

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 numérisa-

tion. 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 de 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 remplacer 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 techniques 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-

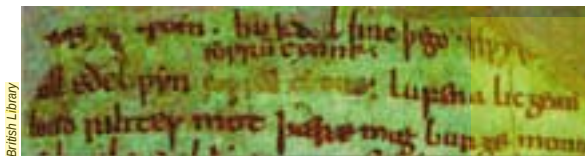
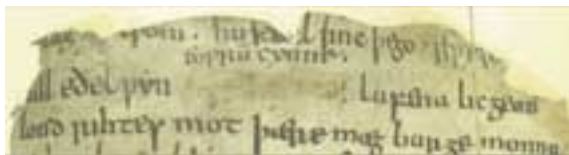
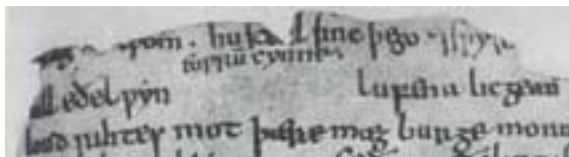


Agence France-Presse



Michel Gauguier, 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. Com-

ment, 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 pho-

tographies, 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 clavistes 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 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.

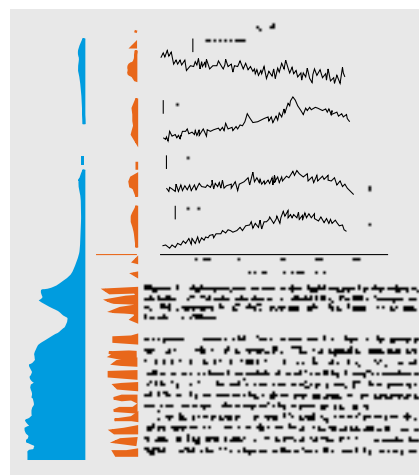
À défaut de ce soutien, des institutions pourraient s'associer pour mettre en commun sur l'Internet leurs collections virtuelles, en se répartissant le coût de création de ces collections. Toutefois, des problèmes d'ordre politique et administratif ont jusqu'ici contrarié les tentatives de coopération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