Action Concertée Incitative

MASSES DE DONNEES

Rapport d'activité final

Rapport Final du Projet M.A.D.O.N.N.E.

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

MAsse de DONnées issues de la Numérisation du patrimoiNE

Equipes associées

Laboratoire L3i - Université de La Rochelle
Laboratoire LORIA - Nancy
Laboratoire IRISA - Rennes
Laboratoire d'Informatique - Tours
Laboratoire L.I.R.I.S - Lyon
Laboratoire P.S.I. - Rouen
Laboratoire CRIP5 - Paris

Porté par Jean-Marc OGIER, Université de la Rochelle

Action Concertée Incitative

MASSES DE DONNEES

Rapport d'activité final

Action Concertée Incitative

MASSES DE DONNEES

Rapport d'activité final

Rapport final
MADONNE : MAsse de DOnnées issues de la Numérisation du patrimoinNE

Ce document propose une synthèse des activités menées dans le cadre de l'ACI Masses de Données MADONNE, débutée en 2003. Cette ACI regroupait initialement un ensemble de partenaires, dont le périmètre s'est considérablement élargi au fil du temps, au plan national, et international, sur des activités très diversifiées : recherche/industrie.
L'ensemble de ces éléments sont recensés sur le portail de cette ACI dont l'adresse est la suivante : http://l3iexp.univ-lr.fr/madonne/

1. Liste des équipes impliquées

Equipes	**Laboratoires**	**Instituts**
Projet SID	Laboratoire d'informatique - image – interaction (L3i)	Université de La Rochelle
Equipe DICo	Laboratoire Perception, Systèmes et Informations (PSI)	Université/INSA de Rouen
Équipe QGAR	Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA)	Université Henri Poincaré de Nancy 1
Equipe Document	Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS)	INSA de Lyon
Equipe Document	Laboratoire Informatique (LI)	Université François Rabelais de Tours
Equipe IMADOC	Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA)	INSA de Rennes
Equipe Document	Centre de Recherche en Informatique de Paris 5 (CRIP5)	Université de Paris 5

2. Liste des participants au 30/09/2006

Equipe SID, L3i, Université de la Rochelle

Nom	Statut	Financement	Période	Taux
Ogier Jean-Marc	Professeur	Ministère de la Recherche	09/03 – 09/06	15 %
Mullot Rémy	Professeur	Ministère de la Recherche	09/03 – 09/06	15 %
Bertet Karell	MCF	Ministère de la Recherche	09/03 – 09/06	15 %
Loonis Pierre	MCF	Ministère de la Recherche	09/03 – 09/06	15 %
Burie Jean Christophe	MCF	Ministère de la Recherche	09/03 – 09/06	05 %
Surapong Uttama	Doctorant	Ambassade Thailandaise	04/03 – 09/06	100 %
Journet Nicholas	Doctorant	Bourse MESR	10/03 – 08/06	100 %
Mouhammed Hammoud	Etudiant de Master	aucun	03/05 – 07/05	100 %
Caldeira Manuel	Etudiant de Master	aucun	03/05 – 07/05	100 %
Karray Ali	Etudiant de Master	aucun	03/06 – 07/06	100 %
Ferré Ghislain	Etudiant de Master	aucun	03/06 – 07/06	100 %
Wilfrid Papin	Etudiant Ingénieur	aucun	03/04 – 09/04	100 %
Herbe Antony	Etudiant Ingénieur	aucun	03/05 – 07/05	100 %
Delalandre Mathieu	Ingénieur	ACI Madonne	10/05 – 08/06	100 %

Equipe DiCO, PSI, Université de Rouen

Nom	Statut	Financement	Période	Taux
Paquet Thierry	Professeur	Ministère de la Recherche	09/03 – 09/06	15 %
Heutte Laurent	Professeur	Ministère de la Recherche	09/03 – 09/06	15 %
Garain Utpal	Post Doctorant	CNRS	03/05 – 12/05	100 %
Bensefia Ameur	Docteur	ATER	09/04 - 08/05	30 %
Nicolas Stéphane	Doctorant	Bourse régionale et ATER	10/02 – 09/06	100 %
Kessentini Yousri	Etudiant de Master	aucun	03/04 – 07/04	100 %
Fouad Slimane	Etudiant de Master	aucun	03/05 – 07/05	100 %
Dardenne Julien	Etudiant de Master	aucun	04/06 – 09/06	100 %
Kessentini Yousri	Etudiant Ingénieur	PSI	03/03 – 06/03	100 %
Khemiri Mahassen	Etudiant Ingénieur	PSI	03/03 – 06/03	100 %
Sellami Mohamed	Etudiant Ingénieur	PSI	03/04 - 06/04	100 %
Torjmen Mouna	Etudiant Ingénieur	ACI Madonne	03/05 - 06/05	100 %

Equipe QGAR, INRIA, Université Henri Poincaré de Nancy 1

Nom	Statut	Financement	Période	Taux
Tombre Karl	Professeur	Ministère de la Recherche	09/03 – 09/06	10 %
Wendling Laurent	MCF	Ministère de la Recherche	09/03 – 09/06	25 %
Tabbone Salvatore	MCF	Ministère de la Recherche	09/03 – 09/06	25 %
Zuwala Daniel	Doctorant	Bourse MESR	10/03 – 09/06	15 %
Salmon Jean-Pierre	Doctorant	Bourse MESR	11/04 – 09/06	75 %
Salmon Jean-Pierre	Etudiant de Master	aucun	02/04 – 07/04	75 %
Victoire Thibaud	Etudiant de Master	aucun	02/04 – 07/04	75 %
Barrat Sabine	Etudiant de Master	aucun	02/05 - 07/05	75 %

Action Concertée Incitative

MASSES DE DONNEES

Rapport d'activité final

Equipe Document, LIRIS, INSA de Lyon

Nom	Statut	Financement	Période	Taux
Emptoz Hubert	Professeur	Ministère de la Recherche	09/03 – 09/06	05 %
Le Bourgeois Frank	MCF	Ministère de la Recherche	09/03 – 09/06	10 %
Eglin Véronique	MCF	Ministère de la Recherche	09/03 – 09/06	15 %
Bali Nadia	Etudiant de Master	aucun	04/04 – 09/04	100 %
El Abed Abir	Etudiant de Master	aucun	04/04 – 09/04	100 %
Gaceb Djamel	Etudiant de Master	aucun	04/04 – 06/05	100 %

Equipe Document, LI, Université François Rabelais de Tours

Nom	Statut	Financement	Période	Taux
Ramel Jean-Yves	MCF	Ministère de la Recherche	09/03 – 09/06	30 %
Romuald Boné	MCF	Ministère de la Recherche	10/04-09/09	10%
Qureshi Rashid	Doctorant	financement SFERE	10/04 – 09/06	50 %
Marteau Hubert	Doctorant	Bourse MESR	10/04-10/05	10%
Leriche Stéphane	Etudiant de Master	Aucun	04/04 - 07/04	100 %

Equipe IMADOC, IRISA, INSA de Rennes

Nom	Statut	Financement	Période	Taux
Coüasnon Bertrand[1]	MCF	Ministère de la Recherche	09/03 – 09/06	15 %
Lemaître Aurélie	Doctorant	Bourse MENRT	10/05 – 09/06	100 %
Lemaître Aurélie	Etudiant de Master	INRIA	02/05 – 06/05	100 %

Equipe Document, CRIP5, Université de Paris V

Nom	Statut	Financement	Période	Taux
Vincent Nicole	Professeur	Ministère de la Recherche	09/03 – 09/06	05 %
Pareti Rudolf	Doctorant		?? – ??	100 %

[1] En détachement INRIA sur un poste de Chargé de Recherche jusqu'à fin août 2005.

3. Changements significatifs intervenus dans le projet

Ce projet s'inscrit dans une démarche de sauvegarde et de valorisation de données patrimoniales. L'aspect « masse de données » était adressé sous l'angle des collections d'ouvrages numérisés, qui constituent d'ores et déjà des entrepôts gigantesques de données, représentés sous forme d'images scannées. La génération de ces entrepôts de données, présentés sous forme de collections de documents hétérogènes faiblement structurés soulève le problème de la recherche d'information et de la navigation au sein de ces corpus.

Dans le cadre de ce projet, il s'agissait de déterminer des indices s'adaptant aux différentes représentations de l'information que l'on peut rencontrer dans ces documents patrimoniaux comme des zones textuelles, imprimées ou manuscrites, des images, des illustrations graphiques. Ces signatures apportent des connaissances spécifiques qui aideront la navigation et la recherche d'informations.

Les indices sont extraits de manière automatique en utilisant entre autres des méthodes propres à l'analyse d'images, étendues et adaptées au cas des masses d'images. La consultation en mode image des documents patrimoniaux suppose leur archivage et exige donc d'examiner également de manière approfondie les possibilités spécifiques de compression de ces masses de documents. Pour cette ACI, les équipes participantes, expertes et complémentaires dans le domaine de l'analyse d'images de documents, apportent des solutions scientifiques et technologiques innovantes à cette problématique liée aux masses de données constituées par des collections numérisées.

Les défis scientifiques soulevés concernent la proposition d'une démarche générique qui permet, à la demande, l'instanciation d'une chaîne de traitements pour la valorisation d'une collection. Il s'agit d'une part de proposer une approche de modélisation des collections compatibles avec la diversité des contenus de ces collections et leur état de préservation. Sur la base de cette modélisation, il s'agit d'autre part de proposer une approche permettant de faire l'adéquation entre un ensemble de techniques de traitement d'images, le modèle ainsi défini et un point de vue de navigation défini par l'utilisateur.

Par rapport à ces objectifs initiaux, le projet MADONNE n'a pas connu de modifications « significatives » sur le plan des thématiques de recherche annoncées. Néanmoins, considérant les ressources attribuées au titre du consortium MADONNE, notamment en terme de ressources humaines, notre réseau de partenaires a proposé de réduire le champs d'investigation annoncé dans le dossier initial aux aspects purement scientifiques, en réduisant la partie consacrée à la notion de plate-forme logicielle. Les équipes ont donc principalement travaillé sur le développement d'algorithmes d'analyse d'image et de reconnaissance des formes pour la mise en place de chaîne d'indexation.

Evoquons à cette occasion le fait que la structuration de l'ACI Madonne a permis à cette communauté de proposer un atelier dans le cadre des programmes société de l'information du CNRS. En effet, dans le cadre de l'appel à projet sur les programmes Société de l'Information du CNRS, en association avec le pilote du Réseau Thématique Pluridisciplinaire Document (RTP Doc- 33, du département STIC), nous avons décidé de proposer une réponse commune qui s'est soldée par un atelier consacré à la numérisation du patrimoine. Cet atelier vise à définir un programme de recherche ambitieux concernant les perspectives de la numérisation, dans un cadre pluridisciplinaire.

Sur un plan plus structurel, le consortium a étendu son réseau de partenaires académiques et industriels comme en atteste le dernier séminaire scientifique organisé à la Bresse, qui a réuni 3 laboratoires supplémentaires et de nombreux industriels.

Cela s'est notamment matérialisé par la présence d'acteurs publics importants lors de la manifestation ANAGRAM, atelier scientifique, que nous avons organisée au mois de Septembre 2006, lors de la semaine du document numérique tenue à Fribourg en Suisse.

Parmi les nouveaux partenaires issus de la communauté STIC, notons depuis le début du projet la présence du laboratoire CRIP5 de l'Université de Paris 5, qui apporte ses contributions en matière d'indexation par le contenu, proposant des approches adaptées au graphique d'une part et au lien « texte-image » d'autre part.

Le Centre d'Etudes supérieures de la Renaissance de l'Université de Tours, dirigé par le professeur Marie-Luce Demonet est également très actif dans le consortium, participant à plusieurs groupes de travail. Les contributions du CESR sont très importantes pour notre réseau, car elles représentent la principale source de données numériques de nos activités d'une part, mais aussi et surtout car le CESR représente la composante associée aux usages, éléments essentiels dans notre dispositif.
Evoquons également la participation d'un laboratoire espagnol, le groupe « Document Analysis Group » du Computer Vision Center de l'Université Autonome de Barcelone. Cette équipe dispose d'une grande expérience dans le domaine de l'analyse automatique de documents et est également confrontée dans son pays à la problématique de la valorisation du patrimoine. Le CVC a d'ailleurs actuellement un partenariat bien établi avec les HP Labs, sur la problématique de la valorisation de données patrimoniales.

Enfin, évoquons la participation récente du laboratoire SIC de l'Université de Poitiers qui investit également sur la problématique de la valorisation du patrimoine, dans le cadre d'un partenariat avec la société RC Soft d'Angoulême.

En terme de ressources humaines, quelques laboratoires ont pris l'option d'orienter un certaines de leur thématique de recherche en direction de la valorisation du patrimoine, du fait de l'existence de la présente ACI. C'est notamment le cas du laboratoire d'historien médiéviste de la Sorbonne le LAMOP, avec lequel des contacts très sérieux de coopération sont engagés.

Dans ce cadre, le potentiel humain a augmenté, sur la base de ces orientations.
Pour d'autres laboratoires, la dynamique recherche sur la valorisation du patrimoine était déjà très fortement engagée, et le potentiel humain correspondant étoffe l'équipe Madonne, même si ces ressources auraient existé en dehors de cette ACI.

4. Résumé des principales avancées

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 surcoûts. 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, vise à concevoir des plates-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

Action Concertée Incitative

MASSES DE DONNEES

Rapport d'activité final

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 lequel 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. Cette synthèse 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.

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.

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 plates-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 [JOUR 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).

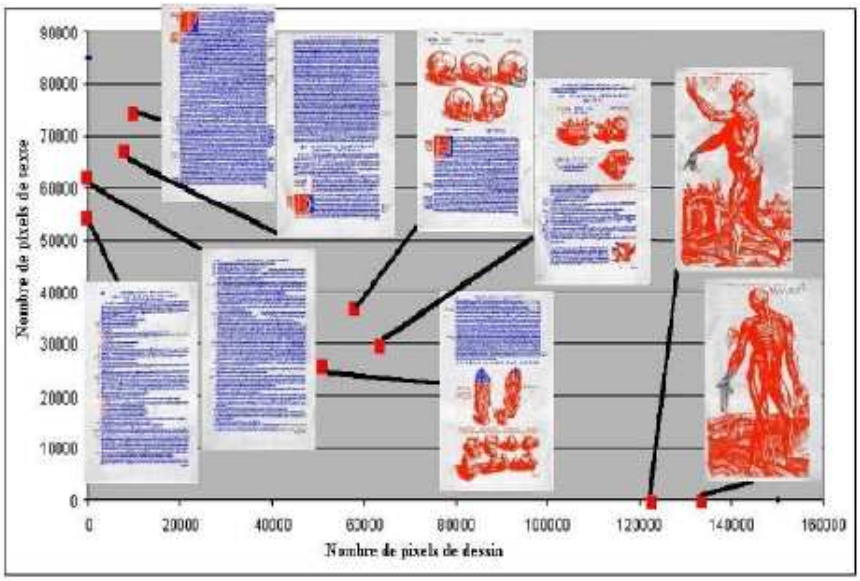


Figure 1 : Catégorisation des images

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 [COU 06] 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 plateforme 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.

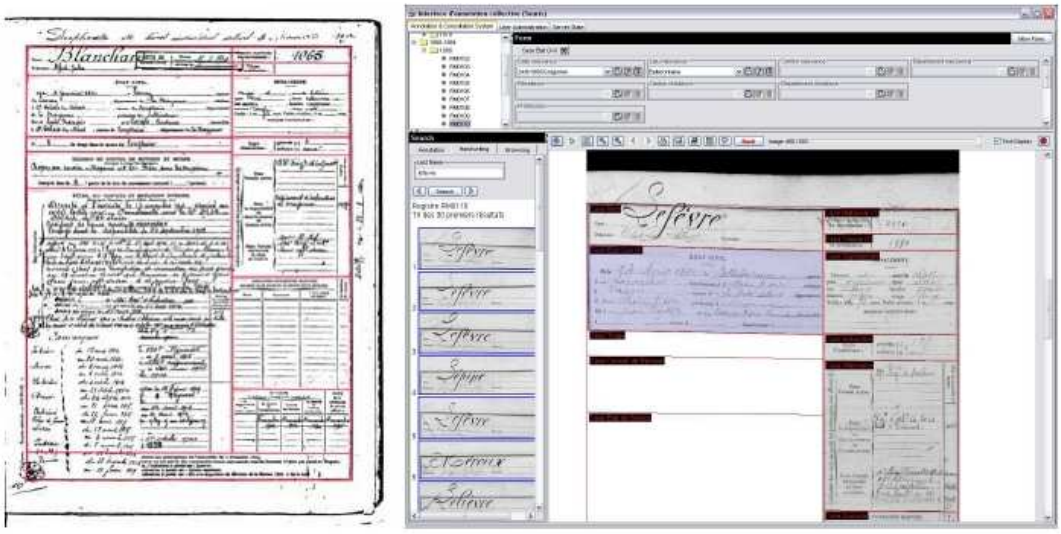


Figure 2 : (Gauche) Extraction de la structures physique
(Droite) Reconnaissance du patronyme manuscrit

Une autre contribution dans ce domaine a été réalisé par J.Y Ramel et al [RAM 06] via la plate-forme AGORA. 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.

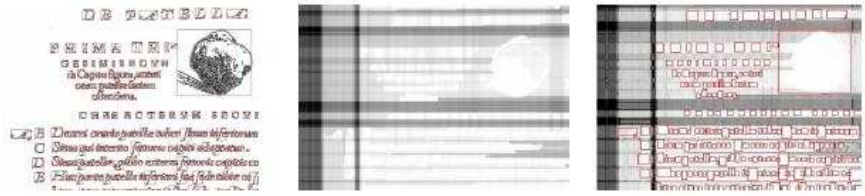


Figure 3 : Cartes du fond et des formes

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.

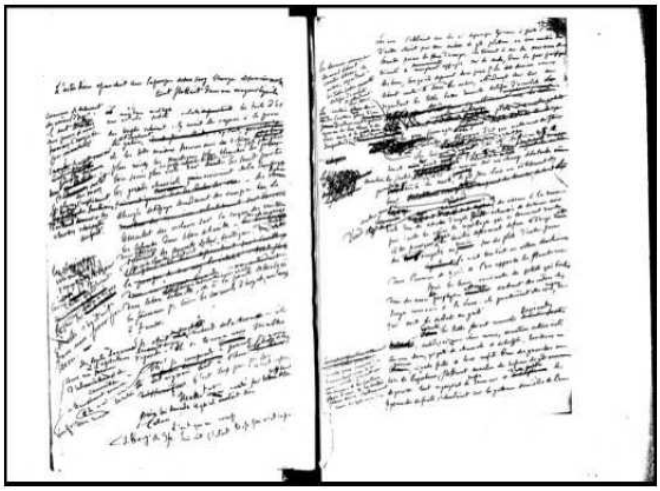


Figure 4 : Document manuscrit avec annotations

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. 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.

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.

Figure 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 multicritères 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.

Figure 6 : Restauration d'images

Sur la base des images prétraitées 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.

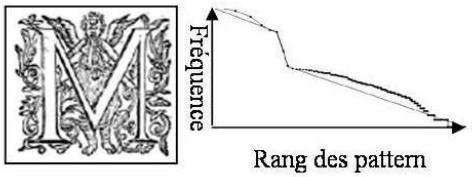


Figure 7 : Recherche sur critère de style

Figure 8 : Recherche sur critère de structure

Figure 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.

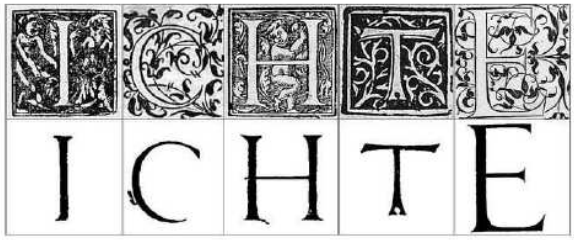


Figure 10 : Lettres extraites des lettrines

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.

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.

Bibliographie

[ABE 04] ABED A. E., Recherche de similarités partielles pour la compression des documents manuscrits du patrimoine, Mémoire de Master, 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, no 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, no 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.

5. Réalisations obtenues dans le cadre du projet

Outre les publications et la coordination en réseau de recherche, les réalisations principales obtenues dans le cadre du projet MADONNE concernent la conception de plates-formes logicielles. Celles-ci sont de deux types :

(1) Tout d'abord différents prototypes recherches ont été développés, ils sont relatés au sein des différentes publications des membres du Consortium.

(2) Ensuite différentes plates-formes opérationnelles ont été mises à disposition par les membres du Consortium. Celles-ci sont aujourd'hui utilisées par les partenaires du projet MADONNE (CESR de Tours, Projet Bovary, Archives des Yvelines, etc.) et proposées au téléchargement sur le site Web. Elles sont détaillées ci-dessous.

AGORA

Le logiciel AGORA permet d'extraire des méta-données des images de documents historiques en fonction d'un scénario défini par l'utilisateur. Pour cela, AGORA repose sur l'analyse de deux cartes de segmentation en blocs de l'image : une des formes et l'autre du fond. AGORA procède alors à une classification des blocs extraits pour la constitution des méta-données. Cette classification opère selon un scénario produit par l'utilisateur au cours d'une phase d'interaction avec AGORA.

DEBORA

DEBORA propose 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. Pour cela DEBORA est basée sur une décomposition des images en objets indépendants qui seront compressés avec des méthodes appropriées. DEBORA propose aussi un format de données hétérogènes, adapté à la navigation dans les ouvrages numérisés compressés, qui permet aussi de les modifier, les annoter ou les échanger sur Internet dans le cadre d'un travail collaboratif.

Action Concertée Incitative

MASSES DE DONNEES

Rapport d'activité final

DOCREAD

DocRead est un générateur automatique de systèmes de reconnaissance de documents structurés. Il est constitué d'un compilateur du langage EPF (permettant de décrire un document à l'aide d'une grammaire), d'un module d'analyse lié à ce langage, d'un module de vision précoce (binarisation et extraction de segments) et d'un classifieur ayant des capacités de rejet. DocRead permet ainsi une adaptation rapide à un nouveau type de document. En effet, il faut simplement définir une nouvelle grammaire (à l'aide d'EPF) qui décrit le nouveau type de document.

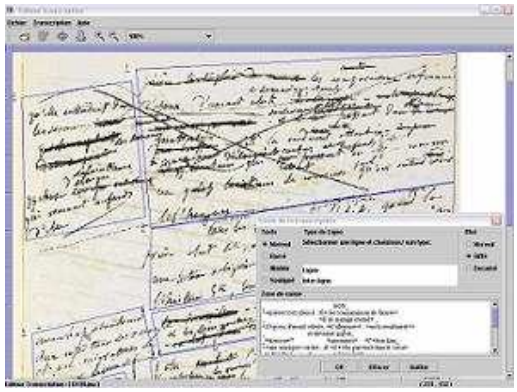
EMMA

L'éditeur EMMA permet de réaliser des transcriptions dites "diplomatiques" d'images de manuscrits. Son principal intérêt est de décharger le transcripteur des problèmes fastidieux de mise en forme des transcriptions. La sauvegarde des données s'effectue dans un format XML baptisé Gustave_ML. Ce dernier facilite les échanges de données en permettant l'enregistrement des transcriptions dans différents formats tel que le HTML pour la publication Web et le PDF pour l'impréssion papier.

Action Concertée Incitative

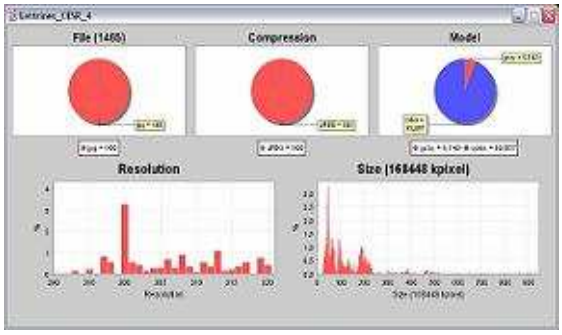
MASSES DE DONNEES

Rapport d'activité final



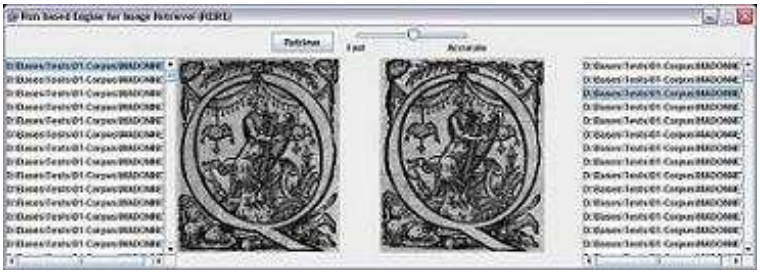
QUEID

QUEID "QUery Engine on Image Databases" est un outil de diagnostic de base d'images numérisées. Il extrait des bases les caractéristiques des images (modèles, formats, résolutions, etc.) afin d'en dresser une analyse statistique présentée sous la forme de graphiques à l'utilisateur. Ce dernier peut dans une deuxième étape utiliser QUEID en mode requête sur les caractéristiques des images. Le but est de naviguer au sein des bases afin d'y identifier les éventuels problèmes de numérisation.



REIRE

REIRE "Run Encoding Image based Retrieval Engine" est un moteur de recherche d'images similaires. Le but de REIRE est le traitement rapide de larges bases. Pour ce faire REIRE exploite une représentation compressée des images à base de plages. Cette représentation est utilisée à différents niveaux au travers d'un mécanisme de recherche perceptif. La recherche est alors affinée successivement afin de limiter l'espace de comparaison. Cette approche permet à REIRE d'effectuer des recherches particulièrement rapides des images au sein de larges bases.



6. Réunions et Conférences organisées dans le cadre du projet

Le tableau ci-dessous résume les différentes manifestations organisées dans le cadre du projet MADONNE tout en pointant vers leurs comptes-rendus et programmes. Outre les réunions du comité de pilotage le projet MADONNE a donné lieu à quatre manifestations principales :

(1) Workshop MADONNE - Tours - 27/06/2004 : Ce Workshop a réuni les membres du Consortium pour une première journée d'échanges scientifiques. Il s'est constitué de 12 présentations organisées en cinq sessions et a regroupé une quarantaine de participant.

(2) Workshop MADONNE - La Bresse - 19/05/2005 : Ce Workshop a réuni les membres du Consortium pour une seconde journée d'échanges scientifiques. Il s'est constitué de 18 présentations organisées en trois sessions et a regroupé une cinquantaine de participant.

(3) Atelier ACI Masse de Données - Tours - 24/01/2006 : Cet atelier a été organisé par l'ACI Masse de données dans le cadre de la conférence RFIA'06[2]. Il s'est constitué de 9 présentations dont quatre effectuées par des membres du Consortium MADONNE.

(4) The International Document Summer School '05Tromsø, Norway June 27-July 3rd, 2005. Cette école internationale sur le document numérique a été organisée par le Réseau Thématique Pluridisciplinaire « Document » (RTP 33). Les membres del'ACI MADONNE y a été conviée au titre d'animateur de l'atelier Numérisation du Programme Société de l'Information « Document Numérique du CNRS ». Cette école a été la base de nombreux contacts scientifiques de dimension inter-disciplinaire, et internationaux.

(5) Atelier « De l'archive à l'open archive. L'historien et internet », organisé par le projet ATHIS, Rome, Mars 2006. Suite à l'école internationale sur le document numérique organisée à Tromsoe en Norvège en 2005, les représentants de l'ACI Madonne ont été invités à cet atelier pour y présenter leur savoir et contribuer à un élargissement de la communauté SHS. Cet atelier était organisé par JP. Genet, directeur de l'UMR LAMOP de La Sobonne, dans le cadre d'un projet ANR-SHS, ATHIS. Cet atelier a été l'occasion de nombreux contacts internationaux, avec l'Italie notamment. Ces éléments seront, nous l'espérons, à la base de projets européens.

(6) Atelier ANAGRAM - Fribourg - 21/09/2006 : Cet atelier a été organisé par le Consortium MADONNE dans le cadre de la SDN'06[3]. Il s'est constitué de huit présentations organisées en deux sessions et a regroupé une trentaine de participant. Il a également donné lieu à des publications avec actes. A l'occasion de cet atelier le Consortium MADONNE a invité plusieurs orateurs de différentes institutions comme la BNF, l'OTAN, le CESR de Tours, la société EVODIA, etc.

Date	Evénement	Lieu	Lien
21/09/2006	Atelier ANAGRAM	Fribourg	programme
23/03/2006	Atelier de l'archive à l'open archive	Rome	Programme
24/01/2006	Atelier ACI Masse de Données	Tours	programme
27/06/2005	International Summer School	Tromsoe	Programme
19/05/2005	Workshop Madonne	La Bresse	programme
07/04/2005	Réunion du comité de pilotage	Paris	compte-rendu
27/06/2004	Workshop Madonne	Tours	programme
16/11/2004	Réunion du comité de pilotage	…	compte-rendu
20/11/2003	Réunion du comité de pilotage	…	compte-rendu
10/09/2003	Réunion du comité de pilotage	Paris	compte-rendu

[2] http://www.antsearch.univ-tours.fr/rfia2006/
[3] https://diuf.unifr.ch/event/sdn06/accueil.html

7. Soutiens obtenus en liaison avec ce projet

Postes chercheurs

Les ressources humaines de Madonne trouvent leur source principalement dans les choix politiques des laboratoires, qui contribuent à la vie scientifique du consortium, Certains laboratoires avaient déjà ces ressources sur des thématiques en relation avec la valorisation du patrimoine, issues de projet ou de financement ministériels et ont permis à leurs chercheurs de participer à la vie scientifiques de l'ACI.

D'autres laboratoires, dont l'activité initiale n'était pas centrée sur la valorisation du patrimoine ont clairement orienté leurs recherches sur cette nouvelle problématique, en proposant des sujets de thèse en relation avec le thème de l'ACI. Il en est de même pour les sujets de Masters Recherche.

Si l'on essaie de faire le point sur l'ensemble des postes de chercheurs recrutés et financés sur des thématiques connexes à MADONNE lors des 3 dernières années, mentionnons les personnes suivantes :

L3I – LA ROCHELLE				
Surapong Uttama	Doctorant	Ambassade Thailandaise	04/03 – 09/06	100 %
Journet Nicholas	Doctorant	Bourse MESR	10/03 – 08/06	100 %
Wilfrid Papin	Etudiant Ingénieur	Madonne	03/04 – 09/04	100 %
Herbe Antony	Etudiant Ingénieur	Madonne	03/05 – 07/05	100 %
Delalandre Mathieu	Ingénieur	ACI Madonne	10/05 – 08/06	100 %
PSI LITIS ROUEN				
Garain Utpal	Post Doctorant	CNRS	03/05 – 12/05	100 %
Bensefia Ameur	Docteur	ATER	09/04 - 08/05	30 %
Nicolas Stéphane	Doctorant	Bourse régionale et ATER	10/02 – 09/06	100 %
Kessentini Yousri	Etudiant Ingénieur	PSI	03/03 – 06/03	100 %
Khemiri Mahassen	Etudiant Ingénieur	PSI	03/03 – 06/03	100 %
Sellami Mohamed	Etudiant Ingénieur	PSI	03/04 - 06/04	100 %
Torjmen Mouna	Etudiant Ingénieur	ACI Madonne	03/05 - 06/05	100 %
LORIA NANCY				
Zuwala Daniel	Doctorant	Bourse MESR	10/03 – 09/06	15 %
Salmon Jean-Pierre	Doctorant	Bourse MESR	11/04 – 09/06	75 %
LI TOURS				
Qureshi Rashid	Doctorant	financement SFERE	10/04 – 09/06	50 %
Marteau Hubert	Doctorant	Bourse MESR	10/04-10/05	10%
IRISA RENNNES				
Lemaître Aurélie	Doctorant	Bourse MENRT	10/05 – 09/06	100 %
Lemaître Aurélie	Etudiant de Master	INRIA	02/05 – 06/05	100 %
CRIP5 PARIS 5				
Pareti Rudolf	Doctorant	Education Nationale	?? – ??	100 %

Action Concertée Incitative

MASSES DE DONNEES

Rapport d'activité final

Postes ingénieurs

Le recrutement d'un ingénieur de Recherche, a été opéré dans le cadre du financement de l'ACI MADONNE, de Septembre 2005 à Septembre 2006. Mathieu Delalandre, Docteur de l'Université de Rouen a été recruté dans ce contexte, et a joué un véritable rôle d'animation scientifique pendant cette année de travaux.

Contrats nationaux

Comme évoqué dans le paragraphe consacré aux évolutions du consortium, un dossier consacré à la numérisation a été déposé dans le cadre d'un projet commun avec les collègues du Réseau thématique pluridisciplinire document du RTP Doc, dans le cadre des projets consacrés au programme Société de l'Information du CNRS. Cet atelier est animé par JM Ogier et vient en complémentarité des activités de l'ACI MAdonne. Son budget de 5000 € a permis de contribuer aux frais de mission des individus dans le cadre des séminaires organisés conjointement par l'ACI et cet atelier.

Contrats internationaux hors CEE

Dans le cadre notre collaboration avec L'Institut de Statistiques de Calcutta, nous avons déposé une demande de financement auprès de l'ambassade de France en Inde (CEFIPRA) pour un projet intitulé :

Automatic Recognition of Degraded Printed Documents

Ce projet vise à proposer une approche robuste pour la production de transcriptions de documents imprimés dégradés. Pour la partie Française ce projet a été explicitement motivé par le projet Madonne et le besoin de produire des transcriptions textuelles dans les cas où les OCR industriels sont inopérants. Le financement de 36 mois de PostDoctorant a été demandé ainsi que les frais de mission nécessaires à la collaboration avec l'équipe Indienne qui développera pour sa part des approches robuste pour les OCR Indiens.

Un autre projet en partenariat avec la Thailande est en cours d'élaboration, en appui sur le consortium.

Contrats industriels

En terme de contrats industriels, plusieurs laboratoires disposent de convention CIFRE en relation avec la numérisation de documents. Cependant, même si ces activités ont des retombées sur la recherche concernant la valorisation du patrimoine, on ne peut pas dire pour l'instant que ces contrats constituent une résultante de de l'ACI.

Par contre, sur un plan national, un rapprochement fort est en cours avec les associations de professionnels dont le coeur de métier est celui de la gestion électronique de document. C'est en particulier le cas avec l'APROGED, (Association des Professionnels de la GEIDE) qui regroupe 80 entreprises de dimension nationale ou internationale. La mise en réseau avec cette association laisse entrevoir de nombreux rapprochements scienttifiques et stratégiques autour de la numérisation

Mentionnons en particulier que l'ACI MADONNE a été invitée à titre gratuit à présenter ses activités dans le cadre du forum de la GEIDE en 2004, 2005, et 2006, forum industriel sur le document numérique. Ces présentations nous ont permis de nouer de nombreux contacts très importants pour les

futures coopérations industrielles. Evoquons également que , plus ou moins directement en connexion avec l'ACI, deux start-up sont nées sur ces questions de numérisation du patrimoine, l'une issue de l'IRISA à Rennes, l'autre au laboratoire LIRIS de l'INSA de Lyon.

Contacts internationaux dans le cadre de ce projet

Nous avons établi de nombreux contact lors de la vie de cette ACI. Nous mentionnons ci-dessous uniquement les partenaires avec lesquels nous avons réellement coopéré, dans le contexte européen ou programme d'action intégré.

- Computer Vision Center, Université Autonome de Barcelone, Espagne
- Laboratoire REGIM, Safx Tunisie
- University of Bern Switzerland
- Universidad de Valladolid Spain
- Hewlett Packard España Spain
- Answare-Tech Spain
- The British Library United Kingdom
- Subdirecció General d'Arxius de la Generalitat de Catalunya Spain
- Escola Superior d'Arxivística i Gestió Documental Spain

Ces contacts internationaux nous ont permis de déposer un projet dans le cadre d'un réseau d'excellence en 2006, sur les mêmes questions, mais le projet n'a malheureusement pas été retenu.Nous ne désespérons pas de re-déposer un projet assez rapidement.
Nous attendons par ailleurs des réponses à des Programmes d'actions intégrées portés par EGIDE.

8. Conclusion générale de cette A.C.I. Masses de Données – Perspectives

A l'issue de ces 3 années de travail en commun, il nous est possible de tirer un ensemble de conclusions structurelles, scientifiques et organisationnelles. Ces conclusions/synthèses sont parfois corrélées les unes aux autres.

La première conclusion de ce type de projet se positionne bien sûr sur des questions scientifiques. Même si l'ACI Madonne n'a pas été très fournie en moyens humains (Un ingénieur directement financé par l'ACI sur un an, ce pour l'ensemble des acteurs, au nombre approximatif de 50 chercheurs), l'ACI a été un véritable catalyseur pour le déclenchement d'activités scientifiques communes.entre les acteurs majeurs de la recherche Française en analyse de documents. En effet, les rencontres organisées par l'ACI a permis une réelle « synchronisation » des actions recherche menées par les acteurs du consortium, ainsi qu'une coordination dans les stratégies élaborées. Ainsi, suite à des rencontres scientifiques (ateliers, conférences, …) des membres de l'ACI, les chercheurs se sont très souvent mis d'accord sur explorations scientifiques parallèles ou complémentaires, souvent en partenariat, au travers de la définition de co-tutelle de thèse. Nous le verrons un peu plus tard, mais ces échanges ont entre autre été à la base d'un projet scientifique ambitieux déposé à l'ANR pour cette année 2006 (projet NAVIDOMASS).

Sur un plan structurel et organisationnel, l'ACI a été une véritable opportunité pour la communauté Française pour se structurer et donner une visibilité nationale à ces activités autour de la numérisation du patrimoine. En effet, au cours des différentes conférences internationales, de nombreuses présentations au titre de l'ACI ont été présentées, et des articles communs dans des revues scientifiques à très fort « impact factor » sont en cours de rédaction, également au titre de l'ACI. De ce point de vue, le bilan est donc extrêmement positif dans le sens où un esprit communautaire s'est réellement installé, en appui sur une très bonne connaissance des compétences/complémentarité de chacun. L'atelier ANAGRAM et ses résultats sont est une excellente illustration de cet aspect lié à la valorisation des activités de recherche française en la matière. Cette ACI a donc joué un rôle très structurant.

Un des résultats importants de cette ACI réside également dans la prise de recul que la communauté sur des questions fondamentales sur lesquelles il est opportun de se pencher aujourd'hui.
En effet, après avoir abordé de nombreuses questions relatives à l'extraction de signature sur des typologies variées d'images de documents hétérogènes (manuscrits, graphiques, structurés, ….) et en masse, les verrous communs que notre communauté a fortement ressenti concernent la problématique de l'indexation multi-dimensionnelle, dans des espaces de dimension élevée, avec des contraintes de masses de données. C'est précisément dans ce contexte recherche que la majeure partie des acteurs de l'ACI Madonne ont proposé de poursuivre cette aventure scientifique et humaine dans le cadre d'un projet ANR, « Masse de Données - Connaissances Ambiantes"

9. Publications obtenues dans le cadre du projet

Articles de journaux et chapitres de livres

[1] B. Coüasnon and J. Camillerapp. Accès par le contenu aux documents manuscrits d'archives numérisés. *Document Numérique*, 7(3-4):61-84, 2003.

[2] V. Eglin, S. Bres, and C. Rivero-Moreno. Hermite and gabor based approaches for patrimonial handwriting processing. *International Journal of Document Analysis and recognition (IJDAR)*, à paraître.

[3] U. Garain, T. Paquet, and L. Heutte. On foreground-background separation in low quality document image. *International Journal on Document Analysis and Recognition (IJDAR)*, 8(1):47-63, 2006.

[4] F. Lebourgeois, H. Emptoz, and E. Trinh. Compression et accessibilité aux images de documents numérisés : application au projet debora. *Document Numérique*, 7(3-4):103-125, 2003.

[5] F. Lebourgeois and H. Emptoz. Debora: Digital access to books of the renaissance. *International Journal on Document Analysis and Recognition (IJDAR)*, à paraître.

[6] J. Ramel, S. Leriche, M. Demonet and S. Busson. User-driven page layout analysis of historical printed books. *International Journal on Document Analysis and Recognition (IJDAR)*, à paraître.

[7] J. Ramel and S. Leriche. Segmentation et analyse interactives documents anciens imprimés. *Traitement du Signal (TS)*, 22(3):209-222, 2005.

[8] J. Salmon, L. Wendling, and S. Tabbone. Improving the recognition by integrating the combination of descriptors. *International Journal on Document Analysis and Recognition (IJDAR)*, à paraître.

Articles de conférences et workshops internationaux

[9] E. Abed, V. Eglin, F. Lebourgeois, and H. Emptoz. Frequencies decomposition and partial similarities retrieval for patrimonial handwriting documents compression. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 996-1000, 2005.

[10] E. Baudrier, G. Millon, F. Nicolier, and S. Ruan. A new similarity measure using hausdorff distance map. In *International Conference on Image Processing (ICIP)*, pages 669-672, 2004.

[11] E. Baudrier, G. Millon, F. Nicolier, and S. Ruan. A fast binary-image comparison method with local-dissimilarity quantification. In *International Conference on Pattern Recognition (ICPR)*, 2006.

[12] E. Baudrier, G. Millon, F. Nicolier, and S. Ruan. Une méthode de comparison d'images binaires quantifiant les dissimilarités locales application à la classification d'impressions anciennes. In *Colloque International Francophone sur l'Ecrit et le Document (CIFED)*, pages 211-215, 2006.

[13] E. Baudrier, G. Millon, F. Nicolier, R. Seulin, and S. Ruan. Hausdorff distance based multiresolution maps applied to an image similarity measure. In *Optical Sensing and Artificial Vision (OSAV)*, pages 18-21, 2004.

[14] E. Baudrier, G. Millon, F. Nicolier, and R. Seulin. Impression virtuelle de tampons en bois gravé anciens. In *Séminaire maquette virtuelle et patrimoine*, pages 25-28, 2003.

[15] J. Camillerapp, L. Pasquer, and B. Coüasnon. Indexation automatique de formulaires anciens par reconnaissance du patronyme manuscrit. In *Reconnaissance des Formes et Intelligence Artificielle (RFIA)*, pages 1493-1502, 2004.

[16] B. Coüasnon, J. Camillerapp, and I. Leplumey. Making handwritten archives documents accessible to public with a generic system of document image analysis. In *Conference on Document Image Analysis for Libraries (DIAL)*, pages 270-277, 2004.

[17] B. Coüasnon and I. Leplumey. A generic recognition system for making archives documents accessible to public. In *International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 228-232, 2003.

[18] M. Delalandre and J. Ogier. Un système pour l'indexation rapide d'image de lettrine. In *Colloque International Francophone sur l'Ecrit et le Document (CIFED)*, pages 253-258, 2006.

[19] V. Eglin, S. Bres, and C. Rivero-Moreno. Biological inspired tools for patrimonial handwriting denoising and categorization. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 59-63, 2005.

[20] V. Eglin, F. Lebourgeois, S. Bres, H. Emptoz, Y. Leydier, I. Moalla, and F. Drira. Computer assistance for digital libraries: Contributions to middle-ages and authors'manuscripts exploitation and enrichment. In *Conference on Document Image Analysis for Libraries (DIAL)*, pages 265-280, 2006.

[21] D. Gaceb, V. Eglin, and S. Bres. Extraction de similarités dans les manuscrits du patrimoine pour la compression des images et la caractérisation des styles. In *Colloque International Francophone sur l'Ecrit et le Document (CIFED)*, pages 229-234, 2006.

[22] D. Gaceb, V. Eglin, and S. Bres. Handwriting similarities as features for the characterization of writer's style invariants and image compression. In *International Conference on Image Analysis and Recognition (ICIAR)*, 2006.

[23] U. Garain, T. Paquet, and L. Heutte. On foreground-background separation in low quality document image. In *International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 585-589, 2005.

[24] N. Journet, V. Eglin, J. Ramel, and R. Mullot. Text/graphic labelling of ancient printed documents. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 1010-1014, 2005.

[25] N. Journet, R. Mullot, V. Eglin, and J. Ramel. Analyse d'images de documents anciens : Catégorisation de contenus par approche text. In *Colloque International Francophone sur l'Ecrit et le Document (CIFED)*, pages 247-252, 2006.

[26] N. Journet, R. Mullot, V. Eglin, and J. Ramel. Dedicated texture based tools for characterisation of old books. In *Conference on Document Image Analysis for Libraries (DIAL)*, pages 60-70, 2006.

[27] N. Journet, R. Mullot, J. Ramel, and V. Eglin. Ancient printed documents indexation: a new approach. In *International Conference on Advances in Pattern Recognition (ICAPR)*, volume 3686 of *Lectures Notes in Computer Science (LNCS)*, pages 513-522, 2005.

[28] N. Journet, J. Ramel, V. Eglin, and R. Mullot. Caractérisation de la mise en page des documents imprimés de la renaissance par une analyse des orientations. In *Colloque GRETSI*, pages 122-129, 2005.

[29] F. Lebourgeois and H. Kaileh. Automatic metadata retrieval from ancient manuscripts. In *Workshop on Document Analysis Systems (DAS)*, volume 3163 of *Lecture Notes in Computer Science (LNCS)*, pages 75-89, 2004.

[30] F. Lebourgeois, E. Trinh, B. Allier, V. Eglin, and H. Emptoz. Documents images analysis solutions for digital libraries. In *Conference on Document Image Analysis for Libraries (DIAL)*, pages 2-24, 2004.

[31] C. MADONNE. Madonne : Masse de données issues de la numérisation du patrimoinne. In *Atelier sur la Numérisation de l'Ecrit Ancien et des GRAndes Masses de données (ANAGRAM)*, 2006.

[32] I. Martinat and B. Coüasnon. A minimal and sufficient way of introducing external knowledge for table recognition in archival documents. In *Workshop on Graphics Recognition (GREC)*, pages 194-205, 2005.

[33] S. Nicolas, Y. Kessentini, T. Paquet, and L. Heutte. Handwritten document segmentation using hidden markov random fields. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 212-216, 2005.

[34] S. Nicolas, T. Paquet, and L. Heutte. Digitizing cultural heritage manuscripts: the bovary project. In *Symposium on Document Engineering (DocEng)*, pages 55-57, 2003.

[35] S. Nicolas, T. Paquet, and L. Heutte. Enriching historical manuscripts: the bovary project. In *Workshop on Document Analysis Systems (DAS)*, volume 3163 of *Lecture Notes in Computer Science (LNCS)*, pages 135-146, 2004.

[36] S. Nicolas, T. Paquet, and L. Heutte. Text line segmentation in handwritten documents using a production system. In *International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pages 245-250, 2004.

[37] S. Nicolas, T. Paquet, and L. Heutte. Un panorama des méthodes syntaxiques pour la segmentation d'images de documents manuscrits. In *Colloque International Francophone sur l'Ecrit et le Document (CIFED)*, pages 237-242, 2004.

[38] S. Nicolas, T. Paquet, and L. Heutte. Complex handwritten page segmentation using contextual models. In *Conference on Document Image Analysis for Libraries (DIAL)*, pages 46-57, 2006.

[39] S. Nicolas, T. Paquet, and L. Heutte. Extraction de la structure de documents manuscrits complexes à l'aide de champs markoviens. In *Colloque International Francophone sur l'Ecrit et le Document (CIFED)*, pages 13-18, 2006.

[40] S. Nicolas, T. Paquet, and L. Heutte. A markovian approach for handwritten document segmentation. In *International Conference on Pattern Recognition (ICPR)*, volume 3, pages 292-295, 2006.

[41] S. Nicolas, T. Paquet, and L. Heutte. Markov random field models to extract the layout of complex handwritten documents. In *International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, 2006.

[42] J. Ogier and K. Tombre. Madonne: Document image analysis techniques for cultural heritage documents. In *International Conference on Digital Cultural Heritage*, 2006.

[43] T. Paquet. Processing of archived historical handwritten documents. In *Workshop on Document Analysis Systems (DAS)*, pages 65-90, 2005.

[44] R. Pareti, S. Uttama, J. Salmon, J. Ogier, S. Tabbone, L. Wendling, and N. Vincent. On defining signatures for the retrieval and the classification of graphical dropcaps. In *Conference on Document Image Analysis for Libraries (DIAL)*, pages 220-231, 2006.

[45] R. Pareti and N. Vincent. Global discrimination of graphics styles. In *Workshop on Graphics Recognition (GREC)*, pages 120-128, 2005.

[46] R. Pareti and N. Vincent. Ancient initial letters indexing. In *International Conference on Pattern Recognition (ICPR)*, pages 756-759, 2006.

[47] J. Ramel, S. Busson, and M. Demonet. Agora: the interactive document image analysis tool of the bvh project. In *Conference on Document Image Analysis for Libraries (DIAL)*, pages 145-155, 2006.

[48] J. Salmon, L. Wendling, and S. Tabbone. Automatic definition of measures from the combination of shape descriptors. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 986-990, 2005.

[49] J. Salmon, L. Wendling, and S. Tabbone. Reconnaissance de symboles graphiques à partir d'une combinaison de descripteurs en intégrant leur comportement sur une base d'apprentissage. In *Colloque International Francophone sur l'Écrit et le Document (CIFED)*, pages 97-102, 2006.

[50] S. Uttama, M. Hammoud, C. Garrido, P. Franco, and J. Ogier. Ancient graphic documents characterization. In *Workshop on Graphics Recognition (GREC)*, pages 97-105, 2005.

[51] S. Uttama, J. Ogier, and P. Loonis. Top-down segmentation of ancient graphical drop caps: Lettrines. In *Workshop on Graphics Recognition (GREC)*, pages 87-96, 2005.

Thèses de doctorat et mémoires de master

[52] E. Abed. Caractérisation de tracés manuscrits et recherches de similarités en vue de la compression par approche fractale. Mémoire de Master, Laboratoire LIRIS, Université de Lyon, France, 2004.
[53] N. Bali. Compression et codage pour la transmission de documents patrimoniaux. Mémoire de Master, Laboratoire LIRIS, Université de Lyon, France, 2004.
[54] E. Baudrier. *Comparaison d'images binaires reposant sur une mesure locale des dissimilarités Application à la classification*. Thèse de Doctorat, Université de Reims Champagne-Ardenne, 2005.
[55] A. Caldeira. Indexation d'images de documents anciens. Mémoire de Master, Laboratoire L3i, Université de la Rochelle, France, 2005.
[56] J. Dardenne. Champs aléatoires discriminants pour l'analyse d'images de documents. Mémoire de Master, Laboratoire PSI, Université de Rouen, France, 2006.
[57] G. Ferré. Segmentation d'images de documents anciens par approche texture. Mémoire de Master, Laboratoire L3i, Université de la Rochelle, France, 2006.
[58] D. Gaceb. Extraction de similarités de formes dans les images de traits. Mémoire de Master, Laboratoire LIRIS, Université de Lyon, France, 2005.
[59] M. Hamoud. Extraction des signatures d'images graphiques pour l'indexation d'images : applications à la valorisation du patrimoine. Mémoire de Master, Laboratoire L3i, Université de la Rochelle, France, 2005.
[60] A. Hassaine. Codage de graphèmes et compression sans perte d'images de manuscrits anciens. Mémoire de Master, Laboratoire LIRIS, Université de Lyon, France, 2006.
[61] N. Journet. *Analyse d'images de documents : une approche texture*. Thèse de Doctorat, Université de La Rochelle, France, à paraître.
[62] A. Karray. Recherche de lettrines par le contenu. Mémoire de Master, Laboratoire L3i, Universités de La Rochelle et de Sfax, France et Tunisie, 2006.
[63] S. Leriche. Segmentation et analyse structurelle interactives de documents imprimées anciens. Mémoire de Master, Ecole Polytechnique, Université de Tours, France, 2004.
[64] F. Slimane. Extraction de la structure de documents à forte variabilité spatiale à l'aide de champs de markov cachés. Mémoire de Master, Laboratoire PSI, Université de Rouen, France, 2005.

Rapports techniques

[65] M. Delalandre. Refonte du site web madonne. Rapport technique, Laboratoire L3i, Université de La Rochelle, France, 2006.
[66] P. Dezanneau, P. Dujon, P. Ferret, D. Fisson, L. Paviot, and A. Verneuil. Réalisation d'un site web pour le projet aci madonne. Rapport technique, Laboratoire L3i, Université la Rochelle, France, 2004.
[67] A. Herbe. Mise en place d'une plate-forme d'indexation d'images par le contenu. Rapport technique, Laboratoire L3i, Université de la Rochelle, France, 2005.
[68] Y. Labarre and J. Selier. Acquisition svg de vérités terrain, application aux livres anciens imprimés. Rapport technique, Laboratoire L3i, Université de La Rochelle, France, 2006.
[69] W. Papin. Projet madonne. Rapport technique, Laboratoire L3i, Université de la Rochelle, France, 2004.