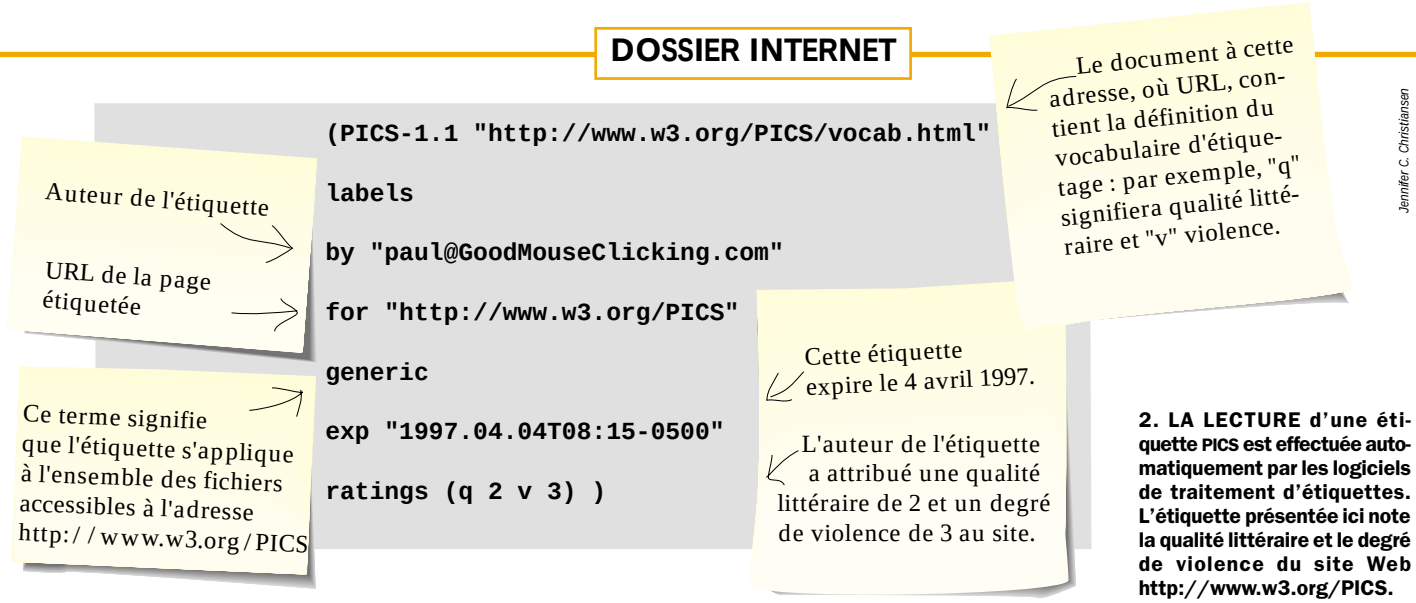


Jennifer C. Christiansen

2. LA LECTURE d'une étiquette PICS est effectuée automatiquement par les logiciels de traitement d'étiquettes. L'étiquette présentée ici note la qualité littéraire et le degré de violence du site Web http://www.w3.org/PICS.

qu'il n'approuve pas les critères qu'elles retiennent. Ainsi, des systèmes tels que *GroupLens* et *Firefly* recommandent des choix de livres, d'articles, de vidéos ou d'œuvres musicales fondés sur les appréciations de personnes partageant des goûts semblables : après que les utilisateurs ont évalué des documents traitant de sujets qui leur sont familiers, le logiciel compare leurs appréciations à celles d'autres utilisateurs et il privilégie ultérieurement les documents approuvés par les personnes qui partagent les appréciations que les utilisateurs portent sur d'autres documents. Les utilisateurs qui ont les mêmes goûts ou opinions n'ont pas à se connaître ; chacun peut participer anonymement à l'étiquetage, préservant ainsi la confidentialité de ses modes d'évaluation et de leur résultat.

L'emploi des étiquettes pose plusieurs problèmes politiques et sociaux. Tout d'abord, qui décide des procédures d'étiquetage et des étiquettes que l'on doit retenir ? En théorie, n'importe qui pourrait étiqueter un site et définir ses propres critères de filtrage. En pratique, l'attribution des étiquettes ou le choix des critères pourraient être imposés par les autorités. Cela s'est déjà produit pour la télévision : sous la pression du Conseil supérieur de l'audiovisuel, les chaînes ont entrepris d'évaluer l'adéquation de leurs émissions à l'âge des téléspectateurs.

L'auto-étiquetage obligatoire ne débouche pas sur la censure, tant que les individus peuvent choisir eux-mêmes quelles étiquettes ignorer. Malheureusement, les gens n'ont pas toujours ce pouvoir. Si un contrôle individuel amélioré est un argument contre un contrôle central, il n'empêche pas sa mise en place. À Singapour et en Chine, par exemple, on expérimente des «garde-fous», matérialisés par des programmes et par des cartes électroniques qui interdisent l'accès à certains forums de discussions et sites du réseau.

Toutefois, même en l'absence d'une censure centrale, la mise en place d'un barème consensuel, quel qu'il soit, encouragera les utilisateurs à faire des choix laxistes qui ne reflèteront pas les valeurs auxquelles ils croient. Aujourd'hui, nombre de parents qui n'approuvent pas obligatoirement les critères de classement des films, interdisent néanmoins à leurs enfants de voir des films «interdits aux moins de 12 ans» ou «interdits au moins de 16 ans» : ils ne prennent pas le temps d'évaluer eux-mêmes ces films.

Les organismes d'étiquetage doivent choisir leur vocabulaire avec discernement, de manière à correspondre aux critères de la plupart. Même ainsi, un vocabulaire unique ne peut pas répondre aux besoins de tous. Les étiquettes mentionnant uniquement le degré de pornographie des sites n'aident pas les utilisateurs préoccupés par le degré de violence verbale. De plus, aucun barème ne peut remplacer une évaluation précise : les comptes rendus de films publiés par les journaux sont bien plus éclairants que les codes parfois utilisés pour classer les œuvres cinématographiques.

De surcroît, les étiquetages, quels que soient les soins apportés à leur conception et à leur application, risquent d'étouffer les informations non commerciales. Un étiquetage bien fait nécessite du temps et de l'énergie : nombre de sites d'intérêt limité ne seront probablement jamais étiquetés. Par souci de sécurité, certaines personnes bloqueront l'accès aux documents dépourvus d'étiquette ou dont les étiquettes ne leur semblent pas fiables. Pour elles, Internet fonctionnera davantage comme la télévision, offrant uniquement accès aux documents suffisamment porteurs d'audience pour qu'il soit financièrement rentable de les étiqueter.

Ce problème inévitable n'est pas dû à l'étiquetage : quel que soit le média utilisé, les gens tendent à éviter l'inconnu lorsqu'il y a des risques, et il est bien plus facile d'obtenir des informations sur des documents destinés à une grande audience que sur des sujets confidentiels.

Si le réseau Internet supprime les barrières techniques de communication, il n'élimine pas pour autant les coûts sociaux de cette communication. Les étiquettes peuvent réduire ces coûts, en aidant les utilisateurs à décider de la confiance qu'ils accordent aux logiciels ou aux sites du réseau. Aux utilisateurs de faire que les systèmes d'étiquetage guident leur exploration sans les empêcher de voyager.

Paul RESNICK, président du Consortium World Wide Web, est informaticien au laboratoire de la Société *AT&T*.

Jonathan WEINBERG, *Rating the Net*, in *Hastings Communications and Entertainment Law Journal*, vol. 19, mars 1997. Disponible sur le Web, à l'adresse : http://www.msen.com/~weinberg/rating.htm

Les spécifications du PICS sont consultables sur http://www.w3.org/PICS

Le multilinguisme

BRUNO OUDET

La multitude de langues parlées à travers le monde résistera-t-elle à l'impérialisme de l'anglais sur l'Internet ?

Ces dernières années, la culture américaine a accru son influence planétaire grâce au commerce international et aux productions cinématographiques. Alors que l'Internet atteint des endroits de plus en plus reculés du Globe, une question se pose : renforcera-t-il cette influence au point que l'anglais deviendra la langue dominante dans les échanges? Ou l'univers de la communication s'enrichira-t-il d'une multitude de langues? Certains observateurs pensent que les langues locales ne survivront pas sur l'Internet et que l'anglais dominera.

Une telle domination aurait des inconvénients. Comme la plupart des utilisateurs de l'Internet ne sont pas de langue anglaise, leur maîtrise de l'anglais est souvent limitée. Pour les discussions complexes, le recours à la langue maternelle s'impose ; mais si l'Internet ne permet pas des conversations multilingues, il ne sera pas le catalyseur attendu des communications internationales. Les erreurs et les malentendus deviendront fréquents, et nombre d'utilisateurs seront privés des fabuleuses potentialités des communications internationales.

Plusieurs forces détermineront la diversité linguistique du réseau. Aujourd'hui, 60 pour cent des ordinateurs reliés à l'Internet sont situés aux États-Unis. Hors d'Amérique du Nord et de quelques pays européens, la plupart des branchements sur l'Internet sont récents et encore limités. Cependant, le nombre de connexions augmente partout. Avec la chute constante du coût d'installation des réseaux de communication, la répartition des utilisateurs approchera bientôt celle des ordinateurs dans le monde.

Grâce à son faible coût et à sa facilité d'emploi, l'Internet permet à certains écrivains – particulièrement ceux qui utilisent l'alphabet latin – de publier leurs œuvres ou d'échanger des messages dans leur propre langue. Certains défenseurs des langues vernaculaires utilisent déjà ce réseau à leur avantage. Par exemple, bien que les Canadiens francophones ne représentent que cinq pour cent de la population francophone, 30 pour cent de toutes les pages publiées en français sur le réseau proviennent du Québec. L'extension mondiale de l'Internet favorise également une langue qui peut être, du moins superficiellement, comprise par le plus grand nombre. Ainsi, je crois que l'Internet véhiculera de nombreuses langues pour les communications locales, tandis que l'anglais servira aux conversations internationales simples.

Les difficultés techniques de la communication dans la plupart des langues de la planète ne sont pas résolues. Les premières machines et les premiers logiciels furent conçus pour traiter des textes en anglais. Les difficultés subsistent même avec les caractères latins standards. Au début d'Arpanet – le réseau qui précéda Internet –, on ne pouvait transmettre que des messages électroniques codés en ASCII (l'acronyme de cette norme américaine de codage des caractères) à sept bits : dans ce code, chaque caractère est défini par une chaîne de sept chiffres binaires, ce qui permet le codage de 2^7, soit 128 caractères. Aujourd'hui le nouveau protocole de transport de courrier (ESMTP) permet de traiter les huit bits nécessaires au codage en ISO-latin, l'une des normes proposées par l'Organisation internationale de la standardisation (ISO) : ce standard autorisant 256 caractères, on peut transmettre et afficher les signes diacritiques (par exemple, les accents sur les lettres) de toutes les langues d'Europe occidentale. Toutefois, comme de nombreux ordinateurs reliés par l'Internet ont des logiciels obsolètes, le huitième bit disparaît parfois, ce qui rend le message quasi incompréhensible. En 1995, sur les 12 000 utilisateurs qui recevaient les

Alis Technologies

1. SUR CE SITE, qui utilise Unicode, une norme de codage des caractères de toutes les langues du monde, l'utilisateur peut choisir la langue.

Accentsoft

2. LA SOCIÉTÉ *ACCENTSOFT* propose des outils multilingues, tel le navigateur Web *Mosaic Multingue*.

nouvelles quotidiennes en français par l'Internet, 8 500 avaient réclamé une version codée en ASCII sept bits, plutôt que la version estropiée en ISO-latin.

Bien que certains logiciels récents puissent afficher leurs textes dans de nombreux alphabets différents, la plupart sont bilingues et traitent seulement une langue locale, par exemple le japonais, plus l'anglais. Les obstacles techniques à l'affichage d'autres alphabets sont-ils temporaires? Unicode (ISO 10646), un codage de la plupart des caractères du monde, permet de coder presque n'importe quelle langue (mais sans résoudre la question de son affichage). Cependant un long chemin reste à parcourir avant que l'Internet ne devienne réellement multilingue et que l'on puisse, par exemple, insérer une citation en grec à l'intérieur d'un texte russe qui sera correctement affiché sur l'écran d'un utilisateur sud-américain. Des logiciels dotés de ces fonctionnalités apparaissent, mais, dans leur concurrence, les grands constructeurs de logiciels produisent sans cesse de nouvelles versions, ce qui ne laisse guère de chances de survie aux petites entreprises qui conçoivent des produits multilingues.

Puisque les interprètes aident à surmonter les barrières du langage, ne pourrait-on pas les utiliser pour traduire les documents échangés sur l'Internet? Le volume et la diversité des échanges limite leur rôle. Seuls les programmes de traduction assistée par ordinateur nous rapprocheront d'un monde – peut-être utopique – où chaque participant d'une conférence virtuelle des Nations unies utiliserait sa langue maternelle, qui serait alors traduite simultanément dans toutes les autres langues.

Malgré 50 ans de recherches, la traduction assistée par ordinateur reste insatisfaisante. Aujourd'hui, les systèmes de traduction ne sont utilisés qu'au Japon, au Canada et en Europe. Généralement, les interprètes électroniques sont uniquement bilingues, et ils doivent être extrêmement spécialisés pour livrer des traductions suffisamment bonnes pour être affinées par un être humain.

Le premier système public fut *Systran*, qui traduisait 14 paires de langues dès 1983, sur le réseau Minitel. Utilisé par la Commission européenne, *Systran* traduit aujourd'hui des centaines de milliers de pages par an. Le système *Meteo*, une autre grande réussite, traduit les bulletins météorologiques canadiens de l'anglais en français et inversement. Il traite quotidiennement 80 000 mots (environ 400 bulletins), avec seulement trois à cinq interventions humaines pour 100 mots.

Plusieurs groupes de recherche élaborent aujourd'hui une procédure à deux étapes dont bénéficiera la traduction multilingue. Dans cette procédure, on analyse d'abord entièrement le texte : on le décompose en ses éléments constitutifs (titre, paragraphe, phrase), et les ambiguïtés potentielles sont levées par un dialogue avec l'auteur. Le texte est alors traduit dans une forme intermédiaire abstraite, qui sert de base pour les traductions dans différentes langues. La dépense occasionnée est justifiée lorsque le texte doit être traduit en plus de dix langues. L'Université des Nations Unies, à Tokyo, a récemment proposé un projet qui devrait, en l'espace de dix ans, concrétiser ce système.

Toutefois, le réseau Internet ne deviendra véritablement multilingue que s'il existe une concertation internationale. Lui donnerons-nous une priorité suffisante? La réponse à cette question est encore floue. Il nous est si facile de nous laisser dériver vers l'anglais en tant qu'unique langue commune.

Bruno OUDET est professeur à l'Université Joseph-Fourier de Grenoble et chercheur au Laboratoire Leibniz de l'IMAG. Il est actuellement président de la Société française de l'Internet.

Les interfaces de recherche

MARTI HEARST

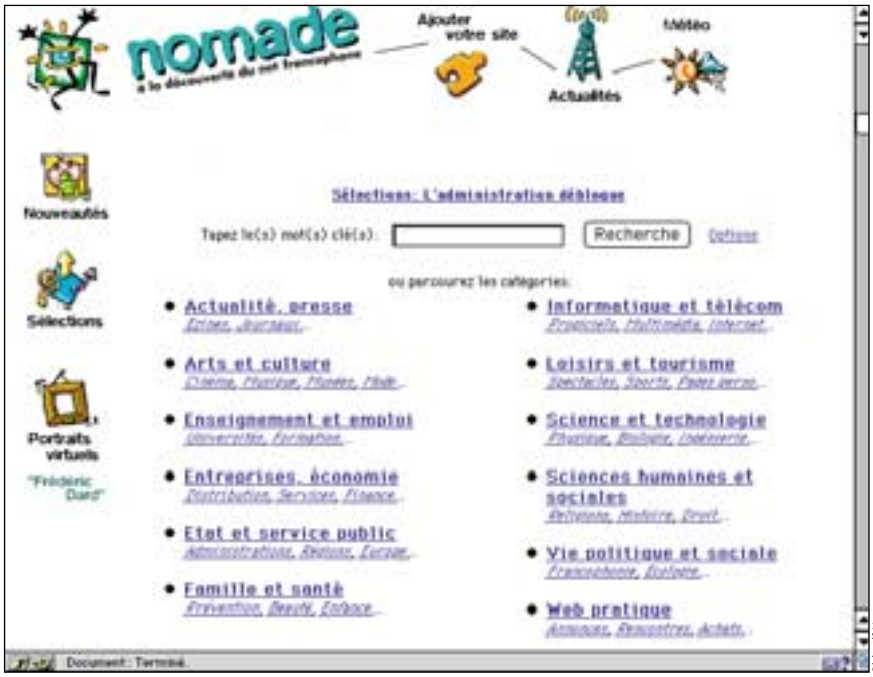
Le rapide développement du réseau mondial gêne sa rationalisation et son organisation. De nouvelles interfaces faciliteraient son emploi.

Comment éviter de se perdre dans l'immensité du réseau mondial? Nombre de sites permettent de retrouver facilement certaines informations courantes et sommaires, tels des numéros de téléphone ou les cours de la Bourse, mais le réseau Internet est remarquable parce qu'il ne nous cantonne pas à un quotidien prosaïque et qu'il fait venir sur nos bureaux des informations extraordinairement variées. Pourtant, par manque d'organisation cohérente, le cyberespace devient de plus en plus confus. Les outils qui servent aujourd'hui à localiser un document dans l'Oregon, un catalogue en Grande-Bretagne ou une image au Japon sont souvent d'une lenteur désespérante.

Quelques algorithmes classent les résultats des recherches d'après leur pertinence, mais le remède à la lenteur du réseau résidera plus probablement dans l'emploi de nouveaux programmes nommés interfaces utilisateur. Aujourd'hui des logiciels d'analyse de textes et de manipulation de données hiérarchisées facilitent la recherche sur le réseau Internet ou dans de grandes bases de données, car la plupart des sites informatiques sont organisés en «pages», une structure simple et familière, mais inutilement restrictive. Dans l'avenir, cette présentation sera remplacée par des structures qui montreront les informations de plusieurs points de vue simultanés. Imaginons qu'Alice, dans la Creuse, se connecte au réseau afin de savoir quels types de bulbes comestibles, tels l'ail ou l'oignon, elle peut planter dans son jardin. Les réponses à sa question se trouvent certainement sur le réseau, mais où?

Alice dispose aujourd'hui de plusieurs techniques pour s'y retrouver, mais aucune n'est parfaitement satisfaisante. Elle peut d'abord demander à des amis s'ils connaissent des sites intéressants. Elle peut aussi feuilleter des index du réseau, qui sont aujourd'hui de deux types : soit des listes construites manuellement, qui classent les sites par type de contenu, soit des moteurs de recherche qui repèrent rapidement des mots clés dans les index du réseau.

Grâce à leurs employés qui répertorient quotidiennement des centaines de documents, des sites, tels *Yahoo* et *Nomade*, proposent aux utilisateurs une table des matières. Pour y rechercher l'information, on choisit dans un menu *(voir la figure 1)* la catégorie qui semble la plus proche de celle que l'on cherche et l'on reçoit un sous-menu plus spécialisé ou une liste de sites dont le personnel de la société pense qu'ils appartiennent à cette catégorie. Toutefois l'emploi de ces tables des matières n'est pas très simple, car les catégories ne sont pas toujours mutuellement exclusives : Alice doit-elle choisir la rubrique «Loisirs», ou bien la rubrique «Écologie», par exemple? Quel que soit son choix, le menu disparaît de l'écran, de sorte qu'en cas d'échec elle doit refaire toute la recherche, et répondre à toutes les questions posées par l'interface de recherche. Si l'un de ses choix n'aboutit pas (si la rubrique «Écologie» ne contient pas l'information cherchée, par exemple), elle doit repartir à zéro et en essayer une autre. Si l'information désirée est enfouie dans les profondeurs de la hiérarchie ou n'existe pas, la procédure peut être fastidieuse, voire exaspérante.



1. LE SYSTÈME de mise en page utilisé sur la plupart des sites du réseau impose des contraintes superflues aux services, tels *Yahoo* ou *Nomade*, qui tentent de structurer le contenu du réseau, afin de faciliter la recherche d'informations. Un utilisateur qui désire connaître les type de bulbes comestibles (ail, oignon, etc.) qu'il peut planter en automne, par exemple, peut choisir une catégorie. Plus il précise sa demande, plus il lui est difficile de se rappeler quelles catégories potentiellement utiles il n'a pas encore explorées.

Marti A. Hearst, Xerox Parc

2. DES PROGRAMMES de recherche facilitent l'exploration du réseau en créant des livres virtuels qui contiennent les pages trouvées, ainsi que des liens vers des sites. Ces livres qui stockent les résultats des recherches sont placés sur des étagères virtuelles, où l'on peut les reprendre ultérieurement. Ils peuvent également guider l'utilisateur pour la recherche de nouvelles informations.

La lenteur de la pensée

Durant la dernière décennie, les recherches sur la visualisation de l'information ont engendré plusieurs techniques efficaces pour l'affichage d'ensembles de données abstraites sous une forme plus propice à l'exploration. L'une des stratégies consiste à transformer le travail demandé à l'utilisateur, en remplaçant les processus lents et intellectuellement exigeants – telle la lecture – par des processus perceptifs plus rapides, comme la reconnaissance de formes. Ainsi on compare plus facilement les barres d'un histogramme que les nombres d'une liste. La couleur, également, facilite la détection des mots ou objets dans un océan de mots ou d'objets.

Une autre stratégie consiste à exploiter l'illusion de profondeur que peuvent donner les écrans : quand l'affichage donne une image bidimensionnelle d'un monde à trois dimensions, la perspective, le recouvrement et les ombrages fournissent des indices perceptifs qui clarifient les relations entre des documents qui formeraient une mosaïque indéchiffrable dans une page à deux dimensions ; on peut mettre au premier plan les informations les plus intéressantes et rejeter à l'arrière-plan ou à la périphérie de l'écran les informations secondaires ou celles qui ont conduit aux informations cherchées. Ainsi, l'utilisateur garde une vision structurée de sa recherche.

Un tel environnement facilite l'exploration. Les utilisateurs peuvent dénicher des résultats partiels qu'ils utiliseront ultérieurement, mieux formuler leurs demandes, reprendre des voies qu'ils avaient jugées inintéressantes ou envisager leurs demandes sous un angle nouveau. Aujourd'hui Alice peut avoir cette démarche si elle prend des notes à mesure qu'elle utilise *Yahoo*, mais un prototype conçu par mes collègues du Centre de recherches *Xerox* vise à accroître l'efficacité de ces activités intelligentes.

Nommée *Information Visualizer*, ce programme dessine un arbre tridimensionnel *(voir la figure 3)* qui relie chaque catégorie à toutes ses sous-catégories. Si Alice cherche l'arbre *Yahoo* pour «jardins», les six zones de *Yahoo* dont «jardin» ou «jardinage» sont une sous-catégorie s'affichent simultanément. En faisant tourner l'arbre, Alice peut amener chacune de ces catégories au premier plan pour découvrir où elle conduit. Lorsqu'un chemin aboutit à une impasse, il suffit de cliquer une fois pour retrouver les voies non explorées.

Lorsque Alice trouve une information intéressante, elle peut utiliser le programme pour la stocker dans un livre virtuel, avec les mots clés qui ont été utilisés ; puis elle peut placer ce livre sur une étagère virtuelle, où il est visible et nettement étiqueté. Le week-end suivant, par exemple, elle peut reprendre sa recherche à l'endroit où elle l'a laissée : il lui suffit d'ouvrir le livre et de relancer la recherche à partir d'une de ses pages.

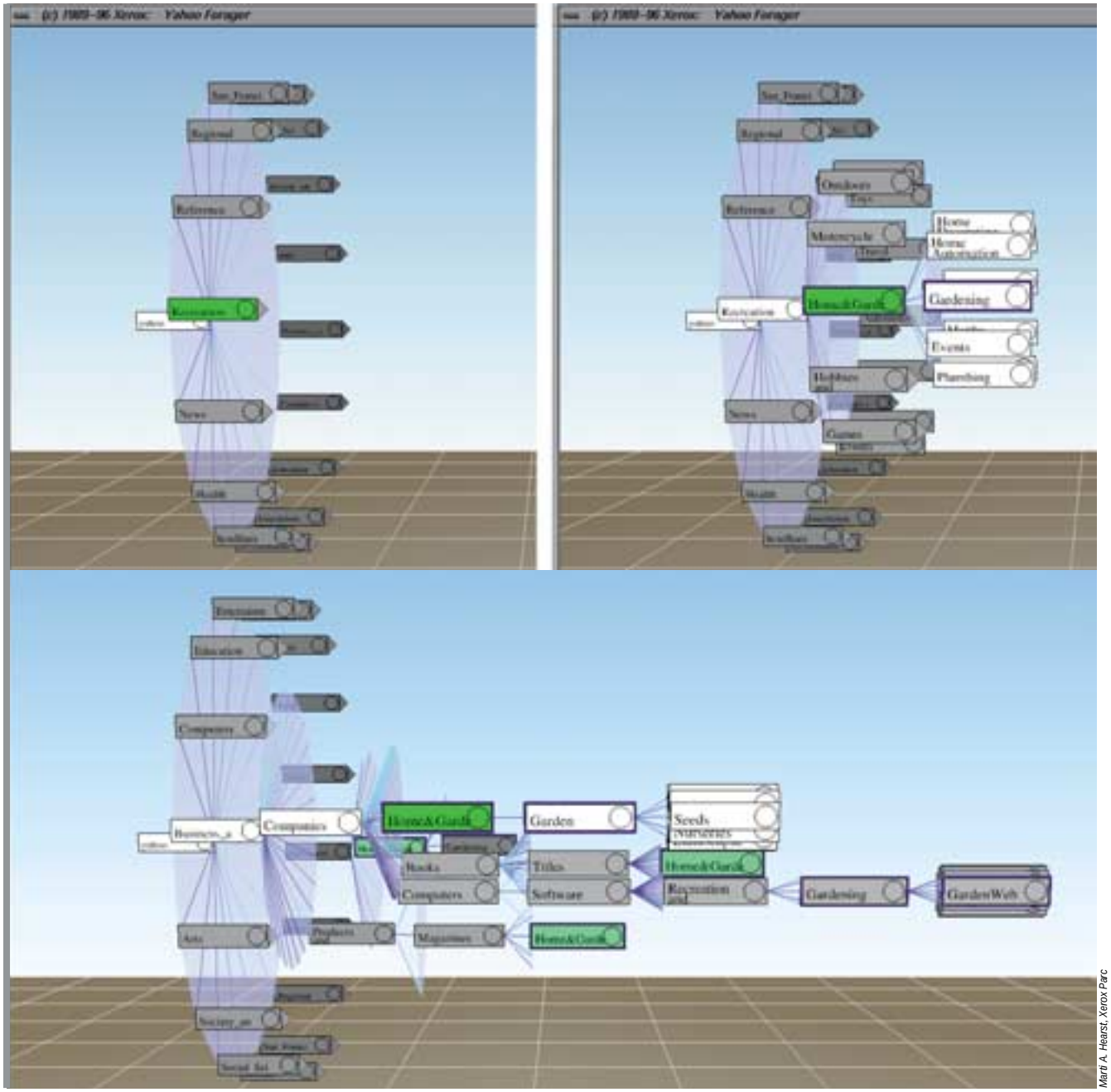
Notre interface ne pourrait servir à structurer tout le réseau : les sites nouveaux apparaissant bien trop rapidement pour être indexés à la main, la proportion de sites répertoriés par *Yahoo* (ou par tout autre service) diminue de jour en jour. En outre, certains sites, tel celui des quotidiens, contiennent des articles sur tant de sujets qu'ils ne sont souvent rangés que dans quelques-unes des catégories possibles.

Les moteurs de recherche tels que *Excite* et *AltaVista* sont bien plus généralistes, mais c'est là leur faiblesse. Si Alice fournissait à *Excite* la chaîne de mots clés «ail, oignon, automne, jardin croissance», elle récupérerait 583 430 pages, ce qui (au rythme de deux minutes par page) lui demanderait plus de deux années ininterrompues de feuilletage. Plus généralement, toute recherche exhaustive procure inévitablement des listes encombrées d'informations superflues, tandis que les recherches plus ciblées risquent de passer à côté de nombreuses pages intéressantes.

Structurer les résultats de recherche

La concision des formulaires d'entrée de la plupart des services de recherche aggrave le problème, parce qu'elle incite les utilisateurs à faire des demandes lapidaires et nécessairement vagues. On devrait encourager l'emploi d'opérateurs logiques tels que AND, OR et NOT pour spécifier les mots qui doivent (ou ne doivent pas) figurer sur les pages affichées. Cependant cette notation booléenne intimide ou déroute de nombreux utilisateurs, et même les spécialistes n'obtiennent que des réponses dont la qualité dépend des termes utilisés dans leur demande. Une meilleure solution consiste à

construire des interfaces utilisateur qui ordonnent les immenses stocks d'informations produites par les recherches sur le réseau. Certains algorithmes regroupent automatiquement les pages en catégories, par exemple, mais ils ne savent pas encore classer un même texte dans plusieurs catégories. Si l'on peut souvent attribuer une place unique aux objets réels dans les taxonomies (un oignon est un type de plante potagère), aucune page du réseau ne traitera probablement que des oignons. Généralement, ces pages contiennent des noms de distributeurs du produit, des recettes de soupe à l'oignon ou présentent des débats opposant les légumes importés aux légumes indigènes, de sorte que la même information peut se trouver à la fois dans deux catégories : par exemple, une recette de soupe à l'oignon peut se trouver référencée à «oignon» ou à «soupe». La seule façon d'éviter cette double catégorisation consiste à créer une catégorie plus spécifique (par exemple, «distributeurs d'oignons»), mais il serait plus approprié de décrire



Marti A. Hearst, Xerox Parc

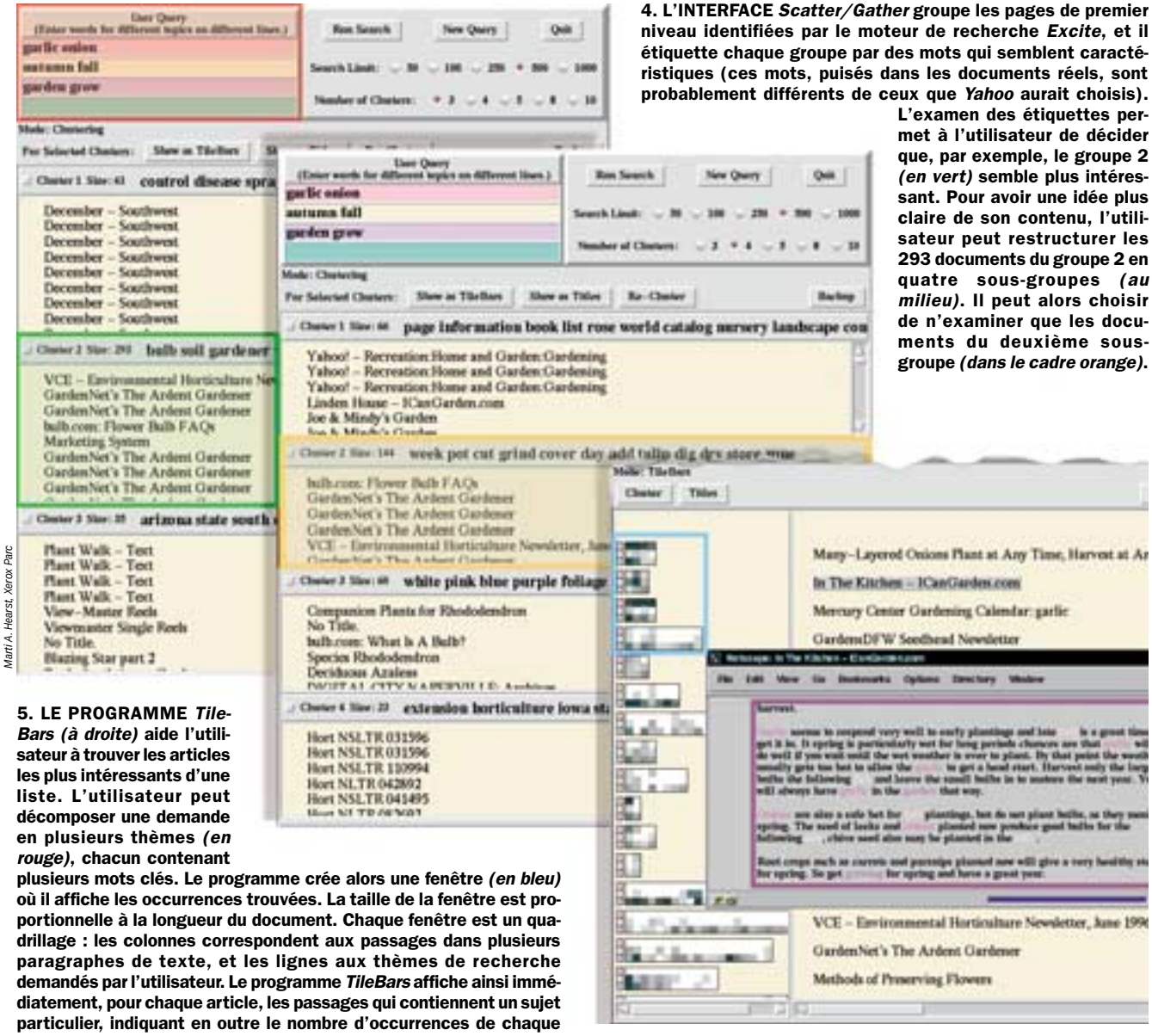
3. POUR CLARIFIER des hiérarchies complexes, telle l'arborescence donnée par le service *Yahoo*, certaines interfaces de recherche proposent des affichages animés tridimensionnels. Les étiquettes des catégories, disposées horizontalement, forment une structure arborescente *(en haut à gauche)*. Une seule étiquette est active *(celle du premier plan)*, mais on fait venir une des étiquettes d'arrière-plan en cliquant dessus : le sous-arbre qui la contient tourne, et montre simultanément toutes les sous-catégories de la catégorie activée. La recherche d'un nom de catégorie (par exemple, «jardin») encadre en violet toutes les étiquettes correspondantes de la hiérarchie *(en bas à droite)*. Contrairement aux pages données par *Yahoo*, cette interface tridimensionnelle permet à l'utilisateur de passer immédiatement à un niveau quelconque de la hiérarchie et de faire des recherches simultanées dans plusieurs catégories.

les documents par la totalité des catégories qui leur correspondent, ainsi que par un autre ensemble d'attributs (par exemple, l'origine, la date, le genre et l'auteur). Les informaticiens de la bibliothèque de l'Université Stanford élaborent actuellement une interface nommée *SenseMaker*, fondée sur ce principe.

Au Centre de recherches *Xerox*, nous avons mis au point une autre procédure pour grouper les pages fournies par un moteur de recherche. Cette procédure nommée *Scatter/Gather* construit une table des matières qui se modifie à mesure que l'utilisateur comprend mieux la nature des documents disponibles et découvre ceux qui sont les plus intéressants.

Imaginons qu'effectuant une recherche à l'aide du programme *Excite*, Alice reçoive les 500 premières pages proposées par le réseau. Le système *Scatter/Gather* examine alors ces pages et les répartit en trois à cinq groupes selon leurs ressemblances *(voir la figure 4)*. Alice peut alors examiner ces groupes et choisir ceux qui lui semblent plus intéressants.

Bien que le comportement des utilisateurs soit variable, les expériences indiquent que les arborescences facilitent l'accès aux documents intéressants. Si Alice juge que le groupe des 293 textes résumés par «bulbes», «sol» et «jardinier» est prometteur, par exemple, elle peut utiliser *Scatter/Gather* pour le parcourir et le diviser en sous-groupes plus petits. Après quelques subdivisions, elle ramène son ensemble de 500 pages, pour la plupart sans intérêt, à un groupe de quelques dizaines de pages qui, elles, sont intéressantes.



Marti A. Hearst, Xerox Parc

4. L'INTERFACE *Scatter/Gather* groupe les pages de premier niveau identifiées par le moteur de recherche *Excite*, et il étiquette chaque groupe par des mots qui semblent caractéristiques (ces mots, puisés dans les documents réels, sont probablement différents de ceux que *Yahoo* aurait choisis). L'examen des étiquettes permet à l'utilisateur de décider que, par exemple, le groupe 2 *(en vert)* semble plus intéressant. Pour avoir une idée plus claire de son contenu, l'utilisateur peut restructurer les 293 documents du groupe 2 en quatre sous-groupes *(au milieu)*. Il peut alors choisir de n'examiner que les documents du deuxième sous-groupe *(dans le cadre orange)*.

5. LE PROGRAMME *TileBars (à droite)* aide l'utilisateur à trouver les articles les plus intéressants d'une liste. L'utilisateur peut décomposer une demande en plusieurs thèmes *(en rouge)*, chacun contenant plusieurs mots clés. Le programme crée alors une fenêtre *(en bleu)* où il affiche les occurrences trouvées. La taille de la fenêtre est proportionnelle à la longueur du document. Chaque fenêtre est un quadrillage : les colonnes correspondent aux passages dans plusieurs paragraphes de texte, et les lignes aux thèmes de recherche demandés par l'utilisateur. Le programme *TileBars* affiche ainsi immédiatement, pour chaque article, les passages qui contiennent un sujet particulier, indiquant en outre le nombre d'occurrences de chaque sujet (plus les carrés sont sombres, plus les occurrences sont nombreuses). Ici, la ligne supérieure de chaque rectangle indique le nombre d'occurrences des mots «ail» ou «oignon» dans le premier thème ; la ligne du milieu correspond au nombre d'occurrences du mot «automne», etc. Cette interface guide les recherches. Les deux articles supérieurs, où les trois sujets figurent dans plusieurs passages, seront probablement plus intéressants que le document inférieur, qui ne contient aucune mention des mots «ail» ou «oignon». En cliquant sur un carreau d'une fenêtre, on affiche une fenêtre *(en mauve)* où s'affiche le passage correspondant à l'intérieur du texte, avec les mots clés mis en relief par des couleurs appropriées.

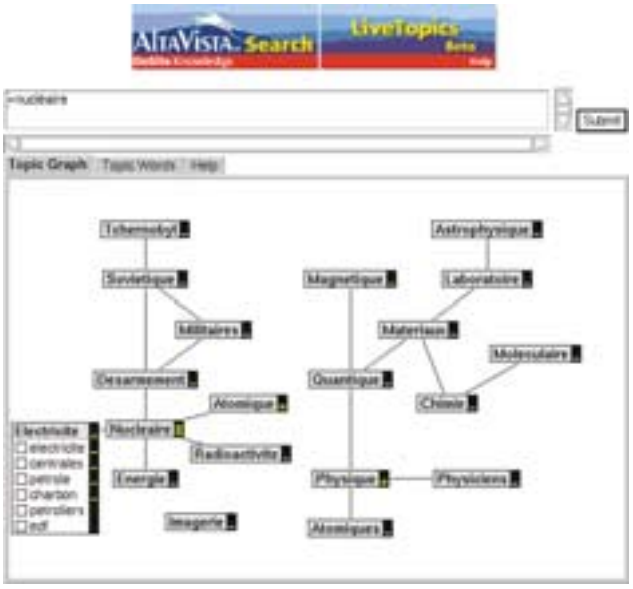
Cartographie thématique

Dans une approche différente de celle des programmes qui aident à visualiser la classification *a priori* des thèmes d'une recherche, tels ceux décrits dans cet article, nous avons conçu un logiciel, nommé *LiveTopics*, qui propose à l'utilisateur, submergé par l'avalanche de documents renvoyés par sa requête, une cartographie thématique et dynamique des documents. À l'aide de cette carte, l'utilisateur appréhende rapidement le contenu des documents. En incluant ou en excluant certains des thèmes de la carte, il affine sa recherche.

Ce logiciel est aujourd'hui proposé sur le moteur de recherche *AltaVista* de la Société *Digital Equipment*. Lorsque l'utilisateur reçoit plus d'une vingtaine de réponses, il peut demander à visualiser une synthèse thématique de l'ensemble des réponses. Par exemple, la recherche du mot «nucléaire» renvoie des thèmes variés où l'on retrouve ce qui concerne la physique, l'énergie, l'électricité et l'astrophysique *(voir la figure)*. L'utilisateur passe ensuite en revue chacun des thèmes et les mots qui s'y rapportent. Il peut alors modifier sa requête d'un simple clic de la souris, en incluant ou en excluant certains mots, et lancer une nouvelle recherche qui pourra de nouveau donner lieu à une analyse thématique.

Fondé sur une analyse statistique du texte intégral des documents, *LiveTopics* n'utilise aucune analyse linguistique ni sémantique, ce qui rend possible l'analyse de documents indépendamment de leur langue, pourvu que cette langue s'écrive avec des caractères latins. Une évolution vers des caractères d'autres langues est à l'étude, ainsi qu'une version de la technique adaptée au monde des réseaux d'entreprises.

François BOURDONCLE, École des mines de Paris



Dans le moteur de recherche *AltaVista* (http://www.altavista.digital.com/), la demande «nucléaire» renvoie 11 877 documents. Lorsque on lance *Live Topics*, une carte des thèmes apparaît. À gauche, on trouve le thème *Nucléaire* relié à *Énergie*, *Électricité*, *Radioactivité*, mais aussi le thème du désarmement, lié à *Soviétique*, *Tchernobyl* et *Militaires*. À droite de la carte apparaissent plutôt les thèmes liés à la recherche : *Quantique*, *Magnétique*, *Physique*, *Chimie*, etc. Notez le thème *Imagerie* qui est isolé du reste. Sur l'écran de droite, on trouve la liste des mots associés aux thèmes de la demande «nucléaire». On peut les exclure ou les inclure pour une nouvelle recherche.

Le groupement de documents n'est pas une panacée, car il n'explique pas comment les documents sont trouvés. En revanche, quand le formulaire d'entrée encourage les utilisateurs à décomposer leurs demandes en plusieurs groupes de mots clés apparentés, une interface graphique peut indiquer les voies de recherche obtenues et les endroits où elles sont apparues dans les documents trouvés. Si tous les sujets se retrouvent au même endroit, le document a plus de chances d'être pertinent, et le programme lui accorde une priorité supérieure. Alice peut n'avoir aucune idée précise des sujets qui doivent apparaître dans le document ou du degré de proximité qui doit les unir, mais elle a des chances de reconnaître ce qu'elle désire lorsqu'elle le voit et peut alors affiner sa demande. Cette technique, que j'ai nommée *TileBars*, peut guider les utilisateurs lorsqu'ils choisissent les documents qu'ils souhaitent examiner et les conduit plus rapidement aux endroits les plus pertinents.

L'étude de ces nouvelles interfaces utilisateur et techniques d'analyse de texte commence seulement. D'autres techniques combinent méthodes statistiques et empirisme pour résumer automatiquement des documents et les insérer dans un ensemble de catégories déjà existant. Elles proposent des synonymes pour des mots clés et répondent à des questions simples.

Marti HEARST est informaticien au Centre de recherche *Xerox*, à Palo Alto.

Ramana RAO, Jan O. PEDERSEN, Marti A. HEARST et Jock D. MACKINLAY et al., *Rich Interaction in the Digital Library*, in *Communications of the ACM*, vol. 38, n° 4, pp. 29-39, avril 1995.

Stuart K. CARD, George G. ROBERTSON et William YORK, *The WebBook and the Web Forager : An Information Workspace for the World-Wide Web*, in *Proceedings of the ACM/SIGHI Conference on Human Factors in Computing Systems*, Vancouver, avril 1996. Présent sur le réseau à l'adresse : http: / / www.acm.org / sigs / sigchi / chi96 / proceedings / papers / Card / skc1txt.html

Les systèmes sécurisés

MARK STEFIK

La mise en place de dispositifs de protection numérique des droits d'auteur permettra aux auteurs de publier sereinement leurs œuvres sur l'Internet.

La banalisation de l'informatique a rapidement conduit les utilisateurs d'ordinateurs à penser que toutes les œuvres numériques – les programmes informatiques, les livres, les revues, la musique et la vidéo – pouvaient être copiées. Certaines figures de l'informatique ont même proclamé que la simplicité de la reproduction marquait la fin des droits d'auteur : «l'information veut être libre», disaient-ils. Autrement dit, il serait impossible de s'opposer à la diffusion de l'information, et tout ce qui peut s'exprimer sous forme de bits pourrait être impunément reproduit.

Cette idée est à la base de la création du réseau Internet : dans la nouvelle ère numérique où nous entrons, l'accès à l'information numérisée serait universel, et l'œuvre de n'importe quel auteur devrait être accessible à n'importe qui, n'importe où et n'importe quand.

Toutefois, par expérience, les utilisateurs du réseau savent que ce dernier ne contient pas de grandes œuvres ni d'informations capitales ; ce qu'on y trouve est souvent de faible valeur. La raison en est que les auteurs ou les éditeurs ne peuvent vivre en faisant cadeau de leur travail. Il suffit aujourd'hui de taper quelques touches d'un clavier pour copier un paragraphe, une revue, un livre, voire l'œuvre d'une vie. La duplication incontrôlée a tant modifié l'équilibre du contrat social entre les créateurs et les consommateurs d'œuvres numériques que la plupart des éditeurs et des auteurs refusent de laisser circuler le meilleur de leurs travaux sous forme numérique.

Toutefois, en coulisse, la technique se prépare à modifier cet équilibre. Au cours des toutes dernières années, plusieurs sociétés, telles *Folio*, *IBM*, *Intertrust*, *NetRights*, *Xerox* ou *Wave Systems*, ont mis au point des programmes et des

Jeff Brice

machines qui permettent aux éditeurs de spécifier les modalités d'utilisation des œuvres numériques et de contrôler leur utilisation. Certains législateurs pensent que ce changement est si radical que les éditeurs auront un pouvoir excessif, et que les consommateurs et les bibliothécaires seront lésés.

Pourtant les besoins des consommateurs seront satisfaits, même avec les nouveaux systèmes : la technique introduisant davantage de sécurité, des œuvres de meilleure qualité pourront être proposées sur le réseau. Bien protégés, des auteurs célèbres auront peut-être envie de publier directement sur le réseau. L'accès à l'information ne sera peut-être pas gratuit, mais il pourrait être bien moins onéreux qu'aujourd'hui, car les coûts d'impression et de distribution sont considérablement réduits par la mise des informations sur le réseau ; les éditeurs pourraient répercuter ces économies aux consommateurs.

Les systèmes «sécurisés» sont la clé de cette mutation technique : ce sont des machines et des programmes dont le fonctionnement se conforme à des règles, nommées droits d'usage, qui définissent le coût d'accès aux œuvres numérisées et les modalités de leur utilisation. Un ordinateur sécurisé, par exemple, refuserait de faire des copies non autorisées ou de transmettre des documents sonores ou graphiques à un utilisateur qui n'aurait pas payé pour les acquérir.

Les types de systèmes sécurisés sont variés : des lecteurs sécurisés peuvent lire des livres numérisés, des magnétophones et des magnétoscopes sécurisés peuvent lire des enregistrements audio et vidéo, des imprimantes sécurisées peuvent faire des copies contenant des signes (des «filigranes») qui indiquent le droit d'utilisation et de reproduction accordé, et des serveurs sécurisés peuvent se charger de vendre des œuvres numériques sur l'Internet. Bien que complexes, les techniques d'élaboration des systèmes sécurisés ont une fonction simple : elles permettent aux éditeurs de distribuer leurs produits – sous forme cryptée –, de sorte que ceux-ci ne soient affichés ou imprimés que par des machines sécurisées. Les premiers systèmes sécurisés équiperaient des imprimantes ou des lecteurs numériques de poche – avec un surcoût pour l'utilisateur, qui accéderait alors à des documents bien plus intéressants qu'aujourd'hui. Puis, comme toujours, l'avènement des nouvelles techniques en ferait chuter les prix. Naturellement les éditeurs pourront encore donner gratuitement l'accès à certaines œuvres : les serveurs sécurisés les transmettront alors sans imposer les protocoles spécifiques des œuvres protégées.

Comment les systèmes sécurisés savent-ils les règles qui doivent être respectées ? La Société *Xerox* et d'autres ont tenté de décrire les montants des droits et les modalités précises d'utilisation dans un langage formel qui sera interprété par les systèmes sécurisés. Le recours à un tel langage est essentiel pour l'édition électronique : les vendeurs et les acheteurs ne pourront négocier et parvenir à des accords que si l'ensemble des actions autorisées ou interdites est explicitement défini. Les droits numériques sont de plusieurs types. Les droits de transport concer-