DjVu d'AT&T Research

une technique révolutionnaire de compression des documents composites numérisés

Mise au point par des ingénieurs du laboratoire de recherches d'AT&T, la technologie DjVu constitue une avancée importante pour la GED et la transmission de documents via internet.

Compressée avec DjVu, cette page numérisée en mode bitonal ne pèse que 10 kilo-octets.

Avec ***DjVu*** (prononcez Déjà Vu), il est possible de compresser des documents numérisés composites et de réduire leur taille à quelques dizaines de kilo-octets tout en préservant la qualité de restitution des textes, des graphiques ou des photographies. Les techniques de compression utilisées permettent de ramener la taille d'un document entre 20 et 90 kilo-octets lorsque le fichier original pèse de 8 à 24 méga-octets. DjVu s'applique aux documents numérisés couleur, noir et blanc et à niveaux de gris. Le taux de compression obtenu est particulièrement intéressant pour les applications de GED ou d'archivage et pour les applications où les documents numérisés doivent être transmis au travers d'internet ou sur un intranet. La taille de fichier obtenue permet de le transmettre rapidement vers un poste distant et de gagner de la place dans le système de stockage; deux points cruciaux dans les systèmes informatiques actuels et à venir. Comme le montre la copie d'écran ci-dessus, une page de formulaire numérisée et traitée avec DjVu ne demande que 12,3 kilo-octets d'espace de stockage contre 21,2 kilo-octets si l'on utilise la compression CCITT G4 et le format TIFF.

La technologie DjVu est disponible sous forme de toolkits destinés aux intégrateurs et aux développeurs désireux de l'implémenter dans leurs programmes et leurs applications. Ce toolkit existe pour les environnements Windows 95/98/NT, Linux, Sun/Solaris et Irix/SGI. Pour des usages sur intranet/internet, AT&T diffuse gratuitement des plug-ins compatibles avec les différents visualiseurs du marché fonctionnant dans les environnements Windows, Unix et MacOS (8.x); il suffit de les télécharger sur le site web de son centre de recherches dont nous indiquons les références à la fin de cet article. DjVu est par ailleurs en cours d'intégration chez plusieurs développeurs de solutions comme ISO Informatique (Aix en Provence) en France. L'éditeur américain de programmes de GED, Feith Systems (voir page 32) l'a déjà implémenté dans ses logiciels. Il en est de même pour CGS et Monarch Imaging Systems qui utilisent DjVu dans leurs solutions de GED et de workflow. Computer Associates le supportera prochainement dans sa gamme de produits Jasmine. Toujours aux USA, la société University Microfilm Inc. (UMI) a choisi DjVu pour mener à bien son projet "Early English Book Online" qui consiste à numériser 22 millions de pages d'ouvrages anciens conservées sur microformes, à les compresser et à les proposer au travers du web. A terme, UMI prévoit de rendre 5,5 milliards de pages accessibles dans le cadre de son programme "Digital Vault" (voir page 33).

Une combinaison de technologies nouvelles pour la compression

DjVu est l'œuvre du département "Image Processing Research" d'AT&T Labs (Red Bank, NY) sous la direction de M. Yann Le Cun, un ingénieur français, docteur en informatique. Son équipe compte plusieurs ingénieurs français qui ont participé à l'élaboration de DjVu, de ses codec et de ses algorithmes. Le produit se compose de deux

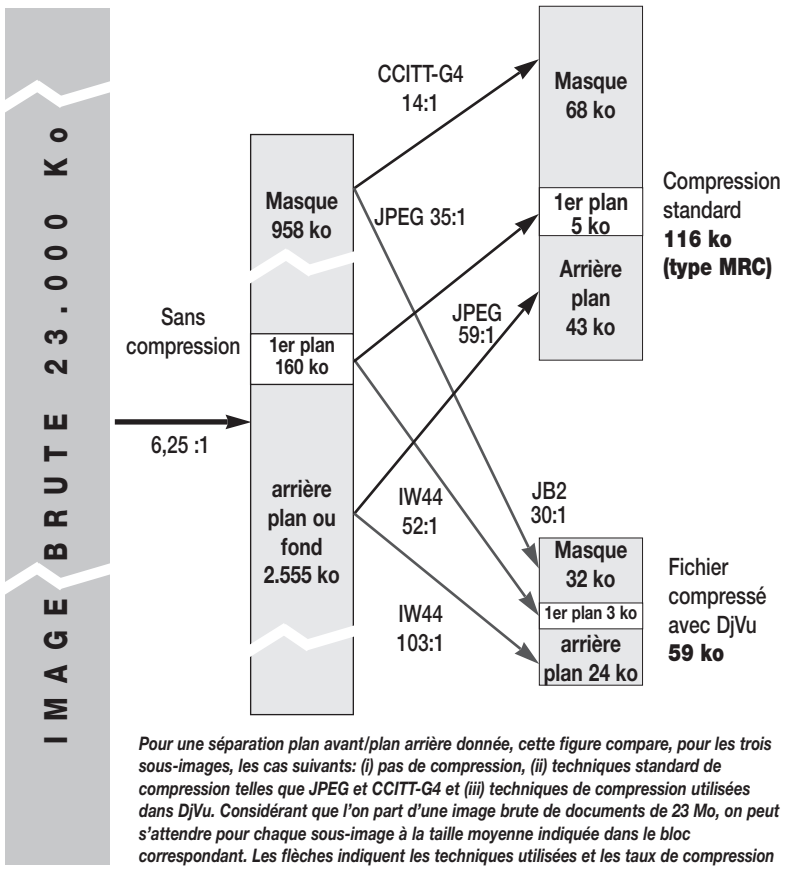
Principe de décomposition des documents numérisés avec DjVu : à gauche une image restituée via un navigateur internet; à droite les différents plans et découpages d'une image 1) le premier plan en couleur, 2) le texte en noir et blanc, 3) le fond en couleur

modules distincts : le compresseur et les plug-ins de décompression et de visualisation des documents. La partie "compression" fonctionne dans les environnements Unix (SunOS, Solaris, SGI/IRIX et Linux pour des bases Intel), Windows 95/98 ou NT ainsi que MacOS. Le compresseur de DjVu accepte des images numérisées couleur codées sur 24 bits, monochromes à niveaux de gris ou de type bitonal et enregistrées aux formats TIFF, PPM, PGM, PBM ou BMP. Sous certaines réserves, il s'accommode de fichiers JPEG ou GIF; cependant les concepteurs de DjVu ne recommandent pas d'utiliser ces formats en raison des pertes de qualité qui pourraient résulter de la compression (avec perte) effectuée par ces techniques. Par contre, il est possible d'utiliser la technologie DjVu pour traiter des images compressées à la norme CCITT G4 au travers de logiciels de conversion. La partie décompression et visualisation se compose de plug-ins ou programmes additionnels des visualiseurs Navigator/Communicator de NetScape et Explorer de Microsoft. Ils sont disponibles pour les versions fonctionnant sous Windows 95, 98, NT, Linux et MacOS 8.x et peuvent être chargés gratuitement sur le site de DjVu. Il est aussi possible de décompresser et de visualiser les fichiers DjVu avec des programmes dans lesquels sont intégrés les codecs du toolkit d'AT&T.

Les techniques mises en œuvre par DjVu pour compresser des images sont multiples. Le programme du compresseur décompose ou segmente le contenu d'un document numérisé afin de localiser et de séparer le fond, les zones de texte, les graphiques et les images (photos ou autres). Un document composite - comprenant du texte et des photos - est ainsi décomposé en trois plans; par contre les documents numérisés en mode bitonal n'utiliseront que le plan "texte" ou noir et blanc de DjVu.

Cette page A4 numérisée à 200 DPI a une taille de 12,3 kilo-octets en utilisant la compression JBIG de DjVu contre 21,2 kilo-octets avec la compression CCITT G4.

Décomposition d'une image d'un document numérisé et les différentes méthodes de compression utilisées



Pour une séparation plan avant/plan arrière donnée, cette figure compare, pour les trois sous-images, les cas suivants: (i) pas de compression, (ii) techniques standard de compression telles que JPEG et CCITT-G4 et (iii) techniques de compression utilisées dans DjVu. Considérant que l'on part d'une image brute de documents de 23 Mo, on peut s'attendre pour chaque sous-image à la taille moyenne indiquée dans le bloc correspondant. Les flèches indiquent les techniques utilisées et les taux de compression obtenus. © AT&T Labs 1998

A chaque plan, le programme applique la technique de compression la plus appropriée pour réduire le plus possible la taille du fichier final. Le fond du document comprenant les photographies, la texture, etc., sont compressés en utilisant une technique par ondelettes appelée IW44 qui convient particulièrement aux images à tons continus. Il en est de même pour le premier-plan qui contient les informations "couleur" des textes ainsi que certaines données complémentaires du fond. Le texte et les graphiques en noir et blanc sont compressés en utilisant le JBIG2, référencé JB2 par AT&T. Interviennent également une nouvelle technique de codage arithmétique appelée ZP-Coder et une technique de masque qui empêche les interférences ou parasitages entre plans. Le fichier est ensuite sauvegardé en utilisant le suffixe ".DJV" ou ".djvu" et peut être visualisé via les plug-ins des visualiseurs internet ou mis à disposition sur un serveur en association avec différentes applications, par exemple, de type GED ou de bibliothèque électronique.

DjVu : optimisation de la perte

Lors du traitement des documents numérisés, DjVu commence par séparer l'image en plans différents pour appliquer à chacun les techniques de compression les mieux adaptées. DjVu conserve la résolution initiale de numérisation (exemple 300 DPI) pour les textes et les graphiques qui sont compressés avec un algorithme JBIG2. Selon la description (1) faite par les auteurs de DjVu, "le principe de l'algorithme de compression JB2 consiste à masquer l'information trouvée dans des caractères rencontrés précédemment pour éviter d'introduire par erreur des substitutions de caractères, problème inhérent à l'OCR. Tout d'abord, l'image de base est segmentée en marques individuelles (composants de pixels noirs connectés). Les marques sont regroupées de façon hiérarchique sur la base de leur similitude en utilisant une mesure de la distance appropriée. Quelques unes sont compressées et sont codées directement en utilisant un codage arithmétique et un modèle

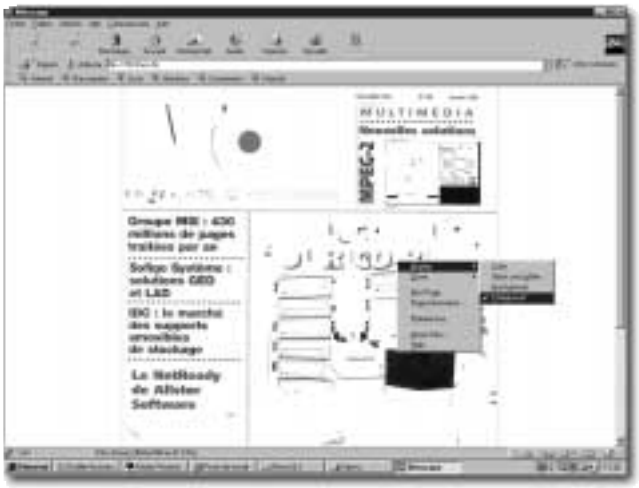
Exemple d'image compressée avec DjVu montrant la décomposition du traitement et les différents plans.

Cette couverture de MOS a été numérisée en couleur à 300 DPI et codée sur 24 bits : taille du fichier 24 méga-octets. Traité avec DjVu, le fichier ne pèse plus que 91,8 kilo-octets.

Plan texte en noir et blanc

Fond couleur de l'image

premier plan en couleur de l'image



statistique. Les autres marques sont compressées et codées indirectement sur la base de marques précédemment codées, en utilisant également un codage arithmétique et un modèle statistique. La marque codée précédemment dont on se sert pour une marque donnée peut avoir été codée directement ou indirectement. L'image est codée en indiquant pour chaque marque, l'index identifiant de la marque et sa position relative vis à vis de la marque précédente".(I)
Le fond de l'image et les photographies sont en général rééchantillonnés à 100 DPI et compressés avec un algorithme utilisant des technologies par ondelettes qui effectue la compression à un taux élevé avec un minimum de dégradation. La technique de AT&T est référencée IW4. Elle s'applique aux éléments photographiques monochromes à niveaux de gris ou en couleurs à ton continu. A ceci s'ajoute une seconde technique de compression appelée ZP-Coder basée sur un nouveau codeur entropique arithmétique binaire adaptif. Le premier plan est également rééchantillonné, par exemple, à 25 DPI sans que cela nuise à la bonne restitution. Il est ensuite compressé avec l'algorithme IW44. La création d'un masque entre les différents éléments ou plans d'un document numérisé fait appel à une technique qui sauve des bits sur les parties de l'arrière plan qui sont couvertes par le texte. Un document technique décrivant les principes de base de ces technologies est disponible en langue anglaise sur le site internet du centre de recherches d'AT&T (voir référence page 33).
La combinaison de ces différentes techniques de compression associées à celle du masquage optimise le résultat du traitement de documents composites, ce à quoi ne peuvent prétendre les systèmes utilisés jusqu'à présent. Le CCITT G4, par exemple, n'est pas adapté aux images à niveaux de gris ou en couleurs et le JPEG n'assure pas un rendu parfait des textes, notamment lorsque la compression est élevée. Comparé au PDF (Portable Document Format) de la technologie Acrobat mise au point par Adobe Systems, DjVu permet une compression plus importante, donc une réduction plus importante du fichier traité. Nous avons testé les deux techniques sur le même fichier (une page numérisée TIFF de 5,8 méga-octets). Avec Acrobat, le fichier PDF a une taille de 130 kilo-octets; avec DjVu, il a une taille de 51 kilo-octets. Avec DjVu, ce fichier est également visualisable sur différentes plates-formes (Windows, MacOS, Unix) au travers des plug-ins disponibles pour les navigateurs internet.
La suite de copies d'écrans page 28 montre la décomposition d'un document traité et compressé avec DjVu. Pour chaque vue, trois plans différents sont accessibles ainsi qu'un fichier descriptif indiquant la taille de chacun en fonction des techniques de compression utilisées. En utilisant le menu "Page information" (à l'aide du bouton droit de la souris sous Windows), il est possible de voir pour chaque page traitée la taille des différents plans de décomposition d'un fichier DjVu. (voir tableau ci-dessous).

Nous avons procédé à différents tests de la technologie DjVu en utilisant le serveur de compression en ligne du centre de recherche d'AT&T ainsi que le programme DjVuer conçu par l'éditeur américain Feith Systems. Le premier porte sur une page A4 numérisée à 300 DPI (12 points/mm) sur 256 niveaux de gris qui a été enregistrée au format TIFF (sans compression). Cette page comprenait du texte et une photographie. La taille de ce fichier était de 5894 kilo-octets; le traitement et la compression réalisés par DjVu ont permis de générer un fichier de 51 kilo-octets. D'autres essais au cours desquels nous avons essayé les options du compresseur ont donné une taille de fichier inférieure, au prix d'une légère dégradation du fond qui reste cependant tout à fait acceptable au stade de la visualisation ou de l'impression.

Notre second document, une page (12 x 16,5 cm) numérisée en couleur sur 24 bits à 300 DPI a généré un fichier de départ de 7980 kilo-octets que nous avons d'une part sauvegardé au format TIFF (sans compression) et d'autre part compressé avec JPEG (2.708 kilo-octets). Nous avons choisi de réduire la taille de cette page A4 lors de la numérisation afin de raccourcir le temps de transfert vers le serveur d'AT&T car sa taille initiale était de 24,5 méga-octets. Dans les deux cas, le fichier généré par DjVu était de 39 kilo-octets après traitement. Le temps de compression

Exemple de document A4 numérisé en noir et blanc sur 256 niveaux de gris à 300 DPI compressé avec DjVu. Le fichier original est de 5,8 méga-octets, une fois traité avec DjVu, il ne pèse que 51 kilo-octets tout en préservant la restitution des photographies et du texte.

Décomposition des différents plans et des méthodes de compression utilisées par DjVu

Type	Compression	Taille
- Informations sur le fichier	———	0,5 ko
- Texte ou graphique N&B à 300 DPI	JB2	39,2 ko
- Fond (Partie 1) à 100 DPI	IW44	2,9 ko
- Premier plan à 25 DPI	IW44	3,8 ko
- Fond (Partie 2) à 100 DPI	IW44	2,3 ko
- Fond (Partie 3) à 100 DPI	IW44	1,8 ko
- Fond (Partie 4) à 100 DPI	IW44	3,1 ko
Total		53,6 ko
Taille du fichier une fois enregistré		51 ko

Pour un fichier original en TIFF de 5.894 kilo-octets. Ce fichier est l'image reproduite ci-dessus dans la copie d'écran.

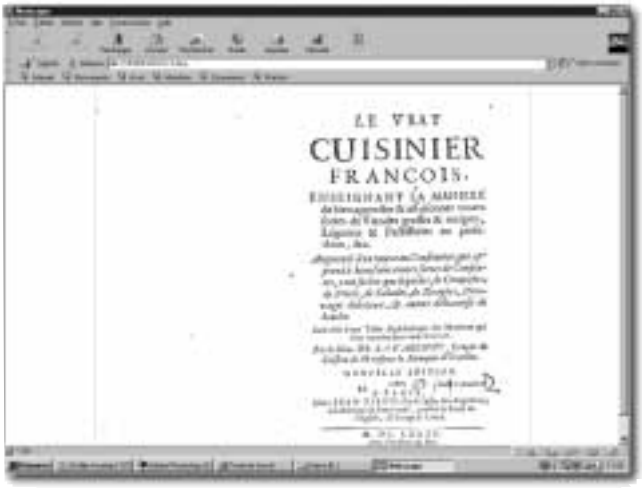
d'une page A4, via le serveur web d'AT&T, est de 40 à 50 secondes à l'aide d'un programme fonctionnant sous Unix. Nous avons également testé DjVu en collaboration avec Iso Informatique qui l'a intégré dans certains de ses toolkits. Les personnes intéressées peuvent faire ce test en ligne en s'inscrivant sur le site web du centre de recherche d'AT&T. Elles recevront en retour, par courrier électronique, un code qui leur donnera accès au serveur de compression en ligne. Il est ensuite facile de transférer les documents vers ce site puis de récupérer dans la minute qui suit les fichiers DjVu que l'on peut sauvegarder en local sur le disque dur de l'ordinateur.

En exploitation, lorsqu'il est ouvert par un programme ou un plug-in adapté, le fichier généré par DjVu est décompressé en local en quelques secondes. Le document s'affiche en plusieurs étapes. Le texte et les graphiques en noir et blanc s'affichent en premier et, en général, dans la seconde qui suit l'ouverture du fichier. S'affichent ensuite le fond et les images photographiques ou les illustrations puis le premier plan. Ces composantes restituent le document numérisé de départ. L'utilisateur peut aussi sélectionner les différents plans afin de ne visualiser que l'un d'eux. Ainsi, on peut imaginer ne visualiser que du texte ou des graphiques en noir et blanc, s'il s'agit de l'information primordiale. A titre d'exemple, nous reproduisons ci-contre un document extrait du site de démonstration d'AT&T. La première copie d'écran restitue la page dans son ensemble avec le fond du livre original tandis que la seconde est une sélection du texte en noir et blanc (auquel s'ajoutent quelques autres éléments reconnus comme des graphiques par DjVu). Selon les concepteurs de DjVu, il est concevable d'effectuer des traitements de cet ordre pour indexer un document en texte intégral, après reconnaissance optique de caractères, extraction du contenu et sauvegarde sous forme ASCII. Certains des membres de l'équipe de recherche d'AT&T à l'origine de DjVu sont également les auteurs de techniques et d'algorithmes de reconnaissance optique de caractères OCR/ICR utilisés par certains industriels. ***suite page 32***

Second exemple de décomposition des plans des images numérisées traitées et compressées avec DjVu. Cette page est en couleur et fait 2136 x 1872 pixels.

Plan texte en noir et blanc

Fond couleur de l'image



Premier plan en couleur de l'image

TAILLE DE LA PAGE COMPRESSE :

33,7 kilo-octets

Un toolkit disponible pour les développeurs

La technologie DjVu est disponible auprès d'AT&T sous forme de licence. Le toolkit est vendu 5.000 dollars (environ 30.000 francs H.T.) aux USA auxquels s'ajoutent des redevances dont le prix varie en fonction des applications. Ce toolkit est destiné aux développeurs et aux intégrateurs désireux de proposer cette technique pour traiter les documents dans des systèmes de GED ou documentaire; mais aussi pour concevoir des sous-programmes de saisie/compression utilisables avec d'autres logiciels. Certaines sociétés comme ISO Informatique en France (voir page 35) proposeront dans quelques mois un toolkit complet facilement intégrable dans les environnements Unix ou Windows. Il comprend la partie compression et des modules de visualisation qui disposent de fonctions de manipulation d'images (rotation, barre de fonctions, etc.).

AT&T propose aussi un ensemble d'utilitaires de compression par lots pour les applications de grand volume. Son prix démarre à 5.000 dollars par poste de saisie, mais peut décroître à l'unité selon le nombre de licence acheté. M. Yann Le Cun d'AT&T Labs nous a fait savoir qu'une bibliothèque de références comprenant les codes sources sera prochainement disponible. Elle permettra de concevoir des modules de visualisation et de manipulation des fichiers DjVu. Cette bibliothèque sera proposée sous forme de licence "logiciel libre"; elle sera intégrable dans des applicatifs, moyennant redevance si le produit est commercial. Le prix de ces "royalties" sera très raisonnable selon M. Le Cun. Une version du compresseur DjVu pour Lunix peut être gratuitement chargée sur le site d'AT&T pour des applications non commerciales.

D'après les tests que nous avons effectués, la technologie DjVu constitue une avancée importante dans la compression de documents composites et pour la GED. Les taux de compression obtenus sont un avantage indéniable dans les applications de type client/serveur en réduisant les temps de transmission des fichiers par réseau intranet/internet. De plus, DjVu restitue des images de meilleure qualité que les techniques utilisées jusqu'à présent, à savoir la compression CCITT G4 et le JPEG dans le cas de documents mixant des images et du texte. De nombreux éditeurs américains et internationaux d'applicatifs prévoient d'intégrer DjVu dans leurs produits dans les mois à venir. Pour sa part, AT&T a pour objectif de diffuser cette technologie le plus largement possible en proposant gratuitement les plug-ins pour visualiseurs internet. Nous ne saurions trop conseiller à nos lecteurs de se connecter sur le site du centre de recherche d'AT&T pour télécharger ces plug-ins, les installer sur leurs ordinateurs et procéder à des tests de visualisation et même de compression de documents numérisés pour vérifier les possibilités des technologies de DjVu.

Feith Systems : des solutions et logiciels intégrant DjVu

L'éditeur américain de logiciels de GED et de workflow Feith Systems propose un programme de compression intégrant DjVu.

Le 8 février, **Feith Systems & Software** (Fort Washington, PA) a annoncé la disponibilité de son programme d'encodage d'images numérisées au format DjVu, le ***DjVuer***. Fonctionnant sous Windows 95/98 et NT, ce logiciel permet de compresser des documents ou des images numérisées en utilisant le format DjVu d'AT&T. Une version de démonstration pour 50 documents peut être téléchargée gratuitement sur le site web de cette société (http:// www.feith.com). Les personnes intéressées pourront ensuite acquérir une licence permettant de traiter 5000 documents au prix public de 195 dollars (environ 1.170 francs H.T.). Feith Systems a également intégré ce nouveau format dans sa solution de GED "Feith Document Database" (FDD). Pour visualiser les fichiers convertis, il suffit de télécharger le plug-in adapté à l'environnement informatique utilisé en se connectant sur le site web d'AT&T Research.

Paramètres de compression du logiciel DjVuer de Feith Systems

Visualisation de fichiers DjVu avec le logiciel DjVex de Feith Systems qui incorpore quelques outils de manipulation des images (rotation, etc.)

Pour les autres solutions intégrant DjVu, voir pages 35/37, la nouvelle offre de la société Iso Informatique (Aix en Provence, France) avec Isimage Java

Des ingénieurs français à l'origine de DjVu

C'est le département "Image Processing Research" d'AT&T Labs-Research dirigé par M. Yann Le Cun qui a conçu et développer la technologie DjVu. Diplômé de l'Ecole Supérieure Electrotechnique et Electronique (Paris), M. Le Cun est également docteur en sciences de l'informatique de l'université Pierre et Marie Curie (Paris). Après un post-doctorat à l'université de Toronto sur la technologie des réseaux de neurones appliqués à la reconnaissance, il a rejoint les laboratoires d'AT&T Bell (USA) en 1988. M. Le Cun est à l'origine de plusieurs algorithmes pour la reconnaissance optique de caractères manuscrits commercialisés par Lucent Technologies et NCR. Il est responsable du département "Image Processing" d'AT&T Labs depuis 1996. Outre M. Le Cun, les principaux co-auteurs de DjVu sont MM. Léon Bottou, Patrick Haffner, Paul Howard. Les deux premiers sont polytechniciens et spécialistes de la conception d'algorithmes de traitement et de compression. M. Léon Bouttou est aussi docteur en sciences de l'informatique, diplômé de l'université Paris XI/Orsay. Il a travaillé dès 1991 pour AT&T Bell Labs (USA) sur des techniques de reconnaissance optique de caractères. Il a ensuite travaillé à l'ONERA de 1992 à 1993 puis assuré la présidence de la société Neuristique (Paris) jusqu'en 1995 avant de retourner dans les laboratoires d'AT&T en tant que consultant puis de rejoindre le département Image Processing en 1996. M. Patrick Haffner est diplômé de l'école Polytechnique (1987) et de l'Ecole Nationale Supérieure des Télécommunications (1989). Il a travaillé en tant que chercheur sur la reconnaissance de la parole au CNET-Lannion et a obtenu un doctorat en 1994. Entre temps il a été "visiting scientist" chez ATR au Japon puis à l'université de Carnegie-Mellon (Pittsburgh, USA) .Il a rejoint AT&T Bell Labs en 1995 puis AT&T Labs-Research en 1997. M. Paul Howard est diplômé du MIT en sciences de l'informatique et en ingénierie. Il a reçu son doctorat en sciences de l'informatique à l'université Brown et a commencé sa carrière professionnelle en travaillant pour la Marine Midland Bank puis différentes sociétés de services et pour des institutions. Il a rejoint AT&T Labs en 1993. **F.P.**

Références:

http://www.djvu.att.com/
http://www.djvu.att.com/plugin/
http://djvuserver.research.att.com/
(1) Article en ligne : "*High Quality Document Image Compression with DjVu*" (juillet 1998) -http://www.research.att.com/~leonb/DJVU/jei/

Gigantesque conversion d'un fonds micrographique vers le numérique

Le fournisseur américain d'informations bibliographiques UMI convertit ses archives sur microfilm en bases d'images numériques.

La société américaine **UMI**, filiale de Bell&Howell, est spécialisée dans la fourniture de bases bibliographiques en ligne, bases constituées par des dizaines d'années de collecte d'ouvrages, de périodiques, de thèses, etc. Ces fonds ont été microfilmés. Ce sont des millions d'écrits: livres, journaux, revues et autres imprimés couvrant 500 ans de civilisation de l'imprimé que UMI conserve sous forme de microformes dans des caves climatisées dans les locaux de son siège à Ann Harbor (Mich. USA). Aujourd'hui, UMI veut franchir le pas du numérique; elle a décidé de numériser ces microfilms et microfiches puis de convertir les images obtenues en utilisant la technologie DjVu d'AT&T Research.

Les caves qui ont abrités et préservés ces microformes ont inspiré à UMI le nom de cette gigantesque entreprise de numérisation: ***Digital Vault Initiative***. La numérisation a débuté en mai 98 et va se poursuivre sur plusieurs années. Il s'agit en effet de saisir 5,5 milliards de pages tout en continuant à absorber les 37 millions de pages d'information actuelles qu'engrange UMI chaque année. Ainsi, les usagers des bibliothèques qui sont clientes de UMI pourront en passant par le logiciel d'interrogation du service en ligne fouiller dans la collection complète, de la littérature du 15ème siècle aux publications les plus récentes en passant par les journaux du 19ème siècle.

En ligne, les livres anglais de la première heure

La première étape consiste à numériser les microfiches de la collection anglaise qui s'étend des premiers ouvrages publiés en anglais depuis l'invention de l'imprimerie jusqu'à la fin du 18ème siècle, de 1475 à 1700. Ce sont ces ouvrages qui ont été les premiers microfilmés à partir de 1938. On y trouve des œuvres telles que les Contes de Canterbury de Chaucer, The English Physician de Culpeper et les premières éditions de 1623 des œuvres de Shakespeare. Dans un second temps, UMI s'attaquera aux énormes collections de journaux et de périodiques que détient UMI, parmi lesquelles se trouvent des éditions complètes de Time Magazine, de The New York Times et d'autres titres d'importance. Ce travail se fera en collaboration avec plusieurs éditeurs. Une fois ce travail achevé, UMI offrira en ligne dans son service ProQuest Direct le contenu le plus étendu et le plus riche des services avec consultation via internet des images. Pour le reste des microformes, la numérisation se fera en fonction de la demande, en particulier des institutions éducatives.

Numérisation en continu

UMI met en œuvre cinq numériseurs de microformes et travaille 24 heures sur 24 par renouvellement des équipes toutes les huit heures. Chaque page est contrôlée par un opérateur qui indexe séparément les illustrations contenues dans le document. Les scanners créent des images de documents qui sont ensuite converties avec DjVu. Ces images ne passent pas à l'OCR; elles sont liées à des références bibliographiques au format MARC qui, elles, seront interrogeables en texte intégral. UMI vend son service de références bibliographique en ligne sur internet sous forme d'abonnement aux universités, aux établissements primaires et secondaires et aux bibliothèques publiques. L'accès aux images des documents d'origine ne coûtera à ces abonnés qu'une fraction supplémentaire du prix de l'abonnement. Les non abonnés paieront le prix de la page et le coût d'accès. **M.C.**