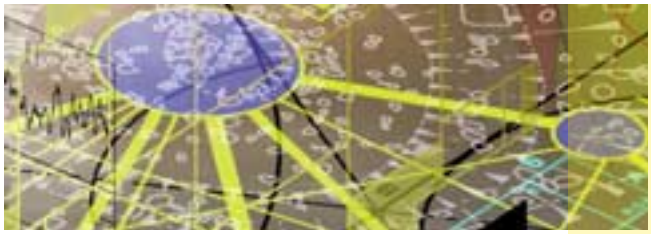
# La recherche d'informations

CLIFFORD LYNCH

*Les bibliothécaires devront s'allier aux informaticiens pour que cesse l'anarchie qui règne sur l'Internet.*


Jeff Brice

Certains prétendent que le réseau Internet est la bibliothèque mondiale de l'ère numérique. Cette idée ne résiste pas à l'examen. L'Internet, et en particulier l'ensemble de ressources multimédia qui constituent le World Wide Web, ou réseau mondial, n'ont pas été conçus pour servir de support à la publication et à la recherche de l'information. C'est plutôt une bibliothèque chaotique de l'ensemble des publications sortant des «presses» numériques de la planète. Cette mine d'informations contient non seulement des livres et des articles, mais aussi des données scientifiques, des menus, des comptes rendus de conférences, des publicités, des enregistrements audio et vidéo et des transcriptions de conversations interactives. L'anecdotique et l'éphémère sont jetés pêle mêle dans l'important et le durable.

En somme, l'Internet n'a rien d'une bibliothèque numérique. Il ne se développera et ne deviendra un nouveau moyen de communication universel que si l'on met au point des services analogues à ceux des bibliothèques classiques pour organiser l'information, la conserver et la retrouver. Même alors, l'Internet ne ressemblera pas à une bibliothèque classique, car son contenu restera dispersé. Les bibliothécaires, qui choisissent et classent les ouvrages, devront être épaulés par des informaticiens, qui mettront au point des techniques d'indexation et de stockage automatisés de l'information. Seule une synthèse de ces deux professions assurera la viabilité de ce nouveau média.

Aujourd'hui, la structuration de l'information sur l'Internet repose essentiellement sur l'informatique. En théorie, les programmes qui classent et indexent automatiquement des données numériques peuvent gérer la surabondance d'information présente sur l'Internet. De surcroît, l'indexation automatique, peu onéreuse, est préférée à l'indexation humaine, lente et coûteuse.

Toutefois, ces programmes classent l'information différemment de nous. En un sens, le travail effectué par les divers outils automatiques d'indexation et de catalogage, les moteurs de recherche, est extrêmement démocratique : il donne un accès uniforme, non hiérarchisé, à toutes les données présentes sur le réseau. Dans la pratique, cet égalitarisme électronique a ses avantages et ses inconvénients. Les «surfeurs» du réseau qui cherchent une information se retrouvent souvent submergés par des milliers de réponses. De plus, les résultats d'une recherche renvoient fréquemment à des sites Web (un site Web est un ensemble de pages enregistrées dans la mémoire d'un ordinateur du réseau) sans rapport avec la demande et omettent d'autres sites contenant des informations pertinentes.

## Arpenter le réseau

Les spécificités de l'indexation électronique apparaissent quand on examine comment les moteurs de recherche, tels *AltaVista*, de la Société *Digital Equipment*, ou *Lycos*, construisent leurs index et cherchent l'information réclamée par un utilisateur. Ces moteurs envoient périodiquement des programmes nommés «robots collecteurs» sur tous les sites qu'ils identifient sur le réseau. Ces robots récupèrent les pages, puis un logiciel en extrait une indexation qui assurera leur description. Selon le moteur, cette dernière procédure va de la simple localisation de la plupart des mots présents dans les pages à l'analyse complexe, assortie d'une identification des mots et des groupes de mots importants. Ces données sont ensuite stockées dans la base de données du moteur de recherche, avec une adresse, une URL (*Uniform Ressource Locator*, soit «localisateur unifié de ressources»), qui localise le document. Pour soumettre ses demandes à la base de données du moteur de recherche, le demandeur utilise un navigateur – tel le programme *Netscape* – qui lui donne une liste de ressources, les URL, sur lesquels il peut cliquer pour se connecter aux sites repérés par le moteur de recherche.

Aujourd'hui, les moteurs de recherche traitent quotidiennement des millions de demandes d'information. Cependant, ces moteurs sont loin d'être des outils idéaux pour s'y retrouver dans la masse d'informations, sans cesse croissante, disponible sur l'Internet. Contrairement aux indexeurs humains, les moteurs de recherche cernent mal les caractéristiques globales d'un document, telles que le sujet ou le genre – poème, pièce de théâtre, voire publicité.

En outre, le réseau mondial ne dispose pas encore de normes qui faciliteraient l'indexation automatique : les documents qu'il contient ne sont pas structurés, de sorte que les programmes ne peuvent en extraire l'information qu'un indexeur humain trouverait par un simple examen superficiel, comme l'auteur, la date de publication, la longueur et le sujet d'un texte (ces infor-

mations sont des «métadonnées»). Un moteur retrouverait certainement le dernier livre de Jacques Martin, par exemple, mais il récupérerait également des milliers d'autres documents dont le texte ou les références bibliographiques contiennent ce nom courant.

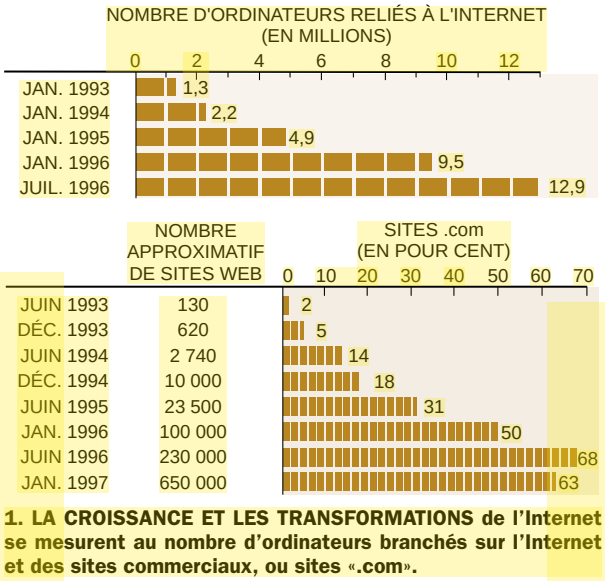
Les éditeurs abusent parfois de ce manque de discernement de l'indexation automatique. Leur site sera plus fréquemment consulté s'ils ont placé dans leurs documents des mots, tels que «sexe», qui sont fréquemment demandés. En effet, les moteurs de recherche affichent en premier les adresses des documents qui mentionnent le plus fréquemment le mot clé recherché. Un individu qui se préoccuperait des thèmes ne se laisserait pas berner.

Les spécialistes de l'indexation savent décrire les contenus de toutes sortes de pages, du texte au document vidéo, et exprimer comment ces composants s'intègrent dans une base de données informative. Ils sauraient, par exemple, regrouper des photographies de la Première Guerre mondiale et des musiques d'époque ou des journaux intimes de soldats. Un indexeur humain sait également décrire les règles qui régissent l'organisation d'un site qui archiverait des programmes pour *Macintosh*, par exemple. L'analyse des objectifs d'un site, de son histoire et de sa politique d'organisation dépasse les capacités d'un moteur de recherche.

De surcroît, la plupart des moteurs de recherche ne reconnaissent que le texte. Or les utilisateurs du réseau y cherchent fréquemment des images, fixes ou animées. Certains programmes reconnaissent aujourd'hui des couleurs ou des motifs présents sur des images *(voir l'encadré des deux pages suivantes)*, mais aucun programme ne sait déterminer le sens et l'importance culturelle d'une image et reconnaître, par exemple, qu'un groupe d'hommes attablés représente la Cène.

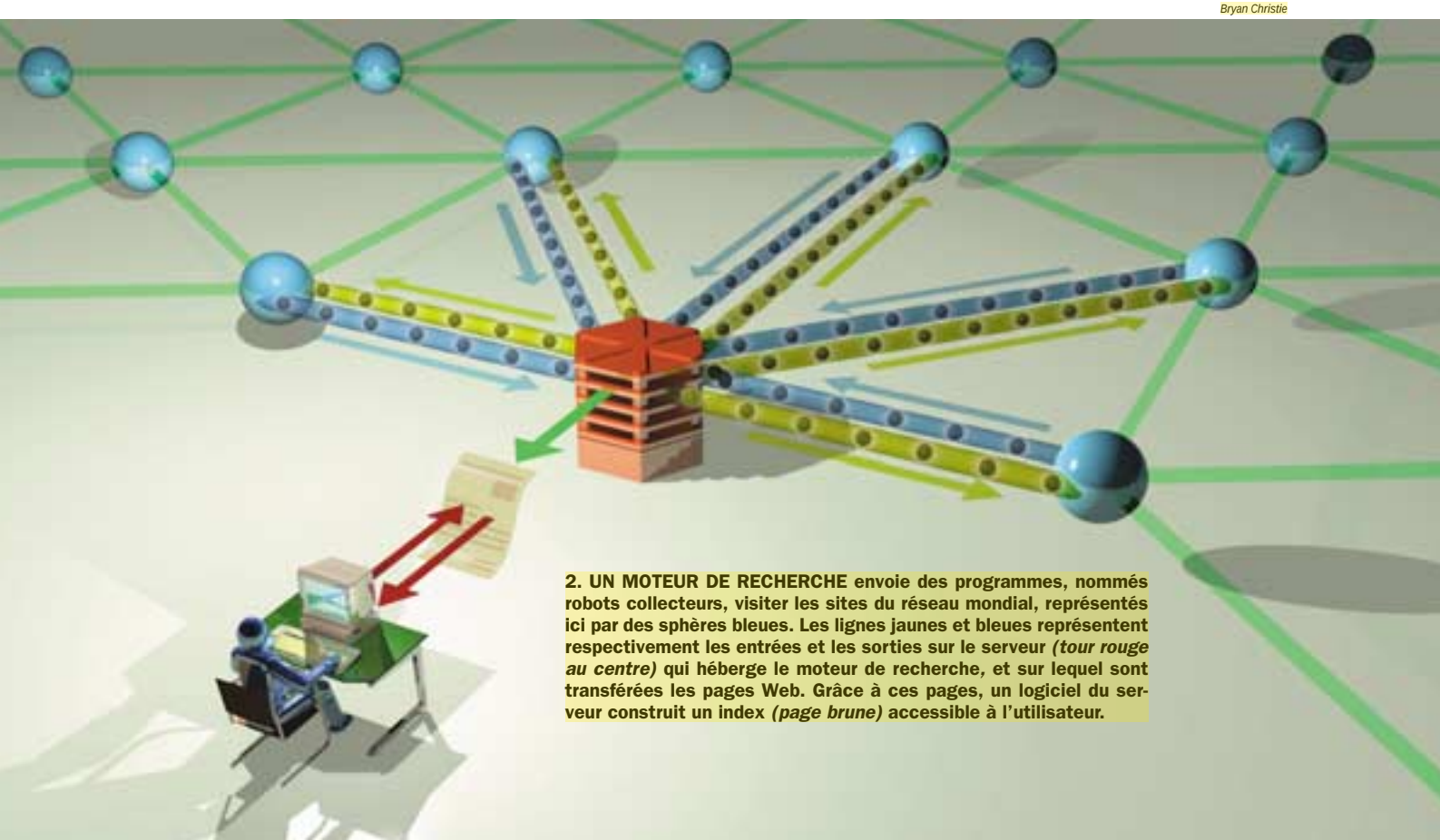
Enfin, de nombreux sites utilisent aujourd'hui une nouvelle structuration des données qui empêche les moteurs de recherche de les analyser : nombre de pages ne sont plus des fichiers statiques que l'on peut analyser et indexer facilement, mais des pages créées dynamiquement lors de l'affichage, en réponse à la recherche d'informations de l'utilisateur. Par exemple, lorsque sur le site d'un journal, un utilisateur demande un article concernant le commerce du pétrole, un programme créera la page avec des cartes, des tableaux et un texte provenant de diverses parties de sa base de données, sans que l'on puisse ni consulter, ni indexer cette base de données.

Un nombre croissant d'informaticiens étudient les méthodes de classification automatisée. L'une des approches




**1. LA CROISSANCE ET LES TRANSFORMATIONS de l'Internet se mesurent au nombre d'ordinateurs branchés sur l'Internet et des sites commerciaux, ou sites «.com».**

Bryan Christie



**2. UN MOTEUR DE RECHERCHE envoie des programmes, nommés robots collecteurs, visiter les sites du réseau mondial, représentés ici par des sphères bleues. Les lignes jaunes et bleues représentent respectivement les entrées et les sorties sur le serveur *(tour rouge au centre)* qui héberge le moteur de recherche, et sur lequel sont transférées les pages Web. Grâce à ces pages, un logiciel du serveur construit un index *(page brune)* accessible à l'utilisateur.**

explorées consiste à attacher des métadonnées aux fichiers, de sorte que les systèmes d'indexation obtiennent directement ces informations. Les travaux les plus avancés sont ceux du programme *Dublin Core Metadata* et du projet *Warwick Framework* – le premier ainsi nommé d'après un atelier de travail qui eut lieu à Dublin, dans l'Ohio, le second d'après un colloque qui s'est tenu à Warwick, en Angleterre. Ces groupes de travail ont défini un ensemble de métadonnées plus simples que celles qui sont classiquement utilisées pour constituer les catalogues de bibliothèques et ont mis au point des méthodes pour intégrer les métadonnées dans les pages du réseau.

Ces métadonnées contiendraient, par exemple, le titre d'un document, son auteur ou son genre (texte, vidéo, etc.). Elles pourraient être produites manuellement ou par des logiciels d'indexation automatique. Par ses annotations précises et détaillées, la description donnée par l'être humain reste bien supérieure à celle qui est obtenue par un programme d'indexation automatique.

Lorsque les coûts sont justifiés, des indexeurs humains compilent laborieusement les bibliographies de certains sites. La base de données *Yahoo*, par exemple, répertorie les sites en fonction de rubriques très générales. Dans un projet de recherche de l'Université du Michigan, on cherche à mettre au point une description de sites contenant des informations spécifiquement destinées aux chercheurs.

## Plus qu'une simple bibliothèque

Les stratégies adoptées pour la recherche et l'indexation (humaine ou automatisée) des documents dépendront des utilisateurs de l'Internet et des perspectives commerciales offertes aux éditeurs. Nombre de chercheurs souhaitent conserver une bibliothèque numériquement structurée, mais, pour d'autres, un média démocratique et libre serait le support idéal pour la diffusion de l'information. Certains utilisateurs, tels les analystes financiers et les espions, souhaitent un accès illimité aux bases de données brutes, libre de tout contrôle. Pour eux, les moteurs de recherche standards offrent de réels avantages, car ils ne filtrent pas les données.

## La recherche d'images sur le Web

Depuis quelques années, le public découvre que le réseau Internet lui offre des photographies, des animations, des graphiques, des sons et des films consacrés à des sujets variés, du grand art à la franche obscénité. Malgré cette surabondance d'informations, la recherche d'un document, parmi des centaines de milliers de sites, repose essentiellement, aujourd'hui encore, sur des index de mots et de nombres.

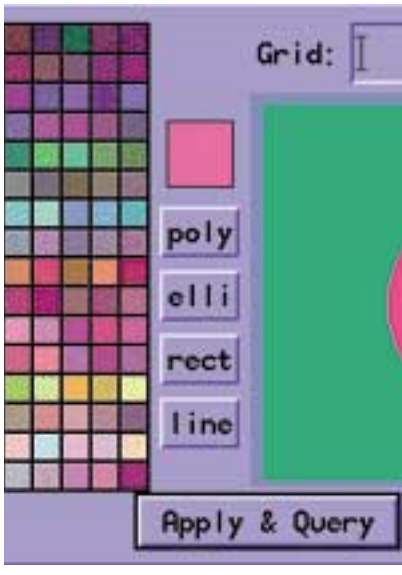
Si l'on recherche «drapeau français» sur le moteur de recherche *AltaVista*, on obtient l'image souhaitée, à condition qu'elle ait été légendée avec ces deux mots. Mais comment quelqu'un qui connaît seulement l'aspect du drapeau peut-il retrouver son pays d'origine?

Le réseau serait utile si les moteurs de recherche permettaient de dessiner ou de numériser un rectangle composé de trois bandes verticales colorées en bleu, en blanc et en rouge, et s'ils trouvaient toutes les images correspondantes stockées sur la myriade de sites. Ces dernières années, des techniques combinant indexation par mots clés et analyse d'images ont ouvert la voie aux premiers moteurs de recherche d'images.

Bien que ces prototypes soient des pionniers de l'indexation graphique, ils montrent à quel point les outils actuels sont rudimentaires et fondés sur l'usage du texte. *WebSEEk*, un logiciel expérimental conçu à l'Université Columbia, illustre le fonctionnement d'un moteur de recherche d'images. Il commence par charger les fichiers trouvés en visitant les sites, puis tente de localiser les noms de fichiers contenant des acronymes, tels que GIF ou MPEG, désignant un contenu graphique ou vidéo. Il cherche également, dans ces noms de fichiers, des mots qui pourraient identifier le sujet des fichiers. Lorsque *WebSEEk* trouve une image, il détermine ses couleurs prépondérantes et leurs positions. À l'aide de ces informations, le programme distingue les photographies, les graphiques, les images en noir et blanc ou en gris. Le logiciel comprime également chaque image et la transforme en une icône miniature qu'il pourra afficher. Pour une vidéo, le programme *WebSEEk* extrait des images de plusieurs scènes.

Un utilisateur commence sa recherche en choisissant une catégorie dans un menu – par exemple, «chats». *WebSEEk* affiche alors un échantillon d'icônes associées à la catégorie «chats». Pour limiter sa recherche, l'utilisateur peut cliquer sur toute icône montrant des chats noirs. Grâce à l'analyse de couleurs précédemment effectuée, le moteur de recherche cherche des correspondances avec des images ayant un profil de couleur similaire. Il présente alors un nouveau lot d'images qui représentent des chats noirs – mais aussi, par exemple, quelques chats jaune et orange assis sur des coussins noirs. L'utilisateur peut alors affiner sa recherche en ajoutant ou en excluant certaines couleurs d'une image avant de relancer la recherche. En excluant le jaune et l'orange, il fait disparaître les chats colorés. Plus simplement, l'utilisateur peut aussi désigner les images qui ne contiennent pas de chats noirs, indiquant par là au programme de ne pas rechercher ce type d'image. À ce jour, *WebSEEk* a transféré et indexé plus de 650 000 images provenant de dizaines de milliers de sites.

Plusieurs centres informatiques et quelques sociétés, telles qu'*IBM* et *Virage*, mettent au point des programmes de recherche d'images sur des réseaux privés ou dans des bases de données. Deux autres sociétés, *Excalibur Technologies* et *Interpix Software*, se sont associées pour fournir des logiciels aux moteurs de recherche *Yahoo* et *Infoseek*.




C'est pourtant l'un des plus anciens programmes de recherche d'images, QBIC d'*IBM*, qui fournit les meilleures correspondances entre les motifs des images. Le logiciel QBIC repère les couleurs dans une image, et il mesure plusieurs paramètres pour en évaluer la texture : le contraste (le blanc et le noir des rayures de zèbre), le grain (la différence entre les galets et les rochers) et l'orientation (la différence entre les poteaux parallèles d'une clôture et l'isotropie des pétales de fleur). Ce programme