CLUSTERS DE RECHERCHE
RHÔNE-ALPES

Allocations Doctorales de Recherche 2011

« Dynamiques sociales et territoriales ; Culture, patrimoine et création ; Sciences/Société : Enjeux, représentations et usages »

1. Titre

Méthodes d'aide à la navigation, la lecture et l'exploitation des images des correspondances manuscrites dan l'Europe du 18ème siècle

2. Cluster concerné

Cluster Culture, patrimoine, création

3. Thématique

(Avant de formaliser votre projet, merci de prendre contact avec les animateurs de la thématique concernée.)

☐ Projet 1 : Risques – Observation sociale et pilotage des politiques publiques
Responsable : Philippe.Warin@iep-grenoble.fr

☐ Projet 2 : Action collective, conflictualité sociale et mobilisation associative
Responsable : Spyros.Franguiadakis@univ-lyon2.fr

☐ Projet 3 : Mobilité, territorialité, marginalité
Responsable : yves.chalas@orange.fr

☒ Projet 4 : Exclusion scolaire, linguistique, apprentissage, socialisation
Responsable : Dominique.Glasman@upmf-grenoble.fr

☐ Projet 4 : Economie sociale et solidaire
Responsable : daniele.demoustier@iep.upmf-grenoble.fr

☐ Projet 5 : Patrimoine et territoire
Responsable : bernard.gauthiez@gmail.com

☐ Projet 6 : Genre et culture
Responsable : Christine.Plante@univ-lyon2.fr

☐ Projet 7 : Editions critiques
Responsable : mckenna@univ-st-etienne.fr

☒ Projet 8 : Corpus numériques
Responsables : thomas.Lebarbe@u-grenoble3.fr, Sylvie.Calabretto@insa-lyon.fr, veronique.eglin@liris.cnrs.fr

☐ Projet 9 : Création
Responsable : Florent.Gaudez@upmf-grenoble.fr

☐ Projet 10 : Les processus de modélisation et la théorie de la science
Responsable : daniel.parrochia@wanadoo.fr et philippe.walter@u-grenoble3.fr

☐ Projet 11 : Sciences, techniques et communication
Responsables : joelle.le-marec@ens-lyon.fr et isabelle.pailliart@u-grenoble3.fr

☐ Projet 12 : Formation scientifique et didactique des sciences
Responsable : sylvain.gravier@ujf-grenoble.fr

☐ Projet 13 : La construction des Interfaces
Responsables : Joelle.Forest@insa-lyon.fr et dominique.vinck@upmf-grenoble.fr

☐ Projet 14 : Politiques scientifiques et politiques publiques : enjeux des sciences sociales
Responsable : renaud.payre@univ-lyon2.fr

4. Auteur(s) de la proposition

Dans le cas où l'action est co-pilotée, indiquez les coordonnées de deux personnes au maximum

Nom : McKenna	**Nom : Eglin**
Prénom : Antony	Prénom : Véronique
Etablissement : Institut Cl. Longeon, U. de Saint-Etienne, 35, rue du Onze Novembre, St-Etienne	Etablissement : INSA de Lyon
Laboratoire : Institut Claude Longeon, Institut d'Histoire de la pensée classique, de l'Humanisme aux Lumières, CNRS UMR 5037	Laboratoire : LIRIS UMR CNRS 5205
e-mail : mckenna@univ-st-etienne.fr	e-mail : veronique.eglin@insa-lyon.fr
Téléphone : 04 77 42 16 71	Téléphone : 04 72 43 60 54

5. Encadrant(s)

Si différents des personnes indiquées dans la section 4

Nom:	Nom :
Prénom:	Prénom :
Etablissement :	Etablissement :
Laboratoire :	Laboratoire :
e-mail :	e-mail :
Téléphone :	Téléphone :

6. Sujet de thèse **(maximum 20 lignes)**

Titre : Méthodes d'aide à la navigation, à la lecture et à l'exploitation des images des correspondances manuscrites dans l'Europe du 18ème siècle.	
ED : Lyon 2	Etablissement gestionnaire : Lyon 2
Résumé du projet : Mise en place de méthodes originales et complètes de valorisation des collections manuscrites du patrimoine. Ces méthodes qui porteront sur le corpus des correspondances clandestines dans l'Europe du 18ème siècle vont conduire au développement de solutions - aujourd'hui inexistantes ou partielles car circonscrites à des collections restreintes de petites tailles - *d'aide* à *la navigation* dans une collection manuscrite, d'*indexation* (des textes et des formes individuelles - graphies) et d'assistance à la lecture par une contribution à la reconnaissance des écritures. Plus généralement cette thèse vise à mettre en place des solutions pour la *caractérisation des contenus* et leur *reconnaissance*. Ce projet pluridisciplinaire possède une composante fondamentale en Histoire de la Pensée classique, il s'intéresse au rôle de la communication manuscrite – lettres et manuscrits savants et philosophiques – dans le développement de la République des Lettres et dans la formation de l'esprit philosophique entre 1685 et 1789. Les corpus électroniques sur lesquels se fonderont les travaux de thèse permettront des recherches originales et fécondes sur le plan de l'analyse historique et philosophique, sur le plan de l'instrumentation électronique et sur le plan du traitement des images des manuscrits. Ce projet se veut généraliste dans le sens où les objets sur lesquels porte l'étude ne peuvent pas être modélisés par des représentations standards du fait de la présence de contenus fortement hétérogènes et composites. Il nécessite donc la mise au point de méthodes flexibles et adaptatives (qui s'adaptent aux particularités de contenus notamment à la grande variabilité des écritures), robustes (peu sensibles au bruit et aux variations de qualité des images) privilégiant une démarche d'analyse et de reconnaissance mixte des écritures (par concurrence d'une modélisation globale des contenus écrits et d'une modélisation par allographes et dictionnaire de formes).	
Membres du comité de thèse :	

7. Présentation

- Contexte

Indiquer le contexte général, préciser le lien avec le cluster Culture, patrimoine, création, expliciter la contribution de la thèse au projet

Depuis une dizaine d'années, l'irruption de l'outil numérique a profondément bouleversé la pratique de la recherche dans les humanités. Dans ce contexte, la demande d'allocation fait suite au projet ANR CITERE (Circulations, territoires et réseaux en Europe de l'âge classique aux Lumières 2008-2011 copiloté par Pierre-Yves Beaurepaire de l'Université de Nice Sophia-Antipolis et Antony McKenna de Université Jean-Monnet Saint-Etienne) ayant pour objectif l'inventaire critique et l'édition en ligne commentée d'un ensemble patrimonial cohérent, d'importance scientifique et culturelle reconnue. Il s'agit du *corpus des correspondances huguenotes et des manuscrits philosophiques clandestins pour la période 1685-1789*.

Ces travaux d'édition actuellement en cours de publication sur support électronique doivent être accompagnés du développement d'une nouvelle instrumentation informatique de recherche et d'édition en ligne annotée et enrichie permettant notamment une recherche avancée de mot en mode image lorsque la transcription textuelle des sources n'est pas disponible ou impossible à produire. L'outil numérique en permettant une plus grande accessibilité des documents, donne ainsi des capacités de diffusion nouvelles qui s'accompagne en même temps d'une création de contenus importante pour la discipline.

La constitution du corpus électronique des correspondances et des textes manuscrits a fait suite à une campagne de numérisation actuellement achevée. Elle avait initialement pour objectif direct de faciliter l'accès à l'information en proposant des outils de navigation simplifiée et en contribuant à une expertise scientifique de haut niveau alliant les spécialistes historiens et littéraires autour des ouvrages de ces documents. Ces nouveaux services adaptés à une exploitation à distance des collections et ces nouveaux usages nécessitent de mettre au point des outils réellement nouveaux. Jusqu'à aujourd'hui, le besoin majeur des bibliothèques était de préserver les collections en les archivant sans se soucier réellement ni de l'exploitation qui pouvait en être faite par la suite (transmission à distance, recherche d'informations dans une collection complète, aide à l'expertise scientifique du chercheur…), ni de la mise en valeur qui pouvait

redonnée vie à certains corpus entiers peu accessibles ou oubliés. C'est donc un environnement de travail par nature pluridisciplinaire qui va accompagner ces futurs travaux de thèse. Pour cela, nous envisageons dans ces travaux plusieurs niveaux de contributions issues du traitement et de l'analyse des images. Nous allons précisément nous intéresser au repérage et à la description des formes présentes dans les textes, à la diversité des écritures, leur identification, leur style, la nature des supports et des mises en page. Il est donc tout naturel que la mise en place des solutions d'aide à l'expertise et à la navigation dans l'ouvrage nécessite de fortes collaborations interdisciplinaires. Les services rendus aux usagers seront au cœur de la problématique et il va de soi que les méthodes d'accès au contenu seront conçues pour faciliter les échanges d'informations, et simplifier la navigation au sein du corpus. Les aspects de navigation personnalisée et thématique seront privilégiés.

Nous avons choisi d'orienter les travaux de cette thèse autour de la mise en place de méthodes originales et complètes de valorisation des collections du patrimoine et d'accès aux contenus. Ces méthodes vont conduire au développement de solutions - aujourd'hui inexistantes ou partielles car circonscrites à des collections restreintes de petites tailles - *d'aide* à *la navigation* dans une collection, à l'*indexation* (des textes et des formes individuelles - graphies) et plus généralement de solutions pour la *caractérisation* des contenus.

Profondément interdisciplinaire, cette thèse constituera une nouvelle expérience d'interactions et de collaborations au sein du Cluster Régional 13 entre des domaines encore trop souvent cloisonnés : les sciences humaines, d'une part, et l'informatique, d'autre part. La pluridisciplinarité constitue un enjeu en soi, une difficulté réelle comme peuvent en témoigner des doctorants dont la thèse a été soutenue ou le sera prochainement au sein du cluster. Aussi, bien plus qu'une aide aux pratiques dans les sciences humaines ou qu'une évolution dans les techniques expertes d'analyse des contenus, l'instrumentation informatisée de l'information qui est envisagée dans cette thèse s'inscrit dans la volonté permettre l'émergence de nouvelles connaissances, favorisant une meilleure accessibilité et disponibilité des manuscrits traités, des recherches inédites, et des capacités de diffusion nouvelles .

Cette thèse accompagnera la mise en ligne de deux bases de données ("Pierre Bayle et la vie intellectuelle des Refuges huguenots" et la base de la "Philosophie clandestine de l'âge classique, 16e-18e siècles") sur le serveur de l'Université Jean Monnet de Saint-Etienne (Institut Claude Longeon, UMR 5037). L'efficacité des approches proposées au cours du travail de thèse pourra être illustrée à partir des documents affichés en ligne, accessibles à tout public.

Il faut donc noter le caractère très novateur du projet dans un contexte national d'investissement d'avenir autour de la valorisation des potentiels de recherche pluridisciplinaire, notamment à travers les différents appels Equipex et Labex de l'ANR.

La réussite de la thèse dépendra de l'ouverture d'esprit dont le doctorant saura faire preuve et des connaissances issues de ces deux domaines de recherche qu'il aura pu mettre en relation en tenant compte de leurs spécificités et des contraintes inhérentes à chacun d'eux.

- Objectifs

Les principaux verrous scientifiques que la thèse devra lever sont liés à des problématiques de traitement des images qui peuvent être résumés dans les points suivants. Soulignons également que la prise en charge des grandes masses de données relatives à cette étude donnera lieu à des réflexions et des développements spécifiques intégrant les dimensionnements conséquents. Les illustrations figurant dans ce document présentent des résultats partiels obtenus sur la base de travaux antérieurs. Les réflexions méthodologiques pourront se fonder en partie sur cet existant

- La visualisation de parties abîmées ou biffées non transcrites (et non transcriptibles) permettant de resituer une partie de la collection des manuscrits dans son contexte originel. La mise au point d'un tel outil requiert une analyse informatique de la structure et des composants de mise en page des textes ce qui constitue un domaine à part entière en analyse d'images des documents.

Les informations de structure et de mises en pages sont considérées comme des métadonnées sources de connaissances supplémentaires des manuscrits. Ces métadonnées pourront être stockées dans des bases de données et rendues accessibles depuis l'interface graphique de visualisation des images. Cela permettra de disposer d'un instrument de travail pour le chercheur et de mettre à long terme à disposition des informations de description très complètes liées à un manuscrit.

- La classification des pages (selon leur mise en forme), des écritures et des styles permettant de retrouver dans la base d'images un ensemble de pages de mise en forme similaire, ou d'écriture visuellement proche sans connaissance a priori sur l'identité du scripteur.

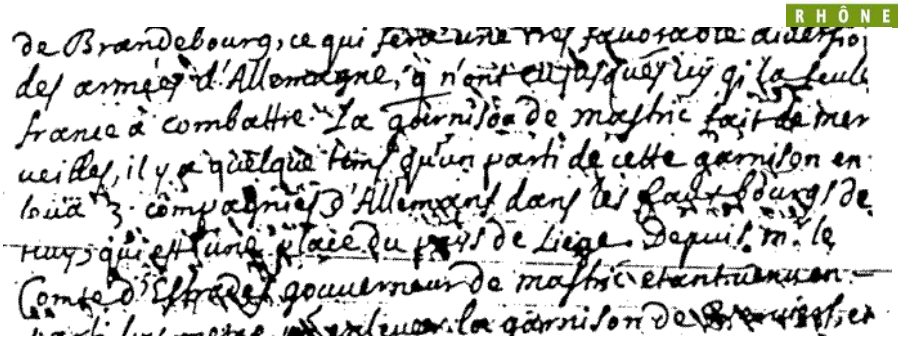
Les différentes approches de caractérisation des formes qui pourront être exploitées permettront de produire un plus grand nombre de descripteurs permettant d'alimenter la classification et qui seront soumis à une validation auprès des experts littéraires seuls juges de leur pertinence. On ne cherchera pas à reproduire le travail des historiens, mais à utiliser une méthode appropriée à l'analyse automatique d'images. Sur le plan méthodologique, nous privilégierons les approches mixtes liées à la définition de descripteurs de formes locaux (reposant sur les variations de graphies des mots et des lettres) et globaux, à la manière de ce qui peut être proposé sur les images naturelles contenant des structures linéaires, des fréquences spécifiques propres aux images de traits. Il sera ainsi nécessaire de définir des critères minimaux pour une bonne mesure des styles individuels des écritures, des mesures robustes aux variations de la qualité de l'image, à la résolution spatiale et les variations de couleur, invariantes à certaines transformations (rotations, translations) et indépendantes aux contenus des textes (et leur longueur).

- L'identification de scripteurs basée sur une sélection de descripteurs discriminants et d'un apprentissage des mains connues dont la période d'intervention est datée et parfois localisée. Les écritures sont reconnaissables par les particularités du tracé mais aussi par les caractéristiques orthographiques ou de vocabulaire utilisé. Les travaux proposés ne visent pas à analyser les particularités des écritures et de poursuivre les travaux existant en grand nombre dans ce domaine (voir. Annexe bibliographique) mais visent à porter une attention particulière à la reconnaissance de mots clés les plus courants employés dans les correspondances dont la signification est reconnue d'un point de vue historique. L'association des deux informations : « forme et contenu » devra être prise en compte dans les mécanismes d'identification.

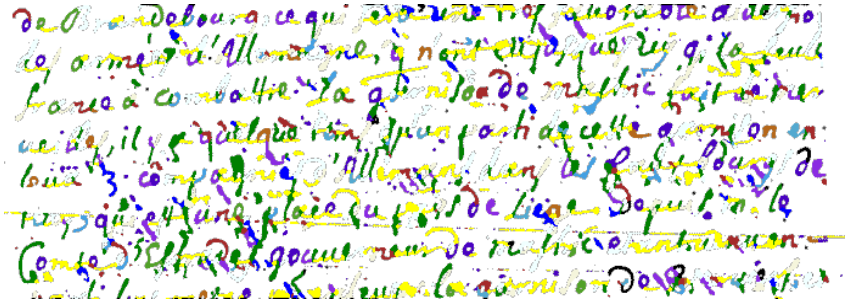
On pourra s'intéresser ici aux formes locales produites à partir d'une analyse grapho-morphologique du tracé qui requerra des opérations morphologiques simples (respectant la topologie des formes : squelettisations et repérage de l'axe médian) pour une décomposition en graphèmes et allographes cohérents avec les caractères utilisés et permettant de construire des dictionnaires de formes, de lettres , de bi et tri grams mais également de termes et de mots clés. Deux grandes approches de classification supervisée et non supervisée seront développées pour produire des classements d'écritures et identifier/authentifier les scripteurs. Un grand nombre de correspondances sont à ce jour anonyme et des recoupements par authentification d'écriture doivent permettre d'aider la datation et la compréhension des processus de communication à cette époque.

- La recherche d'occurrence de mots basée sur une interrogation en mode *texte* ou *image* des manuscrits sera menée en parallèle des actions précédentes. Elle pourra être vue comme une application de l'apport de l'information « *mots clés* » dans la mise en place de nouveaux mécanismes d'identification et d'authentification de mains. En mode *image*, on s'intéresse uniquement à l'image du manuscrit. Le mot est recherché à partir de l'image d'une de ses occurrences. Le terme de « Word-spotting » est couramment utilisé pour décrire ce type de recherche par la localisation de formes similaires. En mode *texte*, l'utilisateur peut directement saisir un mot et provoquer une recherche dans l'image non indexée. Ce processus requiert des opérations de localisation de formes par similarités et constitue une extension possible de la recherche en mode image. Le terme de « Word-retrieval » est employé pour décrire ce type de recherche.

Cette contribution peut ainsi être vue comme une extension de la partie précédente : on s'intéresse cette fois à un échantillonnage de l'écriture à une échelle plus grande. La recherche d'une forme, d'une lettre ou d'un mot (le terme « spotting » est couramment utilisé) pourrait constituer une généralisation de la décomposition morphologique précédente. Cet outil pourra servir à localiser des parties illisibles ou/et à contribuer à produire une transcription partielle assistée.



Extrait original



Décomposition en graphèmes et affichage des graphèmes similaires

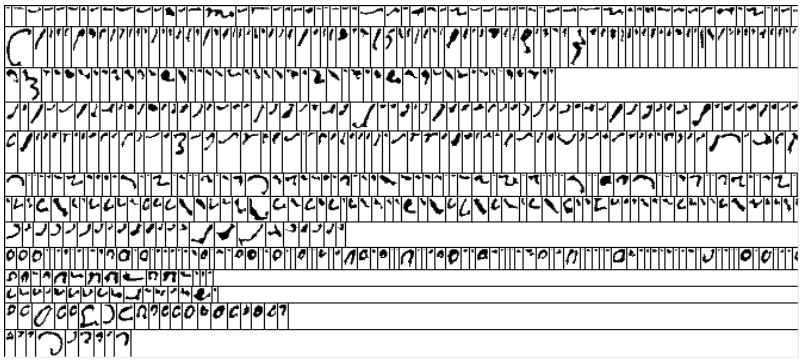


Table de similarités : fréquences d'apparition des graphèmes récurrents

Exemple de décomposition de l'écriture en graphèmes et constitution d'une table de similarités illustrant les fragments d'écritures similaires (extrait des manuscrits clandestins)

- La mise en correspondance des manuscrits et de leur version textuelle établie par les experts permettant de visualiser directement les régions de l'image correspondant au texte de l'édition critique. Pourront être considérées des applications de navigations avancées liées à des stratégies de recherche multicritère (thèmes, mots-clés…).

Le projet s'appuiera notamment sur des techniques de représentation (balisage) des correspondances permettant de créer des interfaces faisant le lien entre les textes et les images, de surimposer les transcriptions aux images et de créer virtuellement un « feuilletage » illustré des manuscrits. Le lien entre images et textes apparaît dans ce projet comme indispensable. Ce travail sera exploitable sur les textes effectivement transcrits et balisés (ces textes seront disponibles suite au projet ANR CITERE 2008-2011 qui précède cette demande d'allocation).

Ces différents objectifs sont très complémentaires. Des travaux en ce sens ont déjà été entrepris au sein de la communauté scientifique depuis plusieurs années. Nous proposons ici de reformuler les mécanismes de recherche et de reconnaissance d'informations manuscrites à partir de la *reconnaissance partielle de mots* (n-grams, mots clés, racines…) qui constitue un support à la transcription automatique des textes manuscrits. Des approches procédant par apprentissage à partir de la génération de dictionnaires de bi-grams, tri-grams ou n-grams formant des caractères ou des mots sera envisagée. Cette contribution est tout à fait novatrice et

inédite dans un contexte aussi ouvert que celui des manuscrits clandestins. Il sera mis au service de la reconnaissance de scripteurs (repérage de vocabulaire fréquemment employé), de la recherche d'occurrence de formes (word-spotting) et de l'aide à la transcription (en alternative au word retrieval).

- Méthodologie

Préciser en particulier les outils, plateaux, observatoires, … qui seront utilisés

Des études récentes portant sur l'analyse des écritures des copistes, dans le cadre de travaux effectués au sein du LIRIS, ont donné lieu à des essais de mise au point d'outils d'analyse de l'écriture et ont déjà fourni des premiers résultats de catégorisation des écritures prometteurs (projet ANR GRAPHEM 2007-2010). Le travail engagé pourra s'appuyer sur cette première contribution qui portait sur des manuscrits très réguliers du Moyen-âge. Il nécessitera des aménagements très nombreux du fait de la grande difficulté à segmenter les textes contemporains selon des règles de décomposition qu'il faudra constituer avec soin. Les recueils manuscrits de Montesquieu, de Flaubert et les correspondances manuscrites européennes à l'époque moderne ont cela de commun, c'est qu'elles présentent des particularités structurelles proches à la fois de mise en page et de graphies dans les écritures: présence de ratures et d'annotations multiples, absence de structure stable et de mise en page régulière, présence d'écritures cursives irrégulières... Il est donc envisager de produire une approche généraliste de la décomposition de l'écriture à échelle variable (graphème, n-grams, mots) rendant compte de la grande diversité des mains présentes (intervenues entre 1685 et 1789) dans le corpus de l'étude. L'étude des singularités des écritures (singularités morphologiques et singularités de termes utilisés) pourra être vue comme une possibilité supplémentaire d'identification des écritures, de datation des feuillets et constituera une aide à la genèse des correspondances.

On étudierait durant la thèse un système pour faciliter les recherches par mots-clefs ou par chaîne de termes (dans la version *textuelle* et surtout dans la version images des correspondances). Il va donc être nécessaire de développer essentiellement trois axes de recherche durant la thèse :

1/ Classification des styles d'écritures et identification des mains

Une grande part du travail de la thèse sera consacrée à l'organisation du réseau de correspondances sur la base de similarités estimées entre les manuscrits anonymes et les documents correctement identifiés comme autographes. La vérification du caractère authentique d'une main à partir de l'analyse automatique de l'image d'un fragment d'écriture constitue un point fort de la partie de traitements informatiques automatisés.

L'enjeu de la thèse est ici de proposer une nouvelle approche de la caractérisation des formes écrites faisant référence aux formes typiques (propriétés grapho-morphologiques des écritures), à leur fréquence d'apparition sur une page et à la dynamique du tracé des écritures. Une extraction d'indices pertinents relevant d'approches perceptives traduisant ce qui est directement accessible à l'œil humain sur ces zones des saillances devra être réalisée sur le corpus de l'étude. Le doctorant s'intéressera également à toutes les autres informations *visibles* qui traduisent une intention précise de l'auteur à travers ses efforts de mises en page (et de structuration des contenus), de mise en relief de certaines données (agencement de données hétérogènes textuelles et graphiques). Dans le cas précis des images de manuscrits clandestins, ces structures peuvent contenir des informations sur le scripteur et sur le fait même que plusieurs scripteurs soient impliqués dans la composition d'une page ou d'un ouvrage.

Pour cette période moderne, il arrivait fréquemment que les philosophes fassent recopier une lettre par un secrétaire et qu'alors, seule la signature soit réellement d'eux. Ce sont d'eux que nous partirons. L'idée ici est d'offrir une aide à l'expertise humaine basée sur une démarche informatisée d'identification automatique des scripteurs qui constitue une aide précieuse de classification des écritures. Il sera également demandé de porter une attention particulière à la qualité des images, la densité suffisante des traits d'écritures dans les régions analysées mais également la relative *lisibilité* des formes écrites qui constituent des conditions nécessaires à une classification robuste. Les outils de classification des écritures des correspondances que le doctorant devra produire se baseront essentiellement sur les informations complémentaires de contours - axe médian, de courbes et de segments linéaires orientés. Le doctorant pourra s'inspirer des travaux déjà réalisés dans le cadre de l'analyse des écritures médiévales régulières (projet ANR Graphem 2007-2010) : approches de la caractérisation des écritures par matrices de cooccurrences, analyse des courbures et des orientations globales par analyse en Curvelets (spécialisation des ondelettes orientées) et décomposition en graphèmes.

L'extraction d'invariants physiques extraits du graphisme de l'écriture constitue le centre de certaines

études récentes. Dans ce type d'approche, une étude de la segmentation en mots, en lettres mais également en graphèmes (fragments de taille plus petite) peut être envisagée en fonction de la nature et de la qualité de l'écriture. L'hypothèse repose sur l'idée que ces invariants représentent les caractéristiques qui sont propres à chaque scripteur. Cette technique est voisine de celle utilisée par les experts en graphologie. Dans ce contexte, les axes de recherche sont donc doubles. Nous avons d'une part un problème de calcul de ressemblance entre les écritures qui permet de rapprocher des écritures ayant des caractéristiques communes ; nous avons d'autre part un problème d'authentification qui permet de décider si un scripteur connu est l'auteur ou non d'un manuscrit. Des techniques d'apprentissage et de reconnaissance, supervisées et non supervisées seront donc envisagés dans ce travail.

On peut noter que les invariants d'écriture (formes fréquemment rencontrées) sont souvent à géométries variables (boucles, barres verticales, etc.) et vérifient des règles de placements topologiques multiples (points convexes, points de croisements, etc.). Certains auteurs ont déjà cerné l'individualité de l'écriture en s'appuyant sur des extractions de caractéristiques telles que la pente ou le nombre de connexions dans le contour (*Rath & Manmatha 2003*). Par conséquent, dans ces travaux de thèse, il sera nécessaire de produire une caractérisation qui mettra en évidence les invariances fréquentes d'une écriture ainsi que ses particularités plus rares qui garantissent bien souvent l'authentification finale.

D'un point de vue méthodologique, il faudra également veiller à lever une difficulté particulière au sein des manuscrits philosophiques et des correspondances clandestines et qui concerne la nature de l'outil d'écriture utilisé qui influence l'épaisseur du trait mais parfois même la dynamique global du tracé. Cette information devra être utilisée comme un paramètre de l'analyse.

2/ Recherche d'occurrences de mots et repérage d'illisibles

Si actuellement on est capable de citer plusieurs travaux phares internationaux en matière de classification des écritures et d'identification des scripteurs, il n'en est pas de même pour les outils d'accès au texte dans le cas des images de manuscrits non contraints et de textes brouillons. Il s'agit ici de relever la problématique générale de la recherche d'occurrences de mots à partir d'une requête image (word-spotting) et/ou d'une requête textuelle (word retrieval). A l'intérieur d'un même manuscrit l'écriture conserve une certaine régularité puisqu'elle est réalisée généralement par le même copiste et utilise la même calligraphie. Grâce à cette relative régularité, il a été montré qu'il était possible de procéder à une recherche de mots. Ce type de recherche fondée sur l'extraction d'invariants de l'écriture devra notamment permettre de repérer des fragments manuscrits illisibles et de faciliter par appariement leur traduction. Une grande partie du travail du doctorant devra porter sur l'élaboration de mesures de similarité fiable portant sur une représentation ou un codage des formes à apparier. Du fait de l'irrégularité attendue dans l'écriture d'un même mot, il sera envisagé des appariements flexibles et élastiques sur la base de modèles de représentation des mots suffisamment souples (points d'intérêt, graphes structurels non planaires…)

3/ Reconnaissance de mots pour une transcription assistée

La transcription automatique des manuscrits brouillons demeure encore à ce jour impossible. La diversité des écritures et des scripteurs rend irréaliste le développement d'une méthode générique de recherche de mots pour tous les manuscrits. Il n'en demeure pas moins que l'aide à la transcription peut être rendue possible par l'exploitation des descriptions produites pour la recherche de mots-images. Malgré le grand nombre des articles au sujet des documents historiques, peu des travaux focalisent sur ces aspects de transcription des documents manuscrits. En effet, la catégorisation et l'indexation de ces documents ont pris plus d'attention surtout avec le progrès des méthodologies de Word Spotting et Word Retrival.

Cette partie de la thèse est exploratoire et pourra conduire à l'élaboration d'un modèle de reconnaissance de fragments de mots ou de caractères conçus à partir des dictionnaires des formes produits lors des étapes 1/ et 2/. Le doctorant pourra notamment s'intéresser aux modèles de Markov qui soutiennent de nombreux travaux de reconnaissance de l'écrit et aux travaux associés à l'exploitation de ce modèle de reconnaissance dans les situations tout à fait inédites qui seront rencontrées lors de l'analyse du corpus des correspondances.

8. Pertinence du projet au regard des priorités régionales

Ce sujet de thèse est en parfaite conformité avec les objectifs du Labex H^2N "Humanités et humanités numériques". On peut notamment souligner qu'au sein de ce laboratoire d'excellence, mais également au sein des projets émergents pluridisciplinaires engagées depuis plusieurs année au sein du Cluster Région « Culture, Patrimoine et Création » il est envisagé de donner à l'outil numérique une nouvelle place dans les Humanités, en considérant sa capacité à conduire à une plus grande accessibilité aux documents, en augmentant les possibilités de diffusion nouvelles, pour finalement conduire à une meilleure exploitation des contenus. Bien plus qu'un perfectionnement des pratiques dans les sciences humaines ou d'une évolution dans les techniques expertes d'analyse des contenus, c'est un contexte de transformation de l'activité en sciences humaines qui est attendu par ces nouvelles collaborations qui se sont concrétisées cette année par la constitution d'un laboratoire d'excellence. Le projet de thèse proposé dans cette demande s'inscrit dans ce contexte riche et ambitieux d'instrumentation informatisée de l'information.

9. Partenaires éventuels (recherche et opérationnel) :

Préciser quels sont les utilisateurs des résultats de la recherche, au niveau de la Région, et à d'autres niveaux éventuellement ; préciser le degré d'implication de partenaires opérationnels dans la définition et le financement du travail.

Le Labex constituera un contexte favorable à cette recherche, à la concertation avec des équipes engagées dans des recherches parallèles et à l'application de l'instrumentation informatique du corpus de l'étude à d'autres corpus numérisés (telle que l'édition électronique des Pensées de Pascal, éd. D. Descotes, L. Thirouin, au sein de l'IHPC, UMR 5037).
Au-delà du cadre régional, ce travail devra permettre de lever les verrous essentiels liés à l'exploitation informatique des images des grandes collections de manuscrits. En visant le développement d'un « outil technique » novateur dont la fonction sera de lever les obstacles matériels et intellectuels qui s'opposent à une véritable valorisation des grands corpus de manuscrits modernes, ce travail pourra trouver des débouchés dans la valorisation des corpus écrits et fonds d'archives. Ceux-ci, faute d'un outil adéquat et suffisamment universel, restent encore, pour la plupart, inexplorés et à l'état de documents illisibles dans les grandes bibliothèques européennes.

Quelques éléments de bibliographie

- *La Correspondance de Pierre Bayle*, édition critique établie sous la direction d' †Elisabeth Labrousse et d'Antony McKenna, en collaboration avec Edward James, Hubert Bost, Wiep van Bunge, Oxford, The Voltaire Foundation, 8 volumes parus (2010).
- *L'Affaire Bayle. La bataille entre Pierre Bayle et Pierre Jurieu devant le consistoire de l'Eglise wallonne de Rotterdam*, Saint-Etienne, Institut Claude Longeon, 2006, texte établi et annoté par H. Bost, Introduction d'Antony McKenna.
- *Les Réseaux de correspondance en Europe (XVIe-XIXe siècle): matérialité et représentation*, dir. Pierre-Yves Beaurepaire, Jens Häseler et Antony McKenna, Saint-Etienne, Presses de l'Université, 2006.
- *La Lettre clandestine, Bulletin d'information sur la littérature philosophique clandestine à l'âge classique*, 1992- : 18 numéros parus (2010) (dir. Antony McKenna).
- R. Manmatha, C. Han et E.M. Riseman, "Word spotting: a new approach to indexing handwriting", dans International Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, USA, 1996, pp. 631–637.
- Marti U.-V. (1) ; Bunke H. (1) ; Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition systems, *Pattern Recognition and Artificial Intelligence,* 2001*, 15*, pp 65-90.
- Bensefia, T. Paquet, and L. Heutte. Handwriting analysis for writer verification. In Frontiers in Handwriting Recognition, 2004. IWFHR-9 2004. Ninth International Workshop on, pages 196–201, 26-29 Oct. 2004.

- Fabio Carrera, “Making History: an Emergent System for the Systematic Accrual of Transcriptions of Historic Manuscripts”, *Eighth International Conference on Document Analysis and Recognition,* 2005, pp 543-449.
- M. Bulacu and L. Schomaker. Combining multiple features for text-independent writer identification and verification. In Proc. of 10th International Workshop on Frontiers in Handwriting Recognition (IWFHR 2006), pages 281–286, 23-26? Oct 2006.
- Writer Identification using Steered Hermite Features and SVM. A. I. Wagan, S. Bres, V. Eglin, H. Emptoz, R.M. Carlos Joel. The 9th International Conference on Document Analysis and Recognition (ICDAR), Brazil, September 23-26, 2007
- Yann Leydier, Frank Lebourgeois, Hubert Emptoz,.”Text search for medieval manuscript mages”. *Pattern Recognition,* 2007, *40*, pp 3552-3567
- Veronica Romero, Alejandro H. Toselli et Enrique Vidal, “Using mouse feedback in computer assisted transcription of handwritten text images”, *10th International Conference on Document Analysis and Recognition,* 2009, pp 96-100.
- Yann Leydier, Asma Oujia, Frank LeBourgeois et Hubert Emptoz, “Towards anomnilingual word retrieval system for ancient manuscripts” *Pattern Recognition,* 2009, *42*, pp 2089-2105
- Vincent Malleron, Véronique Eglin, Stéphanie Dord-Crouslé, Hubert Emptoz et Philippe Régnier, “Un système de mise en relation Image/Transcription pour les documents manuscrits”, *Colloque International Francophone sur l'Écrit et le Document,* 2010.
- Anne-Laure Bianne, Christopher Kermorvant, Laurence Likforman-Sulem, Modélisation de HMMs en contexte avec des arbres de décision pour la reconnaissance de mots manuscrits, CIFED 2010.
- Ancient handwritings decomposition into graphemes and codebook generation based on Graph coloring. H. Daher, DJ. Gaceb, V. Eglin, S. Bres, N. Vincent. Dans International Workshops on Frontiers in Handwriting Recognition (ICFHR), IAPR ed. Kolkata, India. pp. 6-12. 2010.