

personnes atteintes du syndrome de Kartagener. Ces personnes sont plus souvent sujettes aux infections respiratoires, car les cils qui repoussent normalement les micro-organismes hors des voies respiratoires sont absents et, de plus, ils ont souvent un situs

inversus. Les organes des souris porteuses d'une mutation d'une protéine qui compose les cils s'orientent aléatoirement.

En 1998, Nobutaka Hirokawa et ses collègues de l'Université de Tokyo ont observé que les cils des cellules de

nœud de souris tournent, dans le liquide qui entoure l'embryon, en sens inverse des aiguilles d'une montre. Ce mouvement crée un courant qui rassemble vers la gauche du nœud les facteurs importants tels que l'acide rétinoïque et les protéines Nodal et Lefty. Cette accumulation de liquide et de protéines sur le côté gauche correspond peut-être au signal qui déclenche la rupture de la symétrie embryonnaire. La façon dont nos organes internes se développent est donc commandée par une caractéristique architecturale des cellules du nœud : le sens de rotation des cils.

Ce processus dépend probablement de l'asymétrie des molécules qui dirigent le mouvement ciliaire. Cependant, pour la moitié des individus totalement dépourvus de cils au niveau du nœud, les organes s'organisent normalement : les cils ne sont pas nécessaires au développement des organes, mais interviennent plutôt dans leur orientation et dans leur positionnement.

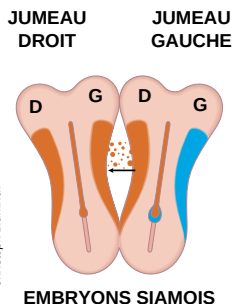
La compréhension de l'orientation gauche-droite au cours de l'embryogenèse a récemment progressé grâce à la découverte des gènes qui agissent de manière asymétrique chez l'embryon précoce. La nature exacte de l'initiation de l'asymétrie dans l'embryon reste inconnue, mais l'identification récente de protéines impliquées dans les événements suivants facilitera la découverte des protéines impliquées dans d'autres aspects du développement. La mise au point de nouvelles méthodes de diagnostic prénatal pour des pathologies liées à ces protéines sera alors possible.

Les siamois

Chez certains siamois, les organes se développent à une place anormale, qui dépend de l'endroit par où ils sont reliés. Pour la moitié des siamois soudés par le côté (siamois bicéphales), telles Abigail et Brittany Hensel (voir la photographie), l'un des membres de la paire, souvent celui qui est à droite, a son cœur à droite et non à gauche. Les siamois bicéphales se développent à partir d'un ovule unique fécondé, qui ne se divise pas complètement : les deux embryons qui se forment sont alors souvent parallèles (voir la figure). Selon l'équipe de Clifford Tabin et ses collègues de Harvard, cette proximité permet aux facteurs sécrétés, qui s'expriment dans un nœud de l'un des jumeaux, d'avoir une influence sur l'autre. La protéine nom-



Steve Weverka, Impact Visuals



mée Activine β B produite du côté droit du jumeau de gauche (*en rouge*), réprime probablement la production de la protéine Sonic Hedge-

hog du côté gauche du jumeau de droite. Celui-ci a donc autant de chances d'avoir un cœur à droite qu'à gauche.



Juan M. Hurle, University of Cantabria

6. LA VITAMINE A RÉGLE le développement du cœur (*en rouge*) de jeunes embryons de poulet. En présence d'une concentration normale d'acide rétinoïque (une forme de la vitamine A), le cœur s'incurve correctement (*à gauche*). Un excès d'acide rétinoïque provoque une déviation inverse (*à droite*). Les embryologistes pensent que l'acide rétinoïque favorise la régulation des gènes, tel le gène Nodal, qui gouvernent l'asymétrie gauche-droite.

Juan Carlos Ispisúa BELMONTE est professeur de biologie du développement à l'Université de Californie, à San Diego.

William B. WOOD, *Left-Right Asymmetry in Animal Development*, in *Annual Review of Cell and Developmental Biology*, vol. 13, pp. 53-82, 1997.

Michael LEVIN et Mark MERCOLA, *The Compulsion of Chirality : Toward an Understanding of Left-Right Asymmetry*, in *Genes and Development*, vol. 12, pp. 763-769, 15 mars 1998.

Ann F. RAMSDELL et H. Joseph YOST, *Molecular Mechanisms of Vertebrate Left-Right Development*, in *Trends in Genetics*, vol. 14, n° 11, pp. 459-465, novembre 1998.

Recherche intelligente sur l'Internet

LES MEMBRES DU PROJET CLEVER

La quantité d'informations en ligne sur le réseau Internet a tant augmenté que les outils actuels de recherche manquent d'efficacité. On améliore les recherches quand on analyse comment les pages du réseau mondial sont interconnectées.

Aujourd'hui, le réseau Internet donne accès à plusieurs centaines de millions de pages et s'enrichit d'un million de pages par jour. Plus d'un milliard de connexions entre les pages, les hyperliens, structurent cet impressionnant volume d'informations. Pour la première fois dans l'histoire, des millions de personnes ont accès, de leur domicile ou de leur bureau, aux productions intellectuelles d'une partie importante de la population de la planète.

Toutefois, en raison de son développement rapide et anarchique, Internet est un réseau d'informations sans organisation ni structure, de sorte que la recherche efficace d'information y est difficile. Le réseau est devenu un bric-à-brac gigantesque : les pages sont écrites par des personnes aux formations, aux intérêts ou aux motivations variés. Certaines pages ne contiennent que quelques caractères ; d'autres plusieurs centaines de milliers. Selon les cas, les informations sont justes, fausses, précises, floues ; la propagande avoisine les documents officiels et les inepties.

Les moteurs de recherche actuels cherchent des mots-clés présents dans les pages, mais ils identifient ainsi des dizaines de milliers de pages, dont la plupart sont inutiles. Comment localiser rapidement l'information voulue, et seulement elle, tout en étant sûr qu'elle soit authentique et fiable ?

Nous avons mis au point un nouveau type de moteur de recherche qui exploite la plus importante des ressources du réseau : ses innombrables hyperliens. Grâce à l'analyse de ces interconnexions, notre système localise

automatiquement deux types de pages : celles qui font autorité et celles qui servent de pivot. Les premières sont supposées être les meilleures sources d'information sur un sujet particulier ; les secondes sont des groupes de liens qui pointent vers les premières. Notre méthodologie permettra aux utilisateurs de localiser rapidement et efficacement la plupart des informations qu'ils désirent.

Les défis des moteurs de recherche

Les moteurs de recherche sont des programmes, accessibles aux utilisateurs du réseau, qui consultent une gigantesque mémoire où est enregistré pour chaque mot-clé présent dans les pages d'Internet, une liste des pages contenant ce mot. Un groupe de telles listes est nommé index. Ainsi, quand un utilisateur cherche des informations sur la revue *Pour la Science*, le moteur de recherche lui communique la liste de toutes les pages qui contiennent ce groupe de mots.

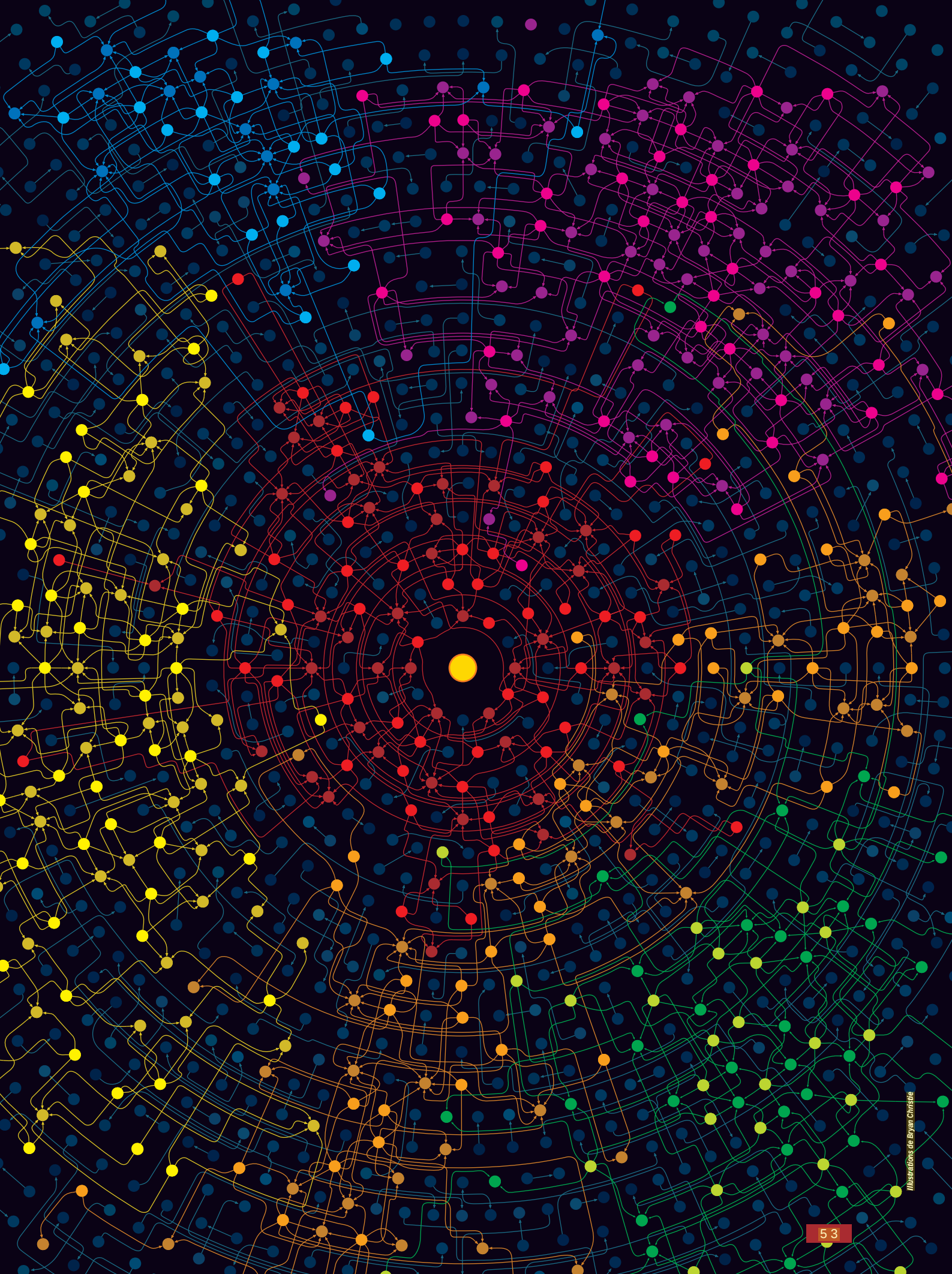
La création et la maintenance de ces index sont des tâches très lourdes (voir *La recherche d'informations*, par Clifford Lynch, *Pour la Science*, mai 1997) et c'est toujours un problème redoutable que de déterminer quelle information retourner en réponse à la question d'un utilisateur. Supposons que quelqu'un

cherche des informations sur la compagnie aérienne népalaise *Nepal Airways*. Sur la centaine de pages qui contiennent aujourd'hui cette expression, comment décider quelles sont les 20 meilleures ? Cette question est difficile, parce qu'on ne sait pas mesurer ce que « meilleure » veut dire : la réponse dépend du demandeur.

Des moteurs de recherche comme *Alta Vista*, *HotBot*, *Lycos*, *Infoseek* et *Excite* utilisent des règles empiriques pour ordonner les réponses. Ces « recettes », nommées fonctions de classement, sont appliquées non seulement aux requêtes précises (*Nepal Airways*), mais aussi à des requêtes plus générales, comme « avion », un mot qui apparaît dans 150 000 pages, ou son équivalent en anglais, *plane*, qui apparaît dans plus d'un million de pages. Comment un moteur de recherche retiendra-t-il seulement 20 pages ?

On obtient un classement très simple en comptant le nombre de fois où le mot demandé apparaît dans les pages. On peut aussi placer en premier les pages où le mot requis apparaît le plus tôt. Toutefois, de telles approches ne donnent pas toujours de bons résultats. Par exemple, la recherche du mot *contribuable* avec de telles règles de classement trouverait que *Polyptote*, extrait des *Exercices de style* de Raymond Queneau, répond très bien à la requête, puisqu'il contient 22 occurrences de ce mot dans un texte

1. LES COMMUNAUTÉS DU CYBERESPACE (montrées ici par des couleurs) cohabitent sur le réseau. L'exploration de la structure du réseau a révélé des communautés sur des sujets aussi variés que les marées noires sur la côte Ouest du Japon, les pompiers d'Australie ou les ressources pour les Turcs vivant aux États-Unis. Le réseau comprend des centaines de milliers de petites communautés de ce type.



de 140 mots environ. Diverses améliorations pondèrent l'importance des mots selon qu'ils apparaissent dans les titres, au début des chapitres ou en caractères plus gros que le corps du texte.

De nombreux sites abusent de ces règles, afin que leurs pages soient bien classées. C'est pour cette raison que le titre de certaines pages est étrangement rédigé, tel «billets avion soldés, billets avion soldés, billets avion soldés». D'autres sites incluent plusieurs fois dans leurs pages des phrases judicieusement choisies, dans

des polices et des couleurs invisibles à l'œil. De telles pratiques, qui inondent les moteurs de recherche avec des informations non pertinentes (le *spamming*), compliquent la tâche des gestionnaires de ces moteurs (le mot anglais *spam* vient d'un épisode de la série télévisée *Monty Python's Flying Circus*, où le mot *Spam*, une marque de jambonneau, couvre les dialogues).

Même sans cette difficulté, le bien-fondé des méthodes de recherche textuelles est douteuse : des pages qui intéresseraient beaucoup un utilisateur

ne contiennent pas le mot qui a été transmis au moteur de recherche, tandis que d'autres pages qui contiennent le mot spécifié sont sans intérêt. Cette seconde difficulté provient de l'abondance de synonymes et de mots polysémiques (des mots qui ont plusieurs sens) : une demande d'information sur les «automobiles» laissera de côté une kyrielle de pages qui évoquent les «voitures». En raison des polysémies, une simple requête sur «jaguar» donnera des milliers de pages concernant, entre autres, l'automobile de ce nom, le félin et une équipe de football américain.

On améliore les recherches en utilisant les relations sémantiques entre les mots. Certains moteurs de recherche utilisent de telles relations, nommées «réseaux sémantiques», identifiées par les linguistes. À la réception de la requête «automobile», un moteur de recherche couplé à un réseau sémantique détermine d'abord qu'«automobile» est un mot équivalent à «voiture», puis il donne la liste des pages qui contiennent l'un des deux mots. Toutefois, cette méthode est à double tranchant : elle résout le problème des synonymes, mais aggrave celui de la polysémie.

De surcroît, l'établissement des réseaux sémantiques est une tâche redoutable, car, surtout sur le réseau Internet, multiculturel, les néologismes fleurissent. Des mots sont apparus, tels que «FAQ» (Foire Aux Questions), «zines» (diminutif de fanzines, de petits journaux diffusés sur le réseau) ou «internaute» ; d'autres mots ont de nouvelles acceptions, tels «surfer» ou «naviguer».

Recherche avec les hyperliens

Le projet *Clever* de la Société IBM voulait résoudre ces difficultés. Très tôt, nous avons compris que l'indexation ou la récupération de pages perdait beaucoup à ne se fonder que sur le texte : plus d'un milliard de liens ne révèlent-ils pas les relations entre les pages. Comment tirer parti de ces informations ?

Lorsqu'ils recherchent «Harvard», par exemple, beaucoup d'internautes désirent en fait des informations sur les équipes sportives qui s'y trouvent. Toutefois, plus d'un million de pages contiennent ce mot, et la page d'accueil de l'Université Harvard n'est pas celle qui l'utilise le plus souvent, de sorte qu'elle ne sera pas la première choisie, ni même jugée importante à l'aide des fonctions de

La «cinétique chimique» sur l'Internet

Tous les moteurs de recherche ne donnent pas de résultats équivalents. Nous avons comparé des moteurs de recherche en analysant les réponses à la requête «cinétique chimique» (l'étude de la vitesse de réactions chimiques).

Plusieurs objectifs se dégagent pour cette recherche : le moyen d'établir un contact entre enseignants pour une mise en commun d'expériences pédagogiques, la recherche des enseignements de troisième cycle dispensés dans cette matière et la recherche de documentation.

Nous avons comparé les résultats donnés par une sélection de moteurs francophones (*Ecila*, *Eureka*, *Excites*, *Francité*, *Lokace*, *Yahoo! France*), à laquelle nous avons ajouté deux moteurs anglophones (*AltaVista* et *GoTo*) et deux métamoteurs, des moteurs qui recherchent dans plusieurs moteurs (*MetaCrawler*, *Savvy Search* et *Trouvez*).

La recherche s'est effectuée uniquement par mot clé, car cette méthode est commune à tous les moteurs. Plusieurs requêtes ont été saisies avec des contraintes de plus en plus fortes. Nous avons d'abord fait la requête *cinétique*, puis *+cinétique +chimique* (qui recherche toutes les pages où les deux termes sont présents) et enfin *cinétique chimique*, qui renvoie les pages où l'expression «cinétique chimique» est présente.

Nous avons classé les moteurs selon deux critères : la quantité de pages indexées et la qualité (renvois

de pages traitant de cinétique chimique jugées pertinentes).

Pour l'étude quantitative, trois moteurs sortent du lot : ce sont *Carrefour*, *Nomade* et *AltaVista*. Ils renvoient 500 à 6 000 pages selon les contraintes imposées. D'autres moteurs donnent de bons résultats (entre 50 et 200 pages) et, enfin, d'autres sont insuffisants.

Pour l'étude qualitative, nous avons prédéfini comme pages pertinentes cinq pages d'universités ou d'écoles qui enseignent la cinétique chimique (un des objectifs précités) : l'Université catholique de Louvain (Belgique), l'École nationale supérieure des industries chimiques à Nancy, l'Université de Pau, l'Université Pierre et Marie Curie (Paris VI) et l'Université de Laval (Canada). Avec la saisie de «cinétique chimique», nous avons classé les différents moteurs en cinq groupes, du meilleur au moins bon, les meilleurs étant ceux qui contiennent les cinq pages d'accueil des institutions précitées parmi les 30 premières pages retournées. Les plus performants sont *Trouvez*, *SavvySearch*, *Metacrawler* et *GoTo*.

Les trois moteurs qui renvoyaient le plus grand nombre de pages ne sont pas les plus performants sur la qualité de l'information. On retrouve ainsi les préoccupations du projet *Clever* : la quantité n'est pas toujours liée à la qualité, d'où l'importance d'un moyen de retrouver une information pertinente.

J. BENAÏM et V. VIOSSAT
Université P. et M. Curie, Paris