Actes du colloque **Vers une nouvelle érudition : numérisation et recherche en histoire du livre**, Rencontres Jacques Cartier, Lyon, décembre 1999.

Numérisation des manuscrits médiévaux : le projet européen BAMBI

Sylvie Calabretto*
Andrea Bozzi**
Jean-Marie Pinon***

* http://liris.cnrs.fr/sylvie.calabretto/ [2008]

**andrea.bozzi@ilc.cnr.it [2008]

*** http://lisi.insa-lyon.fr/pagespersonnes/page28.htm [2008]

Résumé

Le projet BAMBI (*Better Access to Manuscripts and Browsing of Images*) du programme européen LIBRARIES visait à définir des techniques de numérisation de manuscrits médiévaux et à concevoir une station de travail pour les chercheurs en Histoire des Textes (philologues). La station BAMBI est également dédiée aux papyrologues, épigraphistes, paléographes et codicologues, ou plus généralement aux utilisateurs d'une bibliothèque qui souhaitent examiner des sources manuscrites, transcrire et annoter des manuscrits, ainsi que naviguer entre les éléments textuels de la transcription et les portions d'image correspondantes sur le manuscrit scanné.
Mots clés : Manuscrits médiévaux, Bibliothèque Numérique, Philologie, Modélisation de documents, SGML, HyTime

1. Introduction

L'origine du projet BAMBI (*Better Access to Manuscripst and Browsing of Images*) est le projet *Workstation filologica multimodulare* [BOZ 94] développé dans le Laboratoire ILC (*Institute for Computational Linguistics*) du CNR (*Comitato Nazionale della Ricerca*) de Pise : ce projet visait à offrir une station de travail multimédia aux philologues.

Le projet BAMBI [CAL 96], [BOZ 97], [CAL 98] a démarré en janvier 1995 et s'est terminé en avril 1997 : le Consortium était composé de A.C.T.A. (coordinateur, Florence), *Central National Library* (BNR, Rome), *National Research Council* (ILC-CNR, Pise), *Pise Research Consortium* (CPR, Pise), LISI-INSA (Lyon) et *Max Planck Institut für Rechtsgeschichte*

(Frankfurt a. M.). L'objectif principal de ce projet était de concevoir une station de travail pour visualiser, transcrire, annoter, indexer des manuscrits anciens.

La catégorie d'utilisateurs visée par BAMBI est constituée principalement d'étudiants spécialisés en Histoire des Textes : philologues ou éditeurs critiques de travaux classiques ou médiévaux qui sont écrits à la main sur des supports de différents types (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), tout ceux, en résumé, qui sont intéressés par l'étude, l'annotation, ou la transcription de textes contenus dans des documents manuscrits numériques.

Le projet BAMBI s'est appuyé sur cinq phases principales :

- Etude de marché (en particulier pour définir plus précisément les besoins des philologues),
- Sélection et description des manuscrits (méta-données),
- Numérisation de manuscrits anciens à partir de microfilms,
- Modélisation et indexation des documents anciens (modélisation en SGML/HyTime),
- Modélisation et Conception de la base de documents anciens et de l'IHM (Interface Homme/Machine).

Dans la suite de l'exposé, nous évoquons les problèmes rencontrés pendant la phase de numérisation, puis nous présentons les critères de transcription adoptés dans le cadre de ce projet. Ensuite, nous détaillons les fonctionnalités de la station de travail BAMBI. Enfin, nous concluons sur les perspectives du projet BAMBI.

2. Numérisation des manuscrits médiévaux

Un des objectifs principal du projet BAMBI concernait la mise en place d'une chaîne de conversion des manuscrits anciens en format numérique [BON 97]. Les manuscrits ont été numérisés à partir de microfilms 35 mm noir et blanc. Il a été nécessaire de prendre en compte les différentes particularités liées aux microfilms : positif vs négatif, original ou copie de microfilms, pages simples ou multiples, effets de « background » (partie verso visible dans la partie recto), certains manuscrits requièrent 256 niveaux de gris pour représenter leur contenu de manière pertinente.

3. Sélection des manuscrits et critères de transcription

3.1. Sélection des manuscrits pour la validation

Cette phase du projet constituait une des activités clé de la première phase du projet. Au total, 30 manuscrits ont été sélectionnés à partir des fonds anciens de la BNR (Bibliothèque Nationale de Rome) et des archives microfilms du MPI. La qualité des images était très variable. De nombreux manuscrits possédaient un niveau de qualité très faible. Les manuscrits sélectionnés étaient considérés comme représentatifs d'une situation analogue à celle de la plupart des bibliothèques européennes.

3.2. Critères de transcription

Dans BAMBI, les critères de transcription sont basés sur le modèle paléographique [BIS 92], [PRA 79], [TOG 82].

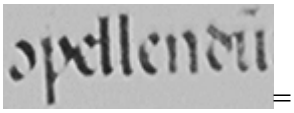
Nous rappelons que la transcription d'un manuscrit est un processus qui vise à noter la prononciation d'une langue donnée à l'aide d'un système de signes d'une langue de conversion. A l'époque qui nous intéresse dans le projet BAMBI, en l'occurrence le Moyen-Age, les abréviations peuvent être divisées en différents types : abréviation syllabique (omission et élision de lettres, par exemple It pour Item), abrévation par suspension (un exemple est donné avec le nom des juristes : ac. = Accurcius, bul. = Bulgarus, ...), et l'utilisation de signes spéciaux (par exemple, le signe graphique qui correspond à la conjonction de coordination et). 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. Les règles sont les suivantes :

- des parenthèses, "(" et ")" sont utilisées pour entourer des caractères qui sont représentés dans le document original par des abréviations syllabiques. Par exemple, e=e(st) ou bn=b(e)n(e) ou m=m(odo) ;
- les symboles, "<" et ">" sont utilisés pour entourer des caractères ayant une abréviation par suspension, ou des initiales de grande taille (comme c'est souvent le cas pour la première lettre de la page). Par exemple, dans le cas des noms de juristes, on a ac. = ac<cursius> ou bul.=bul<garus>;

- des crochets, "[" et "]" sont utilisés pour entourer des caractères qui correspondent à un signe simple ou modifié graphiquement dans l'image :

 = [et]

 = [et]

 = [con]pellendu(m)

4. Fonctionnalités de la Station BAMBI

Les fonctionnalités principales de la station BAMBI (Figure 1) sont :

- La visualisation des images des documents sources (manuscrits) avec une haute résolution,
- La transcription, annotation et indexation du texte contenu dans les images,
- La correspondance automatique entre chaque mot de la transcription et la portion de l'image source dans lequel le mot est trouvé,
- L'exportation des information sur les manuscrits au format SGML/HyTime.

Figure 1

4.1. Recherche d'un manuscrit

La station BAMBI (Figure 1) offre un certain nombre d'outils de recherche qui permettent d'accélérer la sélection de documents. Ces outils sont basés sur une recherche multi-critères par méta-données ou par mots-clés. La recherche multi-critère par méta-données peut s'effectuer à l'aide des méta-données suivantes : dates de création, la langue, le lieu, ... ou par combinaison de celles-ci. Par exemple, la recherche d'un manuscrit peut être basée sur les

trois critères Auteur (*Author*), Date et Bibliothèque (*Library*). D'autre part, à chaque texte est associé un certain nombre de mots-clés qui décrivent le document (sans nécessairement en faire partie). Ces mots-clés peuvent être écrits dans différentes langues (Allemand, Anglais, Français, Italien, Espagnol). Les historiens peuvent combiner ces mots-clés en utilisant des opérateurs logiques (AND, OR, NOT) et des caractères spéciaux (*, ?, #).

Il faut noter que les manuscrits sont stockés sur CDROM ou sur disque et qu'il faut spécifier le chemin d'accès aux manuscrits avant de lancer une recherche.

Lorsqu'un manuscrit a été sélectionné, une fenêtre principale s'ouvre et comporte cinq zones de travail (Figure 1) :

- L'image de la page du manuscrit,
- La transcription correspondante en texte (code ASCII étendu),
- Une liste de marque-pages contextuels, que l'utilisateur estime utiles 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.

4.2. Aide à la transcription de manuscrits

Les historiens peuvent transcrire le texte contenu dans l'image numérique du manuscrit (si cette transcription n'existe pas), suivant les critères de transcription définies en 3.2. Les règles de transcription sont données interactivement aux utilisateurs lorsqu'ils effectuent une transcription avec la station de travail BAMBI. La transcription peut être exportée vers un fichier de type RTF ou SGML, lui permettant d'être réutilisée, soit dans des programmes de traitement de textes standards, soit dans des systèmes de gestion de documents.

Des possibilités de zoom sur l'image et de contrôle sur l'image permettent à l'historien d'identifier plus justement les mots du manuscrit et de lever certaines ambiguïtés (mots acollés, par exemple). L'utilisateur pourra régler le contraste, la luminosité et le facteur de zoom de l'image du manuscrit afin de le rendre le plus lisible possible.

Pour la validation, 1500 images ont été transcrites parmi les 30 manuscrits sélectionnés à la BNR et au MPI.

4.3. Indexation des textes

Lorsque la transcription est complète, l'outil d'indexation génère un *index verborum* et un *index locorum* (Figure 2). 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 (identifiés par un astérisque). Chaque élément de l'index est suivi du nombre de fois où il apparaît dans le manuscrit, de même que dans la page en cours de consultation.

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 chaque 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.

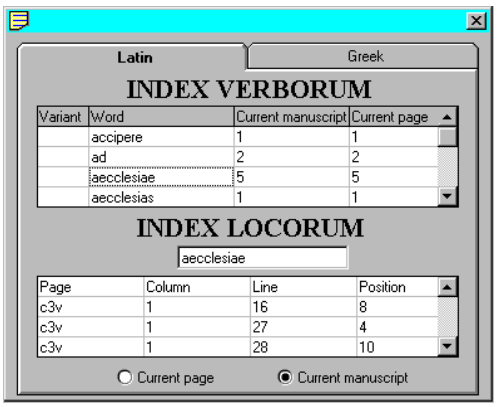


Figure 2. *Index verborum* et *index locorum*

4.4. Annotations des 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).

Une annotation comprend toujours deux champs distincts : un champ pour les commentaires libres (fond), et un champ pour les variantes (synonymes ou corrections de syntaxe) (forme). Pour un groupe de mots donné, chacun de ces deux champs peut être rempli ou non, ils sont complètement indépendants. Cette distinction entre les annotations est nécessaire afin

d'inclure les variantes dans l'*index verborum* du manuscrit, utilisé à des fins de recherche. Dans la figure 1, un exemple simple d'annotation (de type commentaire libre) sur le mot « Item » est donnée : « annotation Item ».

4.5. Correspondance mot/image

La station de travail BAMBI incorpore un outil de génération des correspondances entre les mots de l'image et les mots du texte.

Cet outil se décompose en deux traitements distincts : la reconnaissance automatique des lignes et des colonnes et l'algorithme de correspondance mot/image.

- Reconnaissance automatique des colonnes et des lignes : La représentation numérique de l'image est effectuée automatiquement afin d'identifier les parties qui peuvent représenter le fond de l'image et les parties qui peuvent être interprété comme des parties écrites. L'algorithme identifie également les colonnes, lorsque le texte est composé de deux colonnes ou plus. Il s'agit ensuite d'identifier les parties contenant du texte dans les marges. Ceci est rendu possible en analysant l'histogramme avec la distribution des valeurs des niveaux de gris. Cette méthode est également utilisée pour identifier les lignes : le programme compte les lignes et les numérote progressivement. Ensuite, l'algorithme identifie les mots de chaque ligne.
- Correspondance Mot/image : Un contrôle final de l'image - possible uniquement si la transcription du texte est disponible - produit une correspondance de chaque mot du texte avec la partie de l'image dans laquelle il est inscrit. Le système 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. Plus la définition de l'image numérique est élevée, plus les frontières des régions sont précises. A chaque région (produite par le système) est attribuée une adresse qui permet de maintenir une relation entre les mots correspondants de la transcription. Plus précisément, l'image est organisée comme une mosaïque dans laquelle à chaque région rectangulaire est associé un pointeur sur le mot correspondant du texte transcrit. La segmentation numérique de l'image contrôlée

par la transcription textuelle est mise à jour automatiquement lorsque la transcription est modifiée. Le module montre le texte et l'image segmentée dans deux fenêtres séparées sur l'écran, ainsi l'utilisateur peut corriger les erreurs éventuelles.

Lorsque l'outil de correspondance a terminé ses calculs, l'utilisateur peut cliquer sur un mot dans l'image (It, par exemple) et obtenir le mot correspondant dans la transcription (Item) (Figure 1). Et réciproquement, en cliquant sur un mot de la transcription, on obtient le mot correspondant dans l'image. Cette fonction est très intéressante dans un contexte (les manuscrits anciens) où le nombre d'abréviations est très élevé et dans un objectif de formation des jeunes chercheurs en Histoire des Textes.

4.6. Correction manuelle des résultats de la correspondance automatique

L'algorithme qui fait correspondre mot et image [BOZ 94], bien que très puissant, ne permet pas d'identifier tous les blocs de mots. Certaines caractéristiques du manuscrit (taches, mots acollés, dessins, illuminations, ...) 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.

Le module de correction manuelle montre l'image segmentée dans une fenêtre (Figure 3) et autorise l'utilisateur à effacer des éléments erronés, à réallouer (par exemple à d'autres mots ou lignes) des zones d'image, à modifier la taille de la zone d'image (par exemple, une grande initiale peut être ramener à sa taille réelle).

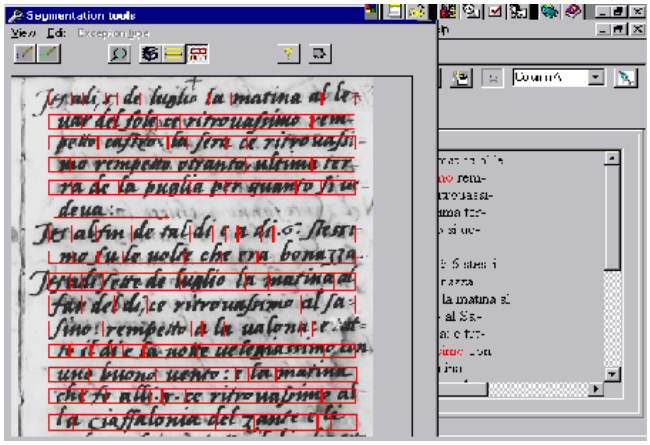


Figure 3. Correction manuelle de la segmentation

4.7. Modélisation HyTime des informations relatives à un manuscrit

Une fonctionnalité intéressante fournie par le système BAMBI est un filtre permettant l'exportation transparente de textes (incluant les données des correspondances texte/image) au format SGML/HyTime [CAL 97a]. Enregistrer et exporter les informations relatives à une page d'un manuscrit (description du manuscrit, annotations, liens mot/image,) dans un format de type SGML constitue une caractéristique très utile, surtout dans la perspective croissante de l'implémentation de la norme internationale TEI (*Text Encoding Initiative*) [VER 95] pour le balisage standard (à travers SGML et HyTime) de textes classiques dans une forme lisible pour une machine. Nous rappelons que SGML (*Standard Generalized Markup Language*) [ISO 86] permet de structurer un document à l'aide de balises. Une balise, constituée d'un tag début et d'un tag fin, délimite un élément (par exemple, <titre> *Diario del viaggio in Terra Santa* 1559</titre>). Une DTD (*Document Type Definition*) permet de définir des balises qui sont utilisées pour structurer un document. Ce qui revient à définir un modèle logique de document.

Nous avons développé une DTD BAMBI, capable de manipuler les principaux « tags » implémentés dans BAMBI : description du manuscrit, annotations, liens, bookmarks, paires de correspondance texte- image, ... Cette DTD a été écrite en SGML, en utilisant certaines caractéristiques avancées de HyTime pour représenter les liens (en particulier, les liens mot/image). Le langage de structuration hypermédia et événementiel HyTime (*Hypermedia/Time-based Structured Language*) [ISO 92] permet de représenter les

informations statiques et dynamiques traitées et échangées par des applications hypermédias. HyTime, qui est une extension de SGML, fournit :

- un langage pour la description de la structure d'un hyperdocument, avec l'intégration de la notion d'hyperlien, d'ordonnancement et de synchronisation;
- des mécanismes normalisés pour décrire les interconnexions (ou hyperliens) à l'intérieur et entre les documents et autres objets d'information, et pour les agencer dans le temps et dans l'espace.

4.8. Bookmarks et marque-pages

- Le bookmark est un outil de raccourci mis à la disposition de l'utilisateur afin de lui permettre d'accéder plus rapidement à une liste de documents privilégiés. Cette liste est personnalisée et l'utilisateur a toute liberté de lui donner l'aspect visuel et hiérarchique qui lui convient le mieux. Cette fonction s'avère particulièrement utile lorsque la quantité de documents déjà référencée est importante, puisqu'elle offre des fonctions de tri et de classement rendant l'accès à l'information plus efficace. Concrètement, le bookmark se présente sous la forme d'une arborescence graphique dont les nœuds sont des dossiers et les feuilles des pages de manuscrits. Il peut être visualisé d'une part sous forme d'arbre pour la mise en forme de la hiérarchie, et d'autre part sous forme de menus et sous-menus offrant un accès plus rapide.
- Les marque-pages constituent une fonctionnalité complémentaire à celle du bookmark. Elle permet à l'utilisateur d'associer jusqu'à cinq autres pages à la page courante par des liens directs. Ces liens sont rappelés par un alias dans la fenêtre de travail (Figure 1) et permettent d'avoir accès plus rapidement à des textes directement en rapport avec le document en cours d'étude. L'utilisateur crée ainsi un contexte de travail autour d'un document.

4.9. Gestion des droits d'accès et des utilisateurs

Le logiciel final devant être prévu pour Windows 95 ou Windows NT, il aurait été à première vue suffisant de s'appuyer sur les droits déjà amplement développés sur ces systèmes d'exploitation. Cependant, pour plus de convivialité, nous avons instauré un système d'identification par mot de passe propre à notre logiciel. Cette couche supplémentaire permet

à chaque utilisateur d'octroyer des droits (consultation, annotation) à certains collègues. Ces droits permettent à différents chercheurs de travailler à tour de rôle sur une même station de travail, en partageant leurs annotations et leurs marque-pages spécialisés. Ce partage peut s'effectuer à deux niveaux : en simple consultation (lecture seule), ou en contrôle total (écriture autorisée).

Le logiciel permet d'ouvrir deux sessions simultanées (la session principale sous un login personnel et la session secondaire sous celui d'un groupe de travail par exemple), afin de pouvoir naviguer rapidement entre les informations strictement personnelles et celles qui peuvent intéresser d'autres chercheurs.

5. Conclusion et perspectives

Le projet européen BAMBI, du programme LIBRARIES, a permis d'élaborer une station de travail pour les Chercheurs en Histoire des Textes permettant aux historiens d'étudier des manuscrits anciens sur tout micro-ordinateur et donc utilisable en tout lieu. D'autre part, l'échange d'informations devient aisé : un manuscrit est associé à une transcription, celle-ci est susceptible de recevoir plusieurs annotations et variantes. L'outil informatique permet un échange facile des données par le travail de plusieurs personnes sur des mêmes versions. De plus, BAMBI propose des outils destinés à aider les chercheurs et à les libérer de tâches fastidieuses (outil d'indexation full-text, outil de correspondance texte/image).

Suite à différentes enquêtes menées auprès de philologues ayant utilisé le logiciel BAMBI, il s'avère que le projet a répondu aux besoins des bibliothèques (numérisation et accessibilité/exploitation de manuscrits anciens) et aux centres de recherche (transcription, annotation).

Mais la station BAMBI est actuellement locale, les images des manuscrits (au format JPEG), les transcriptions et les fichiers de mise à jour se trouvent sur un ou plusieurs CDROM. Des perspectives intéressantes de la station BAMBI visent à concevoir une solution de type Internet ou Intranet. Ces perspectives ont déjà donné lieu à plusieurs études [CAL 97b], dont le projet STEMA (Station de Travail pour l'Etude des Manuscrits Anciens sur le Web) du Programme d'Actions Intégrés (P.A.I) MAE-MENRT Galilée 99 (projet franco-italien entre le LISI et le CNR- Pise). D'autre part, un projet national italien, le projet « Bibliophilo » des Ministères italiens de la Recherche Scientifique et de la Culture dans le cadre du Programme

PARNASO vise à améliorer les fonctionnalités actuelles de la station BAMBI et à produire un logiciel plus robuste.

6. Bibliographie

[BIS 92] Bischoff B., *Paleografia latina. Antichità e medioevo*, Italian edition edited by G.P. Mantovani and S. Zamponi, Padua 1992. pp.218-239

[BON 97] Bonnaterre O., Bozzi A., Calabretto S., Colli V., Maci E., Murano G., Raggioli A., Spotti A., Tariffi F., *Better Access to Manuscripts and Browsing of Images : Aims and results of an European Research project in the field of digital Libraries BAMBI Lib-3114*. CLUEB (Bologne), 1997, 176 pages, ISBN N· 88-8091-569-X.

[BOZ 94] Bozzi A., Sapuppo A., "Biblioteca virtuale e studio dei testi: la multimedialita' al servizio della filologia", *Bollettino d'Informazioni del Centro di Ricerche Informatiche per i Beni Culturali*, IV (1994) n·2, Scuola Normale Superiore, Pisa, 1994, pp. 5-24.

[BOZ 97] Bozzi A., Calabretto S., "Digital Library and Computational Philology : the BAMBI (LIB -3114) project. Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries", Lecture Notes in *Computer Science* N·1324 (Springer Verlag). Eds. C. Peters and C. Thanos. Pisa, Italie. September 1-3, 1997. pp. 269-285. ISBN 3-540-63554-8

[CAL 96] Calabretto S., Sappupo A., Tariffi F., *Modules and functions in the BAMBI software system. European Libraries Project Lib-3114 Report*, Deliverables 4-5. September 10 th 1996. 56p.

[CAL 97a] Calabretto S., Pinon J.M., "Modelling of a medieval manuscript database with HyTime. Proceedings of ICCC/IFIP Conference on Electronic Publishing : EP'97", *New Models and Opportunities*. The University of Kent at Canterbury, Great Britain. April 14-16, 1997. Edited by Fytton Rowland and Jack Meadows. ICCC Press, Washington, pp. 336-345. ISBN 1-891365-00-2

[CAL 97b] Calabretto S., Rumpler B., "WWW access to a philological workstation", *Proceedings of the First East-European Symposium on Advances in DataBases and Information Systems, ADBIS'97*. St. Petersburg (Russia). September 2-5, 1997. pp. 326-330. A paraître dans *LNCS*, Springer Verlag.

[CAL 98] Calabretto S., Pinon J.M., Bozzi A. BAMBI, « Système de Gestion de Manuscrits Anciens pour Historiens », Revue *Document Numérique*. Ed. HERMES, Volume 2, n· 3-4,

Numéro spécial sur les Bibliothèques Numériques, 1998. pp. 31-50. ISBN 2-86601-738-2, ISSN 1279-5127.

[ISO 86] ISO 8879: 1986. *Information Processing-Text and Office Systems, Standard Generalized Markup Language* (SGML), 1986.

[ISO 92] ISO 10744: 1992. *Information Technology, Hypermedia Time-Based Structuring Language* (HyTime), 1992

[[PRA 79] Pratesi A., *Genesi e forme del documento medievale*, Rome, 1979. pp.99-109

[TOG 82] Tognetti G., *Criteri per la trascrizione di testi latini ed italiani*, Rome, 1982

[VER 95] Véronis J., Ide N. *Text Encoding Initiative: Background and Context*, Kluwer Academic Publishers, 1995.