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

Christophe Boudry

Synthèse

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

Christophe Boudry

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

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, ceci, malgré l'existence d'un nombre croissant d'outils de recherche mis à la disposition des internautes.

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 recherche d'informations en 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 on été menées dans ce sens afin d'évaluer la pertinence des outils de recherche destinés au repérage d'information de type "généraliste" [3, 4, 5, 6, 7] ou de type médicale sur internet [8, 9]. Aucun travail de ce type ne semble avoir été mené à l'heure actuelle dans le domaine de la biologie. L'objet de cet article est donc de présenter une étude expérimentale, ayant pour objectif d'évaluer les performances de huit outils de recherche d'informations sur internet, spécifiquement dédiés au domaine de la biologie ou non, pour localiser des informations dans le domaine de la biologie.

Outils de recherche étudiés

De façon assez schématique, il est possible de classer les outils de recherche sur internet en 3 catégories:
- les moteurs de recherche (ou robots), dont les archétypes sont Altavista (http://www.altavista.com) et Google (http://www.google.com) et 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 via une interface unique (site web ou logiciel installé sur un poste client), un nombre plus ou moins élevé de moteurs de recherche différents. Citons par exemple Metacrawler (http://www.metacrawler.com) ou Copernic (http://www.copernic.com).
- les annuaires (ou répertoires), dont les principales caractéristiques sont leur mode d'organisation hiérarchique et la possibilité de recherche en "furetant" dans différentes catégories (citons par exemple Yahoo! (http://www.yahoo.com) ou l'Open Directory Project (http://www.dmoz.org)). Au contraire des pages web présentes dans l'index des moteurs de recherche, les pages répertoriés par les annuaires sont sélectionnées sur des critères de qualité et ne sont pas indexées en texte intégral (ce sont des notices descriptives créées par les indexeurs qui sont indexées).

Les domaines de connaissances couverts par les moteurs et annuaires de recherche sont variables. La couverture peut concerner tous les domaines de connaissance (outils de type généraliste) ou bien concerner spécifiquement un champ scientifique ou disciplinaire donné (outils sélectifs ou spécifiques).

Le nombre d'outils de recherche d'informations disponibles sur internet étant très élevé, une sélection a dû être réalisée pour n'en inclure qu'un nombre limité dans cette étude (8) pour des raisons de faisabilité. Ils ont été choisis dans le but 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, meta-moteur, annuaire) que du point de vue du domaine de connaissance couvert (généraliste ou spécifique).

Moteurs de recherche sélectionnés

Bioview (http://www.bioview.com): cet outil est spécifiquement dédiés aux biologistes avec un index composé de pages web uniquement sélectionnés dans le domaine de la biologie.

Scirus (http://www.scirus.com): ce moteur permet d'interroger l'index généraliste de Fast (http://www.fast.com) limité aux seul 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, l'information proposée dans ce cas n'étant pas en libre accès.

Search4science (http://www.search4science.com): ce moteur est spécifiquement dédiée à la recherche d'informations en sciences. Il offre la possibilité de rechercher dans 2 index différents:

- 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, afin d'élargir ou de restreindre la recherche en cours. S'appuyant sur la technologie de Northern Ligth, il présente la spécificité de proposer les résultats regroupés dans des répertoires pour faciliter leur exploitation.
- et celui de Google lorsque l'utilisateur opte pour l'option "Direct search".

L'option testée pour cette étude a été celle nommée "Dynamic search", l'autre option se rapprochant énormément de l'interrogation directe de l'index de Google via sa propre interface, dont les performances ont été testée également dans cette étude.

Altavista (http://www.altavista.com): probablement dû à son ancienneté (1995) cet outil figure parmi les moteurs généralistes les plus populaires [6] et parmi les plus fréquemment testés [5].

Google (http://www.google.com): apparu relativement récemment (1998), ce moteur généraliste possède à l'heure actuelle l'index le plus important en terme de nombre de pages indexées (plus de 1,3 milliards) [10].

Méta-moteur sélectionné

Copernic (http://www.copernic.com): il s'agit d'un méta-moteur "client" au sens ou son utilisation nécessite l'installation préalable d'un logiciel sur l'ordinateur de l'utilisateur. Deux versions de ce méta-moteur sont disponibles: une version payante offrant des possibilités avancées de recherche par domaine ou type d'informations recherchée et qui est plus spécialement dédiée au public des professionnels de l'information, et une version gratuite offrant la possibilité d'interroger 10 moteurs de recherche généralistes simultanément. Partant du constat qu'une grande majorité d'utilisateurs en biologie n'a pas accès à la version payante, c'est la version gratuite qui a été testée dans cette étude.

Annuaires sélectionnés

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

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

Méthodologie d'évaluation des performances

L'objectif de cette étude a été 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 chacun de ces outils. En d'autres termes, cette étude a consisté à évaluer, par le calcul de paramètres spécifiques, la qualité de l'information retrouvée et proposée à l'utilisateur par chaque outils de recherche, en réponse à chaque requête, formulée via l'interface de recherche.

Le nombre de requêtes testées dans la littérature lors d'études similaires varie d'une étude à l'autre et souffre d'un manque d'homogénéité (3 pour [3], 5 pour [11], 10 pour [4], 25 pour [12] 30 pour [13]). Dans notre cas, chacun des 8 outils testés a été utilisé avec 10 requêtes différentes formulées sous forme d'équations de recherche dans le domaine de la biologie, ce qui représente un nombre théorique de pages web dont le contenu est à expertiser égal à 1400.

Comme c'est le cas dans la majorité des études similaires [5], les 10 requêtes utilisées ont été spécialement construites, avec pour objectif de couvrir différents niveaux de complexité et de possibilités syntaxiques de recherche: mots, phrases, expressions contenant les opérateurs booléens ET, OU, SAUF *(Tableau I)*.

Pour pallier les problèmes de variabilité de syntaxe des différents outils étudiés, chaque requête a été traduite, afin de s'adapter au mieux à la syntaxe de chaque outil testé. Afin de ne pas favoriser ou défavoriser certains outils par rapport à d'autres, l'expérimentation a été effectuée pour chaque requête, dans un délai le plus court possible comme préconisé par [14], les outils testés en dernier étant théoriquement favorisés car ayant plus de chance d'avoir de nouvelles pages web indexées dans leur base de données.

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 ainsi prises en compte dans cette étude. 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: ce paramètre, très largement utilisé dans la littérature [5, 14, 15] correspond au nombre de réponse pertinentes proposées par un outil de recherche à la suite d'une requête donnée divisé par le nombre de réponses étudiées (20 dans notre cas). Une réponse pertinente

a été définie comme une page web présentant la propriété d'être informative par rapport à la requête formulée [16]. 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). 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, correspondant au nombre total de page web pertinentes obtenues pour les 10 requêtes utilisées, divisé par le nombre total requêtes testées, a également été calculé.

- la couverture relative: ce paramètre correspond au nombre de page pertinentes trouvées par un outil de recherche, divisé par le nombre total de pages web 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]: ce paramètre 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 étudiés et indirectement sur la qualité de ces index (plus il y a de liens en erreur, 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 *figure 1*. Il faut d'abord noter l'existence d'une dispersion importante des valeurs pour un outil donné qui peut varier pratiquement du simple au triple, en fonction de la requête considérée (0,25 a 0,67 pour Google par exemple), ce qui semble indiquer, assez logiquement, que les performances des outils présentés dépendent grandement de la requête formulée. Seulement 2 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 3 d'entre eux fournissent une précision moyenne inférieure à 0,1, correspondant à moins de 1 réponse pertinente sur 10 proposées (Bioview, ODP et Infomine).
Les données de la littérature disponibles à titre de comparaison, avec les résultats exposés dans la présente étude, concernent Altavista et font état de valeurs un peu plus élevées (0,46 pour [13]; 0,48 pour [9] et 0,78 pour [4], la différence avec la valeur présentée ici (0,34) étant probablement due à la moins grande spécificité des requêtes testées par ailleurs.
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 quand à eux une très faible proportion des réponses pertinentes sur l'ensemble des requêtes effectuées (Bioview, Infomine et ODP).
Il convient de noter 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).
Si la majorité des outils étudiés présente des pourcentages de liens en erreur voisins de ceux classiquement rencontrés dans la littérature [1, 2, 9] reflétant une mise à jour relativement régulière de leurs 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 concernent les performances des outils de recherche d'informations sur internet et illustrent la difficulté à pouvoir 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). Ces difficultés sont principalement une conséquence de la structure du web, qui provoque une dilution de l'information pertinente (les pages web au contenu "biologique") dans un océan de pages au contenus divers et variés.

Afin de faire face à cette principale difficulté, les concepteurs d'outils de recherche ont élaboré trois stratégies:

- une stratégie consiste à sélectionner de façon interactive les pages web présentes dans l'index interrogé, en se basant sur la qualité de leur contenu. Cela a pour conséquence la création d'index généralement de faible, taille mais contenant des pages dont la qualité de contenu a été contrôlée (cas des annuaires ODP et Infomine). Il semble qu'une sélection de ce type, très draconienne, aboutisse à la sélection de pages relativement généralistes et par conséquent à la formation d'index peu ou mal adaptés a la recherche d'informations spécifiques.
- une 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, à la suite de la saisie de la requête de l'utilisateur, grâce à des algorithmes de tri des résultats permettant leur présentation par ordre de pertinence. Il s'agit de la stratégie des outils généralistes comme Altavista, Google et Copernic. Cette stratégie semble être la plus efficace, car dans cette étude, 2 de ces 3 types d'outils de recherche offrent globalement les meilleures performances. Les 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 sur un calcul de la "popularité" des pages web qui prend en compte le nombre de pages ayant un lien vers chaque page.
- une 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, comme c'est le cas pour 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 requête qui ont été testées dans cette étude (l'index, dont la fréquence de mise à jour est insuffisante, possède une proportion de pages à caractère commercial très importante, 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 adaptées 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 [17] 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 les 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 des méta-données associées aux pages web par leurs créateurs [2, 18] (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.

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 - 109.
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 - 135.
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 Review 1997 ; 21 : 67 - 81.
6. Xie M, Wang H, Goh TN. Quality dimensions of internet search engines. Journal of Information Science 1998 ; 24 : 365 - 372.
7. Wang H, Xie M, Goh TN. Service quality of internet search engines. Journal of Information Science 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 - 932.
9. Wu G, Jie L. Comparing Web search engine performance in searching consumer health information : evaluation and recommendations. Bull Med Libr Assoc 1999 ; 87 : 456 - 461.
10. http://searchenginewatch.com/reports/sizes.html. Page consultée le 23 septembre 2001.
11. Ding A, Marchionini G. Comparative study of web search service performance. Proc ASIS Annu Meet 1996 ; 33 : 136 - 142.
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 Proceedings 1997 ; 49 : 184 - 189.
14. Oppenheim C, Morris A, McKnight C. The evaluation of WWW search engines. Journal of documentation 2000 ; 56 : 190 - 211.
15. Landoni M, Bell S. Information retrieval techniques for evaluating search engines: a critical overview. Aslib Proceedings 2000 ; 52 : 124 - 129.
16. Green R. Topical relevance relationships. Why topic matching fails. Journal of the American Society for Information Science 1995 ; 6 : 646 - 653.
17. Gardner M. A science-oriented search engine could solve problems... Nature 1999 ; 401 : 111.
18. 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 ; 945 - 949.

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

Tirés à part
C. Boudry

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.

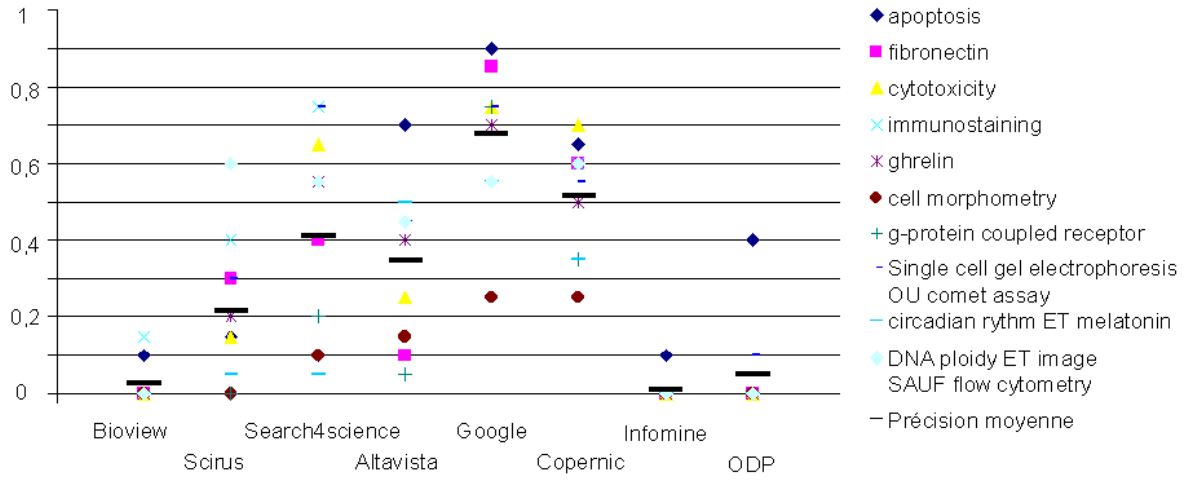

Figure 1. ***Précision des outils de recherche étudiés pour chacune des 10 requêtes testées et précision moyenne***. *La précision correspond à la fraction de réponse pertinentes proposées*

par un outil de recherche en réponse à une requête donnée divisé par le nombre de réponses étudiées (20 dans notre cas).

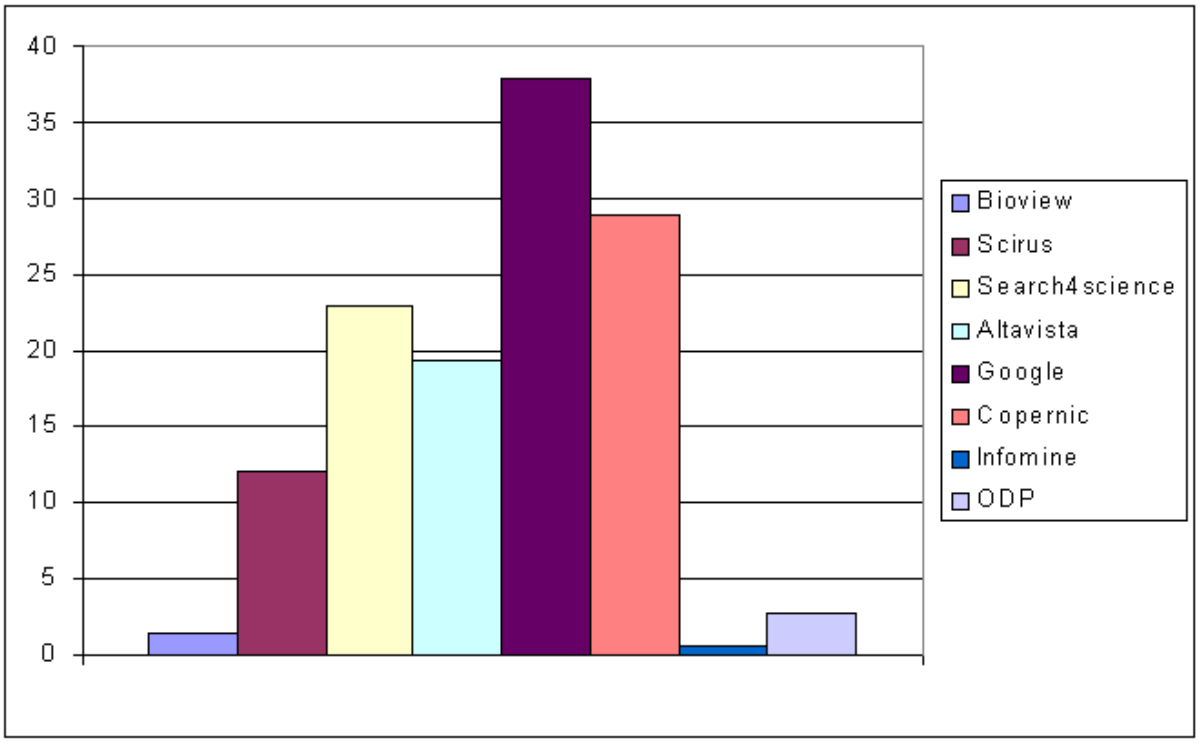

Figure 2. ***Couverture relative moyenne des 8 outils de recherche étudiés***. Ce paramètre correspond à la proportion de réponses pertinentes fournies par chaque outil de recherche pour l'ensemble des requêtes étudiées.

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

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.

	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.