

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 500^e 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.

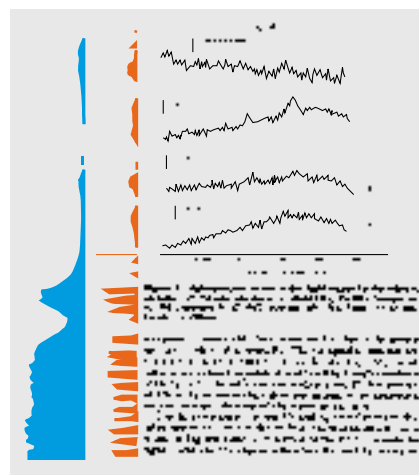
À défaut de ce soutien, des institutions pourraient s'associer pour mettre en commun sur l'Internet leurs collections virtuelles, en se répartissant le coût de création de ces collections. Toutefois, des problèmes d'ordre politique et administratif ont jusqu'ici contrarié les tentatives de coopération : 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



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.

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 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 soudeée, 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 antivirus 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

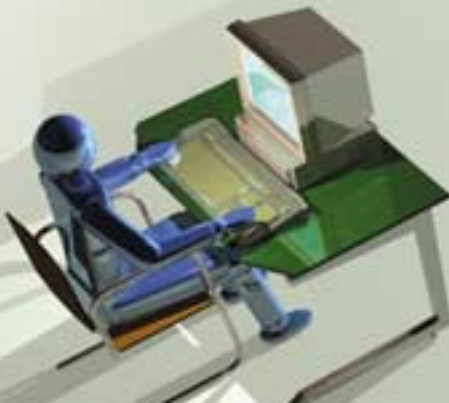
Les étiquettes PICS peuvent décrire toutes les caractéristiques d'un document ou d'un site informatique. Les premières étiquettes repéraient les documents indécents. Par exemple, le Conseil consultatif sur les logiciels de loisirs (RSAC) adapta à l'Internet son système d'évaluation des jeux informatiques. 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.



ferme pas de virus : ces étiquettes indiqueraient les contrôles antivirus effectués sur le logiciel ou le degré de fiabilité du logiciel.

Dans la vie quotidienne, l'information relative à un objet est soit attachée à l'objet, soit regroupée dans un catalogue. Chez un concessionnaire automobile, par exemple, le client lit les caractéristiques et les prix sur des panneaux posés sur les nouveaux modèles, mais il peut également consulter des descriptions de ces modèles dans des revues spécialisées. Il en est de même pour les étiquettes PICS. Un fournisseur d'informations qui propose une description de ses propres productions peut soit placer ses étiquettes dans des documents à part, soit les adjoindre aux documents qui circulent sur le Web. Une tierce personne peut aussi décrire ces documents. Par exemple, le Centre Simon Wiesenthal, qui traque les activités des groupes néonazis, pourrait publier des étiquettes PICS qui signaleraient les pages du Web où figure de la propagande néonazie. Ces étiquettes

seraient stockées sur un catalogue. En abonnant leurs navigateurs à ce catalogue, les internautes éviteraient de visiter les sites en question.

Un logiciel peut être réglé non seulement pour faire apparaître les étiquettes, mais aussi pour les modifier. Plusieurs logiciels, dont *CyberPatrol* et *Microsoft Internet Explorer*, utilisent déjà le standard PICS pour contrôler l'accès aux sites. Un utilisateur qui s'abonne au catalogue présenté par RSACi peut régler ses niveaux maximaux de tolérance à l'égard du langage, de la nudité, du sexe et de la violence. Un utilisateur qui s'abonne à un barème de contrôle antivirus peut choisir les contrôles antivirus appropriés.

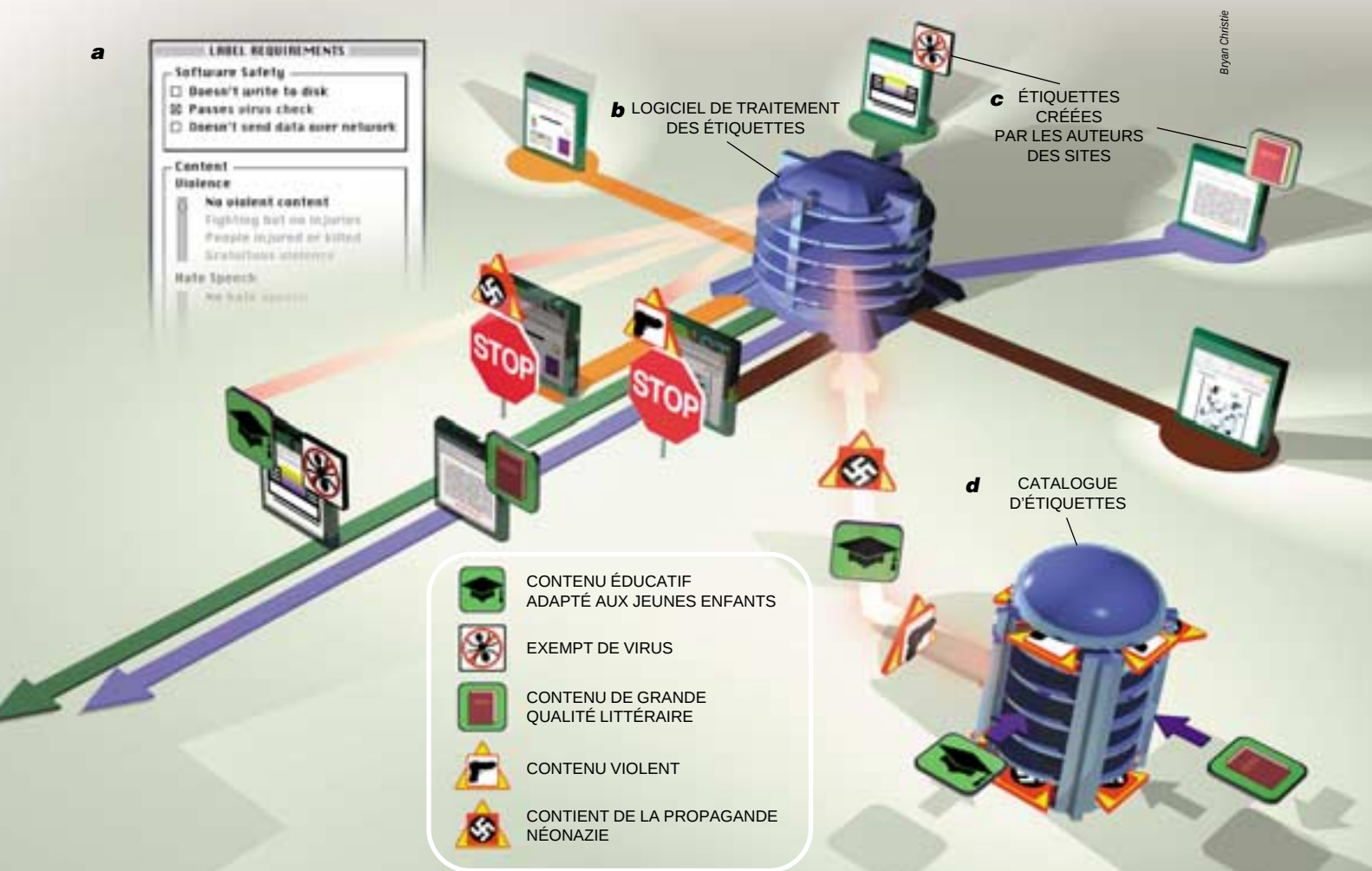
Si le traitement d'étiquettes exclut les documents indésirables, il facilite également la recherche de documents. Sur la foi d'étiquettes idoines, un moteur de recherche peut recommander des œuvres de grande qualité littéraire, et les serveurs Web pourraient trier, pour ceux qui le désirent, les services qui ne recueillent pas d'informations personnelles.

Établir la confiance

Comment savoir si les étiquettes sont fiables ? Le créateur d'un virus peut facilement diffuser une étiquette trompeuse affirmant qu'un logiciel n'est pas infecté. La question de la fiabilité des logiciels n'est que reportée sur la fiabilité des étiquettes. On résoudrait ce problème en utilisant des techniques cryptographiques, qui déterminent si un document a été modifié depuis la création de son étiquette et qui garantissent que l'étiquette provient réellement de l'auteur de l'œuvre numérique.

Toutefois, cette solution déplace à nouveau le problème, car il s'agit alors de savoir, non si l'on peut se fier à une étiquette, mais si l'on peut se fier à son créateur. On devra alors décider de se fier aux étiquettes de quelqu'un qu'on connaît personnellement ou d'une société qui a pignon sur rue, ou bien on s'en remettra à une entreprise de vérification.

Bien sûr, certaines étiquettes sont subjectives. Un usager pourra se méfier de certaines étiquettes, parce



(PICS-1.1 "http://www.w3.org/PICS/vocab.html"

labels

by "paul@GoodMouseClicking.com"

for "http://www.w3.org/PICS"

generic

exp "1997.04.04T08:15-0500"

ratings (q 2 v 3)

Auteur de l'étiquette

URL de la page
étiquetéeCe terme signifie
que l'étiquette s'applique
à l'ensemble des fichiers
accessibles à l'adresse
<http://www.w3.org/PICS>Le document à cette
adresse, où URL, con-
tient la définition du
vocabulaire d'étique-
tage : par exemple, "q"
signifiera qualité litté-
raire et "v" violence.Cette étiquette
expire le 4 avril 1997.L'auteur de l'étiquette
a attribué une qualité
littéraire de 2 et un degré
de violence de 3 au site.**2. LA LECTURE d'une éti-
quette PICS est effectuée auto-
matiquement 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 refléteront 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.

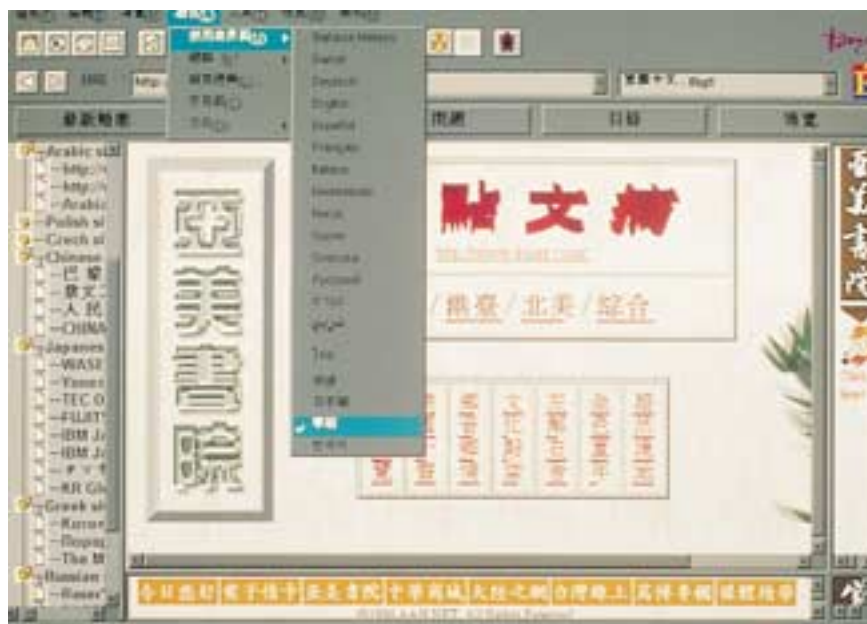
Une telle domination aurait des inconvénients. Comme la plupart des utilisateurs de l'Internet ne sont pas de langue anglaise, leur maîtrise de l'anglais est souvent limitée. Pour les discussions complexes, le recours à la langue maternelle s'impose ; mais si l'Internet ne permet pas des conversations 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



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.