# MADONNE : MAsse de DOnnées issues de la Numérisation du patrimoinNE

Consortium MADONNE

http://l3iexp.univ-lr.fr/madonne/

jean-marc.ogier@univ-lr.fr

**Résumé** : *Cet article présente le projet MADONNE, une initiative française d'utilisation d'approches issues de l'analyse d'image de document dans des buts d'indexation et de navigation au sein des corpus de documents patrimoniaux.*

**Mots-clés** : patrimoine, document, image, indexation, recherche, navigation, masse de données

## 1 Introduction

Aujourd'hui, l'ensemble des experts plaide pour des actions fortes garantissant à l'avenir une conservation durable des ressources culturelles et scientifiques constituant notre patrimoine. L'informatisation sans cesse croissante de nos sociétés nous pousse naturellement à considérer la numérisation comme solution pour mener à bien ce travail. Preuve en est, ces dix dernières années des grandes campagnes ont été entreprises par diverses institutions comme les musées, les cadastres, les bibliothèques, …. De vastes corpus numériques sont désormais disponibles, cependant différents facteurs en empêchent une gestion cohérente et efficace.

En premier lieu les campagnes de numérisation coûtent chères, or on constate différents sur-coûts. La principale raison est le manque de concertation entre les institutions. Il y a absence de politique commune d'où les problèmes de gaspillage de ressources, d'efforts et d'investissements. De plus, le choix des technologies de numérisation est souvent laissé à l'appréciation des instituts sans réelles consultations d'experts du domaine. Or certaines technologies sont d'avantages précaires, elles peuvent devenir rapidement obsolètes conduisant à des renouvellements prématurés des matériels et à la reconduite des travaux de numérisation. Enfin, les droits de propriétés industrielles et intellectuelles soulèvent également différents problèmes financiers. Différentes entités (auteur, institut, gouvernement, …) ont évidemment des droits sur les contenus numériques qui doivent être reconnus et pris en compte. Il y a un fort besoin de solutions communes pour gérer ces droits dans le domaine culturel.

Ensuite, certains corpus numériques produits par les institutions sont parfois non exploitables par des systèmes automatiques de traitement de document. En effet, durant le processus de numérisation plusieurs précautions doivent être prises afin d'assurer la mise en place de tels systèmes par la suite. Un exemple type de manquement est le choix des paramètres lors du stockage au format JPEG des images. Les facteurs de qualité choisis sont souvent trop faibles ce qui détériorent les images au point d'empêcher leur indexation a posteriori. Citons également les résolutions choisies pour la numérisation souvent trop faibles. Elles avoisinent généralement les 200 pp alors qu'elles devraient être dans l'idéal au minimum à 300. Ainsi, les institutions qui ne considèrent pas l'ensemble de ces contraintes produisent des corpus de données peu (ou pas) exploitables dans des optiques d'indexation. Ces différents aspects mettent en évidence la nécessité d'instaurer le dialogue entre les communautés de recherche des sciences humaines et sociales et informatique.

Enfin, la spécificité des corpus patrimoniaux soulève des nouvelles problématiques de recherche. Différents verrous restent à lever pour assurer une exploitation automatique viable des corpus constitués. Comparé à la problématique "classique" de l'analyse d'image de document, la principale évolution concerne les masses de données à traiter. Une autre différence importante est la variabilité des informations contenues dans les images de documents patrimoniaux. De plus, la détérioration importante des images s'ajoute à la liste des difficultés. Finalement, et cela constitue certainement le verrou scientifique le plus important, l'usage fait des documents indexés soulève la question de la structuration de la modélisation des corpus indexés.

Dans ce contexte le projet de recherche MADONNE, financé par le Ministère de la Recherche et de l'Enseignement dans le cadre de l'ACI Masse de Données[1], vise à concevoir des plate-formes d'indexation automatique des bases de documents patrimoniaux. Ces dernières interviennent alors en aval des projets de numérisation fournissant de larges bases d'images faiblement structurées. Le projet MADONNE étudie pour cela des approches issues de l'analyse d'image de document appliquées à l'indexation et la navigation au sein des corpus de documents patrimoniaux.

Le projet MADONNE a commencé fin 2003 et terminera à la fin de 2006. Il regroupe différents partenaires de recherche français au sein d'un Consortium : le L3i (La Rochelle)[2], le LORIA (Nancy), le laboratoire PSI (Rouen), le LI (Tours), le LIRIS (Lyon), le CRIP5 (Paris) et l'IRISA (Rennes). Mentionnons également le CESR (Tours) avec le-

[1] http ://acimd.labri.fr/

[2] Leader du projet

quel le Consortium a eu un partenariat fort durant le projet. La plupart de ces partenaires ont une expérience forte et complémentaire dans le domaine de l'analyse d'image de document ce qui nous a permis d'aborder un éventail large de thématique de recherche. Cet article présente un résumé des principaux résultats obtenus par le Consortium et particulièrement par A. El Abed, B. Coüasnon, M. Delalandre, V. Eglin, N. Journet, S. Leriche, R. Mullot, S. Nicolas, J.M. Ogier, T. Paquet, R. Pareti, J.Y. Ramel, J.P. Salmon, S. Uttama, N. Vincent et L. Wendling.

# 2 Thématiques de recherche

Afin d'assurer une indexation et une navigation pertinente des corpus de documents patrimoniaux il est nécessaire d'exploiter toutes les caractéristiques utiles pour les besoins de recherche. Ceci inclut le traitement de différents types d'informations rencontrées sur ces documents comme les illustrations, les textes, les styles, les symboles, les annotations manuscrites, …. Cela nous a amené à réaliser des collaborations croisées autour de différentes thématiques de recherche en analyse d'image de document. Dans la suite de cet article, nous présentons les principales développées au sein du Consortium MADONNE.

## 2.1 Modélisation des collections

Dans le contexte de "masse de données" on peut observer une forte homogénéité dans la manière dont l'information est structurée au sein d'une même collection d'images (ç-à-d issues d'ouvrages similaires). La modélisation des collections consiste alors à extraire, de la façon la plus automatique possible, des attributs qui caractérisent cette structuration. Le but est de superviser en amont les plate-formes d'indexation par la prise en compte des méta-données décrivant cette structuration. Ceci permet alors le déclenchement d'outils adaptés sur les images, ou parties d'image, en fonction de leur nature (textuelle, graphique, …). Cette problématique relève de la découverte automatique de similarité dans les informations de structuration des collections dans le but d'en construire un modèle pertinent.

Dans le cadre du projet MADONNE, Journet et al [JOU 06] ont proposé une approche permettant une catégorisation des zones de l'image sur des critères d'organisation spatiale des données. L'extraction de caractéristiques décrivant la structure physique des images de document permet alors de constituer le modèle de la collection considérée. Dans cet optique, Journet et al propose une fonction s'appuyant sur le calcul de l'auto-corrélation. Celle-ci a la particularité, lorsqu'elle est estimée sur une zone de texte ou de dessin, de générer une signature unique facilement identifiable. Ce choix permet ainsi de séparer le texte des dessins, tout en minimisant la quantité d'a priori relative aux images traitées. Cette technique doit également permettre de regrouper les pages similaires d'un ouvrage en classes afin de pouvoir appliquer, par la suite, un traitement adapté sur chacune des classes extraites (Fig. 1).

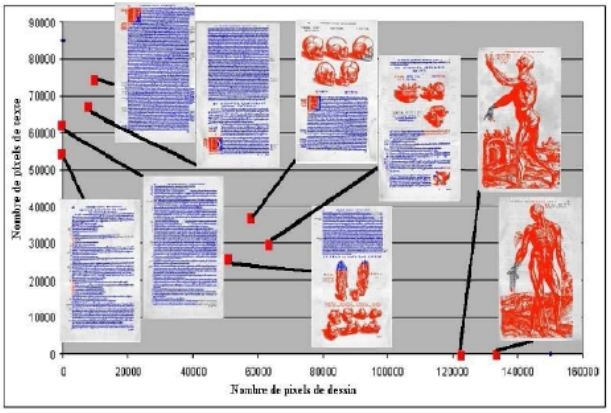


FIG. 1 – Catégorisation des images

## 2.2 Extraction de la structure physique

La structure physique d'un document est habituellement relative à un modèle de présentation. Ceci est d'avantage marqué pour les documents patrimoniaux ou la mise en page est soumise à de fortes contraintes liées aux techniques d'impression utilisées. La recherche de la structure physique peut donc être un excellent support pour l'indexation des images. Par exemple, certains documents possèdent une mise en page tellement spécifique qu'ils se distinguent aisément au sein d'un corpus. Ou encore, une recherche plein texte peut être exécutée uniquement à partir de zones spécifiques préalablement extraites par analyse de la structure physique.

Dans le projet MADONNE Coüasnon et al [COü 03] ont travaillé sur une plate-forme d'annotations collectives de registres militaires du 19° siècle. Celle-ci opère par analyse de la structure physique pour séparer les informations privées et publiques des registres. Les informations publiques sont alors ouvertes aux annotations collectives. Cette plate-forme emploie pour cela une méthode robuste d'analyse des structures, basée sur une grammaire 2D, intégrée dans le système DMOS. Ce système permet de détecter chaque cellule d'un registre donné même en cas de forte détérioration de l'image. En complément de cette détection la plate-forme propose un moteur de recherche opérant à partir des zones patronymiques des registres. Ce moteur procède par appariement des zones sans utilisation d'OCR. La mesure de similarité est basée sur l'extraction de primitives bas niveaux (les graphèmes) dont l'organisation permet de constituer une signature discriminante du patronyme. Ce système a été validé sur une base de 165 000 registres.

Une autre contribution dans ce domaine a été réalisé par J.Y Ramel et al via la plate-forme AGORA [RAM 05]. Cette dernière permet l'extraction de la structure physique d'un document par analyse de deux cartes de segmentation en blocs de l'image : une des formes et l'autre du fond (Fig. 3). AGORA procède alors à une classification des blocs extraits pour la segmentation en zones du document. Cette classification opère selon un scénario produit par l'utilisateur au cours d'une phase d'interaction avec AGORA.

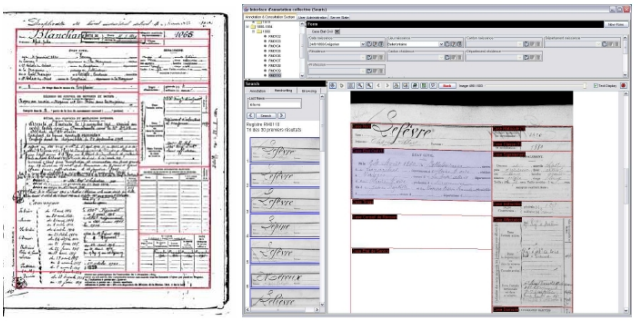

FIG. 2 – Gauche : Extraction de la structures physique
Droite : Reconnaissance du patronyme manuscrit

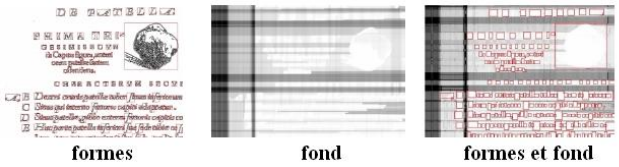


FIG. 3 – Cartes du fond et des formes

## 2.3 Documents manuscrits

Le traitement des documents manuscrits patrimoniaux soulève des problématiques de recherche relativement éloignées de celles habituellement adressées pour les documents contemporains (chèques, enveloppes, formulaires . . .). Le but en est rarement de reconnaître l'écriture mais plus de caractériser et d'identifier les différents scripteurs. La Fig. 4 illustre le type de document auquel nous devons faire face. Au regard du haut niveau de bruit rencontré sur ces documents une indexation à la volée, c-à-d exploitant des indices visuelles sans reconnaissance de l'écriture a priori, semble constituer la manière la plus adéquate de procéder.

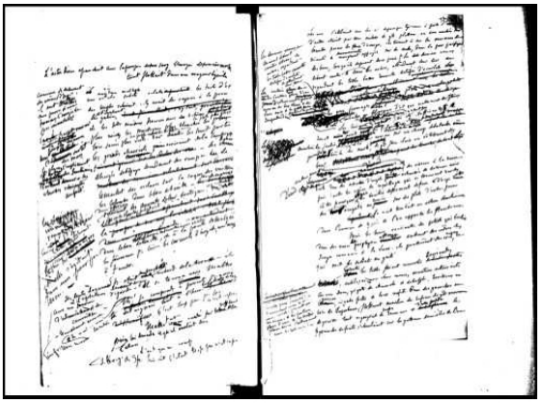

FIG. 4 – Document manuscrit avec annotations

Pour ce faire, dans le cadre du projet MADONNE, S. Nicolas et al [NIC 06] ont proposé un système d'analyse de mise en page appliqué aux manuscrits de Flaubert[3]. Différentes signatures discriminantes sont calculées dans le but de vérifier l'organisation spatiale des primitives manuscrites extraites du document. Cette organisation permet alors de caractériser le style de mise page de l'auteur. Elle permet également de reconstituer le processus genèse du manuscrit au travers de l'analyse des notes de marge. Pour ce faire, les modèles de Markov cachés, aussi bien que la programmation dynamique, sont utilisés pour les processus de segmentation et de modélisation du document.

[3] Auteur français, 1821-1880.

## 2.4 Indexation des parties graphiques

Habituellement les documents sont indexés à partir de l'analyse des parties textuelles. Cependant, de nombreux documents patrimoniaux contiennent également des parties dites graphiques comme par exemple les bandeaux, les figures ou bien les lettrines. La Fig. 5 suivante en donne des exemples.


FIG. 5 – Parties graphiques

Dans le cadre du projet MADONNE différents partenaires ont travaillé sur le problème d'indexation de ces parties graphiques, et plus particulièrement sur les lettrines [PAR 06]. Il y a souvent un a priori sur la façon d'indexer ce type d'image. Nos collaborations avec le CESR ont mis en lumière la diversité des besoins pouvant être exprimés par les historiens. Certains d'entre eux par exemple sont intéressés par l'analyse de la luminance des images de façon à en extraire une datation, d'autres par la recherche d'images similaires afin de retracer l'usage des tampons par les imprimeurs, . . .. À la lumière de ces différents besoins le Consortium a entrepris des travaux parallèles visant à effectuer une indexation multi-critères de ces images.

Un premier problème concerne leur dégradation. En effet, les images de lettrine sont particulièrement bruitées et nécessitent d'être restaurées avant tout traitement d'indexation. Nous utilisons pour cela un filtre procédant par lissage adaptatif. Ce dernier donne de meilleurs résultats que les techniques habituelles au regard des fortes variabilités de bruits présents sur ces images. La Fig. 6 donne un exemple de résultat obtenu par ce filtre.

FIG. 6 – Restauration d'images

Sur la base des images pré-traitées différentes méthodes d'indexation complémentaires ont été proposées par les laboratoires CRIP5 et L3i. La première est basée sur un modèle statistique de la distribution des pixels des images de lettrine utilisant la loi de Zipf [PAR 05]. Ceci permet de classer les images de lettrine en fonction de leur style (Fig. 7). Une autre méthode emploie une approche par segmentation permettant la décomposition en couches des images de lettrine [UTT 05]. Une signature à base de MST est ensuite calculée à partir de l'analyse de couches permettant d'indexer les images selon des critères d'organisation spatiale (Fig. 8). Enfin un dernier système propose un système d'indexation rapide exploitant une représentation compressée des images par encodage en longueur de plages [DEL 06] (Fig. 9). Ce système est alors appliqué à la recherche d'images strictement similaires.

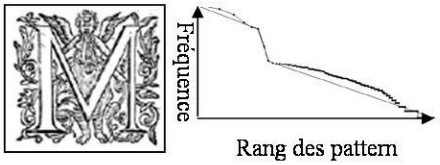


FIG. 7 – Recherche sur critère de style

FIG. 8 – Recherche sur critère de structure


FIG. 9 – Taux de compression par encodage en plage

Salmon et al [SAL 06] proposent eux une nouvelle approche pour la combinaison de descripteurs de formes. Cette dernière permet d'améliorer les résultats de classification, Salmon et al l'applique à la reconnaissance des lettres extraites des images de lettrine (Fig. 10). Elle est fondée sur l'étude comportementale des descripteurs vis-à-vis d'un corpus d'apprentissage. Chaque descripteur est calculé sur plusieurs classes d'objets ou de symboles. Pour chaque échantillon et pour tous les descripteurs un profil type est déterminé. Celui-ci est défini à partir d'un jeu d'apprentissage en prenant en compte les conflits pouvant exister entre descripteurs.

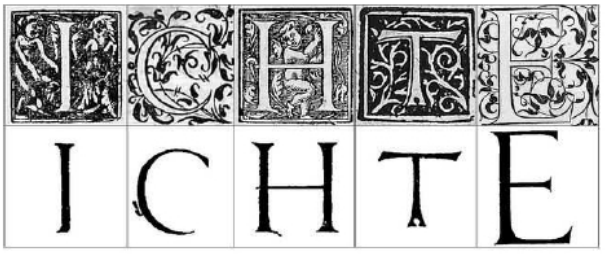

FIG. 10 – Lettres extraites des lettrines

### 2.5 Recherche de similarité par compression des images de document

La numérisation des documents patrimoniaux soulève également la question de leur stockage et diffusion sur les réseaux locaux et Internet. Au regard de la masse de données considérée, et des débits possibles sur les réseaux, seule une compression avec perte peut réduire de manière suffisamment significative la dimension des images. Cependant, la perte engendrée doit évidemment impliquer une altération "raisonnable" des données. De nombreuses méthodes de compression avec perte ont déjà faits leur preuve comme par exemple JPG, DJVU ou DEBORA. Cependant ces méthodes ne sont pas applicables dans le cas ou les images de document contiennent des informations manuscrites. En effet, la complexité des formes manuscrites en empêche une localisation précise altérant ainsi les performances des méthodes de compression existantes.

Dans le cadre du projet MADONNE A. El Abed a proposé une méthode de compression des documents manuscrits [ABE 04]. Cette dernière opère par séparation des couches textuelle et fond des images sur critère de similarité. La détection des redondances est basée sur une décomposition du texte manuscrit en segments orientés avec des points contours invariants. Cette méthode peut être étendue à toutes parties de l'image qui présentent des similarités distribuées.

## 3 Conclusion et perspectives

Comme nous avons pu le voir au travers des différents aspects abordés dans cet article, l'indexation des documents patrimoniaux impose la coopération de différentes approches en analyse d'image de document : pré-traitement

d'images, reconnaissance de l'écriture manuscrite, analyse de documents structurés, analyse de documents graphiques, ….. Ainsi, de nombreuses méthodes peuvent être empruntées à ces domaines et adaptées aux traitements des documents patrimoniaux. Cependant, plusieurs problèmes demeurent et nécessitent à l'avenir de poursuivre des recherches spécifiques à l'indexation des documents patrimoniaux. Ceci concerne principalement l'invariance des méthodes aux dimensions du problème d'indexation. En effet, les documents patrimoniaux introduisent la problématique spécifique de la masse de données, ils sont présents en très grand nombre ce qui impose une grande variabilité de l'information traitée. Un autre problème est celui de la représentation des connaissances spécifiques aux documents patrimoniaux. La prise en compte de telles connaissances au sein des systèmes permettrait alors d'assister les interactions homme-machine dans la constitution de scénarios spécifiques à une problématique donnée.

# 4 Remerciements

Le Consortium MADONNE souhaite tout d'abord remercier Jean-Marc Ogier (L3i, La Rochelle), Karl Tombre (LORIA, Nancy) et Mathieu Delalandre (L3i, La Rochelle) pour leur travail de rédaction de cet article. Le Consortium souhaite également remercier le CESR de Tours pour sa participation active au projet et la diffusion de ses bases d'images.

# Références

[ABE 04] ABED A. E., Recherche de similarités partielles pour la compression des documents manuscrits du patrimoine, Master's thesis, Laboratoire LIRIS, Université de Lyon, France, 2004.

[COü 03] COÜASNON B., CAMILLERAPP J., Accès par le contenu aux documents manuscrits d'archives numérisés, *Document Numérique*, vol. 7, n° 3-4, 2003, pp. 61-84.

[DEL 06] DELALANDRE M., OGIER J., Un système pour l'indexation rapide d'image de lettrine, *Colloque International Francophone sur l'Ecrit et le Document (CIFED)*, 2006.

[JOU 06] JOURNET N., MULLOT R., EGLIN V., RAMEL J., Analyse d'images de documents anciens : Catégorisation de contenus par approche text, *Colloque International Francophone sur l'Ecrit et le Document (CIFED)*, 2006.

[NIC 06] NICOLAS S., PAQUET T., HEUTTE L., Complex handwritten page segmentation using contextual models, *Conference on Document Image Analysis for Libraries (DIAL)*, 2006, pp. 47-56.

[PAR 05] PARETI R., VINCENT N., Global Discrimination of Graphics Styles, *Workshop on Graphics Recognition (GREC)*, 2005, pp. 120-128.

[PAR 06] PARETI R., UTTAMA S., SALMON J., OGIER J., TABBONE S., WENDLING L., VINCENT N., On defining signatures for the retrieval and the classification of graphical dropcaps, *Conference on Document Image Analysis for Libraries (DIAL)*, 2006, pp. 220-231.

[RAM 05] RAMEL J., LERICHE S., Segmentation et analyse interactives documents anciens imprimés, *Traitement du Signal (TS)*, vol. 22, n° 3, 2005, pp. 209-222.

[SAL 06] SALMON J., WENDLING L., TABBONE S., Reconnaissance de symboles graphiques à partir d'une combinaison de descripteurs en intégrant leur comportement sur une base d'apprentissage, *Colloque International Francophone sur l'Écrit et le Document (CIFED)*, 2006.

[UTT 05] UTTAMA S., HAMMOUD M., GARRIDO C., FRANCO P., OGIER J., Ancient Graphic Documents Characterization, *Workshop on Graphics Recognition (GREC)*, 2005, pp. 97-105.