Project : NAVIDOMASS

NAVigation In DOcument MASSes

1er Rapport semestriel d'activité -coordonnateur

Août 2007

Associated Research Teams

Laboratoire L3i - Université de La Rochelle
Laboratoire CRIP5 – Université de Paris 5
Laboratoire LORIA – Nancy
Laboratoire IRISA – Rennes
Laboratoire d'Informatique - Tours
Laboratoire L.I.T.I.S – Rouen
Laboratoire C.E.S.R – Tours

Project Leader : Jean-Marc OGIER
Université de La Rochelle

Rapport semestriel d'activité -coordonnateur
Programme MDCA - Edition 2006

Remarque : le coordonnateur doit fournir un rapport partenaire et un rapport coordonnateur.

Identification

Acronyme du projet	NAVIDOMASS
Numéro d'identification de l'acte attributif	ANR-06-CIS6-
Coordonnateur (société/organisme)	Université de la Rochelle
Période couverte (date à date)	01/01/2007 à 01/09/2007
Période couverte (t0+n mois à t0+m mois)	01/01/2007 à 01/09/2007 ; t0+8 mois
Rédacteur (nom, téléphone, email)	Ogier Jean-Marc, 05 46 45 82 15, jean-marc.ogier@univ-lr.fr
Date	18/07/2007

URL de la page web du projet et date de dernière mise à jour

http://imedoc.univ-lr.fr/tiki-index.php?page_ref_id=5

Activités de coordination des activités du projet

Suivant le calendrier initialement annoncé, ce premier semestre a été principalement consacré à la mise en oeuvre stratégique du consortium NavidoMass. Plusieurs réunions du consortium ont été organisées, afin de :

- travailler le workpackage relatif à la demande utilisateur, et sur l'élaboration des cahiers des charges techniques et fonctionnels pour les utilisateurs.
- définir les axes stratégiques de chaque partenaire
- analyser les interactions scientifiques entre les partenaires
- définir des profils de postes pour les recrutements en fonctions des axes stratégiques et des interactions entre partenaires.
- travailler sur la question de l'inter-opérabilité des logiciels réalisés par chaque partenaire
- démarrer les travaux scientifiques en fonction des axes stratégiques définis.

Ces différentes réunions, dont les dates sont rappelées ci-dessous, ont conduit à des décisions mentionnées dans le chapître « faits marquants ».

- 22 Janvier 2007, Paris : Réunion de démarrage de projet
- 7 Mars 2007,Tours : Présentation des acquis par laboratoires et perspectives sur le projet NAVIDOMASS, présentation des programmes de numérisation du CESR, expression des utilisateurs, présentation du projet iconclass
- 2 Mai 2007, Paris : Point sur la rédaction des projets de recherche pour les ingénieurs en recrutement, échanges sur les partenariats scientifiques
- 4 Juillet 2007 : Point sur les questions d'interopérabilité entre les laboratoires, échanges sur les formats d'inter-opérabilité entre les laboratoires partenaires, pour un accompagnement par un prestataire privé.
- 12 et 13 Juillet 2007 : Workshop Document graphique organisé à La Rochelle

- 12 Septembre 2007 : Séminaire recherche , première contributions, exposés séniors, présentation des axes de recherche des personnes recrutées

Synthèse

Numéro du Partenaire	Conformité des résultats obtenus aux prévisions	Conformité de la consommation des ressources par rapport aux prévisions	Difficultés particulières
1	Conformes aux prévisions	Conformes aux prévisions	Difficultés de recrutement dès le démarrage du projet, considérant la difficulté de trouver des personnes ressources adaptées
2	Conformes aux prévisions	Conformes aux prévisions	Difficultés de recrutement dès le démarrage du projet, considérant la difficulté de trouver des personnes ressources adaptées
3	Conformes aux prévisions	Conformes aux prévisions	Difficultés de recrutement dès le démarrage du projet, considérant la difficulté de trouver des personnes ressources adaptées
4	Conformes aux prévisions	Conformes aux prévisions	Difficulté à sélectionner un prestataire de numérisation compétent compte tenu des contraintes imposées par les marchés (mais la 2e campagne de numérisation prévue en 2008 devrait éviter ces écueils par un cahier des charges plus explicite). Avancement des numérisations : un peu plus lent que prévu (un mois de retard) à cause de la compétence très relative du prestataire (opérateur qu'il a fallu former, contrairement aux contrat), ce qui ne devrait pas sensiblement modifier le calendrier de travail. Pour le reste (recherche sur l'indexation, validation des recherches sur les lettrines par les laboratoires partenaires, modèle de sortie d'OCR sous xml, etc) : conformes.
5	Conformes aux prévisions	Conformes aux prévisions	Difficultés de recrutement dès le démarrage du projet, considérant la difficulté de trouver des personnes ressources adaptées
6	Conformes aux prévisions	Conformes aux prévisions	Difficultés de recrutement dès le démarrage du projet, considérant la difficulté de trouver des personnes ressources adaptées
7	Conformes aux prévisions	Conformes aux prévisions	Aucune

Synthèse	Globalement Conformes aux prévisions	Globalement Conformes aux prévisions	Seules difficultés liées à un démarrage de projet.

Faits marquants

Indiquer les résultats et/ou réalisations marquants. Préciser s'ils peuvent ou non faire l'objet de communications externes par l'ANR et la Délégation ANR-CI.

Pour rappel, 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'agit 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 ce projet, 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.

Suivant le calendrier initialement annoncé, ce premier semestre a été principalement consacré à la mise en oeuvre stratégique du consortium NavidoMass. Plusieurs réunions du consortium ont été organisées, afin de :

- Préparer les contrats de travail, les conventions inter-laboratoires, les accords de consortium
- travailler le workpackage 1.1., relatif à la demande utilisateur, et sur l'élaboration des cahiers des charges techniques et fonctionnels pour les utilisateurs.
- définir les axes stratégiques de chaque partenaire et répartir les tâches en fonction des ressources allouées
- analyser les interactions scientifiques entre les partenaires
- définir des profils de postes pour les recrutements en fonctions des axes stratégiques et des interactions entre partenaires.
- travailler sur la question de l'inter-opérabilité des logiciels réalisés par chaque partenaire
- démarrer les travaux scientifiques en fonction des axes stratégiques définis.
- travailler sur les prestations de services à demander en externe.
- Mettre en place la promiotion des postes à pourvoir et assurer le début de la phase de recrutement des futurs doctorants

Ces différentes réunions, dont les dates sont rappelées ci-dessous, ont conduit à des décisions importantes

telles que :

- le choix d'orientations technologiques, qui seront externalisées à un prestataire de services, afin de permettre aux partenaires de se consacrer réellement aux problématiques scientifiques. Ces orientations ont permis de définir la façon dont les échanges de données et de traitement pourront s'opérer pendant la durée du consortium. Le prestation sera probablement proposée à la société EVODIA localisée à Rennes, dont les compétences semblent très proches de celles requises pour la bonne mise en oeuvre de ce projet scientifique.

- les modalités de recrutement : les financements obtenus pour ce projet étant inférieurs à ceux initialement demandés, les partenaires se sont accordés sur une mutualisation des ressources qui leur ont été attribuées afin de favoriser les interactions entre partenaires. Entre autres, cela se matérialise par des financements conjoints de personnes ressources sous forme d'ingénieur de recherche pour des périodes de 3 ans décomposés en 2 fois 18 mois (chaque tranche de 18 mois étant financée par un partenaire). Cette organisation permettra une réelle interaction scientifique entre les partenaires et permettra d'exploiter au mieux les complémentarités scientifiques de chaque laboratoire. Des thèses seront proposées pour ces travaux, et les sujets proposés font partie de ce rapport. Les recrutements sont arrêtés et les doctorants démarreront leur travaux dès le mois de Septembre. Notons dans ce contexte général que le consortium NAVIDOMASS a bien obtenu 18 mois de financement de salaire supplémentaire grâce à la contribution de la région Lorraine (LORIA).

Le calendrier des réunions et les ordres du jour associés sont les suivants :

- **Journée NAVIDOMASS du 22 Janvier 2007, Paris :**
  ODJ :

  * présentation des ressources, rappels des objectifs scientifiques, discussions sur les axes stratégiques, définitiondes modalités de fonctionnement du consortium NAVIDOMASS : un comité de pilotage est défini, fréquence des réunion techniques, fréquence des réunions scientifiques, analyse des ressources humaines, échanges sur la mutualisation des ressources en fonction des workpackage et des contributions de chaque partenaire au projet.

- **Journée NAVIDOMASS du 7 Mars 2007, Tours :**
  ODJ :

  * travail sur le WP 1 : expression des demandes utilisateurs, programmation de la numérisation, contraintes de numérisation (résolution) suivant la nature des informations à indexer
  * Présentation des compétences de la société EVODIA, partenaire susceptible de prendre en charge les parties technologiques du projet.
  * Présentation du projet européen IconClass susceptible d'être associé au consortium NAVIDOMASS

- **Journée NAVIDOMASS du 2 Mai 2007, Paris :**
  ODJ :

  * travail sur les sujets de recherche proposés aux personnels recrutés (thèse) et répartition des co-encadrements scientifiques. Préparation du workshop scientifique du mois de Septembre.

- **Journée NAVIDOMASS du 4 Juillet, Paris**
  ODJ :

  * travail sur la question de l'inter-opérabilité entre les partenaires du projet, élaboration de cahier des charges travail sur les demandes de chaque partenaire en termes de prestation.

- **Journée Documents graphiques pour l'indexation de masses de documents** : Organisation du « Document Image Analysis Workshop » organisé à La Rochelle les 12 et 13 juillet 2007 faisant le point sur les techniques d'analyse de graphique, entre autres pour le projet NAVIDOMASS. (programme de ces journées joint à ce dossier). Ces journées ont permis aux partenaires d'échanger scientifiquement sur leur savoir-faire en matière d'indexation d'images graphiques, et ont permis de croiser les approches afin d'élaborer un programme scientifique commun pour les 3 années du projet.

Difficultés rencontrées

Difficultés classiques liées à un démarrage de projet.

Difficulté à sélectionner un prestataire de numérisation compétent compte tenu des contraintes imposées par les marchés (mais la 2e campagne de numérisation prévue en 2008 devrait éviter ces écueils par un cahier des charges plus explicite).

Suivi des délivrables du projet

(exemple, le tableau initial est celui contenu en annexe 1)

	Libellé	Nat.	Partenaires Date		07 S1	07 S2	08 S1	08 S2	09 S1	09 S2
T0	**Coordination – Communication - PROJECT MANAGEMENT**									
T0a Deliv 0.1	Website implementation	Web	Tous	Semestriel au minimum	X					
T0b Deliv 0.2	Project coordination : regular meetings, definition of objectives, report redaction	CR	Tous	Plusieurs réunions	X					
T0c Deliv 0.2	Réunion du 22 Janvier, Paris	CR	Tous	trimestrielle	X					
T0c Deliv 0.2	Réunion du 07 Mars, Tours	CR	Tous	trimestrielle	X					
T0c Deliv 0.2	Réunion du 2 Mai, Paris	CR	Tous	trimestrielle	X					
T0c Deliv 0.2	Réunion du 04 Juillet, Paris	CR	Tous	trimestrielle	X					
T0c Deliv 0.2	Réunion du 12 Juillet, La Rochelle		Tous	annuelle	X					
T0c Deliv 0.2	Réunion du 12 Septembre, Paris		Tous	annuelle	X					
T1	**Users needs, participative design and ground truthing**									
T1a : Deliv 1.1	bibliographic synthesis concerning the constitution of representative sets of Mass of documents, integrating the variability in terms of structure, graphic, manuscript. Digitization of the retained documents	R	CESR/ LI / L3i	Nov 2007 comme prévu dans le contrat	En cours A					
T2	**Document Layout analysis and structure based indexing**									
T2a Deliv 2.1	bibliographic synthesis dealing with state of art concerning structure based indexing, and intermediate report concerning scientific orientations and first developments	R	LITIS	Nov 2007 comme prévu dans le contrat	En cours A	R1		X		
T3	**Information spotting**									
T3a : Deliv 3.1	Bibliographic synthesis dealing with state of art concerning the signatures allowing to perform wordspotting, and graphic spotting : signatures, metrics ; intermediate report concerning scientific orientations and first developments	R	CRIP5/ LITIS/LI/LORIA/IRISA/L3i	Nov 2007 comme prévu dans le contrat	En cours A					
T4	**Structuring the feature space**									
T4a Deliv 4.1	Bibliographic synthesis dealing with state of art concerning the structuration of feature space in a context of high dimensional features	R	LORIA/ L3i/CRIP5	Nov 2007 comme prévu dans le contrat	En cours A	R1		X		
T5	**Interactive extraction and relevance feedback**									
T5a Deliv 5.1	bibliographic synthesis concerning interactive extraction, dynamic scenario building, user interactions, relevance feedback, and Interactive content retrieval	R	LORIA/ LI/L3i/ CRIP5/IRISA	Nov 2007 comme prévu dans le contrat	En cours A	R1		X		

Nat. : CR = Compte-rendu, R = rapport, …

X : prévision initiale
A : atteint
R1, R2, … : reprévision

Commentaires

Dans le contrat initial, les membres du consortium avaient prévu des rapports intermédiaires annuels, sur la base des travaux coopératifs et des réunions menées en cours d'année. Sur ce plan, la fréquence des rencontres entre les partenaires étant très élevée, les rapports sont bien avancés et seront disponibles dans les délais. En attendant, plusieurs documents relatifs au plan de route (roadmap) des membres du consortium sont disponibles et joints à ce fichiers, sous la forme des sujets de thèse déposés.

Liste des CDD recrutés par des établissements publics dans le cadre du projet
Lister ici tous les CDD recrutés depuis le début du projet.

Numéro du Partenaire	Nom	Prénom	Qualifications	Date de recrutement	Durée du contrat (en mois)	
1	Coustaty Mikael		**Ingénieur d'études**	**01/09/2007**	**18**	
2	Sidere Nicolas		**Ingénieur d'études**	**01/09/2007**	**24**	**Co-financement propre LI Tours**
3	Jouili Salim		**Ingénieur d'études**	**01/09/2007**	**30**	**Co-financement Région Lorraine**
4	Pedroja Cynthia Dufournaud Nicole		**IGE: Pré-indexation des ouvrages, tests sur la reconnaissance des caractères accentués** **IGE : Anaylyse des sorties d'OCR**	**01-01-07** **01-05-2007**	**6** **2**	
5	Coustaty Mikael		**IGE**	**Recrutement prévu dans 18 mois**		
6	Ait-Mohand		**Ingénieur d'études**	**01/10/2007**	**12**	

Equipements achetés par les partenaires dans le cadre du projet
Lister ici tous les équipements achetés depuis le début du projet

Numéro du Partenaire	Désignation	Date d'achat	Prix d'achat (en Euros)	Part financées par l'aide ANR (en Euros)		
1	2 ordinateurs portables + logiciels		**6000**	**100%**		
2	1 ordinateur portable		**1500**	**100%**		
3	RAS pour l'instant					
4	RAS pour l'instant		**3500**	**100%**		
5	RAS pour l'instant					
6	1 ordinateur portable + logiciels		**3000**	**100%**	**Prévu avant fin 2007**	

Liste des délivrables joints au présent rapport (uniquement pour les rapports de fin d'année)
Les délivrables du projet sont fournis par le coordonnateur.

Numéro du délivrable	Désignation	Forme/Support		

Annexes

RFAI_LI_01 :
Sujet de thèse Navidomass n°1 - LI/RFAI Tours - LI TIS Rouen

Sujet : Vers une indexation structurelle des contenus pour la navigation et l'interrogation de fonds documentaires anciens

Encadrants :
JY Ramel – LI / RFAI Tours – Directeur de thèse
Pierre Héroux LITIS Rouen - encadrant

1. Modalités pratiques de la thèse :

Comment candidater sur ce sujet :
Compétences requises : Le candidat doit avoir, avant le début de la thèse, validé un niveau d'étude équivalent à un Master 2 de recherche dans les domaines de l'informatique, du traitement d'images et/ou de la reconnaissance de formes.

Dépôt de candidature : Envoyer un CV ainsi qu'une lettre de motivation à
- Jean-Yves Ramel (ramel@univ-tours.fr)
- Pierre Heroux (pierre.heroux@univ-rouen.fr)

Financement : CDD de 3 ans soutenu par l'ANR dans le cadre de son appel d'offre Masses de Données et Connaissances Ambiantes sur le projet NAVIDOMASS (NAVigation In DOcument MASSses).

Inscription : Université de Tours – **Date de début prévue :** octobre 2007

Co-encadrement :
- Jean-Yves Ramel, Maître de Conférences au LI, Université de Tours
- Pierre Héroux, Maître de Conférences au LITIS, Université de Rouen

Une mobilité importante sera demandée : le doctorant devra notamment partager son temps entre les 2 laboratoires (2 périodes d'un an et demi par exemple).

2 Description du sujet de thèse
Dans le cadre de cette thèse, nous souhaitons aborder le problème de la recherche d'informations dans de grandes bases documentaires à partir de la caractérisation des structures physique et logique des documents. Les structures ou parties de structures recherchées pourront être représentées sous forme de graphes attribués. Ce type de représentation permet de modéliser dans un formalisme unifié les données ou les concepts et les relations qui existent entre ceux-ci.

C'est particulièrement intéressant en analyse de documents puisqu'il est nécessaire de représenter, à différents niveaux, aussi bien les contenus que les liens existants entre les contenus (le fond et la forme). Les Élements de Contenus présents dans un document, leurs caractéristiques ainsi que leurs positionnements relatifs peuvent être modélisés sous forme de graphe. On peut donc envisager que plusieurs graphes soient attachés à la description d'une page de document et tenter, pour chacune de ces représentations, de mettre en place plusieurs méthodes de comparaisons et/ou mesures de similarité (isomorphisme de graphe, de sous-graphes, ou tout autre calcul de similarité entre graphes. . .) afin d'être capable de les comparer de manière pertinente.
Plusieurs problématiques seront abordées pour apporter de nouvelles solutions: la classification (supervisée ou non) de graphes, l'indexation et la fouille de graphes, la possibilité de procéder par raffinements successifs. La boucle de pertinence pourrait être mise en œuvre pour choisir ou pondérer la ou les

méthodes ou mesures de similarité de manière à renforcer les plus à même de répondre aux besoins de l'utilisateur.
Les travaux fondamentaux réalisés seront validés sur des images de documents anciens numérisés fournis par le CESR de Tours dans le cadre du projet Navidomass. Le doctorant pourra bénéficier de l'expérience et des nombreux outils de construction et comparaison de graphes dont disposent les 2 laboratoires partenaires (LI/RFAI et LITIS).

3 Contexte

Le sujet de thèse proposé se place dans le cadre du projet Navidomass soutenu par l'ANR dans le cadre de son appel d'o_re Masses de Données et Connaissances Ambiantes.
Ce projet regroupe plusieurs partenaires institutionnels et industriels :
_ Laboratoire L3I, Université de la Rochelle
_ Laboratoire CRIP5, Université René Descartes Paris
_ Équipe QGAR, LORIA Nancy
_ Laboratoire d'Informatique LI, Université de Tours
_ Laboratoire LITIS, Université de Rouen
_ Équipe IMADOC, IRISA, Université de Rennes
_ Société EVODIA, Rennes
_ Centre d'Études Supérieures de la Renaissance, Tours

L'objectif de ce projet vise à développer des approches permettant d'améliorer les possibilités de navigation dans les grandes masses documentaires. Dans le cadre d'une valorisation du patrimoine s'adressant aussi bien au grand public, qu'aux chercheurs intéressés par l'étude de ces documents, les corpus visés sont essentiellement des fonds de documents anciens numérisés tels que les nombreux documents dont dispose le Centre d'Études Supérieures de la Renaissance.

De tels documents sont caractérisés par le fait qu'ils disposent souvent d'un contenu hétérogène mêlant texte imprimé avec des fontes anciennes, annotations manuscrites, illustrations diverses (lettrines, schéma. . .). Outre les spécificités des documents (les conventions de mise en page diffèrent des schémas contemporains), l'âge des documents à l'origine de la dégradation du support physique, les conditions d'impression, ou encore les conditions de numérisation sont source d'altération de la qualité des images rendant inopérants les traitements utilisés en analyse d'image de documents plus récents voire contemporains. Le candidat sera amené à une mobilité entre les laboratoires LI Tours et LITIS Rouen.

RFAI_LI_02 :
Sujet de thèse Navidomass n°2 - LI/RFAI Tours - LI TIS Rouen

Sujet : Techniques robustes d'indexation d'images de documents par leur contenu textuel

Encadrants :
Laurent Heutte – LITIS Rouen – Directeur de thèse
Nicolas Ragot – LI/RFAI Tours – encadrant

Comment candidater sur ce sujet :
Compétences requises : Le candidat doit avoir, avant le début de la thèse, validé un niveau d'étude équivalent à un Master 2 de recherche dans les domaines de l'informatique, du traitement d'images et/ou de la reconnaissance de formes.

Dépôt de candidature : Envoyer un CV ainsi qu'une lettre de motivation à
- Nicolas Ragot (nicolas.ragot@univ-tours.fr)
- Laurent Heutte (laurent.heutte@univ-rouen.fr)
- Thierry Paquet (thierry.paquet@univ-rouen.fr)

Modalités pratiques de la thèse :
Financement : CDD de 3 ans soutenu par l'ANR dans le cadre de son appel d'offre Masses de Données et Connaissances Ambiantes sur le projet NAVIDOMASS (NAVigation In DOcument MASSses).
Inscription : Université de Rouen

Co-encadrement :
- Laurent Heutte et Thierry Paquet, Professeurs au LITIS, Université de Rouen
- Nicolas Ragot, Maître de Conférences au LI, Université de Tours

Une mobilité importante sera demandée : le doctorant devra notamment partager son temps entre les 2 laboratoires (2 périodes d'un an et demi par exemple).

Description du sujet :

Contexte
Dans le cadre de la numérisation des fonds patrimoniaux, l'exploitation des composantes textuelles est un élément incontournable. En effet, ces composantes serviront non seulement à la transcription électronique des textes mais elles permettront aussi d'accéder aux informations qu'ils contiennent par l'indexation des masses de collections. D'autre part, l'indexation des composantes textuelles selon leur contenu/représentation graphique permet d'accéder à des connaissances qui jusqu'à présent ont été rarement exploitées, excepté dans le cadre strict de l'analyse de textes imprimés pour détecter des informations telles que fontes, graisses, tailles…
Dans le contexte de masses de documents importantes, l'interrogation des archives fondée sur un mode de description graphique permettrait d'accéder à des informations caractérisant les imprimeurs par exemple. Malgré un certain nombre d'idées reçues sur les performances supposées des systèmes d'OCR actuels, ceux-ci se montrent clairement inadaptés pour traiter de telles masses de documents patrimoniaux. Ces difficultés proviennent de : la variabilité des fontes, graisses, etc. à la fois entre corpus mais également au sein d'un même corpus, variabilité que les OCR ont encore du mal à appréhender ; des nombreuses dégradations présentes lors de la numérisation de ces documents (tâches, défauts d'impression, apparition du verso par transparence…) qui peuvent notamment provoquer une fragmentation des caractères ou au contraire qui les rendent interconnectés ; et enfin de la « liberté » lexicale et orthographique relative à l'écriture des mots. L'ensemble de ces points constitue donc un verrou important quant à l'exploitation, par les systèmes d'OCR actuels, de ces masses de documents anciens.

Objectifs

Dans ce contexte, le sujet de la thèse a pour objectif de développer de nouvelles approches de reconnaissance pour l'indexation de ces masses de documents anciens, approches devant permettre de répondre à la grande variabilité des représentations textuelles qu'il est possible de rencontrer dans ces documents tout en les inscrivant dans une démarche générique de conception visant l'adaptabilité à la demande des techniques d'indexation des images de documents. Les méthodes de reconnaissance développées se doivent donc d'être à la fois robustes vis à vis de la variabilité et de la dégradation des représentations graphiques rencontrées sur un problème particulier, et adaptables à des problèmes similaires mais différents (adaptation rapide à de nouvelles collections). L'approche proposée doit pouvoir répondre de manière générale à tout problème d'indexation par le contenu des composantes textuelles, qu'il s'agisse de textes imprimés de qualité ou dégradés, et permettre ainsi de rechercher, à partir de requêtes « image » ou « texte », des mots, des expressions ou des passages pertinents dans ces documents. Pour cela, l'indexation par le contenu des composantes textuelles devra être envisagée sur deux niveaux de représentation: au niveau graphique (indexation à partir de primitives « image ») et au niveau textuel (indexation à partir de résultats OCR partiels).

En ce qui concerne l'indexation des entités textuelles au niveau graphique, une approche spécifique doit être développée pour prendre en compte à la fois le caractère alphabétique sousjacent des représentations et le caractère graphique (description 2D). Les techniques fondées sur des principes de Recherche d'Informations Visuelles textuelles seront donc plus particulièrement examinées à cet effet. Il s'agira donc de définir, extraire, sélectionner des primitives « image » tolérantes à la variabilité et au niveau de dégradation des documents (caractères dégradés, caractères connectés, espaces inter-mots variables…) et de développer des méthodes élastiques robustes de mise en correspondance des chaînes de primitives permettant l'appariement de mots ou expressions recherchés. En ce qui concerne l'indexation des entités textuelles au niveau le plus haut (à partir de résultats de reconnaissance), l'approche envisagée s'inspirera notamment des approches aujourd'hui bien maîtrisées de la reconnaissance de l'écriture manuscrite pour lesquelles on retrouve cette problématique de la variabilité des styles d'écriture, de l'adaptation à de nouveaux corpus, de sur et sous-segmentation, ainsi que l'utilisation de connaissances lexicales relatives au domaine d'utilisation. D'autre part on se basera également sur des approches statistiques de reconnaissance de formes qui offrent naturellement (par construction) des possibilités d'adaptation des modèles de caractères et de mots (lexiques) à un problème spécifique grâce à des techniques d'apprentissage à partir d'exemples. Deux voies seront plus particulièrement approfondies : elles concernent d'une part l'acquisition des connaissances lexicales liées à un domaine et d'autre part la proposition d'une démarche alternative à l'approche classique dans le domaine de l'apprentissage - la constitution de bases d'apprentissage - étape qui constitue un goulot d'étranglement dans une perspective d'adaptation rapide à de nouveaux problèmes. Sur ce dernier point, l'apport des techniques d'apprentissage semi-supervisé et non supervisé sera plus particulièrement exploré.

En vue de la conception d'un système générique d'indexation par le contenu des composantes textuelles, c'est l'ensemble de ces modèles et méthodes qui devront coopérer de façon à pouvoir s'adapter rapidement à de nouveaux corpus.

Bibliographie :

· A. Bensefia, T. Paquet, L. Heutte. Writer identification and verification: two complementary approaches for the quantitative analysis of handwritten documents. Journal of Forensic Document Examination, vol. 16, pp. 57-76, 2004.

· A. Bensefia, T. Paquet, L. Heutte. Documents manuscrits et recherche d'information. Document Numérique, Hermès, vol. 7, no. 3-4, pp. 47-60, 2003.

· F. Bouteruche, E. Anquetil, N. Ragot, Handwritten gesture recognition driven by spatial context of strokes, in Proceedings of the 8th International Conference on Document Analysis and Recognition, Bob Werner (ed.), Volume 2, Pages 1221-1225, Seoul, Korea, Août 2005.

· U. Garain, T. Paquet, L. Heutte. On foreground-background separation in low quality document images. International Journal on Document Analysis and Recognition, vol. 8, no. 1, pp. 47-63, 2006.

· G. Koch, L. Heutte, T. Paquet. Automatic extraction of numerical sequences in handwritten incoming mail documents. Pattern Recognition Letters, vol. 26, no. 8, pp. 1118-1127, 2005.

· H. Mouchère, E. Anquetil, N. Ragot, Writer Style Adaptation in Online Handwriting Recognizers by a Fuzzy Mechanism Approach: The Adpat Method, International Journal of Pattern Recognition and Artificial Intelligence, special issue on IGS, Vol. 29, N°1, pp.99-116, World Scientific, 2007.

· A. Nosary, L. Heutte, T. Paquet. Unsupervised writer adaptation applied to handwritten text recognition. Pattern Recognition, vol. 37, no. 2, pp. 385-388, 2004.

Sujet de thèse Navidomass-Loria Nancy

Sujet : Indexation de masses de documents graphiques

Encadrant: S. Tabbone (tabbone@loria.fr)

Description du sujet de thèse
Cette thèse s'inscrit dans la problématique de la reconnaissance de symboles dans les document graphiques. Plus précisément, il s'agira d'étudier des méthodes aptes à rechercher de l'information dans de masses de documents représentées sous forme structurelle. Dans les documents graphiques, les méthode structurelles sont adaptées pour la reconnaissance de symboles car elles offrent un cadre riche d description d'un symbole de forme quelconque et des relations structurelles qui peuvent exister au niveau d document. Parmi la multitude des structures de données existantes, les graphes relationnels attribué semblent appropriés pour formaliser les les symboles ainsi que les relations intrinsèques qui le composen Il sera question dans cette thèse d'explorer des problématiques autour de méthodes d'indexation classification et fouille de graphes.

Contexte
La thèse se déroulera dans l'environnement scientifique du projet NAVIDOMASS (NAVigation In Documen MASSes de l'appel à projet ANR MDCA) qui porte sur l'indexation de grandes bases de documents ancien du patrimoine.

Sujet de thèse Navidomass-CRIP5 Paris 5

Sélection et hiérarchisation d'un ensemble de primitives dans un contexte de reconnaissance de formes

Encadrant: Nicole Vincent (nicole.vincent@math-info.univ-paris5.fr)

La multiplication des données qui s'accumulent dans des fichiers images de documents rend nécessaire le développement de méthodes d'extraction de l'information de manière à retrouver des données. Nous nous placerons ici dans le cas d'images et plus précisément de recherche d'images par le contenu. Suivant la nature de la requête, cette extraction est réalisée à l'aide du calcul d'un certain nombre primitives. De très nombreuses primitives sont possibles et qui sont de natures très variées. Suivant le problème étudié, les primitives sont plus ou moins efficaces. L'objectif de la thèse est d'étudier une approche par sélection des primitives les plus efficaces lors d'une étape préalable, cela nécessite souvent de combiner différents indices. La sélection peut être faite de manière directe sur l'ensemble des primitives ou selon des groupements. Les différents modes de sélections devront être étudiés et développer conduisant à une hiérarchisation des primitives.
Le domaine d'application qui permettra de valider les propositions faites de manière théorique est celui des documents anciens numérisés, et en particulier les lettrines qui ont pu être préalablement extraites mais d'autres types d'images sont concernés comme les symboles techniques.

L3i_01 :
Sujet de thèse Navidomass n°1 – L3i / CRIP5/LORIA

Titre : Mise en place d'une plate-forme d'indexation d'image par le contenu : application à la numérisation du patrimoine

Encadrement : Jean-Marc Ogier (L3i), Muriel Visani (L3i), Nicole Vincent (CRIP5), Antoine Tabbonne (LORIA)

Contact :
Bureau n°112 Bâtiment Pascal
Tél.05 46 45 82 15 Mail : Jean-Marc.Ogier@univ-lr.fr

Positionnement dans la recherche au L3i :
Thématique(s) : ISI / DOFIN ; Ce sujet fait partie de l'équipe iMedoc du laboratoire L3i (http://imedoc.univ-lr.fr/tiki-index.php).

Mots Clefs : Reconnaissance des formes, indexation d'images, signatures

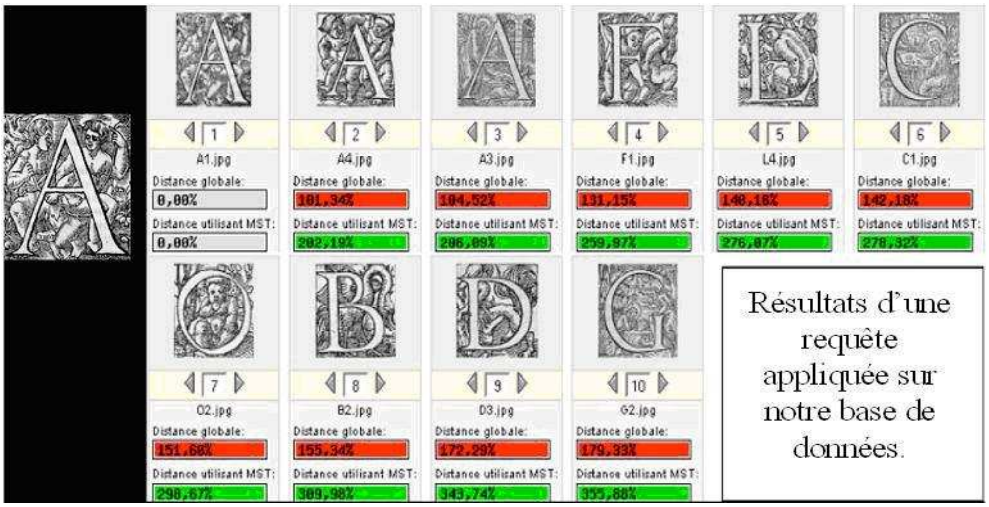
Contexte: projet ANR Masses de Données NAVIDOMASS : http://imedoc.univ-lr.fr/tiki-index.php?page_ref_id=5

Ce projet ANR NAVIDOMASS s'inscrit dans une démarche de sauvegarde et de valorisation de données patrimoniales dont la communauté internationale a pris conscience de l'intérêt, comme en attestent les impulsions prises au niveau européen par les représentants nationaux des grands organismes de gestion du patrimoine. Le contexte de l'étude concerne les collections d'ouvrages numérisés, qui constitueront à très court terme des entrepôts gigantesques de données, représentés sous forme d'images scannées, pour lesquelles les techniques traditionnelles des bases de données sont inopérantes. L'exploitation et la valorisation à venir de ces collections d'images n'a toujours pas trouvé de réponse satisfaisante, du fait même de leur caractère faiblement structuré. 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.

Sujet :

Ce projet concerne le développement d'un moteur de recherche, appliqué à des images. Dans le cadre de projets de recherche intégrant plusieurs laboratoires Français (http://l3iexp.univ-lr.fr/madonne, http://imedoc.univ-lr.fr/tiki-index.php?page_ref_id=5), il s'agit de développer des outils permettant de rechercher permettant de retrouver une image parmi un ensemble d'image stockées dans une base.
Il s'agit donc de développer des signatures caractéristiques des images constituant le « résumé visuel » de ces images, et de développer des métriques, permettant de mesurer le degré de « ressemblance » entre ces images. De nombreuses études ont déjà été développées au L3i dans cette direction, et il s'agit d'accompagner cette démarche, en proposant de nouveaux algorithmes et/ou en expérimentant des jeux de signatures sur des bases d'image de référence.
Les images sur lesquelles seront expérimentés ces algorithmes sont des images de lettrines, issues d'un centre de recherche partenaire.

En particulier, le sujet portera sur l'indexation de lettrines don't nous donnons quelques illustrations sur la figure 1. La difficulté de traitement de ces lettrines est liée aux textures présentes conjointement à des formes très structurées (letter).



RFAI_LI_03 :
Rapport interne sur les demandes de prestations externes

Besoins en prestations RFAI / LI Tours

Besoins liés à la thèse 1 (rédacteur JY Ramel)

- Réalisation et mise a disposition d'une plateforme Navidomass = serveur + Stockage + Interfaces Web d'accès, de navigation, de visualisation et création des métadonnées (annotations) et interrogation de la base d'images sélectionnées pour Navidomass (corpus de grande taille)

- Interface d'ajout de nouveaux jeux de données (images ou autres) avec identification (propriétaire)

- Type d'images désirées => Plusieurs ouvrages complets imprimés (et manuscrits) avec différents niveaux de dégradations, un contenu et une structure diversifiés (texte, graphiques, notes, ...) Pages complètes en couleur résolution >= 300 ppp

- Vérité terrain (avec propriétaire) sur la structure des pages exprimée selon la méthode proposée dans {Breuel2006] (+ svg)

- Version OCRisée des zones textuelles contenues dans les images (on peut rêver !)

- Outils de gestion de plans d'expériences = sélection d'un lot d'images dans le corpus pour application d'une chaine de traitements puis archivage des résultats obtenus

- Outils d'importation de métadonnées vers la plateforme = Réalisation des outils de transcodage des méta données (signatures,...) générées par les algo développés en labo pour les rendre compatible avec la plateforme Navidomass

- Réalisation d'outils d'exportation de métadonnées de la plateforme vers un format spécifique = Sélection

d'un sous ensemble de métadonnées extrait sur un sous ensemble d'images ou morceau d'img du corpus

- Identification, Listage et Activation d'algorithmes Labo (EXE, boite noire) a distance et visualisation des résultats

- outils dévaluation de performance et comparaison de résultats (résultats algo1 vs algo 2 OU résultats algo1 vs Vérité terrain)

- Annotation multi-niveaux : niveau ouvrage, ensemble de pages (chapitre), 1 page complète, 1 zone dans 1 page, des sous-zone dans 1 zone, association de zones, 1 zones dans 2 pages successives, pixel, ...

- Notion de propriétaire (identification du créateur) des annotations

Besoins liés à la thèse 2 (rédacteur N. Ragot)

- Réalisation et mise a disposition d'une plateforme Navidomass (attention problématique : corpus de grande taille) = serveur + Stockage + Interfaces Web d'accès, pour :
- pouvoir naviguer dans les ouvrages (date, auteur, titre, n° page,)
- pouvoir visualiser les pages brutes, zoomer dessus jusqu'au niveau lettre
- pouvoir ajouter/modifier (avec identification) :
  - de nouvelles images : Type d'images désirées => Plusieurs ouvrages complets imprimés (et manuscrits) avec différents niveaux de dégradations, un contenu et une structure diversifiés (texte, graphiques, notes, ...) ; pages complètes en couleur résolution >= 300 ppp
  - pouvoir définir des métadonnées associées à une zone de l'image (mot/caractère) : définir cette zone (rect-englobant ?), donner une valeur à des attributs prédéfinis, définir de nouveaux attributs et leur affectés une valeur ; prévoir éventuellement des valeurs prédéfinie pour certains attributs (année, auteur, style de police, tampon, imprimeur, etc.)
  - pouvoir voir et modifier des méta-données sur les zones de l'image définies (visualisation des rect-anglobants), sur les caractères/mots : étiquette issue d'une reco, vérité terrain, OCR classique (différentes hypothèses possibles ?), attributs sur le style, la police, etc.
- pouvoir faire des requêtes textuelles ou image en se basant sur les méta-données ou bien sur les hypothèses issues de la classif et présenter les résultats de façon ordonnée et lisible (10 premiers resultats par ex. avec l'image du mot correspondant à la requête et les caracs de la page correspondante (n° pp, ouvrage, année) et avec un score de confiance, avec également un accès direct à la visualisation de cette page et du mot dans cette page,
- Avoir des outils de gestion de plans d'expériences =
  - selection d'un lot d'images dans le corpus pour application d'une chaine de traitements
  - définir un protocole de communication entre la plateforme et les applis de traitements et d'indexation/recherche :
    - ♣ récupération de méta-données/ d'informations sur l'image
    - ♣ pouvoir voir les résultats de différents traitements (binarisation, segmentation, composantes connexes, points/traits singuliers/repérés, reconnaissance de lettres, d'un ou plusieurs mots consécutifs ; cette reco pouvant être partielle (différentes hypothèses, qualifiée (score, valeurs d'attributs, etc.)) ou pas) ; taux de rappel, fausse alarme, etc.
  - avoir des outils d'évaluation de performances et comparaison de resultats (resultats algo1 vs algo 2 OU resultats algo1 vs Verité terrain)
  - archivage des resulats obtenus
  - pouvoir exporter les meta-donnees et informations associées aux images ou moceaux d'images

Une personne pour mettre des ouvrages/images dans la plateforme et transcrire/annoter le texte, corriger l'OCR

Traitements standards ??

- binarisation, lissage, morpho, etc.
- sementation en composantes connexes
- rect-anglobant des composantes
- détection de lignes, paragraphes, mots, caractères