des processus qui leur ont donné naissance. Nous donnons, pour illustrer ce propos, deux instantanés de motifs chimiques tels qu'apparus au cours d'états transitoires du système et qui, à long terme, auraient abouti à une structure en réseau hexagonal de taches. Un traitement approprié de la palette de couleurs met en évidence d'étonnantes similarités entre ces motifs chimiques et ceux portés par les peaux du guépard et du léopard *(voir la figure 2)*.

Dans des conditions un peu particulières («loin du seuil d'instabilité», diraient les spécialistes des systèmes non linéaires), bien d'autres motifs surprenants peuvent apparaître. Dans certains cas, les motifs s'étendent par division des premières taches apparues, évoquant curieusement la division cellulaire au cours de la croissance de tissus biologiques. Enfin notre équipe a récemment observé des croissances en forme de doigts, donnant naissance à d'étranges et transitoires «fleurs» chimiques *(voir la figure 6)*.

Analogies et universalité

Aussi extraordinaires que soient les ressemblances entre structures chimiques de réaction-diffusion et certains motifs observés dans la nature, de telles similarités ne permettent nullement de conclure à l'identité des processus qui président à leur développement. De plus, étant donné l'extraordinaire complexité du réseau d'interactions propre à tout système biologique, isoler chez un être vivant des auto-activateurs et des inhibiteurs afin d'en vérifier le comportement *in vitro* tient souvent de la gageure.

Tout au plus, dans certains organismes où apparaissent des structures périodiques, on a réussi à identifier des morphogènes (par exemple le calcium au cours du développement des ramifications de l'algue *Acetabularia*) ou des molécules participant au processus d'auto-activation (au cours du développement du membre supérieur de l'embryon de poulet). Des expériences de biologie confirment aussi que, lors de la formation des motifs pigmentés sur la peau des mammifères, il y a bien développement de prémotifs ; toutefois l'origine de ceux-ci reste mal établie, et l'on sait que ce processus met aussi en œuvre des migrations de cellules et l'adhésion entre cellules.

Quoi qu'il en soit des difficultés inhérentes au milieu biologique, les systèmes biologiques, aussi bien que les systèmes chimiques non linéaires, sont capables d'évoluer d'un état quasi indifférencié vers un nouvel état plus structuré. Cette transition constitue ce que les physiciens appellent une «brisure de symétrie». Ces brisures de symétrie peuvent être décrites par les mêmes équations mathématiques, définissant ainsi des classes d'universalité. Certains systèmes biologiques, tout comme les systèmes chimiques de réaction-diffusion, appartiennent au moins en partie à la même classe d'universalité, et les connaissances que l'on acquiert sur les uns contribuent à élucider les comportements des autres. Si des biologistes de plus en plus nombreux pensent que les processus de réaction-diffusion jouent un rôle au cours de la morphogenèse, il est certain que, outre la réaction-diffusion, d'autres processus physiques contribuent à l'élaboration des formes lors des multiples modifications qui ont lieu au cours du développement embryonnaire. Ce domaine de recherches est en plein essor.

Sur *Archimède*, l'émission franco-allemande de vulgarisation scientifique d'ARTE, P. De KEPPER présente «Les taches du léopard», le 13 mai 1997 à 20 heures.

P. De KEPPER est directeur de recherche et É. DULOS est chargée de recherche au Centre de recherche Paul Pascal, laboratoire du CNRS à Bordeaux.

0scillations and travelling waves in chemical systems, sous la direction R.J. Field et M. Burger, John Wiley & Sons, 1985.

Chemical waves and patterns, sous la direction de R. Kapral et K. Showalter, éditions Kluwer, 1995.

V. CASTETS, E. DULOS, J. BOISSONADE et P. De KEPPER, *Experimental evidence of a sustained standing Turing-type non equilibrium chemical pattern*, in *Physical Review Letters*, n° 64, pp. 2953-2956, 1990.

P. De KEPPER, J.J. PERRAUD, B. RUDOVICS et E. DULOS, *Experimental study of stationary Turing patterns...* in *International Journal of Bifurcation and Chaos*, n° 4, pp. 1215-1231, 1995.

INTERNET

Les promesses seront-elles tenues ?

L'Internet : du chaos à l'ordre?

Le réseau Internet est victime de son succès : en quelques années, ce système, initialement confidentiel, de communications entre scientifiques ou ingénieurs est devenu un carrefour inévitable, emprunté par des personnes du monde entier, dans des secteurs aussi variés que l'analyse financière ou la haute couture. Pour remédier à la saturation du réseau, le gouvernement américain a même annoncé son intention de construire un réseau réservé aux scientifiques : l'Internet II.

Comment mettra-t-on le réseau existant au service du grand public? Le courrier électronique, voire les vidéoconférences, sont déjà répandus, mais ces applications sont bien en deçà de ce que le réseau mondial, le World Wide Web, peut offrir : la connexion, par le téléphone ou par des lignes spécialisées, des divers ordinateurs du monde a engendré la constitution d'une gigantesque mine d'informations et d'analyses, la base de données de toutes les bases de données. Les inquiétudes sur l'avenir d'Internet, qui portent habi-

tuellement sur les temps de chargement prohibitifs et sur les limitations d'accès dus à la surcharge du réseau, ne sont que des «douleurs de croissance», qui se résoudront. En revanche, on n'utilisera bien le réseau que si l'on parvient à structurer, à stocker et à classer l'information qui y circule.

Dans les pages qui suivent, des informaticiens expliquent comment l'organisation du savoir rendra l'Internet réellement efficace. À partir de points de vue variés, ils cherchent comment simplifier la recherche de l'information (oui, de nouveaux moteurs de recherche faciliteront l'exploration du réseau). Ils examinent les meilleurs moyens de structurer et d'afficher des données, de sorte que chacun (y compris les aveugles) accède facilement à l'information, de toutes les manières imaginables. Les solutions techniques qu'ils proposent ne sont peut-être pas celles qui seront finalement adoptées, mais leurs idées susciteront certainement des débats.

La structuration d'un milieu fondamentalement désordonné comme le réseau mondial contribuera peut-être à réaliser le rêve des encyclopédistes français du XVIII^{e} siècle : réunir en un seul lieu toutes les connaissances du monde. Deux siècles après eux, Vannevar Bush, qui dirigea le bureau américain de la recherche scientifique et technique durant la Seconde Guerre mondiale, proposa le «memex», un poste de travail comprenant un lecteur de microfilms et quantité de films qui auraient été l'équivalent d'une bibliothèque de recherche complète ; le lecteur du memex aurait pu lier entre elles et annoter des informations provenant de films différents. Les idées de Bush influencèrent Ted Nelson, l'inventeur du système de l'hypertexte, qui fut finalement introduit sur le réseau. Le dynamisme intellectuel qui court, des encyclopédistes à T. Nelson, se retrouve dans les articles qui suivent.

Leurs auteurs y esquissent les grandes orientations techniques qui pourraient conduire à la constitution d'un fonds complet du savoir humain. Le réseau Internet deviendrait ainsi un lieu où les rêves d'Homère, de Shakespeare et de Lao-tseu seront à un clic de souris des menus des cantines scolaires ou de l'ordre du jour du prochain conseil municipal. Toutes les activités humaines seront ainsi représentées, des plus nobles aux plus communes.

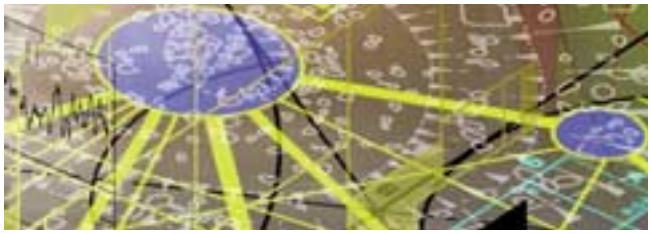
Retrouvez les articles de ce dossier avec des informations complémentaires et des liens sur notre site Web http://www.pourlascience.com

Jeff Brice

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

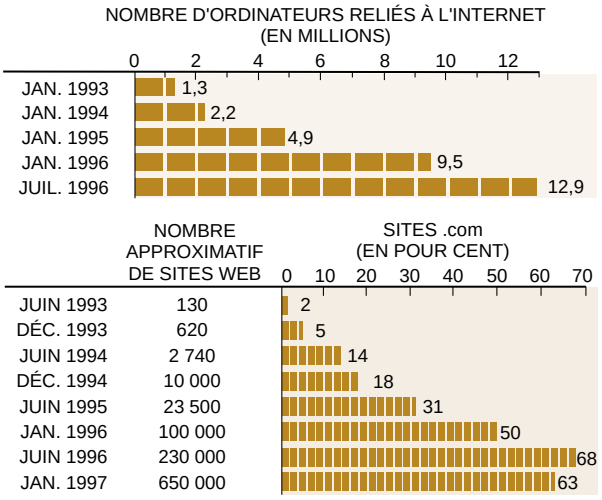
Les spécificités de l'indexation électronique apparaissent quand on examine comment les moteurs de recherche, tels *AltaVista*, de la Société *Digital Equipment*, ou *Lycos*, construisent leurs index et cherchent l'information réclamée par un utilisateur. Ces moteurs envoient périodiquement des programmes 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.



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».

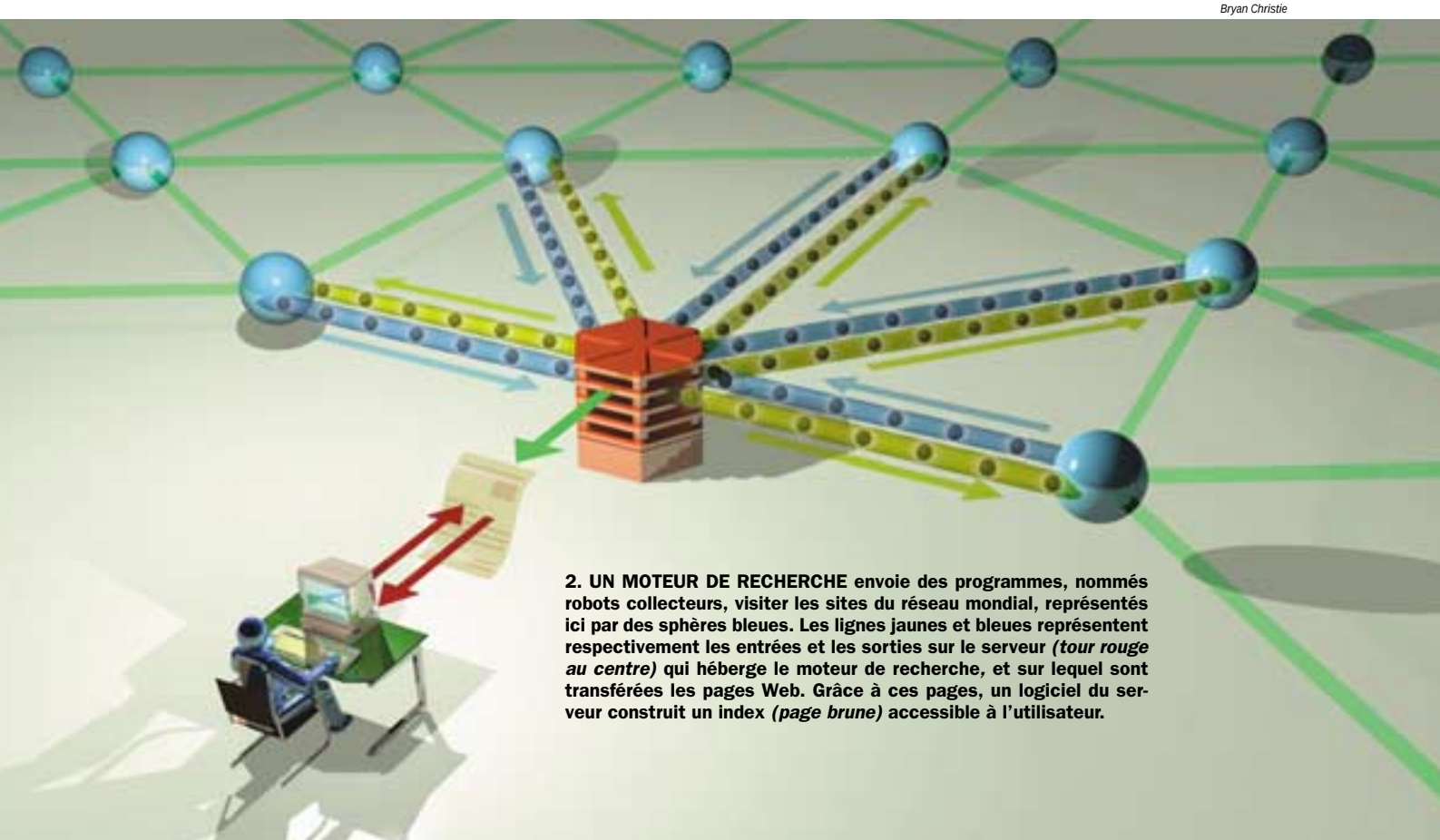
Les spécialistes de l'indexation savent décrire les contenus de toutes sortes de pages, du texte au document vidéo, et exprimer comment ces composants s'intègrent dans une base de données informative. Ils sauraient, par exemple, regrouper des photographies de la Première Guerre mondiale et des musiques 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

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.

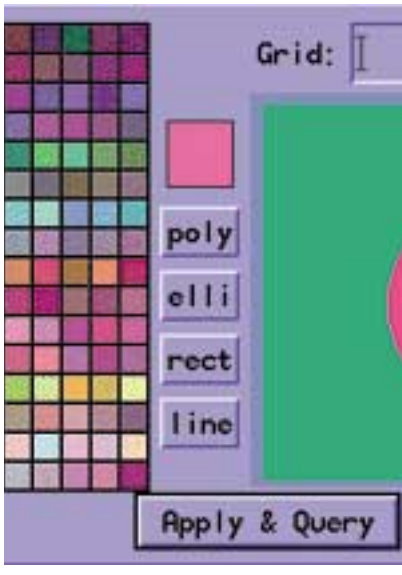
Si l'on recherche «drapeau français» sur le moteur de recherche *AltaVista*, on obtient l'image souhaitée, à condition qu'elle ait été légendée avec ces deux mots. Mais comment quelqu'un qui connaît seulement l'aspect du drapeau peut-il retrouver son pays d'origine?

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.

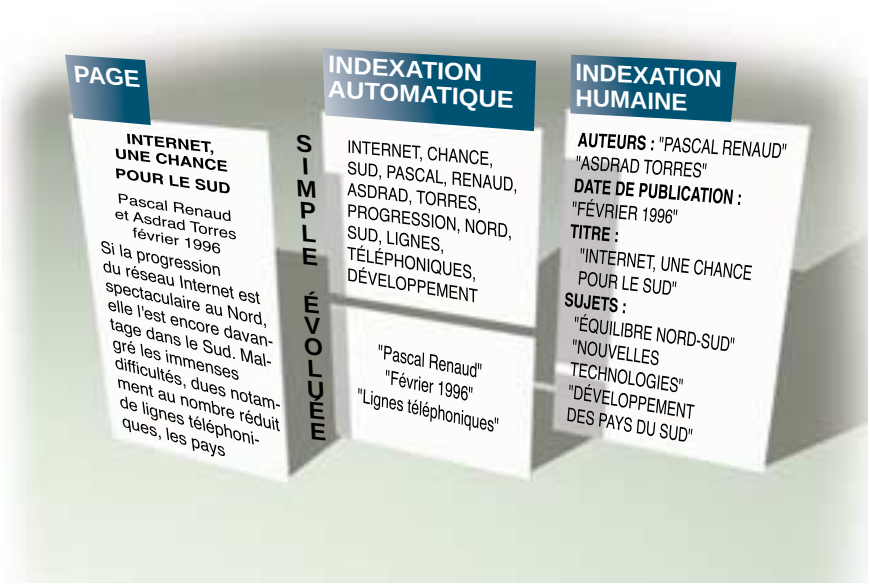
Un utilisateur commence sa recherche en choisissant une catégorie dans un menu – par exemple, «chats». *WebSEEk* affiche alors un échantillon d'icônes associées à la catégorie «chats». Pour limiter sa recherche, l'utilisateur peut cliquer sur toute icône montrant des chats noirs. Grâce à l'analyse de 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

L'information circulant sur l'Internet est bien plus variée que celle des bibliothèques classiques. Une bibliothèque ne classe pas, par ordre de qualité, les ouvrages qu'elle contient. Toutefois, comme l'Internet comprend un volume d'information supérieur, ses utilisateurs veulent utiliser au mieux le temps limité dont ils disposent pour effectuer une recherche. Ils veulent connaître, par exemple, les trois «meilleurs» documents traitant d'un



Bryan Christie

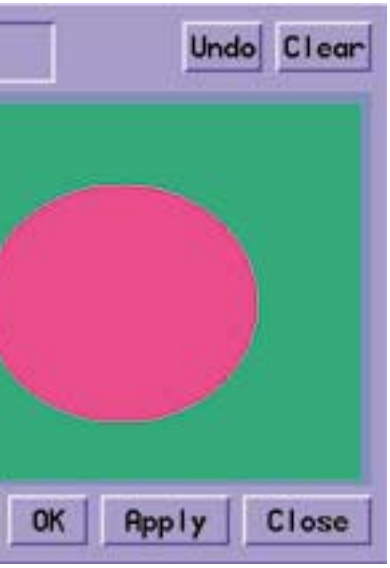
3. L'INDEXATION AUTOMATIQUE effectuée par un moteur de recherche *(panneau de gauche)* conduit, à partir d'une page d'un article du *Monde diplomatique*, à une liste de mots *(panneau central, en haut)* ou de groupes de mots simples *(panneau central, en bas)*. L'indexation humaine *(panneau de droite)* donne des indications plus structurées.

peut également rechercher des formes simples dans les images. Si on lui demande de rechercher des images similaires à un point rose sur un fond vert, il renvoie des fleurs et d'autres photographies contenant des formes et des couleurs analogues *(voir ci-dessous)*. Ce logiciel possède plusieurs applications, allant du choix des motifs de papiers peints à l'identification policière de criminels par le type de vêtements.

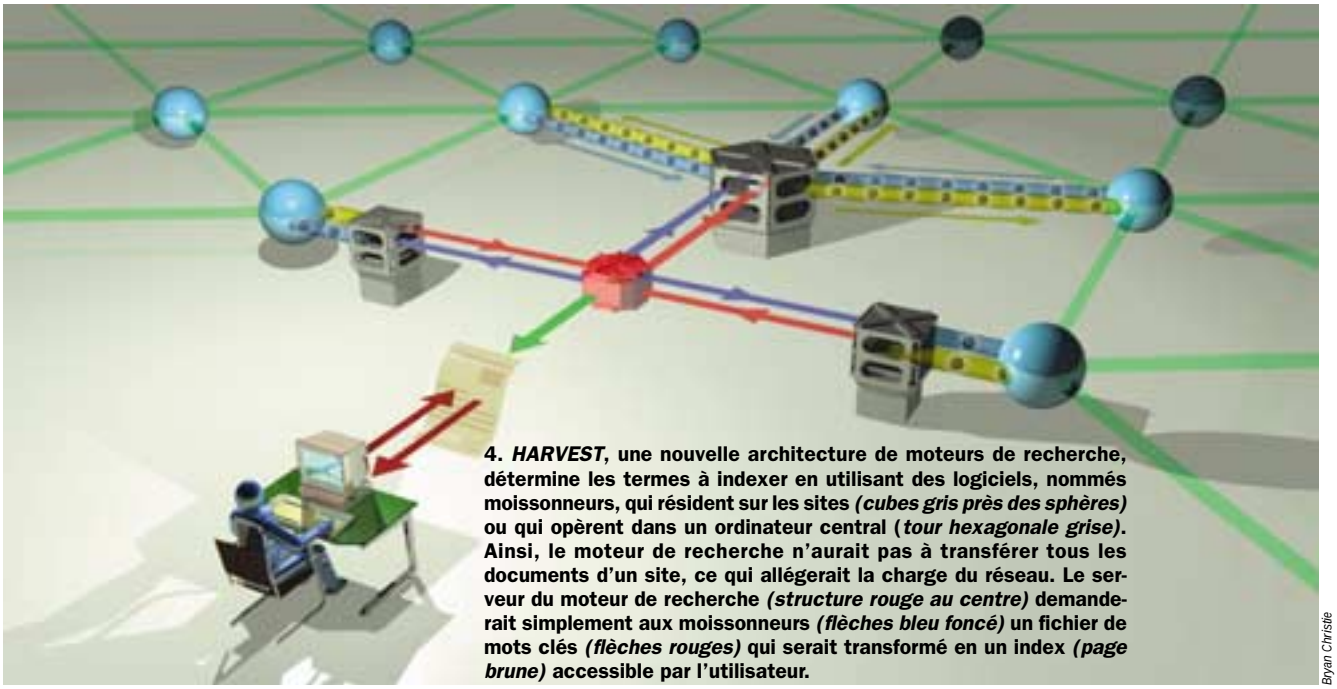
Tous ces programmes cherchent un motif analogue au motif demandé. Ils requièrent encore un observateur humain – ou un texte accompagnateur – pour confirmer si un objet est un chat ou un coussin. Depuis plus de dix ans, des chercheurs en intelligence artificielle essaient de donner aux programmes la possibilité d'établir directement l'identité d'objets présents sur une image. Ils tentent de corréler les formes d'une image avec un modèle géométrique d'objets du monde réel. Les programmes déduisent ensuite, par exemple, qu'un cylindre rose ou brun est un bras humain.

Considérons, par exemple, le logiciel mis au point par David Forsyth, de Berkeley, et par Margaret Fleck, de l'Université d'Iowa, qui recherche des photographies de personnes nues. Il commence par analyser la couleur et la texture d'une photographie. Lorsqu'il trouve des zones de couleur chair, il enclenche un algorithme qui recherche des zones cylindriques pouvant correspondre à un bras ou à une jambe. Il recherche ensuite d'autres cylindres couleur chair, disposés selon certains angles, qui confirmeraient la présence de membres. Lors d'un essai réalisé à l'automne 1996, le programme a retrouvé 43 pour cent des 565 personnes nues présentes sur un total de 4 854 photos – ce qui constitue un bon résultat pour une analyse d'images aussi complexe. En outre, le programme n'a produit que quatre pour cent de fausses réponses pour un ensemble de 4 289 images qui ne contenaient aucune personne nue.

Une dizaine d'années de recherches sera encore nécessaire avant que les utilisateurs du réseau puissent chercher des images, des corps nus, des chats ou des drapeaux aussi facilement qu'ils recherchent du texte. Les informaticiens espèrent qu'ils mettront au point, lors de cette recherche, des programmes qui collecteront l'information en comprenant ce qu'ils voient.



IBM Corporation/Romtech/Corel



4. *HARVEST*, une nouvelle architecture de moteurs de recherche, détermine les termes à indexer en utilisant des logiciels, nommés moissonneurs, qui résident sur les sites *(cubes gris près des sphères)* ou qui opèrent dans un ordinateur central (*tour hexagonale grise)*. Ainsi, le moteur de recherche n'aurait pas à transférer tous les documents d'un site, ce qui allégerait la charge du réseau. Le serveur du moteur de recherche *(structure rouge au centre)* demanderait simplement aux moissonneurs *(flèches bleu foncé)* un fichier de mots clés *(flèches rouges)* qui serait transformé en un index *(page brune)* accessible par l'utilisateur.

Bryan Christie

sujet donné, sans payer les services d'êtres humains chargés de juger la myriade de sites Web. Une solution, qui fait à nouveau appel à l'être humain, consiste à partager entre utilisateurs l'évaluation des sites. Des systèmes d'évaluation numériques récemment mis au point permettent aux utilisateurs de donner leur sentiment sur des sites particuliers (voir *Le filtrage d'information*, par Paul Resnick, page 52).

Certains moteurs explorent l'Internet et, de surcroît, séparent le bon grain de l'ivraie. Il faudrait toutefois de nouveaux programmes pour décongestionner le réseau lors de l'examen répété des sites. Certains gestionnaires de sites affirment en effet que leurs ordinateurs consacrent bien plus de temps à fournir leurs informations aux moteurs qu'aux personnes qu'ils espèrent séduire par leurs offres.

Pour résoudre ce problème, Mike Schwartz et ses collègues de l'Université de Boulder ont écrit un programme, nommé *Harvest*, qui permet à un site de compiler des données d'indexation de ses propres pages et de transmettre, à la demande, ces informations aux différents moteurs de recherche. Ainsi, grâce à *Harvest*, les moteurs n'ont plus à exporter la totalité des sites. Alors que les moteurs de recherche téléchargent un exemplaire de chaque page visitée dans leurs sites d'origine pour en extraire les termes qui constituent un index, ce qui engorge le réseau, *Harvest*, le «programme moissonneur», n'envoie qu'un fichier de termes indexés. En outre, comme il transmet uniquement l'information concernant les pages qui ont été modifiées depuis sa dernière visite, il évite d'encombrer le réseau et les ordinateurs qui lui sont connectés.

Les programmes moissonneurs pourraient également aider les éditeurs à limiter l'information qui quitte leurs sites. Ce contrôle se révèle nécessaire, car le réseau n'est désormais plus un simple support de libre circulation de l'information ; il devient progressivement un moyen d'accéder, contre paiement, à des informations protégées, dont la lecture n'est pas forcément autorisée pour les utilisateurs. Les moissonneurs, eux, distribueraient uniquement l'information que les éditeurs diffusent gratuitement, tels les sommaires et des extraits de l'information stockés sur leurs sites.

Les besoins des utilisateurs détermineront les méthodes de collecte d'information qui seront mises en œuvre. Il est encore trop tôt pour dire si l'Internet ressemblera alors à une bibliothèque, avec des collections structurées, ou s'il restera anarchique, avec un accès fourni par des systèmes automatisés.

Si les utilisateurs acceptent de payer pour protéger le travail des auteurs, alors les éditeurs, les indexeurs et les critiques privilégieront l'information structurée du type bibliothèque classique. En revanche, si l'information est fournie gratuitement ou subventionnée, l'indexation numérisée, peu coûteuse, l'emportera très probablement, et l'environnement anarchique qui caractérise aujourd'hui une grande partie de l'Internet subsistera. Ainsi, plutôt que des problèmes techniques, ce seront surtout des problèmes économiques et sociaux qui détermineront la forme que prendra, dans l'avenir, la recherche de l'information sur l'Internet.

Clifford LYNCH est informaticien à l'Université de Californie.

C.M. BOWMAN et al., *The Harvest Information Discovery and Access System*, in *Computer Network and ISDN Systems*, vol. 28, nos 1-2, pp. 119-125, décembre 1995.

Des informations sur le programme *Harvest* sont disponibles à l'adresse Web http://harvest.transarc.com

Lorcan DEMPSEY et Stuart L. WEIBEL, *The Warwick Metadata Workshop : A Framework for the Deployment of Resource Description*, in *D-lib Magazine*, juillet-août 1996. Les comptes rendus de cette conférence sont disponibles sur http://www.dlib.org/dlib/july96/07contents.html

Carl LAGOZE, *The Warwick Framework : A Container Architecture for Diverse Sets of Metadata*, ibid.

Les bibliothèques numériques

MICHAEL LESK

La quantité d'informations disponible sur l'Internet sera très inférieure à celle des futures bibliothèques. Toutefois, la construction de ces géants ne sera pas aisée.

À Paris, au bord de la Seine, quatre tours dominent le quartier de Tolbiac : avec leurs 395 kilomètres de rayonnages, elles pourront recevoir jusqu'à 22 millions de livres. La bibliothèque François-Mitterrand pourrait être la dernière et la première du genre. La dernière, car la plupart des grandes villes n'ont plus les moyens de construire des bâtiments publics aussi ambitieux. La première, car son installation s'achèvera avec le branchement de centaines d'ordinateurs, qui donneront au public un accès électronique immédiat aux textes de 100 000 ouvrages couvrant une grande part de l'histoire et de la culture françaises.

Dans le monde entier, des bibliothèques se sont lancées dans la tâche herculéenne de numériser des livres, des images et des sons, qui forment la mémoire culturelle de l'humanité. Les utilisateurs des bibliothèques devraient bientôt disposer, sans quitter leur bureau, d'une quantité d'informations bien plus considérable que celle qui est aujourd'hui présente sur l'Internet. Cependant, les nombreux obstacles légaux, économiques et techniques rendent cette perspective encore incertaine.

Les bibliothécaires voient trois avantages évidents à la numérisation. Premièrement, elle leur permettra de préserver des documents fragiles et rares sans en restreindre l'accès. La British Library de Londres, par exemple, détient le seul manuscrit existant du *Lai de Boewulf* (un poème épique anglo-saxon, d'auteur inconnu, écrit entre les VII^{e} et X^{e} siècles) ; seuls des spécialistes étaient autorisés à le consulter jusqu'à ce que Kevin Kiernan, de l'Université du Kentucky, ne numérise ce manuscrit sous trois lumières différentes (révélant des détails qui n'apparaissent pas à l'œil nu) et ne mette ces images sur le réseau Internet. De même, la bibliothèque du Parlement japonais est en train réaliser des clichés numériques, extrêmement détaillés, de 1 236 blocs de bois imprimés, rouleaux et autres matériels qu'elle considère comme des trésors nationaux, afin que les chercheurs puissent les examiner sans manipuler les originaux.

Le deuxième avantage est la commodité : les usagers accèdent aux informations en quelques secondes, au lieu de perdre des heures dans les bibliothèques. Plusieurs personnes peuvent lire simultanément le même livre ou regarder la même image. Les bibliothécaires n'ont plus à replacer les ouvrages sur les rayons.

Le troisième avantage des copies électroniques est d'occuper un espace de quelques millimètres sur un disque magnétique – au lieu de mètres sur des rayonnages. La construction des bâtiments d'une bibliothèque est de plus en plus coûteuse : la bibliothèque François-Mitterrand coûtera plus de sept milliards de francs ; le stockage de chaque volume reviendra en moyenne à 300 francs (dans des bâtiments moins ambitieux, le coût moyen de stockage varie autour de 150 francs). À titre de comparaison, le prix du stockage sur disque revient à environ dix francs pour une publication de 300 pages et ne cesse de diminuer.

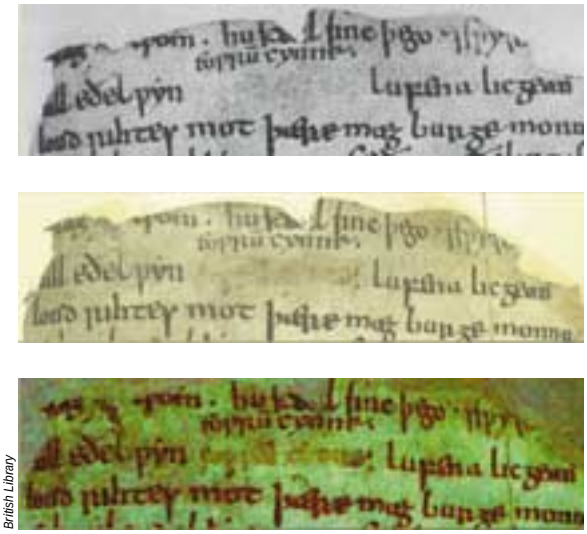
Des compromis techniques

La numérisation sera progressive, car chacune des technique de numérisation est un compromis entre la conservation, la commodité et le coût. La numérisation des pages en vue d'une création d'images numériques est l'opération la moins coûteuse. Elle revient à environ 200 francs par volume, la majeure partie de cette dépense servant à payer les salaires des personnes qui utilisent le scanner. La technique du *photoCD* de *Kodak*, qui numérise automatique-

Agence France-Presse

Michel Gangne, Agence France-Presse

1. LES TOURS de la bibliothèque François-Mitterrand, dans le 13^{e} arrondissement de Paris *(à gauche)*, ont une capacité de 22 millions de livres. Cependant, 100 000 de ces ouvrages, stockés sous forme d'images numérisées, n'occuperont qu'une place infime et seront instantanément accessibles sur des ordinateurs installés dans les salles de lecture.



British Library

2. LE *LAI DE BEOWULF* est éclairé en lumière naturelle *(en haut)*, par l'arrière *(au milieu)* et en lumière ultraviolette *(en bas)*. Ces images numérisées sont aujourd'hui disponibles sur l'Internet (http://www.uky.edu/%7Ekiernan). Elles révèlent des détails qui ne sont pas immédiatement perceptibles sur l'original, conservé à la British Library de Londres.

ment les négatifs photographiques, abaisse considérablement ce coût. Grâce à la *photoCD*, l'Université Harvard a numérisé 80 000 affiches de son fonds juif au rythme de 1 000 par jour et pour seulement dix francs par affiche. Les bibliothèques étant constamment à court d'argent, la plupart des documents qui seront numérisés – notamment 100 000 livres de la bibliothèque François-Mitterrand et cinq millions de titres de la bibliothèque du Congrès américain – le seront de cette façon.

Cette méthode est un bon moyen de conservation, car elle rend fidèlement tout l'aspect du livre et toutes les notes manuscrites qui s'y trouvent ; en outre, sur des images à haute résolution, elle révèle des aspects de fabrication imperceptibles à l'œil nu. En contrepartie, on doit stocker des fichiers informatiques encombrants : une carte de 40 centimètres sur 80 centimètres, avec des caractères de seulement un millimètre de haut ne reste lisible que si l'image comporte 4 500×9 300 pixels, ce qui correspond à un fichier de 125 méga-octets, qui saturerait un ordinateur personnel. On résoudra cette difficulté en utilisant des ordinateurs puissants et de nouveaux formats d'image, qui permettront la visualisation à des résolutions inférieures lorsque la haute résolution n'est pas nécessaire.

Toutefois les textes des images ainsi numérisées ne peuvent être traités : seule leur lecture est possible. Comment, alors, retrouver facilement une citation ou un passage enfouis dans un livre numérisé ? En outre, ces images sont si grosses qu'on ne peut ni les transmettre par courrier électronique ni les «copier-coller» dans un document de travail. Enfin ces images ne sont pas manipulables par les non-voyants.

Dans de nombreux cas, le texte électronique – manipulable et indexable par des programmes de traitement de texte – est plus pratique que des images numérisées. Les logiciels de reconnaissance de caractères transforment le texte imprimé en texte électronique. Ces logiciels, qui utilisent des techniques de reconnaissance de formes pour extraire, dans des images numérisées, des mots lettre par lettre, sont devenus bien moins coûteux et un peu plus précis que naguère. Aujourd'hui, les meilleurs d'entre eux identifient correctement plus de 99 pour cent des caractères d'images tests standards, ce qui signifie qu'ils font encore une dizaine d'erreurs par page. Cette proportion d'erreurs est acceptable si les textes convertis sont seulement indexés ; quand on doit corriger les erreurs (à la main), la numérisation revient au moins aussi cher que la numérisation au scanner.

La Fondation Andrew Mellon a utilisé cette dernière approche pour la numérisation de dix grandes revues d'histoire et d'économie. Bien qu'il en coûte deux francs par page numérisée, corrigée à la main et traitée par un programme de reconnaissance de caractères – soit 600 francs pour un livre de 300 pages –, la Fondation pense rentrer dans ses fonds grâce aux coûts réduits du catalogage et du stockage : sous la forme de fichiers textes, les livres occupent dix fois moins de place que les fichiers images correspondants. En outre, cet investissement permettra aux lecteurs de retrouver bien plus rapidement les articles qui les intéressent.

Pour numériser les livres illustrés, les bibliothèques doivent numériser les textes, d'une part, les dessins et photographies, d'autre part. Les illustrations provoquaient des erreurs dans la plupart des programmes de reconnaissance de caractères. Cependant, de nouveaux logiciels identifient aujourd'hui automatiquement les illustrations et permettent de les conserver en tant qu'images, qui sont ensuite replacées aux endroits appropriés du texte électronique. Cette technique, utilisée par la Société américaine de chimie pour extraire environ 400 000 diagrammes et figures des revues de chimie, repose sur les différences de densité de l'encre sur la page, selon qu'elle correspond à du texte ou à une image : le texte obscurcit la page de manière uniforme et prévisible, tandis que les illustrations apparaissent généralement moins foncées et irrégulières *(voir la figure 3)*.

La méthode de numérisation d'un texte la plus lente est sa saisie sur un clavier. C'est également la méthode la plus coûteuse, même si des entreprises asiatiques emploient des milliers de claavistes sous-payés pour saisir des ouvrages volumineux. La saisie, surtout lorsqu'elle est effectuée par une personne ignorant la langue qu'elle copie, reproduit généralement les fautes d'orthographe et les particularités de l'original, qui sont souvent «corrigées» à tort par le programme de reconnaissance de caractères. Quant aux ouvrages où plusieurs polices de caractères ont des sens spéciaux, la saisie directe est le seul moyen de les numériser. En raison de son coût, la généralisation de cette dernière n'est pas envisageable : la saisie d'un texte standard de 300 pages revient à 3 000 francs ; la description de sa mise en page par le langage hypertexte HTML (*Hypertext Markup Language*, le langage informatique utilisé sur le réseau mondial) ou par l'un de ses nombreux cousins, atteint 4 500 francs, soit 30 fois plus que la simple numérisation des pages.

Aujourd'hui, la plupart des nouveaux documents achetés par les bibliothèques proviennent d'ordinateurs. Aussi les bibliothèques négocient-elles l'obtention des données sur des CD-ROM ou sur d'autres supports informatiques, afin d'éviter le recours à la numérisation. L'Institut des ingénieurs en électricité et en électronique (IEEE), par exemple, programme les 62 revues qu'il publie en langage HTML pour une lecture à

distance. Comme les usagers utilisent plus souvent les documents récents des bibliothèques que les documents anciens, je pense que, vers l'an 2000, la moitié du matériel accessible dans la plupart des grandes bibliothèques sera numérisée. Dix ans supplémentaires seront probablement nécessaires pour que les bibliothèques plus petites suivent le mouvement.

Les incertitudes de l'avenir

Cette transition n'est pas sans risques. Dans les années 1980, quand nombre de bibliothèques entreprirent de numériser leurs ouvrages, les bibliothécaires observèrent que les usagers ignoraient progressivement les titres qui n'étaient pas numérisés. Ainsi, bien qu'elle ait pour ambition de favoriser la culture, la numérisation risque de l'étouffer temporairement.

Le problème des droits d'auteur est également épineux. Lorsque la Société *IBM* prépara un CD-ROM pour célébrer le 500e anniversaire du voyage de Christophe Colomb, elle dut dépenser cinq millions de francs pour obtenir les droits de reproduction nécessaires. Jusqu'ici, la plupart des bibliothèques ont évité les difficultés en ne numérisant que des documents publiés avant 1920 (et donc entrés dans le domaine public). La bibliothèque du Congrès américain, par exemple, a numérisé des milliers de photographies de la guerre de Sécession, des documents du Congrès continental et des discours de la Première Guerre mondiale, mais pas *Autant en emporte le vent*. Lorsque, pour aider les pays en développement, l'Université Cornell décida de numériser les principaux titres de la littérature agronomique parus entre 1850 et 1950, elle évita soigneusement tous les livres protégés par des droits d'auteur.

Si les lois n'autorisent pas les bibliothèques à manipuler des œuvres numérisées aussi facilement que les livres et les revues, les usagers disposeront, sous forme numérique, de tous les documents anciens qu'ils désirent, mais ils ne trouveront les livres publiés entre 1920 et 1990 que sur les rayonnages. Comment, alors, justifier les demandes de crédits de numérisation ? En 1996, *Autant en emporte le vent* a été beaucoup plus consulté par le public américain que n'importe quel discours de la Première Guerre mondiale ; numériser d'abord les ouvrages les plus lus serait une bonne manière d'obtenir un soutien populaire.

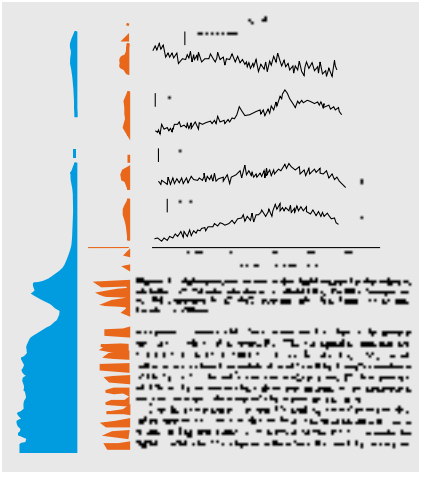
À défaut de ce soutien, des institutions pourraient s'associer pour mettre en commun sur l'Internet leurs collections virtuelles, en se répartissant le coût de création de ces collections. Toutefois, des problèmes d'ordre politique et administratif ont jusqu'ici contrarié les tentatives de coopération : quels doivent être les services offerts par les bibliothèques aux gens qui en sont éloignés ? Comment répartir équitablement les frais d'achat et de stockage des documents entre différentes institutions ?

Lorsque ces obstacles seront surmontés et lorsque seront numérisés des millions de livres, d'images et d'enregistrements, nos enfants pourront-ils les trouver, les voir, les explorer, les imprimer ? Le problème n'est pas tant la durabilité des matériaux physiques que l'obsolescence technique (voir *L'archivage des documents informatiques*, par Jeff Rothenberg, *Pour la Science*, n° 209, mars 1995). Tous les quatre ou cinq ans, les bibliothécaires devront adapter leurs collections aux formats de nouveaux dispositifs. La copie de fichiers d'un dispositif à un autre ne devrait guère poser de difficultés – un bit est un bit –, mais, face à l'abondance et à la constante évolution des formats, on peut être rapidement dépassé.

En outre, certains formats ne sont pas traduisibles en d'autres formats sans perte d'information. Certes, les formats du texte et des images sont de plus en plus standardisés, mais certains de ces formats, tel SGML *(Standardized General Markup Language)* pour le texte, ont une définition si approximative qu'il n'existe aucun programme capable d'afficher correctement toutes ses caractéristiques.

Outre la bibliothèque François-Mitterrand, les bibliothèques municipales de Lyon et de Valenciennes ont également un important programme de numérisation. Les bibliothèques ne sont pas les seuls dépositaires de l'information qui peut être numérisée. Leurs problèmes sont également ceux des organismes publics, tels que les Archives nationales, les archives des services médicaux, les entreprises et même les Mormons lorsqu'ils numérisent les archives des naissances et des mariages à travers le monde.

Les travaux de numérisation dureront plusieurs dizaines d'années et coûteront des milliards de francs. De même qu'aujourd'hui nous écoutons des fugues de Bach jouées aussi bien sur un piano que sur un clavecin, que nous lisons les pièces de Molière sur du papier ou les récitons à voix haute, et que nous regardons un film sur une bande vidéo ou au cinéma, un jour viendra où, grâce aux réseaux, nous serons en mesure d'apprécier, avec un confort inégalé et pour une dépense minime, la richesse de la créativité humaine.



Michael Lesk

3. L'EXTRACTION DES IMAGES d'un texte numérisé s'effectue automatiquement grâce à un logiciel conçu pour la Société américaine de chimie. Ce logiciel mesure la densité de pixels noirs sur chaque bande verticale de la page entière pour identifier les limites des colonnes. Il mesure ensuite la densité de pixels des lignes horizontales *(en orange)* et déduit, à l'aide d'un algorithme, l'espacement régulier séparant les lignes de texte. Le logiciel construit ensuite une fonction *(en bleu)* qui attribue des valeurs élevées au texte uniformément espacé et des valeurs inférieures aux illustrations. En sélectionnant les régions associées aux faibles valeurs, le système isole les images avec une bonne fiabilité.

Michael LESK est informaticien aux Laboratoires *Bellcore*, dans le New Jersey.

Building Large-Scale Digital Libraries, numéro spécial de *Computer*, vol. 29, n° 5, mai 1996.

Michael LESK, *Practical Digital Libraries : Books, Bytes and Bucks*, Morgan Kaufmann (à paraître).

Sur le site de la Bibliothèque nationale de France (http://www.bnf.fr), vous trouverez des expositions virtuelles et, en particulier, 1 000 enluminures de l'époque de Charles V.

Le filtrage d'information

PAUL RESNICK

Grâce aux étiquettes, les utilisateurs des réseaux peuvent vérifier si les logiciels et les sites qu'ils découvrent sont intéressants et exempts de virus.

On lit parfois que le réseau Internet est un village planétaire : ses utilisateurs formeraient une communauté gigantesque, mais soudée, partageant des valeurs et des expériences. Cette idée est trompeuse, car si de nombreuses cultures coexistent sur l'Internet, elles sont parfois en conflit. Sur les forums d'Internet, les utilisateurs interagissent, commercialement et socialement, aussi bien avec des étrangers qu'avec des connaissances ou des amis. L'expression «village planétaire» devrait être remplacée par «ville planétaire», qui fait état de plus de possibilités, mais aussi de plus de dangers.

Pour éviter les voisinages les plus gênants ou les plus dangereux, les utilisateurs disposent de procédures de filtrage qui identifient facilement des risques définis. L'une d'elles est l'analyse des services disponibles : par exemple, les logiciels antiviraux recherchent des fragments de programme qu'ils savent être courants dans les virus informatiques, et certains services, tels *AltaVista* et *Lycos*, mettent en avant ou interdisent l'accès des documents qui contiennent des mots particuliers. Avec mes collègues, nous proposons une autre technique de filtrage, fondée sur la création d'étiquettes électroniques qui sont ajoutées aux sites pour décrire les œuvres numériques. Ces étiquettes indiquent des caractéristiques subjectives (telle page est drôle ou agressive) ou donnent des informations qui ne sont pas directement évidentes à la lecture des pages, telle la politique du site sur l'utilisation ou la vente de données confidentielles.

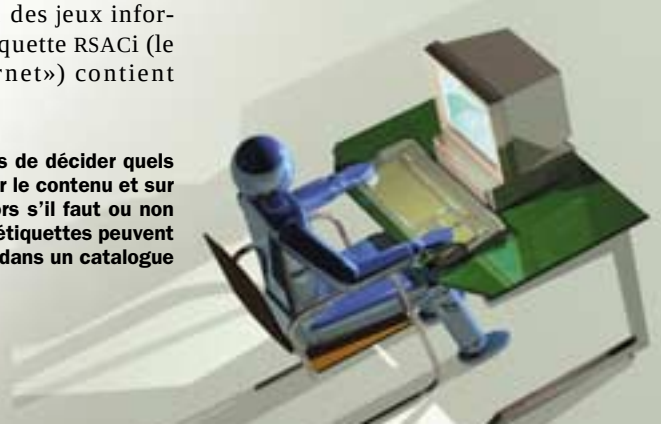
Le consortium World Wide Web, dont l'antenne européenne est située à l'Institut national de recherche en informatique et en automatique (INRIA), a défini un ensemble de critères techniques, nommés PICS (l'acronyme de *Platform for Internet Content Selection*, soit «plate-forme de sélection du contenu d'Internet») qui permettent aux utilisateurs de diffuser électroniquement ces étiquettes sous une forme simple. Les ordinateurs, traitant l'information de ces étiquettes en tâche de fond, protègent automatiquement les usagers contre des documents indésirables ou, à l'inverse, attirent leur attention sur des sites particulièrement intéressants. Initialement, ces étiquettes PICS devaient permettre aux parents et aux enseignants de masquer les documents qu'ils estimaient choquants pour des enfants. Plutôt que d'interdire purement et simplement les sites choquants, elles permettent aux utilisateurs de filtrer ce qu'ils reçoivent.

Les dessous des étiquettes

Les étiquettes PICS peuvent décrire toutes les caractéristiques d'un document ou d'un site informatique. Les premières étiquettes repéraient les documents indécents. Par exemple, le Conseil consultatif sur les logiciels de loisirs (RSAC) adapta à l'Internet son système d'évaluation des jeux informatiques. Chaque étiquette RSACi (le «i» signifiant «Internet») contient quatre nombres qui indiquent respectivement les degrés de violence, de nudité, de sexe et de vulgarité. Une autre organisation, *SafeSurf*, a défini un barème à neuf critères.

Toutefois, les étiquettes permettraient de traiter d'autres problèmes que l'indécence. Un barème de confidentialité, par exemple, décrirait les pratiques des sites, indiquant quelles sont les informations confidentielles collectées par ces sites et quel est l'usage ultérieur de ces informations (fichiers commerciaux, revente, etc.). D'une manière analogue, un barème de propriété intellectuelle décrirait les conditions sous lesquelles on peut regarder ou reproduire un document (voir *Les systèmes sécurisés* par Mark Stefik, page 62). Divers services d'indexation du Web pourraient créer des étiquettes indiquant le sujet des documents offerts par les sites ou la fiabilité des informations qu'ils diffusent.

Ces étiquettes pourraient même protéger les ordinateurs contre les virus. Aujourd'hui, les utilisateurs téléchargent couramment des modules de logiciels informatiques, voire des logiciels complets à partir de sites Internet. Les utilisateurs admettent généralement que les logiciels ainsi récupérés n'introduisent pas de virus ; ils amélioreraient la sécurité de leur ordinateur si le téléchargement était précédé d'une lecture d'étiquettes qui garantiraient que le logiciel ne ren-



1. LE SYSTÈME DE FILTRAGE du réseau mondial permet aux utilisateurs de décider quels types de documents ils veulent recevoir. Ils précisent leurs exigences sur le contenu et sur la sécurité *(a)* ; le logiciel de traitement d'étiquettes *(b)* détermine alors s'il faut ou non interdire l'accès à certaines pages (précédées d'un panneau STOP). Ces étiquettes peuvent être attachées aux documents par l'auteur du site Web *(c)* ou stockées dans un catalogue *(d)* par un organisme d'évaluation.