite P, Harman D. Results and challenges in web search evaluation. *Comput Networks* 1999; 31: 1321-30.
19. Gardner M. A science-oriented search engine could solve problems. *Nature* 1999; 401: 111.
20. Shon J, Musen MA. The low availability of metadata elements for evaluating the quality of medical information on the world wide web. *Proc AMIA Symp* 1999; 1945-9.

TIRÉS À PART

C. Boudry