# Compression et accessibilité aux images de documents numérisés

## Application au projet DEBORA

**Frank Le Bourgeois – Hubert Emptoz – Eric Trinh**

*LIRIS-RFV de l'INSA de Lyon*
*20 bd A. Einstein, 69621 Villeurbanne Cedex*
*Tel : 33 4 72 43 80 96*
*Fax : 33 4 72 43 80 97*
*{flebourg,emptoz,trinh}@rfv.insa-lyon.fr*

RÉSUMÉ. *Les bibliothèques numériques en-ligne en mode image se heurtent à un certain nombre de problèmes techniques comme le manque de méta-données détaillées pour une recherche plus précise des informations, le volume important qu'occupent les images qui limitent leurs transmissions sur le réseau et l'absence d'un format de données hétérogènes pour la navigation. Dans ce contexte, nous proposons des méthodes d'analyse et d'interprétation du contenu des images pour à la fois réaliser une compression plus efficace et extraire automatiquement des méta-données utiles à l'indexation par le contenu. Notre approche est basée sur une décomposition des images en objets indépendants qui seront compressés avec des méthodes appropriées. Nous proposons aussi un format de données hétérogènes, adapté à la navigation dans les ouvrages numérisés compressé, qui permet aussi de les modifier, les annoter ou les échanger sur Internet dans le cadre d'un travail collaboratif.*

MOTS-CLES : *Bibliothèque numérique, Compression d'images de documents, analyse d'images, segmentation, format de données, méta-données.*

ABSTRACT. *On-line digital libraries using documents images meet some technical problem like the lack of detailed metadata for fine queries, the space used by images which limit their transmission through the network and the missing of a file format suited for the navigation among heterogeneous data. In this framework, we propose methods for digitised documents contents understanding in order to improve image compression and to retrieve automatically useful metadata for documents indexing. Our approach is based on the partition of the images into independent objects that we compress with appropriate methods. We also propose a file format suited for the query and the consultation of compressed books which allow to edit , annotate and exchange data and metadata for a collaborative work.*

KEYWORDS: *Digital library, documents images compression and understanding, file format, metadata*

## 1. Le cadre du projet DEBORA

DEBORA est l'acronyme de Digital AccEss to BOoks of the RenAissance (Accès numérique à des livres de la Renaissance). Il s'agit d'un projet européen de la fin du 4ème PCRD (N° LB 5608/A) dont l'objectif a été de concevoir un ensemble d'outils permettant l'accès distant et collaboratif à des livres numérisés du XVIème siècle sans passer obligatoirement par les bibliothèques dépositaires des originaux. DEBORA met donc à la disposition des lecteurs, des ouvrages rares qui ne pouvaient être consultés que par un petit nombre de spécialistes et d'érudits, dans les fonds anciens des bibliothèques elles-mêmes. Le XVIème siècle a été privilégié, en raison de l'existence de ressources plus nombreuses qu'au XVème, parce que c'est au XVIème siècle que le livre imprimé a commencé à prendre sa forme moderne. Ce projet était centré sur une étude des usages, des méta-données et des fonctionnalités attendues par les utilisateurs des outils de consultation. Il a aussi spécifié les outils de travail collaboratif requis par les utilisateurs ainsi qu'une étude des coûts de la numérisation.

Notre apport en terme de traitement d'images à ce projet a été transversal, dans un premier temps, nous nous sommes intéressés à la numérisation physique des ouvrages et à la mise en place des protocoles de numérisation ; dans un deuxième temps, nous avons étudié une solution de mise en ligne, de consultation et d'édition partagée de nos ouvrages. Cet article présente les résultats des travaux sur le développement des outils informatiques pour améliorer l'accessibilité aux documents numérisés du XVI , favoriser le travail distant et collaboratif sur les données et enfin inventer de nouvelles méthodes pour extraire automatiquement ou semi-automatiquement les méta-données nécessaires à l'indexation. Le projet DEBORA, nous a fait prendre conscience du manque de réels d'outils de consultation de documents numérisés. Nous nous sommes alors intéressés dans un deuxième temps au développement d'une plate-forme de consultation et de travail sur ces documents numérisés. Cette plate-forme a été conçue pour permettre à plusieurs utilisateurs de pouvoir consulter et interagir à distance sur ces documents.

Les usagers des bibliothèques numériques en-lignes désirent un certain nombre de fonctionnalités qui ne sont actuellement pas offertes par les sites existants. Les usagers cherchent d'abord une plus grande accessibilité aux ouvrages, aux images et à leurs contenus. L'accessibilité signifie à la fois une consultation en-ligne ou hors-ligne facile et rapide ainsi qu'une interrogation plus avancée des contenus pour retrouver une information désirée. Ce besoin peut être satisfait par une compression efficace des images et une redéfinition d'un format de données qui permet une interrogation de leurs contenus. Ce format doit permettre de conserver localement une copie des données pour un travail hors-ligne indépendamment d'une architecture classique client/serveur. Le format doit aussi pouvoir gérer tous les contenus d'une bibliothèque numérique qui sont par définition des données hétérogènes (textes/images/ structures/annotations par des objets multimédia/ hyperliens…).

Une recherche efficace nécessite une description fine des documents et de leurs structures. Celle-ci peut être obtenue soit par analyse d'images soit par la saisie manuelle de méta-données lors d'un travail collaboratif d'un groupe d'experts ou d'usagers expérimentés. L'enrichissement de collections par des méta-données nécessite des outils de travail collaboratif, basés sur l'annotation d'un élément quelconque (structure, une page, le contenu d'une page, une image, une région d'une image, une annotation) avec n'importe quelles données (texte, image, vidéo, liens internes, URL...). Le format de données approprié aux usages doit pouvoir gérer l'historique des annotations, les droits des auteurs et les autorisations aux accès. Un document n'a pas de limitation physique, il peut être constitué d'un simple mot, d'une image, une ligne de texte, une page entière, un livre complet ou bien une collection d'ouvrages. Le format doit pouvoir gérer des quantités très variables de données et proposer des outils pour l'édition du contenu des images, des structures et des annotations.

Enfin, les utilisateurs ont besoin d'images de qualités variables suivant si les documents sont simplement destinés à la lecture sur un écran, imprimés en haute résolution ou dotés d'un niveau de détail suffisant pour une expertise. Les besoins des usagers étant trop différents, la compression d'image et le format de donnés doivent autoriser le téléchargement partiel des détails à la demande de l'utilisateur.

Dans ce contexte, nous proposons à la fois des outils d'analyse et d'interprétation du contenu des images pour à la fois extraire automatiquement des méta-données et comprimer plus efficacement les images. Enfin nous proposons un format de données semi-structuré adapté à la représentation des livres numérisés qui permet une meilleure navigation et interrogation. Dans une première partie, nous décrirons les méthodes de compression des images que nous avons développé pour ce projet européen. Dans la seconde partie, nous présentons les outils d'analyse d'image pour l'extraction automatique des méta-données. Enfin, dans la dernière partie, nous définirons le format adapté à nos données ainsi que les fonctionnalités qu'il permet d'offrir aux usagers.

## 2. La Compression des images de texte

### 2.1. *La compression JPEG*

Beaucoup de bibliothèques numériques utilisent la compression JPEG sur leurs images de documents. Cette méthode de compression psycho-visuelle retire l'information non perçue par l'œil humain mais qui affecte l'analyse automatique des images. De plus la compression JPEG n'est pas adaptée aux images de documents. Le taux de compression de 1:10 est trop faible et les déformations visuelles trop importantes affectent même la lisibilité du texte. Cela s'explique par le fait qu'une image de texte est constituée principalement de caractères de formes très complexes présentant des contours trop détaillés. La compression JPEG qui

réduit la redondance des couleurs adjacentes par un filtrage fréquentiel ne peut pas fidèlement reproduire les contours des caractères.

Nous avons mesuré l'impact des effets de la compression JPEG sur les images de textes du point de vue de la reconnaissance automatique et de la lisibilité en fonction du facteur de qualité fixé par l'utilisateur en pourcentage. Les résultats montrent que la compression JPEG modifie de plus de 50% la qualité de l'image dès que nous diminuons le facteur de qualité de 100% à 90% seulement. Donc même avec un taux de qualité maximum, la compression JPEG modifie de façon importante le contenu des images. En utilisant le facteur de qualité de 70%, fixé par défaut dans tous les logiciels, seule 35% de l'image est préservée intégralement. A ce niveau de compression suivant la résolution des caractères, l'OCR ne fonctionne plus correctement.

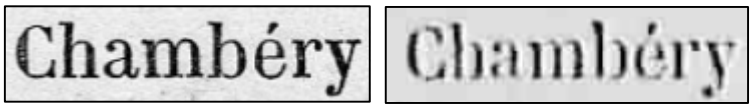



Image d'origine (800Ko) Image JPEG (500Ko)

**Figure 1**. *Effets de la compression JPEG*

L'avenir de la compression d'images semble maintenant appartenir au nouveau standard du Joint Picture Expert Group, JPEG2000. Ce nouveau format se base sur une transformation en ondelettes qui s'appuie sur une approche pyramidale multi-résolution plus pertinente dans le cas de signaux 2D tels que les images. La compression JPEG2000 offre un meilleur rendu visuel pour un taux de compression similaire à celui de JPEG. Par contre les images compressées grâce à JPEG2000 souffrent d'un effet de flou particulièrement accentué si le taux de compression est élevé. Ce phénomène peut rendre alors plus difficile l'application de certains traitements telle que la binarisation ou la localisation précise des contours des caractères.

Il n'existe pas de filtre inverse qui corrige les effets de la compression JPEG. Même si certains artifices comme le filtrage des contours permettaient de restaurer les images, l'information perdue ne pourra jamais être retrouvée. Ainsi, les fonds documentaires déjà numérisés et stockés sous ce format ou dans des formats qui utilisent de façon cachée la compression JPEG, comme le format PDF par exemple, resteront difficilement interprétables par ordinateur. La solution passe par des moyens de compression adaptés aux images de document et à la conservation des informations essentielles pour la lecture et le traitement informatique telles que nous allons développer dans la partie suivante.

## 2.2. *Les méthodes de compression adaptées aux images de textes*

### *2.2.1 Compression des images binaires*

Les images de textes sont des images particulières qui ne peuvent pas être traitées comme des images naturelles ou aléatoires car ce sont des images dont les formes de base sont des traits. Des algorithmes ont été développés pour la compression des images binaires de texte, par exemple TIFF ou JBIG, mais il ont peu de chance d'être supportés par les principaux navigateurs Web, limitant ainsi considérablement la diffusion de ces formats auprès du grand public. Le nouveau standard émergent appelé JBIG2 [HOW98] a été créé pour compresser les images binaires de texte en acceptant un taux de perte variable. JBIG2 utilise pour la première fois une connaissance a priori (model-based coding) du contenu d'un document avec la séparation texte/graphique et la redondance des formes des éléments de la partie textuelle. Cette approche constitue une très bonne alternative, car elle est adaptée à la répétition des formes de caractères ou de symboles qui apparaissent fréquemment dans les images de texte. La compression par redondance des formes génère une image artificielle où toutes les formes redondantes sont substituées par une forme générique unique. Bien que restant lisible, l'image décomprimée laisse apparaître des erreurs de substitution entre caractères, dues à la méthode d'appariement des formes. Pour retrouver une image proche de l'original, il faut donc conserver la trace résiduelle des informations perdues par la substitution des formes de caractères. JBIG2 propose plusieurs stratégies pour comprimer l'information résiduelle. Elle opère un lissage des contours des caractères qui permet de simplifier l'image résiduelle qu'elle comprime avec perte par un codage arithmétique avec contexte. JBIG2 a spécifié le format et les méthodes, mais ne propose pas de mise en œuvre qui peut être différente d'un logiciel à l'autre. L'efficacité de la compression JBIG2 va donc reposer essentiellement sur l'implémentation des méthodes et surtout sur la segmentation de l'image du document, la séparation image/texte et la qualité de la comparaison des formes.

### *2.2.1 Compression des images en niveaux de gris ou en couleurs*

Pour les images de documents en niveaux de gris ou en couleurs, nous avons vu précédemment que la compression JPEG affiche des taux de compression très faible de 1:10 qui restent très insuffisants pour la transmission d'images en haute résolution sans dégradation visible. De grandes entreprises privées ont pourtant développé des solutions intéressantes pour la compression des images couleurs de textes comme la compression DjVu [BOT98] d' AT&T et TIFF-FX de Xerox. Ces nouvelles approches permettent d'atteindre des taux de compression supérieurs à 1:100 en utilisant des informations sur la spécificité des images de documents. La

compression des images couleur de texte est effectuée en séparant l'arrière-plan ou « fond » (texture du papier) de l'avant-plan ou « forme » (image de traits comportant les caractères, dessins, éléments graphiques…) et en appliquant sur chaque plan une compression adaptée. L'avant-plan contient principalement des informations textuelles qui peuvent être binarisées et sur lesquelles on peut appliquer une compression de type JBIG2. L'image couleur de l'arrière-plan, débarrassée des traits des caractères et des illustrations peut être compressée de façon plus efficace avec des algorithmes classiques comme JPEG ou JPEG2000.

### *2.3. La solution développée pour DEBORA*

Nous allons présenter une variante des dernières méthodes de compression des images de textes imprimés qui utilise l'analyse de la structure du document pour décomposer les images en objets élémentaires qui seront comprimés avec des techniques adaptées. Cette compression centrée sur l'analyse permet aussi d'extraire le plus d'information possible sur le contenu des images pour enrichir les méta-données et offrir de nouvelles fonctionnalités aux usagers. Nous avons aussi choisi d'appliquer la compression sur un ouvrage entier voire une collection d'ouvrages dans sa globalité pour améliorer les performances de la compression par redondance et bénéficier des informations sur la similarité des formes pour d'autres services comme la transcription manuelle assistée, l'étude des étapes de fabrication d'un ouvrage pour les documents anciens.

Nous proposons de diviser l'image du document en quatre plans que l'on va compresser différemment avec un modèle de compression adapté pour chacun d'eux.

Le plan textuel (image binaire): *Il ne comporte que des images de caractères qui ont la propriété d'être fortement redondants sur l'ensemble du livre. Les plans textuels d'un livre entier sont compressés efficacement en utilisant l'appariement des formes des caractères sur le livre entier. Les formes de caractères du dictionnaire nécessaires à la reconstruction et leurs positions suffisent à reconstituer rapidement l'image du texte.*

Le plan graphique (image binaire): *Il contient tous les éléments graphiques qui ne sont pas des caractères et qui se caractérisent par une très faible redondance de forme. Nous proposons de compresser sans perte le plan graphique à l'aide d'un codeur adapté aux images binaires aux traits comme celui du CCITT-G4*

L'arrière-plan (image couleur): *il représente l'image du support papier privé des caractères et des éléments graphiques. Cette image, qui apporte peu d'information, est fortement compressée en utilisant JPEG.*

Plan compensatoire (image binaire): *Ce plan contient les différences entre l'image décompressée et l'image d'origine. Il permet de reconstituer l'image originale exacte. Il est aussi compressé sans perte par une approche spécifique*

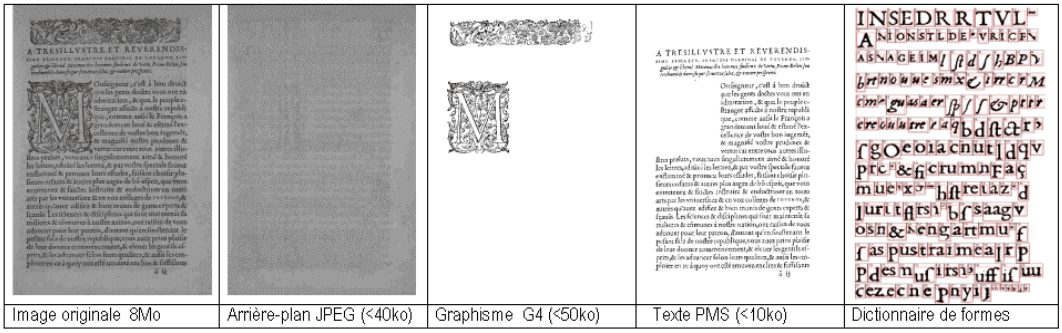



**Figure 2.** *Différents plans de décomposition d'une image de document*

Notre approche se distingue des méthodes précédentes comme DjVu ou TIFF-FX sur les points suivants :

- Exploitation de la redondance des formes de caractères sur toutes les pages d'un même livre et non indépendamment sur chaque page.

- Une compression différenciée entre les objets graphiques et les caractères afin d'éviter de surcharger le compresseur basé sur l'appariement de formes avec des éléments graphiques a priori non redondants.

- Comme l'étude des redondances n'est réalisée que sur des images de caractères, il devient donc possible de faire une transcription assistée par ordinateur en identifiant manuellement les formes de caractères du dictionnaire pour obtenir un texte intégral complet d'un livre en quelques heures. Cette analyse de la redondance des caractères sur l'ensemble de l'ouvrage permet donc la recherche par mots, la création automatique de liens entre l'image du texte et sa transcription et l'analyse de la typographie d'un ouvrage.

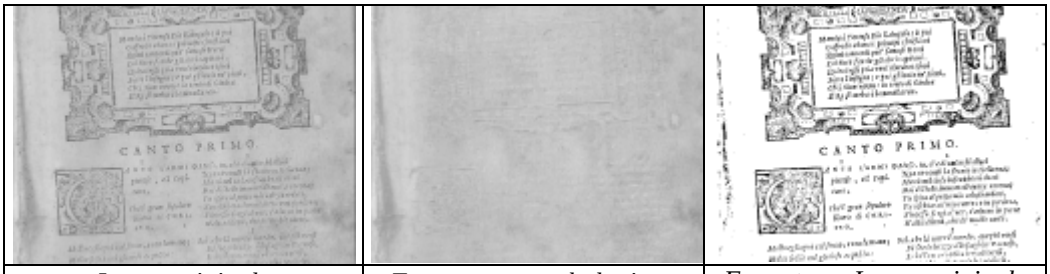
Pour réaliser cette séparation en plans, nous avons besoin d'effectuer préalablement une segmentation des images (séparation fond/forme, texte/graphique) et l'analyse de la structure physique (paragraphes, lignes, mots, caractères…).

### *2.4. Séparation initiale entre avant et arrière-plan*

Nous avons utilisé pour la séparation fond/forme des algorithmes plus sûrs et susceptibles de fonctionner avec le plus grand nombre de cas sur des documents très divers. Pour cette raison, nous avons privilégié des approches dites ascendantes (data-driven) où l'on recherche une interprétation avec peu de connaissances a

priori contrairement aux approches descendantes (model-driven) qui nécessitent des connaissances sur la forme et la localisation des zones de texte, ce qui est impossible sur un large spectre de documents divers et variés. Cependant l'analyse de la structure physique ne peut fonctionner que sur une image binaire résultant elle-même d'une séparation fond/forme. Nous avons choisi donc de procéder par étapes successives en conservant tout au long de la chaîne de traitement, l'image originale. Dans un premier temps nous appliquons une séparation fond/forme assez générique sur laquelle l'analyse ascendante donnera la position des zones de texte et des illustrations. A partir des zones différenciées, nous appliquons à nouveau des algorithmes de séparation fond/forme adaptés à chaque zone à partir de l'image originale conservée.

La correction des variations de l'illumination et la première séparation fond/forme sont réalisées simultanément par un algorithme de transformation morphologique de type «chapeau haut de forme» qui consiste à dilater l'image par un élément structurant de taille définie par l'utilisateur pour éroder les traits de l'impression jusqu'à les faire disparaître et ne garder que la texture du papier. Cette opération est suivie d'une érosion avec le même élément structurant pour obtenir une fermeture morphologique de l'image. L'image corrigée des variations de la luminosité et de la couleur du support papier est obtenue à partir de la soustraction entre l'image originale et sa fermeture. La fermeture de l'image représente le support papier qui sera conservée et comprimée pour définir ce que l'on appellera l'arrière-plan. L'avant-plan sera calculé grossièrement par un seuillage calculé automatiquement et appliqué globalement sur l'image corrigée des variations globales de luminance. Cependant si l'image de l'arrière-plan est très représentative, celle de l'avant-plan comporte beaucoup d'erreurs provoquant des variations de l'épaisseur des traits des caractères et une perte plus ou moins fréquente de la conservation des formes des caractères. Cette segmentation grossière ne sera utilisée que pour effectuer une analyse temporaire de la structure physique qui définira les zones de texte par rapport aux zones d'illustration. Après une analyse de la structure physique décrite dans le paragraphe suivant, nous proposons une seconde séparation avant-plan/arrière-plan particulière, plus adaptée aux zones de texte lorsque le verso apparaît sur le recto.




| *Image originale* | *Fermeture morphologique* | *Fermeture- Image originale* |
|---|---|---|

**Figure 3.** *Séparation morphologique entre l'avant et l'arrière-plan*

### *2.5. Analyse de la structure physique des pages*

C'est une méthode classique ascendante purement agrégative qui a été utilisée pour la segmentation de la structure physique et qui consiste à agglomérer suivant les normes typographiques, les éléments du plus bas niveau (simples connexités) en éléments successifs définissant les caractères, les mots, les lignes de texte, les paragraphes, les colonnes. Cette méthode agrégative a été assouplie pour pouvoir construire des lignes légèrement courbées ou déformées. Un problème sérieux a été néanmoins rencontré sur certains ouvrages anciens pour lesquels la segmentation des mots est impossible à partir des seules informations sur les espaces entre caractères. La fabrication artisanale d'un livre à cette époque obligeait inévitablement les typographes a resserrer les espaces entre les mots pour insérer un nouveau mot dans un ligne. La séparation texte/image s'effectue à partir de règles de classification des connexités.

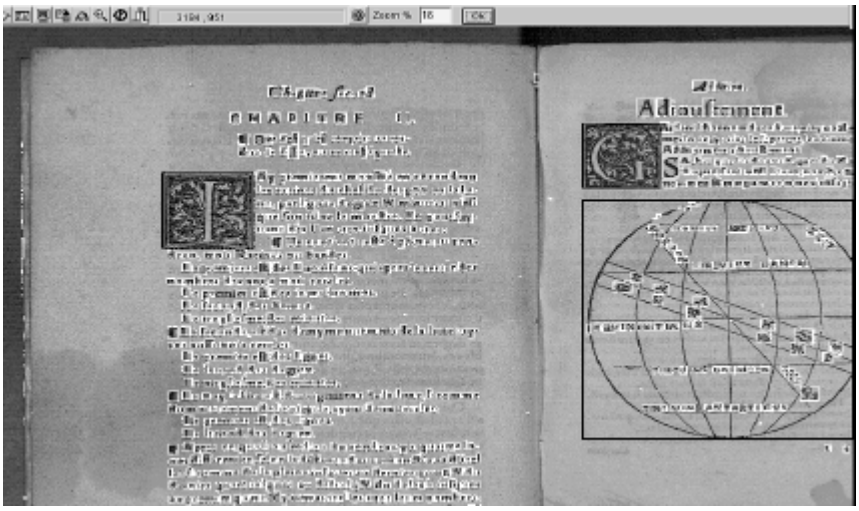


**Figure 4.** *Séparation entre les caractères et les éléments graphiques*

Toutes connexités alignées, de tailles homogènes, et régulièrement ordonnées sont considérées comme des caractères d'une ligne de texte. Ces connexités reconnues comme éléments de texte seront ensuite classées respectivement dans les mots et les paragraphes et les colonnes. Les connexités non classées comme texte sont considérées comme des éléments de décoration ou d'illustration et seront agglomérées, de proche en proche, pour former des zones connexes d'images. Les méthodes ascendantes agglomératives permettent un contrôle total du processus de segmentation par l'utilisateur qui peut en modifier le comportement sur certains ouvrages. Par exemple, la figure 4 montre que la présence de texte dans les zones d'illustration peut être autorisée par l'algorithme de segmentation si l'utilisateur le souhaite. Les paramètres concernent les règles d'agglomération, les règles de classification et l'activation de certaines procédures du processus de segmentation.

### *2.6. La séparation entre le recto et le verso*

Nous proposons une seconde séparation avant-plan/arrière-plan plus adaptée des zones de texte à partir des résultats de l'analyse précédente. Lorsque le verso apparaît sur le recto, un seuillage fixe de l'image corrigée des variations lumineuses ne donne pas de résultat satisfaisant. Nous avons appliqué une méthode de binarisation adaptative à partir d'un critère de qualité des formes de caractères. Nous avons constaté que la séparation des caractères, la régularité des épaisseurs des traits et la conservation de la topologie peut se mesurer directement à partir de calcul de la redondance des formes. Nous appliquons donc un seuillage adaptatif de type Niblack [NIB86] sur les zones de texte avec un jeu limité de paramètres (taille de la fenêtre et facteur k) jusqu'à obtenir un taux maximal de redondance de formes de caractères.

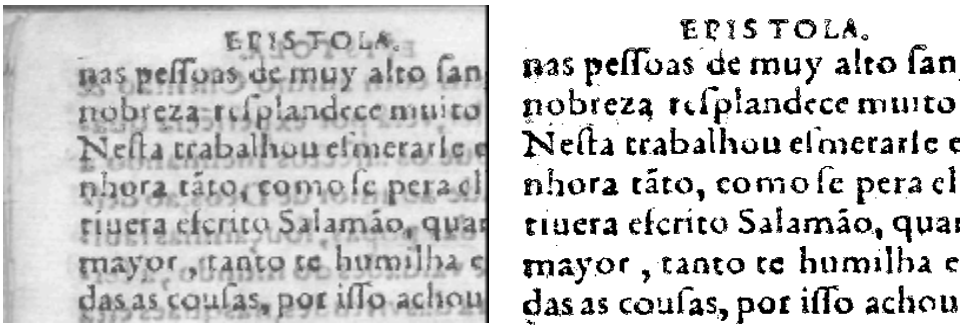



**Figure 5.** *Suppression du verso en transparence*

### *2.7. Les problèmes de la compression des caractères par redondance de formes*

La compression par redondance des formes appelée « Pattern Matching and Substitution » (PMS) ou « token-based compression », est une compression basée sur un dictionnaire des formes redondantes. Elle permet de transmettre l'adresse d'une forme stockée dans un dictionnaire à la place de la forme elle-même. Cette compression a un long passé historique puisque c'est une extension de l'algorithme de compression à base de dictionnaire créé par Lempel Ziv et Welch (LZW) appliqué aux formes et qui a été déjà imaginée par Asher et Nagy au début des années 70 [ASH74], puis développée quelques années plus tard par IBM [WON82][MOH84]. Cette forme de compression est à la base une compression avec perte, car tous les caractères sont remplacés par la même image, arbitrairement choisie dans le texte et qui ne représente pas exactement la forme originale. Cette méthode n'a jamais vraiment été exploitée jusqu'aux travaux de Witten [WIT94][ING94] dans les années 90 qui a décrit une compression sans perte basée sur la redondance des formes en codant l'image résiduelle de différence

entre l'image originale et l'image décompressée. Il y a trois problèmes dans compression d'images de textes imprimés par redondance de formes :

*La comparaison des formes de caractères*

Compte tenu du bruit des images binaires, un même caractère est représenté par plusieurs images légèrement différentes. Les artefacts des images binaires sont dus à l'action conjointe de la position aléatoire de la grille d'échantillonnage lors de la numérisation et de la binarisation [SAR98]. Ce bruit est accentué par la mauvaise qualité de l'encrage et du support papier et par l'irrégularité de l'impression pour les ouvrages plus anciens. Si la comparaison est plus rigoureuse alors la taille du dictionnaire augmente. Afin de conserver un niveau de compression satisfaisant il faut augmenter le taux de redondance en effectuant un comparaison approximative des formes en regroupant plusieurs caractères différents mais visuellement similaires dans une même classe. Les erreurs de substitution seront corrigées avec le codage de l'image résiduelle (plan de compensation). C'est les choix d'AT&T et de Xerox qui réalisent une classification approximative des caractères représentés par des vecteurs de caractéristiques ou bien qui utilisent la distance de Hausdorff. Nous avons choisi un taux d'erreur de substitution nul dans l'appariement des formes de façon à pouvoir utiliser les informations sur la redondance des formes de caractères pour une transcription assistée. La comparaison entre les formes de caractères doit donc être à la fois rigoureuse et très tolérante aux déformations importantes des caractères.

*La limitation de la taille du dictionnaire*

Le codage séquentiel des formes redondantes ne permet pas de connaître la taille finale du dictionnaire ; celui-ci peut grossir indéfiniment jusqu'à atteindre une taille excessive qui annulerait les avantages de cette compression. Pour éviter ce problème, on réinitialise généralement le dictionnaire lorsque sa taille devient trop importante. Ainsi les formats existants reconstruisent un dictionnaire pour chaque page perdant ainsi les avantages de la redondance des formes sur la totalité d'un livre. Nous proposons donc une compression d'un ouvrage entier plutôt qu'une compression page par page pour mieux exploiter la redondance des formes de caractères pour à la fois améliorer la compression et permettre la transcription assistée par ordinateur.

*Le codage des résidus*

Le plan résiduel est essentiel pour retrouver la forme exacte des caractères d'imprimerie après la substitution de tous les caractères identiques par une même forme choisie au hasard. L'image résiduelle (c), qui correspond à la différence entre l'image d'origine (a) et l'image générée avec les formes de caractères (b), est très difficile à comprimer car elle est constituée de séquences de pixels localisées aléatoirement le long des contours des caractères. Ce problème a été la cause principale de l'échec des méthodes de compression par redondance des formes. Les méthodes de compression des résidus proposés jusqu'à maintenant sont tous des

codages avec perte d'information [HOW96] [KIA97]. Nous proposons donc un codage sans perte des résidus à partir d'un modèle de prédiction sur leur


localisation à partir de la théorie de la numérisation.

**Figure 6.** a) *Image d'origine, b) image décomprimée, c) l'image résiduelle*.

### *2.8. Comparaison des formes de caractères*

Cette comparaison entre les formes doit donc être à la fois rigoureuse et très tolérante aux déformations importantes des caractères. Le calcul de la similarité entre deux formes s'effectue à partir des différences symétriques entre la nouvelle forme et son modèle. La théorie de la numérisation [GRO99] stipule que tous les résidus connexes au contour sont le produit du déplacement aléatoire de l'échantillonneur durant la phase de numérisation et doivent donc être ignorés. Par conséquent nous ne comptabilisons que les pixels de différences qui sont distants de plus d'un pixel du contour, qui forment des régions d'épaisseur supérieure à un pixel et que nous comparons au périmètre total. Cette mesure de similarité entre deux formes est comparée à un seuil, défini par l'utilisateur, qui dépend de la tolérance aux déformations. Ce seuil varie en fonction de l'époque du document et de sa qualité de conservation. Comparé à la distance de Hausdorff qui mesure en réalité l'épaisseur maximale des régions de différence, avec des valeurs comprises entre 0 et 5 pour un seuil d'acceptation en général de 1, notre mesure de similarité est graduée sur une échelle plus fine et permet de tolérer des variations infimes ou importantes des formes tout en évitant les erreurs de substitution.

### *2.9. Compression des résidus*

Notre objectif est donc de comprimer le plus efficacement possible les résidus sans aucune perte d'information pour conserver une image fidèle à l'originale. La carte des résidus, calculée par la différence symétrique entre les formes substituées et les formes originales, dépend entièrement des critères d'appariement des formes décrits précédemment. Ces résidus présentent donc une organisation prévisible qui peut être modélisée pour la compression. Nous avons effectué un certain nombre d'observations qui constituent la base de notre approche :

**Conséquence de l'appariement et de l'effet de l'échantillonnage**

*A1 : Les pixels résiduels sont toujours connexes au contour*

**Conséquences du calcul des résidus par différence symétrique**

*A2 : Les résidus alignés dans une même direction ont fréquemment la même distribution,*

*A3 : Dans une direction donnée, les pixels résiduels se trouvent exclusivement soit complètement à l'intérieur soit strictement à l'extérieur de la forme.*

L'observation A3 permet de définir complètement la localisation d'un résidu par rapport au contour à l'aide d'une seule valeur K (figure 7) codé sur 3 bits pour 8 configurations possibles. Ce codage exploite la corrélation des configurations des résidus selon les directions orthogonales aux contours. A partir des observations A1 et A2, on exploite la redondance des séquences successives de valeurs de K le long des contours. Cette suite de valeurs $K_1..K_n$ est très efficacement comprimée par un codage arithmétique sans perte avec un contexte limité aux deux valeurs précédentes

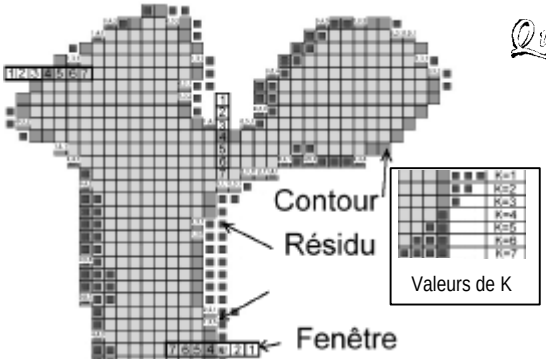



**Figure 7**. *Codage des résidus*

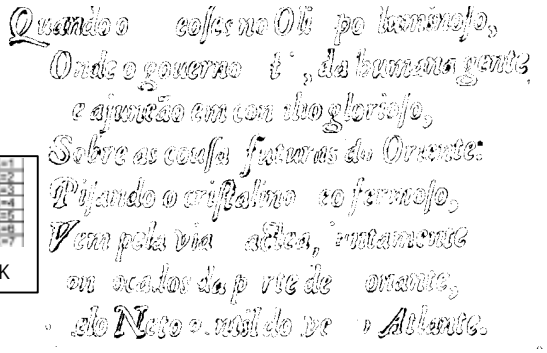


**Figure 8.** *Résidus d'un document*

### *2.10. Résultats de la compression*

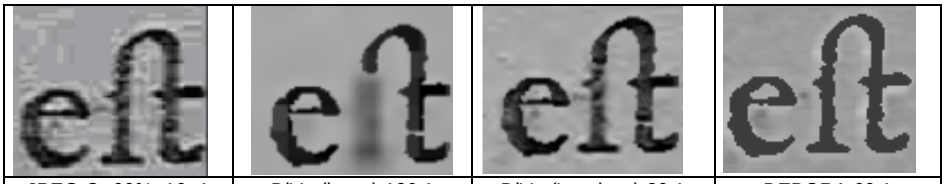
Les résultats de la compression de six ouvrages provenant de différentes bibliothèques sont décrits dans le tableau présenté figure 9. Ce tableau récapitule par livre, le nombre de pages, la taille des images avant et après la compression, le poids du dictionnaire de formes de caractères ainsi que le nombre d'éléments de ce dictionnaire pour évaluer le temps de transcription.

| Cote de l'ouvrage | Nombre de pages | Taille des images non compressées (A) | Taille des images (B) compressées | taille du dictionnaire | nombre de caractères dictionnaire | Taux (A) : (B) |
|---|---|---|---|---|---|---|
| BML373177 | 98 | 1224412 Ko (1,1 Go) | 18625 Ko | 592 Ko | 5127 | 65 : 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| BML341145 | 452 | 3679280 Ko (3,5 Go) | 68914 Ko | 1440 Ko | 13458 | 53 : 1 |
| BML355882 | 93 | 1034112 Ko (0,9 Go) | 15456 Ko | 495 Ko | 4588 | 66 : 1 |
| COI-RB-23-6 | 173 | 1508673 Ko (1,4 Go) | 27507 Ko | 1580 Ko | 12939 | 54 : 1 |
| COI-RB-32-4 | 373 | 1015200 Ko (0,9 Go) | 21226 Ko | 1183 Ko | 11176 | 47 : 1 |
| Côte 15992 | 363 | 770633 Ko (734 Mo) | 10537 Ko | 2025 Ko | 17565 | 73 : 1 |

**Figure 9.** *Résumé statistique des résultats de la compression des ouvrages*

Les taux de compression sont en moyenne de 50:1. Les informations à télécharger pour avoir un texte lisible à l'écran varie entre 3 et 8 Ko par page suivant le nombre de pages de l'ouvrage. Plus le nombre de pages est élevé, plus le taux de compression du plan textuel par redondance de formes augmente.



| JPEG Q=60% 10: 1 | DjVu (lossy) 120:1 | DjVu (Lossless) 20:1 | DEBORA 60:1 |
|---|---|---|---|

**Figure 10.** *Comparaison visuelle entre les différents algorithmes*

Notre taux moyen de compression sans perte se situe à mi-chemin entre les deux réglages extrêmes de la compression DjVu c'est à dire entre le mode sans perte (lossless) et celui avec perte (lossy). DjVu montre que l'on peut donc doubler le taux de compression en acceptant une certaine perte raisonnable sur l'image de l'arrière-plan ainsi que sur les pixels résiduels de la compensation. En effet plus de la moitié de la taille de nos fichiers est occupée par le codage des résidus. Nous avons constaté qu'un simple filtrage des résidus par un lissage des contours ou l'élimination systématique des pixels résiduels isolés permettait de doubler le taux de compression. La compression avec perte pose toutefois le difficile problème du niveau d'acceptation des dégradations visuelles par les utilisateurs et remet en cause les possibilités d'utilisation de la compression pour une conservation pérenne des images. Il s'agit maintenant de réfléchir sur l'information résiduelle inutile qu'il faut éliminer sans altérer la fidélité avec l'image originale.

Le taux de compression obtenu est au minimum dix fois supérieur à celui obtenu par la compression JPEG pour un meilleur rendu visuel. Avec ce niveau de compression, nous pouvons enregistrer plus d'une dizaine de livres numérisés par Cdrom alors que sans compression, plusieurs Cdroms sont nécessaires pour stocker un seul livre. La compression permet de transmettre progressivement les informations pour l'affichage d'une page en commençant par le texte, les zones graphiques puis l'arrière-plan et les informations sur la couleur. Les utilisateurs peuvent sélectionner les informations qu'ils désirent afficher suivant leurs usages et l'ordre de réception.

## 3. Extraction automatique des méta-données

Les méta-données extraites pour DEBORA sont de deux natures :

- ***La structure physique et typographique des livres :*** L'analyse de la structuration physique des pages, réalisée lors de la compression (2.5) contient les positions des zones graphiques, des paragraphes, des lignes de texte, des mots, des caractères et leurs redondances. Cette structure va définir à la fois les méta-données sur la mise en page et les informations nécessaires à la détection automatique des lettrines et des ornements. Les objets graphiques sont ainsi classés, en fonction de leurs positions et de leurs tailles, en trois catégories : les lettrines (formes carrées situées en haut à gauche des paragraphes et entourées de textes sur la droite), les ornements (formes situées dans les zones hors texte, en haut comme les bandeaux, ou en fin de paragraphe comme le cul-de-lampe), et enfin les illustrations et gravures. L'information sur la redondance des formes de caractères sur l'ensemble de l'ouvrage constitue aussi une méta-donnée importante pour les historiens du livres qui peuvent analyser la régularité de la construction d'un livre en étudiant les changements de police de caractères et donc une rupture de la redondance des formes de caractères sur certaines pages qui trahissent des insertions de nouvelles pages. Les typographes peuvent aussi constituer un catalogue des informations typographiques détaillées sur les formes de mise en page et sur la typographie des caractères et leurs défauts pour authentifier l'éditeur d'un ouvrage inconnu, comparer les versions différentes d'un même ouvrage ou analyser les rapports entre les différents imprimeurs d'une ville.

- ***La transcription du texte assistée par ordinateur :***

Les systèmes de reconnaissance optique de caractères (OCR) ne peuvent pas fonctionner sur les ouvrages imprimés du XVI$^{ième}$ à cause de la typographie particulière des caractères (police Fraktur et Civilité…), de la présence de caractères disparus (« s » droit, « et » collé) de la mauvaise qualité de conservation des ouvrages, d'une mise en page obsolète et de l'absence de dictionnaires des mots courants de cette époque. La modification des OCR commerciaux serait aujourd'hui trop coûteuse pour un marché aussi restreint. La solution passe par la transcription assistée par ordinateur. Elle consiste à saisir manuellement une seule fois les caractères qui présentent une forme différente des autres caractères, l'ordinateur pouvant aider l'opérateur à ne pas ressaisir les formes similaires. Les taux de redondances de formes de caractère sur les ouvrages du XVI, mesurés par l'algorithme de compression, dépassent 98% avec un taux d'erreur presque nul, ce qui permet de faire la transcription du texte avec la saisie de moins de 2% des formes de caractères. La transcription d'un ouvrage entier est réalisée en quelques heures par la saisie du texte correspondant aux différentes formes de caractères codées dans le dictionnaire. Cependant la moitié du dictionnaire concerne des formes rares de caractères généralement dégradés. Par conséquent, la transcription de la moitié du dictionnaire concernant les formes de caractères les plus fréquentes,

permet d'obtenir la transcription de plus de 80% du texte sans erreur, très suffisante pour une recherche par mots.

## 4. Format d'échange de données hétérogènes

Les solutions actuelles en terme de format ne sont pas satisfaisants pour une gestion complète de documents numérisés en terme d'images, de texte (transcription/annotation), et de structures (structure physique ou logique du texte, hiérarchie des annotations…). En effet aucun des formats existant parmi les trois grandes familles de formats n'apportent de solution définitive à la représentation d'une collection de documents numérisés :

**Les formats d'images** (TIFF, PNG, GIF, JPEG…) ne proposent pas de solution efficace sur les images de texte. Aucun de ces formats d'images n'a été conçu pour la description d'ouvrages et de leurs structures

**Les formats d'édition** (RTF/Word, LateX, SGML/TEI, XML…) sont plus adaptés à la représentation des textes que des images mais ils offrent une description structurée. En tant que tels, ils ne gèrent pas la séparabilité des données de façon à ne télécharger que les informations nécessaires par le réseau. Chaque format a ses inconvénients (séparation physique texte/image, absence d'édition dynamique ou du suivi des modifications pour XML, absence de structuration pour word etc..).

**Les formats d'impression** (PS, PDF…) plus adaptés à la visualisation et l'impression des documents, n'offrent pas la possibilité d'édition ni une gestion optimale sur le réseau (pas de téléchargement partiel à la demande). Ces formats généralement propriétaires ne peuvent pas évoluer et utilisent une compression des données très rudimentaire (généralement JPEG sur des images de textes scannés et Zip sur les autres données).

Les bibliothèques numériques ont été contraintes de choisir des solutions techniques inadaptées à l'accès aux ouvrages numérisés ce qui limite actuellement leurs développements sur Internet. Nous avons constaté que la plupart d'entre elles ont choisi de décrire les documents entiers en PDF ou en HTML en mode page avec des pointeurs sur la transcription et sur l'image en basse résolution compressée JPEG pour limiter le volume de transfert sur un réseau bas débit. Ainsi il n'y a pas de possibilité de faire le lien entre le texte et son image et les possibilités de consultations en mode image sont souvent limitées par la faible qualité des images due à la compression et au niveau de résolution des images. Les possibilités de recherche par le contenu sont limitées seulement aux textes (transcriptions) non structurés.

La solution pourrait venir du nouveau standard XML qui s'imposera rapidement comme le format d'échange de documents dans un avenir très proche. Le problème actuel d'une gestion XML des documents numérisés réside dans la

séparation physique entre donnée textuelles et les données non textuelles. Cette séparation entre les méta-données et les données pose un problème de finesse de description et de gestion des liens entre la description et la zone de l'image associée. Les tentatives de description des ouvrages numérisés par un schéma XML sont actuellement en cours de développement [METAe].

### *4.1 Proposition d'un format binaire de données hétérogènes pour DEBORA*

Nous avons développé une solution qui soit une synthèse des propriétés des formats d'image, qui ne nous permettent notamment pas de structurer de manière suffisante notre document, et des formats de données hétérogènes tels que SGML ou XML, pour lesquels, leur adaptation à notre problématique aurait nécessité la mise en œuvre de certaines pirouettes conceptuelles [SKO02][MIC98]. Contrairement à ce que propose XML, nous conservons la notion de fichier unique et cohérent pour tout un document voire une collection de documents. XML se trouve plus particulièrement adapté en tant que format de structuration et de présentation des données alors que notre besoin se porte vers une solution, indépendante des mécanismes ou normes externes (XSLT, XSL, TEI…), de structuration, d'échange, de manipulation et de stockage (avec support des BLOBs).

Notre proposition permettra à l'utilisateur de ne récupérer que les éléments du document qu'il souhaite. Ce choix imposera alors au format d'être séparable, c'est à dire sécable en plusieurs morceaux indépendants pouvant être apportés à l'utilisateur à tout moment. Ainsi, un document correspondant à une vue partielle du document original pourra être reconstitué sur le poste client. Nous choisirons donc de gérer toutes les informations dans un seul fichier décomposable et modulaire qui permet de représenter différents documents avec leurs informations associées (description physique et logique, méta-données, images…). De plus, nous introduirons la gestion de plusieurs représentations physiques d'une même entité logique (multi-présentation**)**. Pour traiter efficacement l'édition et les modifications sur les documents volumineux, nous avons intégré à ce format des mécanismes d'édition dynamique de toutes les composantes de celui-ci. Mais aussi, une gestion multi-utilisateur incluse dans notre format permettra de fixer des groupes de personnes qui ont accès en lecture ou en écriture et d'assurer le suivi des modifications.

Nous avons ainsi défini un format pivot nous permettant de gérer toutes les opérations nécessaires au fonctionnement efficace de systèmes collaboratifs (consultation à la demande, authentification des utilisateurs…). Ce format peut être considéré comme un parent de XML, il est ouvert et semi-structuré. Tout comme lui, il ne définit pas le formatage d'un fichier mais seulement les règles de construction. Ces dernières nous permettront d'élaborer une implémentation particulière pour DEBORA. Des mécanismes d'édition dynamique permettant de

manipuler efficacement les fichiers volumineux. Ces processus évitent une regénération longue et systématique du fichier et permette à l'usager de travailler sans interruption perceptible dans son travail. A l'instar des formats tels que HTML ou SGML, notre format est construit autour de la notion de balises (usuellement appelée tags). Ces dernières décrivent les différents éléments du format comme l'identification des informations, leurs contenus et leur place dans la structure. Ces balises portent toutes les informations permettant les manipulations sur le contenu (type de la donnée, date de création, auteur…).

La forte densité de la structuration des informations (figure 11) qui permet de représenter des méta-données très détaillées et leurs liaisons avec les données d'une collection complète jusqu'au niveau de la page, explique notre choix d'utiliser un format compressé binaire et non ASCII comme XML. Les données stockées dans nos ouvrages et le nombre de liaisons internes avec les méta-données auraient nécessité plusieurs centaines de méga octets de description au format XML-METS pour un encombrement qui serait 10 fois supérieur à celui qu'occupe actuellement nos images. Notre format binaire fortement compressé permet de conserver une finesse de description des méta-données et leurs liens avec les images compressées pour un poids qui reste très raisonnable par rapport à celui des images. La taille des balises et de leurs contenus pour décrire un livre numérisé complet avec la transcription, les différentes structures et le détail du contenu des pages occupe environ 5% de la taille des fichiers soit 10Ko par page environ.

Une collection peut être soit l'ensemble des ouvrages numérisés soit une collection personnelle de quelques pages, réalisée par l'utilisateur. Cette collection est une liste d'ouvrages qui comporte des méta-données de niveau 1 (fiche documentaire du livre : auteur, date, éditeur, titre…) , des méta-données de niveau 2 (structure logique de l'ouvrage) , des méta-données de niveau 3 (contenu d'une page : transcription, structure physique et typographique) et des méta-données de niveau 4 (annotations). Seules les méta-données de niveau 3 sont extraites automatiquement par analyse d'image, les autres sont manuellement saisies grâce à l'interface du poste client/producteur. On remarquera que les informations issues de notre compression sont étroitement liés aux méta-données. Ainsi le dictionnaire de forme de caractères se situe bien au niveau de l'ouvrage.

Une page peut être représentée par plusieurs images de qualité différentes avec ou sans compression ; certaines pages contenant que des illustrations rares peuvent ne jamais être comprimées. Pour les pages qui sont comprimées au format DEBORA, l'usager accède directement aux méta-données de niveau 3 qui autorisent la recherche d'un mot dans la transcription, la localisation des lettrines, des ornements et des illustrations et la recherche d'une mise en page particulière.

On notera l'intérêt d'avoir un format structuré qui gère aussi des liens entre les différents niveaux de l'arborescence. Il existe un champ «transcription » associé au caractère qui peut être différent de celui associé au prototype de forme de caractère.

Ce dédoublement de l'information est nécessaire si l'usager a identifié une erreur toujours possible de substitution d'une forme de caractère par une autre.

Lors de la transcription, l'usager renseigne les champs «Transcription » de chaque prototype de formes de caractère. Ceux-ci sont automatiquement dupliqués dans les champs «Texte » associés à chaque caractère. Ce sont les transcriptions associées aux caractères qui sont affichées, en respectant la mise en page, dans le champs « Texte de la page » et qui seront éventuellement corrigées par l'usager lors d'une phase de validation avec l'image. Les corrections locales éventuelles sur le texte de la page s'effectuent toujours au niveau des caractères, la transcription des prototypes de formes reste inchangée.

Les propriétés du format proposé sont optimales dans une architecture client-serveur pour une consultation distante des ouvrages. Le système que nous proposons

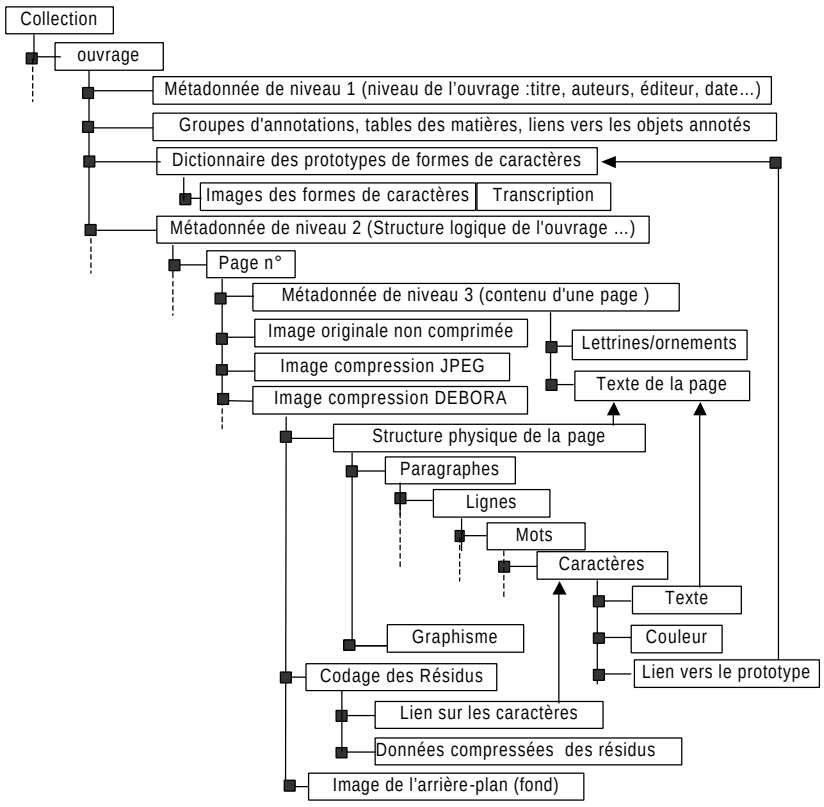



**Figure 11.** *Partie de la structuration des informations du format DEBORA*

repose sur un serveur de livres. Celui-ci gère un ensemble de documents décrit selon notre format. D'un autre coté, le poste client peut exécuter des requêtes

d'interrogation au serveur, ce dernier, grâce aux propriétés de séparabilité de notre format, pourra envoyer uniquement les informations pertinentes pour l'utilisateur. Ainsi une vue partielle du document pourra ensuite être reconstituée localement sur le poste client, ce dernier sera aussi décrit par le format proposé. Un serveur d'index optionnel pourra servir pour gérer plusieurs serveurs de livres de manière transparente pour l'utilisateur final. De plus, la reconstruction locale d'une image partielle du document permet à l'utilisateur d'effectuer des requêtes sans solliciter le réseau et éditer localement des modifications et des annotations sans avoir à être connecté en permanence [TRI02][DEB01].

### *4.2. Fonctionnalités du poste client DEBORA*

Dans le cadre de Debora, nous avons pu développer une structuration adaptée aux documents que nous manipulions, intégrant la structure d'une collection d'ouvrage, les méta-données, des textes, des annotations sur les éléments logiques ou sur une zone d'image. Mais aussi, grâce à des procédés de traitement d'image et de reconnaissance de formes nous avons pu mélanger les images avec les informations textuelles et travailler sur ces images presque comme sur le texte.

Nous avons développé un poste client de démonstration utilisant le format décrit [DEBO01]. Celui-ci nous a permis d'intégrer aisément de nombreux services demandés par les usagers:

- Créer un projet personnel regroupant n'importe lequel des éléments des ouvrages téléchargés et autorisant les annotations privées
- Enrichir les documents numérisés par des annotations
- Editer, suivre les annotations d'un travail public ou privé
- Partager les travaux (liste de pages de texte, d'images, d'annotations… )
- Télécharger à la demande les ouvrages et travailler localement
- Editer et consulter les méta-données de l'ouvrage

Deux types de requêtes sont désormais possibles :

- **Requêtes textuelles :** Rechercher un mot dans les annotations, les notes, les fiches documentaires ou la transcription d'un ou plusieurs ouvrages.
- **Requêtes par image :** Rechercher une forme très similaire de caractères ou bien de rechercher des éléments qui ont été reconnus comme du graphique (enluminures, lettrines, bandeaux, esquisses…).

## 5. Conclusion

Nous avons décrit une compression adaptée aux images de documents en comparaison avec d'autres approches et développé un format de représentation des

documents (par nature hétérogènes) lequel permet aux utilisateurs un meilleur accès aux contenus. Le format inspiré des concepts d'XML et des formats d'images, représente les documents par leurs images, leurs contenus textuels et leurs méta-données (description de l'ouvrage, de sa structure et les commentaires...). Ce format décrit les ouvrages à l'aide d'une structure libre, autorise l'édition et l'annotation tout en conservant les droits des usagers afin de pouvoir assurer la propriété intellectuelle et gère plusieurs représentations avec différents média pour afficher plusieurs interprétations d'un même texte dans plusieurs langues ou plusieurs qualités différentes d'images. Cette représentation associée à une compression intelligente des images de documents qui conserve les liens entre chaque élément de l'image et le texte, est adaptée à une transmission progressive des données sur réseau. Un certain nombre de logiciels ont été développés pour démontrer la validité des idées et la faisabilité de mise en œuvre d'un format qui gère les données hétérogènes sous forme d'objets indépendants. Une maquette opérationnelle de poste client développé à partir d'une étude des usages permet l'édition et le partage des annotations ainsi que l'interrogation par le contenu. Il convient d'insister sur le fait que nos travaux ont été influencés par la prise en compte des besoins des usagers, par les récents travaux sur la compression des images de documents et par les langages structurés de description comme XML. Le projet DEBORA a permis de faire cette synthèse entre les usages et les besoins de plusieurs disciplines (sciences humaines, sciences de l'information et informatique).

## 6. Bibliographie

[ASH74] R. ASHER, G. NAGY, "A means for achieving a high degree of compaction on scan-digitized printed text". IEEE Trans on computers, 23:1174-1179, 1974

[BOT98] L. BOTTOU and al., "High quality document image compression with DjVu", electronics imaging, 7(3):410-428, 1998

[DEB01] DEBORA : http://debora.enssib.fr/ , Logiciels de démonstration: http://rfv6.insa-lyon.fr\debora\client.htm

[GRO99] **A. Gross, L. J. Latecki**, '*Digital geometric methods in document image analysis'*. Pattern Recognition 32 (1999) pp. 407-424.

[HOW96] P. HOWARD, "lossless and lossy compression of text images by soft pattern matching", Proc. of the IEEE Data compression Conference pp. 210-219, 1996.

[HOW98] P. HOWARD and al., "The emerging JBIG2 standard", IEEE trans. on Circuit and Sys. for video tech., vol 8, n°5, 1998.

[ING94] S. INGLIS, I. WITTEN, "Compression-based template matching", IEEE data compression conference, pp. 106-115, 1994.

[KIA97] O. E. KIA, "Document Image Compression and Analysis", PhD of the university of Maryland, 1997, 191 p.

[LEB01] F. LEBOURGEOIS and al., "Networking Digital Document Images". 6 ed. ICDAR. 2001, Seattle USA. pp. 379-383

[LEB02] F. LEBOURGEOIS and al., "Compression de documents imprimés numérisés". CIFED'2002, Hammamet, Tunisie, 21-23 oct. 2002, pp. 195-204.

[METAe] Projet Européen METAe, http://meta-e.uibk.ac.at/

[MIC98] A. MICHARD, "XML : langage et applications", 1998, Paris: Eyrolles,. xi, 361 p .

[MOH84] K. MOHIUDDIN and al. "lossless binary image compression based on pattern matching", Proc. of the Int. Conf. On computers, systems and signal processing, pp. 447-451, 1984.

[NIB86] Niblack Wayne, *An introduction to digital image processing*, 1986, Hemel Hempstead: Prentice Hall, 216 pages

[SAR98] Sarkar Prateek, Nagy George, Zhou Jiangying et Lopresti Daniel, *Spatial Sampling of Printed Patterns*. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 20, 1991, pp. 344-351

[SKO02] A. SKONNARD and al., Essential "XML quick reference : a programmer's reference to XML, Xpath, XSLT, XML Schema, SOAP, and more", 2002, Addison-Wesley. 402 p.

[TRI02] E. TRINH and al., "Un format de document ouvert et optimise pour une exploitation collaborative", CIFED'2002, Hammamet, Tunisie, 21-23 oct. 2002, pp. 113-120.

[WIT94] I. WITTEN, "textual image compression : two stage lossy/lossless encoding of textual images", IEEE, 82:878-888, 1994.

[WON82] K. WONG and al., "Document analysis system", IBM journal of research and development, 26:647-656, 1982

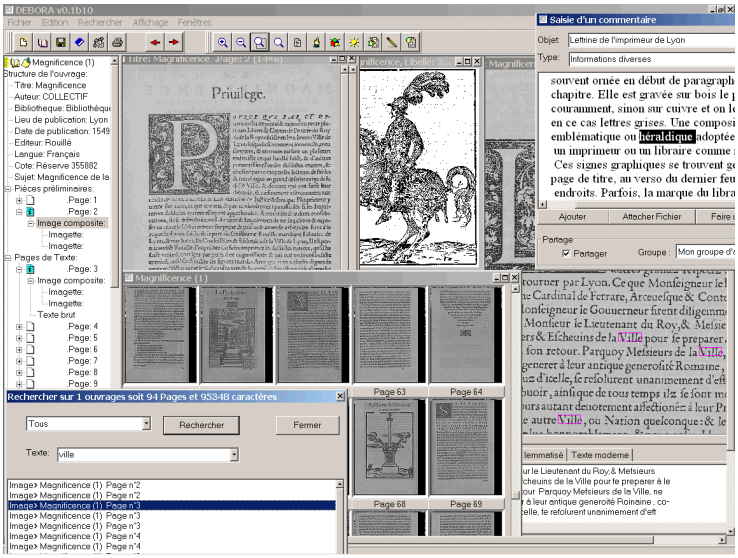


**Figure 12.** *Interface de consultation du poste client DEBORA*

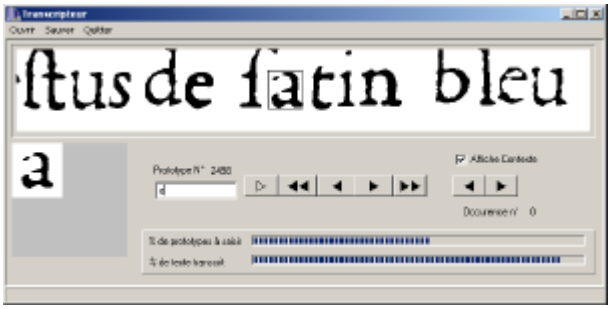



**Figure 13.** *Logiciel de transcription*