



Université Claude  
Bernard - Lyon 1

## LES LOGICIELS DE NUMERISATION DES LIVRES ANCIENS



**Par livres anciens, on entend les livres imprimés ou édités avant 1801 et les publications postérieures éditées artisanalement.**

### **Sauvegarder le patrimoine**

Le sentiment que le livre est le moyen le plus sûr de transmettre un héritage, un savoir, remonte à des époques lointaines. Or il existe quelque 25 millions d'ouvrages et documents anciens conservés dans les bibliothèques et archives françaises qui sont menacés de disparaître, rongés par l'acidité du papier et par l'humidité. Il est capital de préserver ce patrimoine, mémoire du millénaire qui s'achève.

Avec Gutenberg, l'humanité a vécu la révolution de l'imprimerie. Microformes et photographies ont permis depuis des années d'accéder à la consultation. La numérisation constitue désormais une alternative, et ouvre de nouveaux horizons pour le livre et l'écrit.

Plusieurs bibliothèques aujourd'hui, pour des raisons de conservation et d'utilisation, sont en train de produire des archives numérisées, mais ces initiatives sont toutefois isolées et ne suivent pas toujours des règles homogènes.

C'est à noter aussi la permanence d'un certain *manque d'intérêt pour des logiciels innovateurs* de la part des personnes auxquelles ces logiciels sont adressés.

Chaque système qui, par exemple, réalise seulement l'indexation et les concordances d'un texte, représente *un outil partiel et lointain* des besoins technologiques réels d'un chercheur. De même, un excellent programme de navigation dans les archives des images numérisées, avec la possibilité d'effectuer des agrandissements, des variations de luminosité ou des contrastes en temps réel, est considéré comme avantageux par rapport aux limites d'un lecteur traditionnel de microfilms. Ce grand progrès de la technologie numérique pourrait paraître toutefois stérile s'il n'est pas complété par un système qui lie entre eux le texte et les images de manière à faciliter les processus d'interprétation par le paléographe ou par le codicologue ou par le philologue.

Ce document s'intéresse aux logiciels de numérisation des livres anciens, produits très peu commercialisés et qui se développent que ces deux-trois dernières années.

On trouve relativement peu d'article sur ce thème. Les informations sur l'état de la recherche se trouvent dans des revues scientifiques spécialisées (sur les documents numérisés et les bibliothèques, sur les documents anciens), des rapports de recherche ou compte-rendu de congrès, et les informations de type commercial sont diffusées par les producteurs ou distributeurs de logiciels via Internet.

### **1. Quelques éléments techniques**

Des méthodes se mettent en place pour la création et la réalisation des logiciels correspondants pour extraire les informations nécessaires au fonctionnement de la banque de données d'images. Ces données seront extraites des informations stockées (entités numérisées en mode d'image et informations complémentaires). Il ne s'agit pas de créer des logiciels pour l'OCR ou pour la dématérialisation mais des outils permettant une sorte de premier accès aux textes. Ils pourront requérir l'assistance de l'homme pour des tâches non triviales et non répétitives.

Le logiciel doit être capable de faire une analyse de bas niveau des images permettant : de séparer les composants des documents (texte, illustration, lettrines, etc.) ; de segmenter les textes en lignes ; de segmenter les lignes en graphèmes (mots).

L'analyse des formes, des images de mots (et de groupes de lettres) par ces logiciels permet : d'élaborer des distances et des ressemblances entre les mots dans le contexte des documents anciens ; de retrouver un mot par son image ; de déterminer une liste d'index ; d'accéder partiellement au texte à partir de la table des matières (ou de la table des index) et l'analyse et la comparaison des composantes non textuelles : lettrines, bandeaux, ornements, etc.

(cela pour le livre imprimé).

Pour les manuscrits médiévaux les difficultés se multiplient. Même si l'informatique a fait d'énormes progrès en reconnaissance des caractères manuscrits, il reste difficile de faire lire à l'ordinateur les documents médiévaux. Les caractères des manuscrits (et des incunables) médiévaux calligraphiés (ou typographies) ont une variabilité de formes qui reste importante et ne permet pas de bien résoudre aujourd'hui les problèmes de segmentation (de lignes en mots, de mots en caractères) et de reconnaissance de caractère. Lorsque la lisibilité du manuscrit n'est pas suffisante, le travail de paléographe (la transcription) qui se fait au crayon, est nécessaire. L'outil de correspondance texte/image se décompose en deux traitements distincts : la reconnaissance automatique des lignes et des colonnes et l'algorithme de correspondance mot/image.

## 2. Les applications

(Chaque programme de numérisation -> son logiciel propre)

Il existe plusieurs projets de numérisation des livres anciens, qui pour l'instant ne touchent que les bibliothèques les plus importantes des pays industrialisés.

Au niveau de logiciel, chaque bibliothèque « intègre » son propre logiciel pour la numérisation.

Les applications dans le domaine des livres anciens sont très diverses et intéressantes. Chaque « grande » bibliothèque ou programme de numérisation dans ce domaine a pu adapter son « propre » logiciel, selon ses besoins et ses intérêts.

### L'expérience de numérisation des manuscrits à la bibliothèque Vaticane

- La première série d'application dérive de la plus grande quantité de points concentrés sur une même surface : c'est celle des *agrandissements*.

- L'application aux manuscrits : ils arrivent à faire pâlir, voire à *faire disparaître une tâche d'humidité* absorbée par le papier et qui rend illisible l'original, cependant que les mots de l'original apparaissent de manière très claire sur l'écran de l'ordinateur.

- Un autre cas est celui de *l'encre trop acide* qui a traversé la feuille de papier en sorte que l'écriture apparaît sur le verso et se confond avec celle de la page suivant : avec un peu d'habileté et beaucoup de compétence, un opérateur peut « nettoyer » la page, laissant bien visible la seule écriture qui l'intéresse.

- Des bons résultats sont obtenus dans *la lecture de mots effacés*, qui jusqu'à présent n'était pas possible à lire à l'aide de la lampe de Wood. Une application exceptionnelle se trouve dans le domaine des *palimpsestes*, où il n'était pas possible avant de réussir à déchiffrer l'écriture inférieure. Les techniques d'élaboration électronique des images si raffinées ont permis d'obtenir des résultats extraordinaires et de montrer à l'écran l'écriture inférieure pratiquement dans l'état où elle se trouvait dans le manuscrit avant d'être effacée pour en permettre le emploi.

## 3. L'utilisateur

Qui s'intéresse de la numérisation des documents anciens ?

- L'utilisateur général d'une bibliothèque qui souhaite examiner des sources manuscrites ou imprimés anciens.
- L'étudiant spécialisé en histoire des textes : philologues ou éditeurs critiques de travaux classiques ou médiévaux qui utilisent différents types de support: papier, papyrus, pierre. Ceci inclut, de fait, des étudiants en textes anciens comme les papyrologues (spécialistes dans l'étude des papyrus), les épigraphistes (spécialistes de l'étude scientifique des inscriptions – appelées Incipit – placées en tête d'un livre, d'un chapitre), les paléographes (spécialistes en science des écritures anciennes), et les codicologues (spécialistes étudiant le support des manuscrits).
- Les chercheurs qui mènent des études de philologie (étude historique d'une langue par l'analyse critique des textes) ou d'histoire en général.
- Tout ceux, qui sont intéressés par l'étude, l'annotation, ou la transcription de textes contenus dans des documents anciens numériques.

## 4. Caractéristique des produits

4.1 - Les principaux produits : Panorama de l'offre logicielle actuelle

### • **TransVision**

#### Les éléments constitutifs d'une base TransVision

Il propose deux modes d'accès à l'information : un mode d'accès de type arborescent, et un mode d'accès par questionnement textuel sur les éléments de la base.

Un *objet-image* est l'élément de base (élément documentaire) qui contient de une à quinze images et une référence texte qui est composée des champs définis au moment de la création de l'entité. Chaque élément de la référence texte renseigné par le producteur est indexé afin de permettre une consultation par questionnement.

#### Saisie et mise à jour des données

TransVision propose un certain nombre de concepts introduisant des notions telles que la production répartie d'informations sur un réseau, la navigation arborescente dans une banque d'images, l'interrogation multi-critère des éléments qui la composent, et la gestion des droits d'accès à l'information directement assurée au niveau des producteurs de la base. Une des particularités de l'application est qu'elle gère une mise à jour quasi temps réel de toutes les informations produites ou modifiées.

#### Application serveur

L'application TransVision est composée de trois serveurs. Un serveur d'administration, un serveur de gestion de la banque image et un serveur d'indexation. Les serveurs sont développés en langage C sous Unix.

#### Les clients « natifs »

Sont les applications clientes spécialement développées pour la production et la gestion des banques image TransVision. Ces applications ont été développées sur des postes Macintosh et PC (Windows) et forment une interface graphique entre l'utilisateur et le serveur TransVision.

Le client TransVision permet d'intégrer **différents formats d'images fixes ou d'images animées**. Le choix des formats reconnus par l'application est directement lié aux formats les plus utilisés dans ce domaine, tels que JPEG, GIF, TIFF version 6, KODAK (en lecture seulement) pour les images fixes, et MPEG 1, MPEG 2 et les séquences QuickTime pour les images animées. Il intègre des fonctionnalités de traitement d'images.

#### • **BAMBI**

Cette station permet d'enrichir les informations associées à l'image en entrant manuellement la transcription (si elle n'existe pas) et des annotations. L'utilisation de la composante hypermedia assure des possibilités intéressantes de navigation. Les liens entre images et textes représentent une fonction très innovante de cette application. Tous ces éléments confèrent à BAMBI son originalité par rapport aux produits existants sur le marché.

#### Description de la station BAMBI

Le système (BAMBI) examine d'abord les valeurs des niveaux de gris pertinents pour chaque ligne, et les évalue le long d'un axe vertical ; plus précisément, le système essaie d'identifier les séparations entre les mots ; il exploite la transcription textuelle à partir de laquelle il est possible d'extraire le nombre exact de mots pour chaque ligne, pour contrôler la segmentation de la ligne de l'image en zones de mots. Les utilisateurs souhaitent accéder à un certain nombre de services comme la possibilité de transmettre l'image d'un manuscrit et de recevoir en retour la transcription (ou une partie de la transcription) réalisée par OCR (*Optical Character Recognition*).

#### Recherche d'un manuscrit

La station BAMBI offre un certain nombre d'outils de recherche qui permettent d'accélérer la sélection de documents. Ces outils sont basés sur une recherche multi critère par méta données ou par mots-clés.

Lorsqu'un manuscrit a été sélectionné, une fenêtre principale s'ouvre et comporte cinq zones de travail :

- l'image de la page de la manuscrit,
- la transcription correspondante en texte (code ASCII étendu),
- une liste de marque-pages contextuels, que l'utilisateur estime utile à son travail (contenu ou sujets similaires),
- des annotations sur des mots ou des groupes de mots (phrases par exemple) de la transcription,
- un verborum des mots du manuscrit.

La transcription d'un manuscrit pour Bambi est une opération manuelle qui doit respecter un certain nombre de règles et de conventions afin d'être interprétée et utilisée correctement par l'application BAMBI. Elle peut être exportée vers un fichier de type RTF ou SGML, lui permettant d'être réutilisée, soit dans des programmes de traitement de textes standard, soit dans des systèmes de gestion de documents. Des possibilités de zoom sur l'image rendent plus lisible le manuscrit.

#### Indexation de transcriptions

Lorsque la transcription est complète, l'outil d'indexation génère un *index verborum* et un *index locorum*. L'index verborum contient tous les mots apparaissant dans la transcription (sans les caractères( , [ ] et < > ) ainsi que les mots corrigés par l'utilisateur avec la fonction de variante de texte.

L'utilisation de plusieurs alphabets dans le même manuscrit (Grec et Latin par exemple) nécessite la création d'index locorum pour chaque alphabet. L'index locorum permet de visualiser les positions où mot apparaît dans le manuscrit.

La référence à un mot donné prend la forme d'une liste contenant le numéro de page, le numéro de colonne, le numéro de ligne et la position du mot dans la ligne. La technique d'indexation utilisée est l'indexation *full-text*.

#### Annotations sur les transcriptions

Des annotations peuvent être ajoutées aux travaux des historiens sur les manuscrits. D'une part, ces annotations permettent d'une part d'ajouter des commentaires personnels sur un groupe de mots dans le texte et, d'autre part, de pouvoir ajouter des corrections ou des synonymes (variantes de texte).

#### Correction manuelle des résultats de la correspondance automatique

L'algorithme qui fait correspondre mot et image, bien que très puissant, ne permet pas d'identifier tous les blocs de mots. Certaines caractéristiques du manuscrit (taches, mots accolés, illuminations, etc.) nécessitent des interventions manuelles afin d'être correctement analysées. Une fois que la correspondance automatique est terminée, l'utilisateur peut agir librement sur le résultat afin de corriger ces erreurs. Les corrections effectuées seront conservées dans la base de données et appliquées à chaque nouveau lancement de l'algorithme.

4.2 – Des versions du même produits (DigiBook)

	OUVRAGES ACCEPTES		NUMERISATION		COMPRESSION ET STOCKAGE D'IMAGE
<b>DigiBook 5400 et 5600</b> Stations de numérisation haute productivité pour ouvrages à reliure rigide	Formats	DIN A5, DIN A4, DIN A3 x 2	Type de capteur	CCD linéaire N&B 5000 pixels	JPEG pour les images en niveaux de gris TIF G4 pour les images binaires
	Hauteur et largeur maximum numérisables	500 mm (h) x 840 mm (w) 600 mm (h) x 840 mm (w) – DIN Aa2 x 2 en option	Type de numérisation	256 niveaux de gris ou binaire	
	Epaisseur maximum	120 mm	Resolution optique	200 à 600 dpi	
	Poids maximum (équilibre type Roberval des plateaux)	15 kg	Résolution numérique	400 à 600 dpi (interpolation)	
			Réglage format / résolution	automatique	
<b>DigiBook (a) 2000 et (b) 3000</b> Station de numérisation haute productivité pour ouvrages à reliure souple	Format	du DIN A5 au DIN A2 x 2	Type de capteur	CCD linéaire N&B 500 pixels	(a) JPEG (b) JPEG ou TIFF
	Hauteur et largeur maximum numérisables	600 mm (h) x 840 mm(w)	Type de numérisation	(a) 256 niveaux de gris (b) 256 niveaux de gris ou binaire	
	Epaisseur maximum	4 cm	Résolution optique	(a) 200 dpi sur tous les formats (b) 200 dpi (h>300 mm) 400 dpi(h<300 mm)	
	Epaisseur maximum avec option porte livre	(b) 8 à 12 cm	Résolution numérique	(a) 300 et 400 dpi (interpolation) (b) de 300 à 600 dpi (interpolation)	
	Poids maximum	15 kg	Réglage format / résolution	(a) Position fixe (b) 2 positions - manuel	

<b>Book Restorer</b>	CONSULTER	RESTAURER	SAUVEGARDER ET PUBLIER
<p>Le logiciel <i>Book Restorer</i> intervient dans la vérification des documents numérisés, dans leur restauration, et dans leur édition.</p> <p>Il peut être installé sur tout PC fonctionnant sous environnement Windows NT.</p> <p>Ainsi, un atelier de restauration mono ou multipostes en réseau.</p>	<p>Permet d'afficher et de consulter l'ensemble des images d'un livre numérisé avec <i>DigiBook</i>.</p> <p>Avec lui est possible de réorganiser très simplement l'ensemble des fichiers</p>	<p>Permet de retravailler à la demande les pages déjà numérisées par <i>DigiBook</i></p> <ul style="list-style-type: none"> <li>• en définissant une ou plusieurs zones de formes quelconques, optimiser le contraste, tramer, gommer des taches</li> <li>• peut redresser le texte,</li> <li>• corrige la courbure de la page soit automatiquement, soit manuellement, dans le cas d'ouvrages très altérés.</li> </ul>	<p>Lors de la phase de sauvegarde, il peut rectifier et améliorer la mise en page : centrer différemment le texte, modifier les marges, changer le format de sauvegarde...</p> <p>Les images sont compressées en TIF G4 pour les images binaires et en JPEG pour les images en niveau de gris.</p>

4.3 - Produits « concurrentes »

Le projet ou la station	La société	Caractéristiques
<p><b>DEBORA</b> (Digital accEss to Books of the RenAissance) 1/1/1999 -&gt; 30 mois Projet européen qui développe des outils permettant l'accès, a partir de postes de consultation distants, à des collections de documents du 16<sup>ème</sup> siècle</p>	<p>Xerox -En association avec I2S (Bordeaux) (Le scanner Digibook)  SGBI Entreprise (Transvision)</p>	<p>-Logiciels d'amélioration et de restauration des images -Logiciels pour extraire les informations nécessaires au fonction de la banque de données d'images -Logiciels d'analyse des formes, des images de mots -Logiciels permettant d'accéder partiellement au texte à partir de la table des matières -Logiciels pour l'analyse et la comparaison des composantes non textuelles : lettrines, bandeaux, ornements, etc.</p>
<p><b>Philectre</b> (PHILologie ELECTRONiquE)  Projet conduit de 1994 à 1997 et qui avait pour but d'explorer les techniques informatiques nécessaires aux chercheurs en sciences littéraires, notamment les généticiens et les médiévistes.</p>	<p>Projet dans le cadre de l'appel d'offre « Mutation de l'édition induites par le livre électronique » du GIS Sciences de la cognition du CNRS et qui a réunis des chercheurs en histoire des textes et des chercheurs en informatique</p>	<p>1-Extraction de lettrines et traits filiformes  <ul style="list-style-type: none"> <li>• Extraction de la graphie</li> <li>• Extraction des colonnes et des lignes</li> <li>• Extraction des lettrines</li> </ul>                 2-Etude de versions et variantes  <ul style="list-style-type: none"> <li>• Aider à trouver des variantes ou des portions communes,</li> <li>• Montrer la chronologie des variantes ou versions</li> <li>• Montrer en parallèle une ou plusieurs versions</li> <li>• Travailler sur elle, indépendamment ou en relation avec les autres...</li> </ul> </p>
<p><b>Initiale</b>  Banque de données d'images numériques des enluminures des manuscrits depuis 1997..  -Format de visualisation est le format Scopus</p>	<p>Base de données , produit de l'IRHT (Institut de recherche et d'histoire des textes); gérée sous Taurus (logiciel documentaire aujourd'hui de la société EVER)  -(développé par la société Avelem)</p>	<p>Les images proviennent, d'une part, de la numérisation rétrospective des diapositives existantes et, d'autres part, depuis janvier 1998 de la photographie numérique directe des manuscrits.  -Il permet 5 niveau d'images (vignette, imagette, image plein écran, puis quart d'image et seizième d'image : soit un cinquième niveau de très haute définition).</p>
<p><b>BAMBI</b> (Better Acces to Manuscript and Browsing of Images)  Une station de travail pour historiens travaillant sur des textes anciens</p>	<p>Basé sur un projet développé au CNR (Comitato Nazionale della Ricerca) de Pise dans le Laboratoire ILC (Institute for Computational Linguistics).</p>	<p>La station BAMBI permet de  <ul style="list-style-type: none"> <li>• visualiser l'image d'un document source (un manuscrit) avec une haute résolution</li> <li>• transcrire, annoter et indexer le texte contenu dans les images</li> <li>• visualiser la transcription et l'Index Locorum dans une fenêtre adjacente à celle du document source</li> <li>• faire correspondre automatiquement chaque mot de la transcription avec la portion de l'image source dans lequel le mot est trouvé</li> <li>• exporter des informations sur les manuscrits au format SGML/HyTime</li> </ul> </p>

<p><b>Arkhênum</b></p>	<p>développées par la société bordelaise <u>I2S</u></p>	<ul style="list-style-type: none"> <li>-Ouvrages posses seur des plateaux autocompensés</li> <li>- Pas de vitre</li> <li>- Lumière froide sans UV, tout ceci dans un esprit de conservation de l'intégrité de l'ouvrage (reliure, papier, encres)</li> <li>- L'absence de vitre ; la correction de la courbure naturelle des feuilles est réalisée par <i>traitement logiciel</i></li> <li>- Le logiciel employé permet également de détourer, d'éliminer les tâches (réglages sur le contraste), d'effectuer des réglages sélectifs dans la page</li> </ul>
------------------------	---	--

Certains des produits présentes ci-dessus (Debora & Initiale), sont en cours d'élaboration, il est donc difficile de dire aujourd'hui lequel est le mieux adapte ou le plus performant .

Concernant Bambi et Philectre, il faut se rapporter a la rubrique « Les principaux produits » du tableau ci-dessus.

### Diffusion d'informations

<p>TransVision</p>	<p>TransVision permet de diffuser les entités qui auraient été rendues publique par leur créateurs vers des applications clientes largement répandues telles que les clients W3 (<i>Netscape, Microsoft explorer...</i>).</p>
<p>BAMBI</p>	<p>La station BAMBI est actuellement locale, les images des manuscrits (au format JPEG), les transcriptions et les fichiers de mise à jour se trouvant sur un ou plusieurs <i>CD-Rom</i>. (Une ouverture vers une solution de type Internet ou Intranet est envisagée.</p>
<p>Banque d'image de l'IRHT</p>	<p>L'archivage des fichiers est fait au format TIF sur des CD-Roms (650 MO par CD-Rom pour environ 250 images).</p>

## 5. Bibliographie sommaire et sites web

### Documents écrits

**CALABRETTO,S.; PINON,J-M.; BOZZI,A.** BAMBI : système de gestion de manuscrits anciens pour historiens. *Document numérique*. Vol. 2, n. 3-4/1998, p. 31-50

**Gusnard de Ventabert.** Représentation et exploitation électroniques de documents anciens (textes et images) : à propos d'expériences du projet Philectre. *Document numérique*. Vol. 3, n. 1-2/1999, p. 57-73

**LALOU, Elisabeth.** La numérisation des manuscrits médiévaux à l'Institut de recherche et d'histoire des textes. *Document numérique*. Vol. 3, n. 1-2/1999, p. 29-38

**PIAZZONI, Ambrogio M.** Vers une paléographie électronique ? : l'expérience de numérisation des manuscrits à la bibliothèque Vaticane : quelques réflexions. *Gazette du livre médiéval*, n. 33, automne 1998

### Sites web

<http://debora.enssib.fr> pour le programme européen DEBORA

<http://gutenberg21.com/pages/ecrit/b3-b.htm> pour la station de numérisation DigiBook

<http://www.arkhenum.com/technique.htm> pour la station de numérisation Arkhênum

<http://www.cei-sgbi.insa-lyon.fr/SGBI.htm> Pour SGBI Entreprise et le logiciel TransVision

<http://gallica.bnf.fr> Pour le serveur expérimental de la BnF (l'édition des livres imprimés au XVIème siècle)

<http://www2.echo.lu/libraries/en/projects/master.html> pour le programme européen MASTER