

La numérisation des documents graphiques

La bibliothèque numérique est un des grands débats qui agitent aujourd'hui le monde de l'information. Une profonde mutation est en cours, inéluctable, et les bibliothèques se doivent de négocier avec succès ce virage décisif.

Numérisation de documents n'implique pas seulement texte imprimé, mais également image animée, son, et image fixe que l'on rencontre dans les bibliothèques sous la forme des documents dits « spécialisés », c'est-à-dire gravures, estampes, photographies, documents cartographiques, partitions, monnaies, médailles, manuscrits enluminés ou non, divers objets, etc. Le traitement numérique de ces documents soulève de nombreux problèmes, tant politiques et juridiques que techniques, ce qui peut amener une bibliothèque à s'interroger sur l'évaluation de sa mission en ce qui concerne le traitement numérique de l'image et sa mise à disposition du public sur place et à distance.

Pourquoi numériser des collections spécialisées ?

L'ordinateur est aujourd'hui un outil de recherche reconnu et largement employé et pas seulement dans les milieux universitaires ou de recherche. Bien qu'en de toutes autres proportions, il pénètre chaque jour plus largement les foyers domestiques, accompagné d'un environnement multimédia, voire d'une connexion à l'Internet. Il est même aujourd'hui possible de se connecter à l'Internet sans ordinateur puisqu'un simple modem relié à la prise péritel d'un téléviseur permet de « surfer sur le Web » en famille. La demande des utilisateurs est ainsi elle aussi en pleine mutation puisque la bibliothèque n'est plus forcément le lieu où l'on doit physiquement se rendre pour trouver une information, mais le lieu qui possède l'information que l'on recherche et qu'on souhaite se voir communiquer à distance. C'est pourquoi

par Pierre-Yves Duchemin

Bibliothèque nationale de France

le monde des bibliothèques se doit d'offrir à tous ces utilisateurs potentiels des fonds numérisés composés aussi bien de textes que d'images.

Si la numérisation de textes imprimés est souvent implicite et considérée comme « normale », la numérisation de documents graphiques, bien que correspondant à un besoin, n'est pas considérée comme aussi « naturelle ». Il est vrai que les problèmes posés par la numérisation d'images sont techniquement et financièrement à une autre échelle que ceux posés par la numérisation de textes.

- Les pages de texte issues d'ouvrages imprimés sont généralement plus faciles à gérer, à la fois par la manipulation du document, même s'il possède une reliure précieuse, et par l'utilisation de l'image numérique que l'on peut en obtenir. Les problèmes colorimétriques posés par une page de texte imprimé sont, dans l'immense majorité des cas, réduits à leur plus simple expression : une page de texte n'est composée que de points blancs et de points noirs ; de plus, un ouvrage présente le plus souvent une grande cohérence dans la couleur et la texture du papier.

- Par ailleurs, les documents les plus intéressants, les plus rares, les moins connus, etc. proviennent souvent des collections spécialisées et ils sont en outre un excellent choix en ce qui concerne la conservation, la communication, la valorisation et les publications.

- Ensuite, les collections spécialisées et, en particulier en raison de leur taille, les documents cartographiques, certains documents relevant du domaine de l'estampe et de l'image fixe en général, certains manus-

crits, soulèvent des problèmes tant pratiques que techniques bien plus difficiles à résoudre que l'imprimé pour pouvoir obtenir une bonne qualité et une utilisation efficace des images numériques.

- Enfin, les collections importantes de documents spécialisés (cartes, images fixes, partitions, etc.) ne sont pas nombreuses en France.

Définir une politique documentaire

Il n'est pas facile de concevoir et de maintenir une politique cohérente à l'intérieur de la bibliothèque quand on doit traiter des documents aussi nombreux et aussi différents dans le cadre d'un projet de numérisation. C'est pourquoi il est nécessaire de n'entamer un programme de numérisation d'images que si l'on a, au préalable, soigneusement défini une politique documentaire cohérente. Une campagne de numérisation n'est pas « gratuite », à tous les sens du terme...

Selon ses besoins, sa vocation, ses missions... et ses ressources, la bibliothèque doit mettre en œuvre une politique documentaire cohérente pour la numérisation de ses documents : il ne s'agit pas de s'adonner à l'attrait d'une modernité vaine ou d'une technologie à la mode. La numérisation doit répondre à un besoin et la sélection des documents à numériser doit ainsi s'effectuer selon plusieurs critères principaux :

- la valeur documentaire, patrimoniale et parfois esthétique des collections ;
- l'aspect préservation et conservation des originaux, notamment en ce qui concerne les documents de grand format ou les

documents très fréquemment consultés (le recours à l'original d'un document numérisé devient ainsi exceptionnel);

- la mise en valeur des collections, notamment les plus rares, en les ouvrant à un plus large public, grâce à des expositions « virtuelles » ou des éditions sur support numérique;
- la complétude et la cohérence de la reproduction photographique, ainsi que la présence de notices bibliographiques informatisées, sur fiches ou sur catalogues imprimés.

Si la politique documentaire de la bibliothèque est clairement définie, il n'y a pas besoin d'effectuer une sélection de documents dans une collection donnée : la collection elle-même est considérée comme un ensemble qui conserve ainsi son intégrité documentaire. Un ensemble documentaire cohérent peut être à la fois utilisé comme un projet à caractère scientifique ou artistique, et éventuellement, dans le cadre d'une exposition ou d'une recherche particulière, comme complément d'information pour des ouvrages imprimés. Cette méthode permet d'éviter l'inconvénient qui consiste à offrir des ensembles thématiques incomplets. Il va enfin de soi que la politique documentaire de la bibliothèque ne doit pas être figée, mais être capable d'évoluer selon la demande et l'évolution des techniques.

But de la numérisation

Mettre en œuvre une campagne de numérisation suppose que plusieurs questions importantes ont trouvé une réponse, notamment : pourquoi la bibliothèque souhaite-t-elle proposer des images numériques à ses utilisateurs ? De quelle sorte et quelle qualité d'imagerie numérique la bibliothèque a-t-elle besoin ?

Parmi d'autres finalités, les images numériques sont créées pour :

- permettre la préservation et la conservation du document original;
- améliorer les possibilités d'accès au document en proposant un substitut électronique à l'utilisateur final;
- permettre une communication plus importante et/ou plus savante du document grâce à l'utilisation de stations de travail spécifiques. Ces postes de travail permettent à l'utilisateur d'effectuer des zooms et d'obtenir une finesse de détails qu'il ne pourrait espérer à partir de l'original;

- permettre un accès multiple au document;
- permettre la communication du document hors de l'établissement grâce à des médias optiques ou électroniques et des réseaux en fournissant un accès à distance;
- favoriser la mise en valeur de collections prestigieuses de documents rares et de grande valeur, parfois non publiés, en éditant des disques optiques compacts (DOC), des disques compacts interactifs (CD-I), des Photo-CD ou toute autre sorte de disque optique;
- permettre des tirages de qualité, à l'intérieur de l'établissement;
- offrir à l'utilisateur final une copie électronique du document sur un outil de travail spécifique dans un nouvel environnement technique.

Support photographique ou original ?

Dès que l'on aborde la numérisation des documents proprement dite, un nouveau problème apparaît : la numérisation doit-elle être effectuée à partir du document original ou d'un substitut photographique ? Globalement, ce n'est pas un problème technique puisque, dans la plupart des cas, il est possible de numériser un document à partir de l'original, même si les tests montrent que la numérisation d'un support photographique de haute qualité donne les mêmes résultats que celle d'un original de taille moyenne. Le problème est différent en ce qui concerne les documents de grand format dont la taille ne permet pas toujours une numérisation directe. Le problème inverse existe avec les documents de très petit format : il est évident qu'une prise de vue « pleine page » sur un plan film permet la numérisation d'un document dont la taille est supérieure à celle de l'original, ce qui offre à l'utilisateur un confort d'utilisation accru.

En ce qui concerne les photographies anciennes, les tests montrent qu'on obtient un meilleur résultat si la numérisation est effectuée à partir d'un substitut photographique de bonne qualité plutôt qu'à partir du cliché original : ces photographies sont fragiles et cassantes, certaines ont viré au brun ou au gris clair et un substitut photographique de bonne qualité améliore le contraste et la définition du cliché lui-même ; cela permet

d'offrir à l'utilisateur un document numérique dont la lisibilité est supérieure à celle de l'original.

Par des tests de vieillissement accéléré réalisés en laboratoire, on sait qu'un inversible couleur réalisé il y a cinq ans n'est plus complètement parfait et qu'un inversible couleur réalisé il y a vingt ans peut être inutilisable, même s'il a été conservé dans un environnement protégé et favorable. On sait également que la durée de vie d'une microfiche noir et blanc peut atteindre une centaine d'années. Par ailleurs, les reproductions photographiques existantes dans les bibliothèques, remontant parfois à quelques dizaines d'années, n'ont généralement pas été réalisées selon de très hautes normes de qualité : leur rôle était plutôt de procurer à l'utilisateur un substitut à un document incommunicable, et les reproductions photographiques n'étaient pas censées atteindre la qualité qu'il est possible d'attendre de fac-similés. C'est pourquoi, malgré le coût supplémentaire engendré par l'opération, il ne faut pas hésiter à refaire une campagne photographique si l'on désire une cohérence colorimétrique et une haute qualité lors de la campagne de numérisation.

Dans cette optique, lors de la sélection des documents à numériser, la nature du support, la couverture et la qualité des reproductions photographiques existantes, ont bien évidemment été un important critère de choix : il serait stupide de vouloir numériser des cartes à partir de diapositives 35 mm 24 x 36 ou de microfiches.

La numérisation

La numérisation de documents graphiques doit-elle être réalisée en mode « texte » ou en mode « image » ? Le mode « texte », largement plus coûteux, offre toutefois de sérieux avantages. Grâce à des taux de résolution adaptés à chacun des problèmes spécifiques à traiter, il garantit un niveau de qualité élevé et permet ultérieurement à l'utilisateur d'« entrer » dans l'image numérique, de sélectionner des mots dans l'image, d'effectuer un zoom sur un détail artistique, géographique, toponymique ou textuel. Cette façon de faire, si elle garantit une grande qualité, coûte beaucoup plus cher,

occupe un espace mémoire notablement plus important qu'une numérisation en mode « image » et nécessite des moyens techniques plus performants. Les taux de numérisation requis peuvent couvrir un éventail allant, par exemple, de 1 000 x 1 500 points par pouce ou 1 500 x 2 000 points par pouce pour des photographies noir et blanc 20 x 30 cm à 2 000 x 3 000 points par pouce pour des photographies en couleur 50 x 60 cm, voire 4 000 x 6 000 points par pouce pour des cartes ou des manuscrits. Pour mémoire, une page de texte est généralement numérisée au taux de 300 points par pouce, voire, pour quelques cas difficiles, à 400 points par pouce.

Dans la plupart des cas, une bibliothèque obtiendra toute satisfaction avec une numérisation en mode « image » : même si la taille des fichiers est impressionnante, les normes de compression actuellement existantes permettent de réduire les fichiers à des tailles raisonnablement exploitables. Par ailleurs, le mode « image » est techniquement plus facile à réaliser, moins coûteux et donne d'excellents résultats. Il permet notamment de numériser des dossiers de presse, par exemple dans le cas d'un fonds local, et de sauvegarder ainsi la mise en page et l'apparence de l'original, tout en s'affranchissant d'un support papier de mauvaise qualité dont la pérennité est très compromise.

Contrôle de qualité

Le contrôle de la qualité est une opération très importante de la chaîne de travail et l'utilisation de normes, à la fois pour la numérisation, l'exploitation et la communication, permet d'obtenir de bons résultats. Le contrôle de qualité s'effectue en comparant l'image numérique au document numérisé selon une charte colorimétrique. Si la numérisation a été effectuée à partir d'un support photographique, c'est ce support qui doit être le critère de contrôle, et le recours au document original ne doit être que très exceptionnel. Dans le cas de numérisation d'une série « cohérente » de documents (collection de photographies, coupures de presse, enluminures d'un même manuscrit, etc.), le contrôle de qualité peut s'effectuer par sondages statistiques : en effet, le contrôle qualité est une opération très longue et minutieuse qu'il est

Bibliotheca Universalis

par Sonia Zillhardt

Bibliothèque nationale de France

Direction du développement scientifique et des réseaux

Le rôle essentiel des bibliothèques dans la société de l'information a été affirmé par le G7 lors de la conférence de Bruxelles en février 1995, par l'adoption d'un projet spécifique sur la bibliothèque électronique pour illustrer l'avènement de la société de l'information.

« Bibliotheca Universalis » a pour objectif de donner accès aux principales œuvres du patrimoine culturel et scientifique par le biais des technologies multimédias et des réseaux de télécommunication.

Il s'agit donc de constituer à partir des programmes de numérisation existants une collection universelle des savoirs de l'humanité disponible pour le plus large public. Piloté par la Bibliothèque nationale de France, la Direction du livre et de la lecture et la Bibliothèque nationale de la Diète du Japon, ce projet rassemble les bibliothèques nationales des pays partenaires du G7, la Bibliothèque du Congrès aux États-Unis, la British Library au Royaume-Uni, la Deutsche Bibliothek en Allemagne, la Bibliothèque nationale

du Canada et la Discoteca di Stato en Italie.

Ce concept de bibliothèque virtuelle mondiale a franchi une première étape avec la présentation en mai 1996 à Midrand (Afrique du Sud), lors de la conférence du G7 « Société de l'information et développement », d'un prototype offrant l'accès par Internet à des échantillons numérisés – textes, images et sons – des bibliothèques partenaires. Le thème « Afrique, culture et civilisation » est venu appuyer le nécessaire échange d'informations entre pays du Nord et du Sud sur les inforoutes.

Afin de s'assurer de la convergence des stratégies menées par les partenaires pour aboutir à la création d'une « Bibliotheca Universalis » à l'échelon mondial, une enquête stratégique, tant en termes de constitution de collections numérisées que de choix technologiques ou juridiques a été lancée début 1997. Cette concertation renforcera la présence des bibliothèques nationales sur les réseaux d'information internationaux, garantissant ainsi une diversité tant linguistique que culturelle.

ainsi possible de réduire dans des proportions raisonnables.

Tout au long de la chaîne de travail, on doit également se préoccuper des différentes sortes de support physique : depuis le support d'enregistrement, le support de transfert, le support servant à la fourniture du document numérique, le support de stockage pour la consultation jusqu'à, enfin, le support final utilisé pour la conservation des images numériques. Le problème de la longévité de ces supports dans le temps ne connaît pas encore de réponse satisfaisante. Mais le point qu'il faut toujours garder à l'esprit est que les données seront toujours plus importantes que les supports physiques : le support peut évoluer, mais les données restent, notamment si elles possèdent une structure normalisée et sont conservées dans un environnement favorable.

Catalogage et indexation

Il est vain de proposer des images numériques sans notices bibliographiques associées. Mais doit-on cataloguer les images numériques comme des documents spécifiques ou comme des substituts à l'original, au même titre qu'une microfiche ou un microfilm ? ou doit-on les cataloguer à l'unité ou bien par lots ?

La logique veut que des projets qui concernent l'image fixe répondent généralement au principe qui détermine que chaque image est une entité documentaire propre et nécessite donc une notice bibliographique et une indexation spécifiques. Il ne viendrait à l'idée de personne de cataloguer l'ensemble de l'œuvre de Balzac ou de Hugo en lots. Pourtant, pour un plus grand confort d'utilisation par l'utilisateur, certains projets particuliers

peuvent être organisés en lots, grâce à des notices de regroupement.

Cela semble évidemment une solution idéale que de cataloguer chaque document comme une unité documentaire séparée et de créer des liens entre eux afin d'obtenir un ensemble « virtuel » dans la base de données, mais cette solution s'avère évidemment plus longue. Si les documents originaux ont déjà été catalogués à l'unité dans la base de données, le problème est plus simple : il suffit d'ajouter une adresse logique aux descriptions bibliographiques et de créer un lien réciproque afin d'obtenir aussi bien l'image numérique à partir de la notice de description bibliographique dans la base de données que la description textuelle à partir de l'image. Le résultat est encore meilleur quand le système est capable de gérer des notices analytiques ; il est alors possible de lier précisément une ou plusieurs images numériques à une notice textuelle contenue dans un ensemble plus important. Une conversion rétrospective d'un fichier ou d'un catalogue imprimé, par numérisation et reconnaissance optique de caractères ou par saisie « classique », est une solution élégante pour obtenir un catalogage à la pièce.

Il est également possible de procéder autrement si les documents originaux n'ont pas été catalogués au sens propre, mais seulement inventoriés. Dans certains cas, le choix d'un catalogage par lots pourra s'avérer une solution suffisante si chaque document est au moins identifié par une courte légende (le titre du document) et une adresse logique dans la base de données. Dans ce cas, l'utilisateur effectuera sa recherche dans la base de données par les critères disponibles dans la notice de lot, obtiendra une mosaïque des images numériques constituant le lot, identifiées par leur légende et devra les faire défiler pour trouver celle qu'il souhaite.

Les ISBD, les normes AFNOR, les formats MARC et les listes d'indexation matières (LCSH, RVM ou RAMEAU) restent les éléments nécessaires à la structuration des données d'interrogation même si la norme SGML, l'hypertexte et l'interrogation en langage dit « naturel », sont actuellement au centre de la discussion. L'étude d'une DTD (définition de type de document : *document type definition*) pour les formats MARC dans une structure SGML est actuellement menée à la Bibliothèque du Congrès.

Il est possible de numériser en même temps, non seulement évidemment l'image elle-même, mais également la « légende », généralement le titre extrait de la notice bibliographique, la référence du cliché qui permet la commande directe d'une reproduction photographique, ainsi que la clé de la notice bibliographique dans la base de données, utilisée comme lien hypertexte.

Problèmes juridiques

En France, la situation juridique autour de l'utilisation des images numériques n'est pas simple : les lois de propriété et de protection intellectuelle sont très strictes sur un plan général et on doit toujours rechercher d'éventuels ayants droit. C'est la raison pour laquelle il est sage de veiller à ce que la majorité des projets de numérisation concernent des documents anciens (plus de 70 ans), pour lesquels n'existent pas de problèmes de communication ou de reproduction, puisqu'ils appartiennent à la bibliothèque depuis des années, voire des siècles. Mais se cantonner à numériser des documents appartenant au domaine public peut toutefois empêcher la bibliothèque de numériser des documents répondant à sa politique documentaire. Dans le cas contraire, en ce qui concerne les documents de moins de 70 ans non encore tombés dans le domaine public, la solution consiste souvent en la signature d'une convention spécifique entre la bibliothèque et les ayants droit.

La bibliothèque numérique montre ainsi un gigantesque gouffre juridique au plan international : les intérêts divergents des éditeurs, des ayants droit et des bibliothèques freinent actuellement le développement des projets de numérisation d'images. Ce ne sont pas les idées qui manquent aux bibliothécaires, ni même les moyens... Ce qui fait le plus défaut est un cadre juridique et légal qui permette à la bibliothèque d'offrir l'accès à distance à ses documents numérisés et qui, en même temps, la protège elle-même puisqu'elle devient ainsi diffuseur d'information.

L'accès à distance

D'un point de vue pratique, la bibliothèque se doit de définir clairement ses besoins et ses objectifs, notamment :

- combien d'utilisateurs sont susceptibles de se connecter simultanément ?

- quelles images doit-on rendre disponibles pour l'accès à distance ?
- combien d'images doit-on rendre disponibles pour l'accès à distance ?
- quelle qualité d'image numérique doit être disponible sur le réseau ? Il est tout à fait possible de consulter un document à un taux de définition très élevé à l'intérieur de la bibliothèque et d'offrir une image numérique de qualité inférieure sur le réseau ; cette question est déterminante et structurante pour les flux de données et le temps de réponse.

Le développement récent de l'Internet soulève de nouveaux problèmes : les accès et communications à longue distance ne sont pas faciles à gérer d'un point de vue technique, pratique et juridique. Pourtant, c'est souvent ce que souhaite l'utilisateur final. L'accès à distance à des collections de documents numérisés introduit un profond changement dans les relations que la bibliothèque entretient avec son lectorat ; dans ce cas précis, l'assiduité des lecteurs ne se mesure plus à l'occupation physique des places dans la salle de lecture, mais au nombre d'accès au site Web de la bibliothèque qui, bien entendu et compte tenu des décalages horaires, doit être disponible pratiquement 24 heures sur 24. La bibliothèque doit ainsi se préoccuper des protocoles de communication TCP/IP, de la norme Z39.50, de l'affichage HTML, etc. Ces considérations techniques ne doivent pas être négligées : en effet, une page d'accueil bien construite est souvent le gage d'une consultation importante. L'efficacité, l'esthétique et l'ergonomie du serveur Web sont des garanties importantes du succès de l'entreprise.

Enfin, pour parachever la mise au point du serveur Web de la bibliothèque, il est bon de rechercher d'autres sites proposant des collections de documents de nature équivalente et d'établir des liens hypertextuels vers ces sites extérieurs. C'est sur ce plan que se situe la profonde mutation en cours dans le rôle du bibliothécaire : il ne doit pas se sentir « court-circuité » parce qu'il offre l'accès à distance aux documents qu'il conserve à des utilisateurs qu'il ne connaît pas, mais au contraire s'attacher à offrir le meilleur service à distance possible, par un serveur d'utilisation simple et agréable et régulièrement mis à jour.