A N N E X E S

ANNEXE N° 1

PRESENTATION DU RESEAU THEMATIQUE PLURIDISCIPLINAIRE AUTOUR DU DOCUMENT NUMERIQUE (RTP-DOC) DES DEPARTEMENTS STIC ET SHS DU CNRS

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Département Sciences et Technologies de l'Information et de la Communication

RTP 33 : Documents et contenu : création, indexation, navigation

Objectifs thématiques et Finalités

Le traitement numérique et la mise en réseau transforment en profondeur les relations aux documents pour les individus, les communautés et les sociétés, avec notamment comme enjeux l'accès partagé à l'information, la gestion des connaissances pour les organisations, une valeur ajoutée pour les services, la transformation des industries culturelles. Il s'agit donc de :

Analyser les documents, les médiations et leur relation avec l'activité humaine et ses limites.

S'interroger sur les outils à construire, sur les modalités de leur construction et sur les conséquences de leur utilisation.

Croiser l'informatique, la linguistique, les sciences cognitives, et les « sciences de l'information ».

Discuter la notion de « document », renouvelée sinon contestée par le développement numérique.

Construire un programme de recherche ambitieux qui fédère des laboratoires français, européens et étrangers.

Activités développées

L'effort du réseau porte sur l'établissement de relations, notamment entre **STIC** et **SHS**, sans contrarier les approfondissements en cours. A cette fin, les activités se développent dans deux directions :

- L'exploration du domaine par le lancement d'ateliers sur des problématiques particulières où l'accent est mis sur les croisements disciplinaires (cf. ci-dessous).
- Des rencontres et synthèses transversales autour d'occasions et d'évènements forts, relayés ou suscités (cf. ci-contre).

Ateliers en cours (2004-2005)

A1 - Numérisation *(JM Ogier, L3i, La Rochelle)*
A2 - Convergence multimédia *(N. W. Lund, Univ-Tromso, Norv.)*
A3 - Annotation, coopération *(M Zacklad, TechCICO, Tours)*
A4 - Points de vue et confrontation *(A. Iacovella, ENS-LSH, Lyon)*
A5 - Lectures du numérique *(C. Bélisle, LIRE CNRS Lyon2)*
A6 - Mesures de l'internet *(Eric Guichard, ENS)*
A7 - Auteur et autoritativité *(E. Broudoux, Paragraphe, Paris 8)*
A8 - Théorie pour le document *(JM. Salaün, ENSSIB, Lyon)*
A9 - Interaction, bib. num. *(P. Cubaud, CEDRIC, CNAM, Paris)*
A10 - PLEXIR : Ling. et corpus *(M. Boughanem, IRIT, Toulouse)*

Actions réalisées (2002-2004)

AS32 - Web sémantique (J. Charlet, P. Laublet, Ch. Reynaud)
AS34 - Structuration de ressources terminologiques à partir de textes (Nathalie Aussenac, Anne Condamines)
AS95 - Les temps du document numérique (Sylvie Calabretto, Geneviève Lallich-Boidin, Florence Sèdes)
AS96 - Numérisation et Valorisation des Collections (Abdel Belaïd, Hubert Emptoz, Georges Vignaux)
AS103 - Modèle(s) de publication sur le Web (Ghislaine Chartron, Franck Rebillard)
AS137 – Document et organisation (Jacques Labiche, Maryvonne Holzem)
EP – Du partage de corpus documentaires structurés à la confrontation de points de vue (Andréa Iacovella)

Organismes partenaires

Caractéristiques

Date de lancement : Mai 2002

Responsable

Jean-Michel Salaün : salaun@enssib.fr

Site Web

http://rtp-doc.enssib.fr

Mots clés

Document multimédia, document structuré, hypermédia, reconnaissance de forme, numérisation, métadonnée, terminologie, ontologie, bibliothèque virtuelle, édition numérique, web sémantique, gestion des connaissances, recherche d'informations, gestion des contenus, industrie des contenus, navigation sémantique et sociale, lecture sur écran.

Comité de pilotage

Membres : Ch. **Licoppe** (STIC), B.**Victorri** (SHS), N. **Aussenac** (IRIT), B. **Bachimont** (INA), A. **Belaid** (LORIA), JB. **Berthelin** (LIMSI), D. **Boullier** (UTC), Ch. **Brouard** (CLIPS), A. **Condamines** (ERSS), S. **Chaudiron** (Min. Rech.), J. **Charlet** (AP-HP), G. **Chartron** (INRP), C. **Chrisment**, (IRIT), Ch. **Dessaux** (Min. Culture), J. **Ducloy** (INIST), H. **Emptoz** (LIRIS), C. **Fluhr** (CEA), M. **Gaio** (LIUPPA), P. **Gros** (IRISA), A. **Iacovella** (ENS-LSH), P. **Laublet** (LaLICC), J. **Madelaine** (GREYC), **JP. Metzger** (URSIDOC), R. **Mullot** (L3I), V. **Quint** (INRIA), JM. **Salaün** (URSIDOC), E. **Trupin** (PSI), G. **Vignaux** (LCP).

Laboratoires impliqués

- **CEA-DIST** / Saclay
- **CLIPS-IMAG** / UMR5524 / Grenoble
- **COSTECH** / EA2223 / Compiègne
- **CRIS** / EA1738 / Paris
- **DYALANG** / UMR6065 / Rouen
- **DDS /** Tromso, Norvège
- **ERSS** / UMR5610 / Toulouse
- **GREYC** / UMR6072 / Caen
- **IRISA** / UMR6074 / Rennes
- **IRIT** / UMR5505 / Toulouse
- **ISTIT** / FRE 2732 / Troyes
- **LaLICC** / UMR8139 / Paris
- **LCP** / UPR36 / Paris
- **LIRE** / UMR 5611 Lyon
- **LORIA** / UMR7503 / Nancy
- **LIMSI** / UPR3251 / Paris
- **LIRIS** / FRE2508 / Lyon
- **LIUPPA** / EA3000 / Pau
- **L3I** / EA2118 / La Rochelle
- **MEI** / EA1858 / Lyon
- **Porphyry** / ENS-LSH / Lyon
- **Paragraphe** / Paris 8
- **PSI** / FRE2645 / Rouen
- **RST** / ENS Paris
- **Tech-CICO** /FRE2732 / Troyes
- **URSIDOC /** EA3718 / Lyon

Faits marquants

Roger T. Pédauque : Un travail de fond sur la notion de document est mené sous cette signature collective.

Conférences Web : Organisées sur l'année 2005, elles permettent à chaque atelier (voir ci-contre) d'exposer ses travaux en direct sur Internet. Elles sont ensuite enregistrées et mises en ligne.

Plateformes : Recherche d'information (http://www.irit.fr/**PLEXIR**). Numérisation du patrimoine européen (http://l3iexp.univ-lr.fr/**madonne**)

Semaine du document numérique : l'édition 2004 de La Rochelle a réuni les chercheurs du document numérique autour de manifestations internationales (CIFED, CIDE, CIFT). Une session est envisagée pour 2006. http://sdn2004.univ-lr.fr.

ANNEXE N° 1 bis

LES PRINCIPALES COMPETENCES REPEREES EN FRANCE AUTOUR DU DOCUMENT NUMERIQUE (HORS NORMANDIE)

En dehors des compétences scientifiques présentes en **Normandie** et développées dans le rapport (pour la Basse-Normandie) et dans les annexes n° 5 et 6 pour la Haute-Normandie, il a paru opportun d'enquêter sur les équipes de recherche qui travaillent en France autour du document numérique dans les domaines des Sciences des Technologies de l'Information (STIC) et des Sciences Humaines et Sociales (SHS), deux domaines qui travaillent étroitement dans le cadre des thématiques scientifiques en question.

De manière probablement non exhaustive, il convient de citer les principaux établissements et laboratoires qui ont pu nous être signalés à l'occasion de nos investigations ou que nous avons simplement repéré après une recherche sur Internet ou au fil de nos lectures.

En France, en dehors de la Normandie, les compétences autour du document numérique sont principalement concentrées en Ile de France (Paris), en Région Rhône-Alpes (Lyon et Grenoble), en Provence-Alpes-Côte d'Azur, en Lorraine (Nancy), en Bretagne (Rennes), en Aquitaine (Bordeaux), Midi-Pyrénées (Toulouse), Languedoc-Roussillon (Montpellier), Poitou-Charentes (La Rochelle)

On trouvera à cet égard en annexe n°1 la liste des équipes de recherche ayant participé aux travaux du réseau thématique pluridisciplinaire RTP-DOC mis en œuvre en 2002 sur l'initiative conjointe des départements STIC et SHS du CNRS. Il rassemblait des laboratoires de recherche dont les travaux portent sur le traitement numérique du document avec, pour enjeux, l'accès partagé à l'information pour les communautés et les individus, la gestion des connaissances dans les organisations et la notion de "valeur ajoutée" pour les services.

● **Paris et l'Ile de France** se distinguent assez logiquement par une importante concentration de compétences diverses et variées autour du document numérique. Certaines thématiques du pôle de compétitivité francilien Cap Digital consacré aux contenus numériques et aux technologies de l'information sont d'ailleurs, dans l'absolu, concernées par le document et les flux numériques.

Parmi les équipes concernées, il convient de citer l'Université **Paris VIII - Vincennes** à Saint-Denis où il existe depuis 2004 le **groupe de recherches "Document numérique & Usages"**, lui-même partenaire scientifique de l'**Equipe Sémiotique, Communication et Nouvelles Technologies de l'Information** (ESCoM) de la Fondation de la Maison des Sciences de l'Homme de Paris (FMSH) qui, en outre, collabore avec la Maison de la Recherche en Sciences Humaines de l'Université de Caen sur le thème des nouvelles formes d'écriture.

Au sein de l'Université de **Paris V - Descartes**, le Département "Information-Communication" de l'Institut Universitaire de Technologie propose une option **"Métiers du Livre et du Patrimoine"** qui intègre une formation aux nouveaux supports numériques. On y trouve également le Laboratoire Systèmes Intelligents de

Perception du Centre de Recherche Informatique de Paris 5 (CRIP5 - SIP), Pôle "signal, parole, image, réseau", équipe "analyse d'image et imagerie biomédicale" très impliquée dans le traitement d'images de document.

Le **Laboratoire d'informatique de l'Université Paris VI** (LIP6) - UMR CNRS 7606, par l'intermédiaire de son département "Données et Apprentissage Artificiel" entretient une collaboration avec le Pôle de compétitivité francilien Cap Digital dans le domaine du traitement du document numérique plus particulièrement textuel. Citons notamment l'équipe MALIRE (Machine Learning & Information Retrieval) du département précité.

A l'Université de **Paris X - Nanterre**, dans le domaine des sciences de l'information et de la communication, le **CRIS (Centre de Recherche sur l'Information Spécialisée)** est composé de deux équipes : l'une mène des recherches centrées sur la "Culture populaire, médias et politique : évolution des formes, supports et contenus", l'autre sur les "Industries électroniques du savoir" touchant à de nombreux sujets en lien direct avec le document numérique.

A l'Ecole Normale Supérieure de Paris, l'**Institut des Textes et Manuscrits Modernes (ITEM)** est une Unité Mixte de Recherche CNRS (Centre National de Recherche Scientifique) / ENS (Ecole Normale Supérieure) - UMR 8132 qui se consacre à l'étude des manuscrits d'écrivains pour élucider les processus de la genèse des textes.

Le **CEDRIC** (Centre d'Etudes et De Recherche en Informatique du CNAM) regroupe l'ensemble des activités de recherche en informatique menées au **Conservatoire National des Arts et Métiers**. Les membres du CEDRIC font partie du Département STIC du Conservatoire. Au sein de l'équipe Multimédia & Interaction Homme-Machine (MIHM), un axe "Interaction et Bibliothèques Numériques" (IBN) mène des travaux sur les questions de l'interaction avec le document et de l'apport des technologies 3D dans ce domaine. Le CNAM a été à l'origine du site ABU (http://abu.cnam.fr) qui fut, en 1993, le premier site Web de bibliothèque numérique en France. Depuis 2000 s'est développée la bibliothèque numérique du CNAM (http://cnum.cnam.fr), site de référence pour l'histoire des techniques en France.

Au sein de **l'Institut NAtional des Langues et Civilisations Orientales** (INALCO) de Paris, l'Equipe de Recherche en **"Textes, Informatique, Multilinguisme"** (ER-TIM) réunit des linguistes et des informaticiens travaillant sur des problématiques communes : lexicologie computationnelle, "dictionnairique", "traductique", formation aux langues en ligne, filtrage et détection de contenus automatiques, recherche d'information, etc. en contexte multilingue. C'est une Equipe d'Accueil (EA), créée en 2006, dont les axes de recherche principaux sont "Les enjeux du multilinguisme" et "Le document numérique".

Citons également l'**Institut de Recherche et d'Histoire des Textes (IRHT),** unité propre de recherche du CNRS dépendant du département scientifique SHS. Il a pour mission la recherche fondamentale sur le manuscrit médiéval et la transmission des textes de l'Antiquité à la Renaissance. Il est géré administrativement par les Délégations **Paris A** du CNRS et le **Centre-Poitou-Charentes**. Il compte cent membres, chercheurs, ingénieurs, techniciens, répartis sur **cinq sites** (au siège de l'avenue d'Iéna à **Paris**, à **Orléans-La Source**, dans les locaux du **Collège de France**, de **l'Institut quai de Conti**, de la **Sorbonne**) qui abritent quatorze sections et cinq services.

● Une autre forte concentration de compétences scientifiques autour du document numérique au plan national est recensée en **région Rhône-Alpes** et notamment sur Lyon autour de **l'Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques** (ENSSIB) qui a pour mission à la fois de former les cadres des bibliothèques et de l'information (conservateurs, bibliothécaires) et propose ainsi de nombreux masters et de développer la recherche en sciences de l'information, en bibliothéconomie et en histoire du livre. L'ENSSIB diffuse en outre le Bulletin des Bibliothèques de France (BBF) très axé sur les technologies numériques autour du document.

L'**Equipe de recherche de Lyon en sciences de l'information et de la communication** (ELICO) rattachée à l'Université Lyon 2 présente, parmi ses thématiques de recherche, un axe " Documents et Société" qui intègre des travaux autour des sujets comme "Documents et connaissances" ou encore "Socio-économie du document numérique".

Citons également l'**Institut des Sciences du Document Numérique en Rhône-Alpes** (ISDN), "laboratoire hors les murs" qui ne vise pas à se substituer aux équipes de recherche existantes, mais dont l'objectif est de faciliter le travail collaboratif dans le domaine du document numérique. Un site Web[1] hébergé par l'ENSSIB favorise la coopération et l'information mutuelle, tout en servant de vitrine à la recherche sur ce thème en Rhône-Alpes. Une structure permanente légère assure les tâches de gestion et de coordination. Cette structure peut, selon les besoins, agir au nom des équipes concernées pour négocier et gérer les contrats de recherche.

Pour répondre aux recherches interdisciplinaires effectuées dans plusieurs laboratoires de la région Rhône-Alpes, conduisant à l'émergence d'une filière technologique qui se structure depuis quelques années autour du document numérique.

En application de son Schéma Régional de l'Enseignement Supérieur et de la Recherche, la Région Rhône-Alpes a souhaité mettre en place une politique de soutien à la création de 14 "Clusters de recherche"[2]. Le **cluster 13** a pour thème : **"Culture, Patrimoine, Création"**, lui-même décliné en 5 projets :

- "Patrimoine et territoire",
- "Genre et culture",
- "Editions critiques",
- "Corpus numériques",
- "Création".

Les projets sont détaillés sur le site http://cluster13.ens-lsh.fr/. Notons que des travaux sont notamment conduits en lien avec **Grenoble**, autour des manuscrits de Stendhal au titre du projet 3 "Editions critiques".

La constitution du projet "Corpus numériques" s'est appuyée sur les travaux autour du développement des méthodes et des outils permettant de mieux valoriser

1 http://isdn.enssib.fr/institut/institut.html

2 Le Cluster est un programme de recherche regroupant des projets. Il est constitué d'un réseau de laboratoires ou d'équipes travaillant à la réalisation de ce programme scientifique commun, et sa mission est d'animer et de coordonner cette communauté scientifique régionale pendant une durée d'environ 5 ans. Le Cluster est un lieu d'animation de la communauté scientifique (séminaires, journées d'échange…). Il est le garant de la pluridisciplinarité.

et exploiter les objets et contenus patrimoniaux via les thématiques "Bases de données multimédia", "Corpus linguistiques" et "Numérisation et reconnaissance des documents". Dans ce cadre**, le Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS)**[3] né début 2003 à la suite du regroupement de plusieurs laboratoires de recherche lyonnais est fortement impliqué dans cette thématique et propose de mettre en place en 2008 un atelier de huit journées de travail sur les problèmes de numérisation, traitement des images de document, reconnaissance dans ces images et accès au contenu textuel.

Par ailleurs au sein du **Laboratoire d'Informatique de Grenoble**, le Groupe d'Etude en Traduction/Traitement des Langues et de la Parole est organisé autour de 4 thèmes de recherche : "Traduction Automatique", "Reconnaissance automatique de la parole, des locuteurs et des sons", "Ressources lexicales et corpus", "Dialogue, communication et émotions".

Il convient de signaler également l'existence du Réseau d'Information Scientifique Interdisciplinaire des Documentalistes en régions Rhône-Alpes et Auvergne et bibliothécaires **ISIDORA** en partenariat avec le CNRS, les professionnels travaillant dans une bibliothèque ou gérant la documentation d'un (ou plusieurs) laboratoire(s), CNRS ou rattachés au CNRS. Il s'agit d'un réseau pluridisciplinaire qui aborde des problématiques appliquées à la gestion documentaire (formations, mutualisation des ressources…).

● En **Midi-Pyrénées**, l'**Institut de Recherche en Informatique de Toulouse (IRIT)**, se distingue par certaines thématiques de recherches autour du document numérique comme "Analyse et synthèse de l'information" et "Indexation et recherche d'informations". Des travaux pointus sont menés en termes d'indexation du texte, de l'image fixe et animée et du son.

● **En région Poitou-Charentes,** le **Laboratoire Informatique, Image et Interactions** (L3I) de l'Université de **La Rochelle** mène des travaux de recherche sur l'image et le comportement. Ses activités sont réparties autour de 3 projets scientifiques :

- I-Médoc : "**Image, Média et Documents Numériques"** dont les points d'entrée scientifiques portent sur l'imagerie du visible à l'invisible, les séquences d'Images (de la pellicule au flux vidéo) et les systèmes d'informations documentaires (du Patrimoine au document numérique),
- ImagIN : **"Images et Interactivité Numérique"** dont les points d'entrée scientifiques portent sur l'image et le comportement avec contraintes temporelles, l'analyse des comportements implicites ou explicites et le Génie Logiciel pour Systèmes Interactifs,
- Sido : **"Sémantique et Intermédiation de Données"**, qui s'intéresse aux Métiers, aux organisations, ontologies associant par ailleurs les aspects dynamiques.

● A **Nancy**, le **Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA)**, est une Unité Mixte de Recherche - UMR 7503 - commune à plusieurs établissements (CNRS, Institut National Polytechnique de Lorraine, Institut National de Recherche en Informatique et en Automatique, Henri Poincaré, Nancy 1

[3] UMR 5205 CNRS/INSA de Lyon/Université Claude Bernard Lyon 1/Université Lumière Lyon 2/Ecole Centrale de Lyon.

et Université Nancy 2). Au sein des compétences scientifiques du laboratoire, des travaux sont conduits autour des grammaires formelles, la déduction automatique, les représentations structurées et formalisation du raisonnement, la fouille de données… Des recherches ont notamment comme applications les réseaux et l'Internet, la sécurité des systèmes informatiques et la réalité virtuelle. Citons de manière non exhaustive deux équipes particulièrement bien impliqués dans le document numérique : l'équipe READ (Reconnaissance de l'écriture Et Analyse de Document) et l'équipe QGAR (Navigation dans les documents graphiques par l'analyse et la reconnaissance-Projet INRIA-LORIA).

Egalement à Nancy, l'unité mixte de recherche **ATILF (Analyse et Traitement Informatique de la Langue Française),** créée en 2001 suite au rapprochement de l'Institut National de la Langue Française (INALF - CNRS) et de LANDISCO (Langue Discours Cognition - Université Nancy 2), voit son excellence reconnue tant au plan national qu'international.

S'appuyant sur l'UMR ATILF de Nancy, le **Centre National de Ressources Textuelles et Lexicales** (CNRTL) a été créé en 2005 par le CNRS, en vue de fédérer, au sein d'un portail unique, un ensemble de ressources linguistiques informatisées et d'outils de traitement de la langue. Il intègre le recensement, la documentation (métadonnées), la normalisation, l'archivage, l'enrichissement et la diffusion des ressources. En d'autres termes, le CNRTL a été mis en place afin de fournir un service à la communauté nationale de recherche pour la création, gestion et diffusion de ressources textuelles et lexicales. Cette structure bénéficie du soutien du CNRS, de l'INRIA et de la Région Lorraine.

● Au sein de l'Université de **Montpellier 3**, le **Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM)** a développé un axe **"Bases de Données et Systèmes d'Information"** qui intègre des travaux autour de l'intégration et la fouille de données.

Le Département de Mathématiques et d'Informatique de l'Université de Montpellier 3 s'est distingué par la mise en œuvre d'**ARCANE**, système informatique développé à l'usage des étudiants et des chercheurs en sciences humaines et sociales, pour produire individuellement ou collectivement d'importantes bases de documents électroniques à forte valeur ajoutée, et les publier sous forme de livres papier et/ou de livres électroniques : éditions savantes et pédagogiques, thèses, bases documentaires personnelles de recherche. C'est aussi un instrument de production générique de documents balisés.

Il convient de signaler les travaux autour des corpus et bases de données textuelles produits avec ARCANE autour de l'édition électronique qui a donné lieu à des collaborations entre des chercheurs de plusieurs équipes universitaires (Montpellier, Saint-Etienne et Lyon notamment) sur des thèmes qui intéressent tout particulièrement les historiens.

● Au sein de l'**Ecole des Mines de Saint-Etienne**, l'une des thématiques de recherche du Département "Réseaux, Information, Multimédia" (RIM) est dénommée "Création, Organisation, Classification et Recherche d'Information" (COCRI). Il s'agit de travaux autour d'une organisation et d'une structuration explicite lors de la mise à disposition des documents sur les serveurs Web, et sur la façon d'utiliser au mieux ces nouvelles données pour améliorer les processus d'appréhension de l'information

par les utilisateurs, que ce soit pour le butinage, la navigation ou la recherche d'information.

● En région **Provence-Alpes-Côte d'Azur**, au sein de l'Ecole des Hautes Etudes en Sciences Sociales de Marseille, l'**équipe "Sociologie, Histoire, Anthropologie des DYnamiques Culturelles (SHADYC)** rattachée au CNRS mène des travaux autour de trois domaines : "Édition électronique. Écriture et lecture électronique : méthodes, modèles d'édition, outils logiciels", "Analyse formelle des contenus textuels, en particulier des ego-documents (correspondances, journaux intimes, journaux de voyage, livres de raison, etc.)" et "la correspondance manuscrite à l'époque moderne : espaces et réseaux relationnels"

Le **Laboratoire des Sciences de l'Information et des Systèmes (SIS) de l'Université du Sud Toulon-Var** conduit des travaux autour des documents XML, l'Interrogation du Web sémantique, la recherche d'information et la modélisation d'images sur le web… Des travaux communs sur le XML sont conduits avec le laboratoire LIUPPA (équipe Sécurité des Systèmes Communicants) de l'**Université de Pau.**

● En sciences humaines et sociales à **l'Université de Bordeaux 3**, l'équipe TELEM - TExtes, Littératures, Ecritures et Modèles mène des travaux de recherches autour de la sémiotique, la sémantique lexicale et la textualisation.

● Plus proche de nous, **en Bretagne,** l'**équipe IMADOC** (IMAges et DOCuments) est une équipe de recherche de **Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA)** de l'Université de Rennes 1 rattachée à l'Institut National de Recherche en Informatique et Automatique (INRIA) qui conduit des travaux autour de l'interprétation et la reconnaissance d'images et de documents.

Les recherches menées au sein du projet IMADOC concernent l'écrit et le document sous toutes leurs formes (manuscrit, imprimé, image, graphique, multimédia, etc.) ainsi que les activités qui y sont liées, notamment la production de nouveaux documents, la transformation sous forme électronique élaborée de documents papier existants et leur traitement "intelligent" ainsi que l'Interaction Homme-Document. De manière plus générale, les centres d'intérêt du projet IMADOC touchent à la communication écrite sous un triple aspect : synthèse de documents, analyse de documents, interaction homme-document.

● Dans la région **Pays de la Loire**, l'équipe "**Image et Video Communications"**" de l'Institut de Recherche en Communication et Cybernétique de Nantes (IRCCyN - UMR CNRS 6597), structure rattachée à l'Ecole Centrale de Nantes, l'Université de Nantes et l'Ecole des Mines de Nantes est compétente en segmentation d'images de documents et en reconnaissance de l'écriture manuscrite.

● Enfin, en région Centre, le **Laboratoire d'informatique de Tours** LI EA 2101 a une équipe "Reconnaissance des formes et analyse d'images" qui s'intéresse aux images de documents qu'ils soient textuels, graphiques ou photographiques.

Il convient enfin de citer parmi les centres de ressources en France le **Centre d'Etudes Supérieures de la Renaissance (CESR)** de l'Université François RABELAIS de **Tours** qui propose à la fois des formations supérieures spécialisées (Masters) et des recherches consacrées à l'étude de la civilisation de la Renaissance.

ANNEXE N° 2

PRESENTATION DES TRAVAUX DE RECHERCHE DES EQUIPES DU GREYC PRINCIPALEMENT CONCERNEES PAR LE DOCUMENT NUMERIQUE

L'équipe ***Données, Document, Langue (DoDoLa)*** se trouve directement au cœur du sujet. Ses activités et compétences concernent la production, l'interrogation, l'extraction et la restitution de l'information contenue dans les données, que celles-ci soient sous une forme structurée ou non (texte, image…). De manière plus précise, le thème *"découverte de connaissances dans les bases de données"* porte sur la découverte d'information pertinente au sein de grands ensembles de données. Des projets ambitieux en la matière concernent la création des futurs moteurs de recherche d'informations dans les "masses de données".

La thématique sur le *document numérique composite* porte un intérêt particulier pour le document géographique, le traitement automatique des langues, l'extraction et la découverte d'information géographique.

Le groupe *Sémantique et Traitement Automatique des Langues* réunit les activités de recherche de l'équipe DoDoLa autour de la composante "langue". Plus spécifiquement, ses études concernent la sémantique, les notions de sens et d'interprétation, avec comme thèmes linguistiques l'expression de l'espace, l'expression de la temporalité, la notion de référence, de figures de style, la structure du discours, thématique ou rhétorique. La langue représente un aspect majeur dans ce contexte et les techniques de sémantique et de Traitement Automatique du Langage (TAL) sont ainsi utilisées pour l'interrogation des données. Inversement, la mise à disposition de grandes masses de données textuelles et de divers types de ressources linguistiques (dictionnaires électroniques, analyseurs…) rend possible un nouveau type d'études linguistiques reposant sur des observations et expérimentations sur corpus à grande échelle.

Les défis relevant de ces problématiques et touchant plus particulièrement le document sont multiples. Citons par exemple la recherche fine d'information (c'est-à-dire extraire d'un ensemble de documents des passages, de l'ordre de quelques phrases, contenant précisément l'information recherchée), la veille d'un flux de documents (par exemple : dépêches d'alertes médicales, appels d'offres) pour en extraire une synthèse à intervalles réguliers, le routage automatique de message électroniques à partir d'une analyse fine des demandes et plus généralement la conception, l'interrogation et la restitution de l'information contenue dans les documents numériques. En ce qui concerne les aspects plus fondamentaux de la recherche, il s'agit de concevoir de nouveaux modèles sur la sémantique de la langue et l'analyse du discours, la conception et la réalisation de cadres génériques et d'outils logiciels pour la découverte d'information dans les données et notamment les textes.

Les activités de l'équipe se nourrissent d'échanges entre recherche et applications. Ces allers-retours signifient non seulement le développement en termes d'usages et d'applications de modèles élaborés en amont, mais aussi que des problèmes concrets et des applications sont sources de travaux et enrichissent une réflexion plus fondamentale. Par exemple, certains résultats théoriques sur la

sémantique de l'espace et l'analyse de textes ont été motivés par des collaborations sur le document géographique et la génomique. Ces activités se traduisent aussi par le développement de plates-formes d'expérimentation afin de faciliter l'utilisation des méthodes et technologies. Citons à cet égard LinguaStream, une plate-forme générique pour le traitement automatique des langues naturelles, fondée sur l'enrichissement incrémental des documents électroniques http://www.linguastream.org (cf. annexe n° 3).

L'équipe attache ainsi une grande importance aux collaborations, que ce soit avec d'autres chercheurs (pouvant provenir de domaines et de communautés de pensée différentes comme l'ingénierie des systèmes, l'interaction homme-machine, la linguistique, le traitement d'image) ou dans le cadre de partenariats industriels (exemples : Orange Labs, Facompo, Asterion) ou avec des organismes publics ou para-publics (exemples : Caisse des Dépôts, Communauté Urbaine de Cherbourg). Ce positionnement conduit aussi l'équipe à être impliquée dans plusieurs projets nationaux (relevant d'appels d'offres du type "ACI" (Action Concertée Incitative) et maintenant de l'ANR (Agence Nationale de la Recherche) ainsi qu'un projet européen inter-disciplinaires (Espace Manche Développement Initiative).

On notera enfin que les chercheurs de l'équipe DoDoLa participent aux travaux du réseau national RTP-DOC présenté dans l'annexe n° 1.

L'équipe ***Interaction sémiotique, langues, diagrammes (ISLanD)*** a développé des compétences dont les objectifs sont d'une part, de mieux connaître les processus d'interaction sémiotique et les objets supports (étude des textes, des dialogues, des diagrammes de conception ou de description) et d'autre part, de produire des objets intermédiaires supports de nouveaux usages (dialogue homme-machine, mondes virtuels, tables de catégorisation, coloriages de documents…).

Syntaxe et Rhétorique est un axe de recherches sur la modélisation de la structure des textes à partir de leur mise en forme matérielle et d'indices de surface. Le discours collectif et le rapport du texte et de l'illustration sont les principaux thèmes abordés. Deux directions de recherche appliquée se sont affirmées : la recherche d'information sur Internet avec ressources minimales d'une part et la fouille de données textuelles d'autre part. Dans les deux cas, l'accent est mis sur la couverture et la robustesse des traitements, pour permettre l'analyse rapide de collections de textes, soit éphémères (presse internationale), soit archivées (bibliothèques numériques).

La thématique *Diagrammes, Information et Communication* étudie les propriétés des schémas qui favorisent le partage, la communication et la production d'information. Il s'agit aussi d'identifier les types de raisonnements qui guident l'interprétation des schémas dans un processus de conception dialogique (auteur / lecteur). Une meilleure connaissance des propriétés sémiotiques des schémas doit permettre de mieux spécifier les rôles et les fonctions des participants dans les activités interactives qui visent à créer et expliciter des connaissances. La traduction de ces propriétés dans la modélisation informatique des interfaces, permettrait d'améliorer l'efficacité de la coopération homme - machine.

Le thème *Modélisation du langage et de l'activité langagière* (dialogue, documents textuels) développe des logiciels d'étude pour la modélisation du langage et de l'activité langagière en observant des propriétés du matériau linguistique et de l'interaction homme-machine pour en déduire des instrumentations légères. Ces

outils sont conçus selon le point de vue de l'utilisateur et de la tâche qu'il envisage. L'étude et la modélisation partent de l'analyse de corpus attestés. Les outils développés partagent un modèle interactionniste du sens et une visée anthropocentrée. Ils utilisent des formats d'échange de données communs et constituent ainsi une plate-forme logicielle homogène d'outils d'étude linguistique.

L'équipe ***Algorithmique*** porte des travaux sur la sécurité et la protection des données informatiques : construction d'algorithmes, modélisation, cryptographie…. Elle est plus spécifiquement concernée par les problématiques de transactions électroniques sécurisées mais peut rejoindre également des préoccupations dans le domaine de l'authentification et la sécurisation des documents.

La réflexion autour du document numérique doit avoir une approche multimodale et en ce sens, il convient d'aborder l'apport de l'image dans la thématique "document numérique". Le document numérique est, rappelons-le, un médium capable de transmettre de l'information. Aussi, les compétences autour de *l'équipe* ***Image*** sont à intégrer dans une acception large du "document". Ses activités de recherche s'articulent autour de trois thèmes théoriques : "les approches géométriques pour le traitement d'images", "l'estimation, la détection et la reconnaissance des formes", et "l'ingénierie des connaissances pour le traitement d'images". L'équipe s'appuie sur la pluridisciplinarité, la variété des compétences lui permettant d'aborder l'analyse du contenu des images selon plusieurs aspects. Les applications principales concernent le secteur médical, le multimédia, la sécurité, l'imagerie astronomique et le contrôle non destructif, thématiques qui font l'objet de nombreux partenariats industriels.

L'apport des recherches autour de l'image revêt un caractère stratégique dans le cadre d'une approche multimodale du document numérique L'articulation avec les recherches autour du multimédia ouvre ainsi les thématiques du document numérique au-delà du texte.

Dans certains domaines de l'image et du multimédia (vidéo, sécurité par l'image, biométrie…), la Basse-Normandie apparaît bien positionnée sur l'échiquier national. Des compétences sont en développement autour de la recherche d'image par le contenu ouvrant des projets de collaborations avec des entreprises.

Avec l'arrivée d'un nouveau Professeur (anciennement chercheur à l'INRIA), de nouvelles compétences sont développées au sein du GREYC. Elles concernent la recherche de sens dans l'image, problématique complexe, pour laquelle des travaux novateurs portant sur les vocabulaires visuels ont été proposés. Il s'agit de décomposer les images en structures élémentaires, les mots visuels, et d'interpréter les images comme s'il s'agissait de textes.

Il est à noter que dès janvier 2008, un projet ANR de trois ans, sous la responsabilité du GREYC et incluant la société EXALEAD et l'INRIA (centre de Rocquencourt) va impliquer le GREYC sur un périmètre concernant l'indexation de document contenant texte et image. Le programme consistera à extraire des informations sémantiques à partir de photographies et à les combiner à celles extraites des textes, dans le but d'améliorer les performances des moteurs de recherche. Un autre projet ANR, d'une durée de deux ans, portera sur la recherche de vidéos sur Internet avec comme objectif la détection de vidéos ayant des copyrights (à des fins de facturation des droits d'utilisation).

ANNEXE N° 3

LA PLATE-FORME LINGUASTREAM DEVELOPPEE PAR LE GREYC

Développée au GREYC depuis 2001, LinguaStream est une plate-forme générique pour le traitement automatique des langues naturelles, fondée sur l'enrichissement incrémental des documents électroniques. Elle permet la conception et l'évaluation de chaînes de traitement complexes, par assemblage de modules d'analyse de types et de niveaux variés : morphologique, syntaxique, sémantique, discursif ou encore statistique. Ainsi, chaque palier de la chaîne de traitement se traduit par la découverte et le marquage de nouvelles informations, sur lesquelles pourront s'appuyer les analyseurs subséquents. En fin de chaîne, différents outils permettent de visualiser les documents analysés et leurs annotations.

La plate-forme propose différents mécanismes d'élaboration des composants de traitement : règles morphologiques, grammaires d'unification, transducteurs, lexiques sémantiques, règles de production, etc. La plupart d'entre eux s'appuient sur des formalismes déclaratifs, certains étant couramment utilisés en Traitement Automatique du Langage (TAL). Chaque composant d'analyse est réutilisable immédiatement dans d'autres chaînes de traitement et peut être remplacé par un autre composant fonctionnellement équivalent. Une interface graphique prend en charge les différents aspects de l'élaboration d'une chaîne de traitement complète. En outre, grâce à une API (Application Programming Interface)[4] Java et à l'utilisation systématique des normes et outils XML, la plate-forme ainsi que les traitements élaborés avec son aide sont facilement extensibles et intégrables.

LinguaStream a pour ambition de faciliter la réalisation d'expériences sur corpus non triviales en TAL, ainsi que le cycle d'évaluation/ajustements qui en découle. Les disciplines liées au TAL conçoivent ou utilisent aujourd'hui des systèmes dont la réalisation informatique est souvent très complexe, pour lesquels il n'est plus envisageable de procéder systématiquement à une implémentation ad hoc : le coût de développement induit par chaque nouvelle expérience devient en effet un frein considérable à l'approche expérimentale. D'autre part, l'importance des compétences informatiques mises en jeu impliquera souvent le recours à un spécialiste, augmentant encore la complexité du processus expérimental. Pour répondre à cette problématique, LinguaStream facilite la mise en œuvre de procédés complexes tout en requérant des compétences informatiques minimales.

Source : http://www.linguastream.org/whitepaper.html

[4] Interface de programmation.

ANNEXE N° 4

L'IMPLICATION DES PRESSES UNIVERSITAIRES DE CAEN DANS L'EXPLOITATION DE STANDARDS DE STRUCTURATION DES DONNEES

Pour décrire un modèle de document en XML, il convient de définir de manière rigoureuse une structure ainsi que les différents types de données qui pourront être intégrées.

La norme XML définit ainsi une définition de document type appelée DTD Définition de Type de Document) standard de balisage permettant de structurer le texte. Une DTD décrit les règles qui régissent la structure d'un document notamment au standard XML. Elle énumère les éléments, attributs et entités d'un document, et détermine les relations entre les différents éléments et attributs.

Plus précisément, la DTD (utilisée est le standard de balisage, de notation et d'échange intitulé TEI[5] (Text Enoding Initiative), très utilisée pour le marquage des corpus et développé dans l'univers des sciences humaines et sociales. Il permet le codage des contenus et l'échange de tout document. Les PUC ont travaillé sur l'exploitation de ce standard, utilisée à la fois en édition et dans le domaine de la recherche tournée vers l'édition. La TEI est reconnue dans le monde de la recherche en SHS.

Une réflexion est en cours au sein des PUC sur la prise de brevet dans le domaine public en lien avec l'Association des éditeurs de la recherche et de l'enseignement supérieur. L'outil prend un flux XML-TEI et le met en page.

Les PUC commencent également à travailler sur une autre DTD qui concerne l'archivage en lien notamment avec l'IMEC : l'EAD (Encoded Archival Description), standard qui permet de créer des inventaires normalisés, à structure hiérarchisée, et dont le contenu peut faire l'objet de recherches approfondies grâce aux techniques de balisage utilisées.

Des travaux concernent également une troisième DTD, appelée "ONIX", standard d'échange d'information entre éditeurs et diffuseurs-distributeurs qui comprend une partie des référencements des métadonnées et informations commerciales. Cette initiative est menée en collaboration avec la Maison des Sciences Humaines de Paris qui vient d'ouvrir le comptoir des Presses d'Universités sur Internet[6] consacré à la promotion et à la vente de la production scientifique française et francophone des éditeurs de la recherche et de l'enseignement supérieur (catalogue de 8 000 ouvrages). Les PUC font partie des premières presses universitaires en France à être intégrées dans ce système. Notons enfin que le catalogue des PUC lui-même est alimenté par des données ONIX.

---

[5] L'objectif du consortium TEI est de définir un format d'échange, de création et de stockage de textes annotés, ce qui implique à la fois de définir un jeu de balises standardisées et de rendre ce schéma indépendant des matériels, spécificités de réseaux, diversité des jeux de caractères, etc., ce que permet XML. Il serait envisageable que l'Université de Caen entre dans le Consortium TEI.

[6] www.lcdpu.fr

ANNEXE N° 5

LA COOPERATION AVEC LA HAUTE-NORMANDIE DANS LE DOMAINE DU DOCUMENT NUMERIQUE

Le Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes (LITIS) en Haute-Normandie au cœur des complémentarités avec la Basse-Normandie autour du document numérique

Le Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes (LITIS) est un laboratoire commun aux universités de Rouen et du Havre et de l'INSA de Rouen. Il présente, parmi ses travaux de recherche, un axe fort autour du document numérique et plus particulièrement de la numérisation de documents répartis au sein de trois équipes :

- *"Document et Apprentissage"* qui mène des travaux sur la numérisation, la forme et l'imagerie documentaire et développe parallèlement des techniques d'apprentissage par les machines,
- *"Modélisation et Usages"* avec deux axes : modélisation cognitive et modélisation linguistique conduisant à des recherches autour de l'interaction de l'usager avec les systèmes d'information,
- *"Connaissances en santé"* qui repose sur les travaux menés au CHU de Rouen autour du système d'information de santé (CISMEF) comprenant un travail d'indexation et de catalogage de documents médicaux.

Les activités autour de la numérisation constituent des domaines où les transferts vers l'industrie sont les plus nombreux et notamment en ce qui concerne le traitement automatique de l'écrit (du papier à l'image – de l'image au texte – du texte à l'information).

Le traitement de l'écrit franchit plusieurs barrières : l'imagerie, la reconnaissance de formes (titres, paragraphes, mots, caractères) puis l'intelligence artificielle (post-traitement linguistique, cohérence logique…). La maîtrise de ces différentes technologies repose sur des travaux pluridisciplinaires et permet d'aboutir à des systèmes de lecture assez opérationnels. Les résultats permettent, à partir de documents papier, de produire des documents électroniques dans lesquels l'information est encodée, comprenant des métadonnées, et donc interrogeable par l'utilisateur. En d'autres termes, on va apporter du traitement supplémentaire à l'image accessible au plus grand nombre.

Les travaux sur la numérisation (du papier à l'image) concernent les systèmes de lecture de l'écriture manuscrite[7], y compris dans de mauvaises conditions, le traitement de l'image dans les manuscrits anciens (élimination du fond, reconnaissance du vieux français…), la reconnaissance de formes notamment (par exemple reconnaissance d'un montant sur un chèque) ce qui pose la difficulté de l'apprentissage par la machine de toutes les formes de chiffres manuscrits possibles

[7] Les techniques OCR classiques sont inopérantes dans ce domaine.

en évitant les confusions. Des systèmes de visions reposant sur les réseaux de neurones artificiels sont construits.

L'étape ultime consiste à extraire des informations du texte et de détecter dans un document textuel des faits spécifiques.

Le LITIS est depuis plus de 20 ans totalement intégré dans une démarche de valorisation de ses travaux concrétisée par de nombreux projets industriels en partenariats avec des entreprises dans les thématiques suivantes :

- la reconnaissance de caractères imprimés (machines de lecture de tri postal, lecture de chèques),
- le traitement des documents structurés (formulaires, ouvrages imprimés de bonne qualité, magazines),
- la lecture d'adresses manuscrites (bureau distributeur),
- la lecture de chèques (montant numérique et littéral),
- le traitement automatique du recensement,
- le papier numérique,
- le traitement des courriers entrants,
- le traitement des formulaires couleur,
- le traitement des documents graphiques,
- la numérisation des collections pour les bibliothèques numériques.

Le LITIS est en outre impliqué dans des programmes de recherche nationaux. Citons ainsi :

- l'identification biométrique par la modalité écrite (programme ITEM-CNRS, programme société de l'information),
- l'Action Concertée Incitative (ACI) MADONNE - Programme ANR Navidomass (Développement d'outils d'analyse d'image pour la navigation dans des Masses Documentaires) en partenariat avec le L3I de La Rochelle, le LORIA de Nancy, l'Université Paris V, l'IRISA Rennes, le Centre d'Études Supérieures de la Renaissance (CESR) de Tours,
- le projet BOVARY, programme pluridisciplinaire de numérisation et d'édition hypertextuelle de l'ensemble du corpus des manuscrits de cette œuvre de FLAUBERT. Une coopération associe la Bibliothèque municipale de Rouen, le laboratoire CEREDI et le laboratoire LITIS,
- le projet OPTIMA (Outils pour le Traitement de l'Information dans les Manuscrits Modernes) en partenariat avec l'Institut des Textes et Manuscrits Modernes du CNRS (ITEM), MSH (Maison des Sciences de l'Homme), la Bibliothèque Nationale de France, le LIPN (Laboratoire d'informatique de l'Université de Paris Nord). Ce programme vise à développer des outils informatiques pour l'édition génétique des grands corpus (Proust, Flaubert, Valery, F. Braudel).

Le LITIS est également impliqué dans des projets européens comme le programme INTERREG III MEDDRAW en collaboration avec l'Université du Kent sur le thème : "Rééducation et Diagnostic Assistés par Ordinateur de Tâches de Dessins". Citons également le Projet CISMEF (Catalogue et Index des Sites Médicaux Francophones) en partenariat avec le CHU de Rouen.

ANNEXE N° 6

LES FORMATIONS SUPERIEURES SPECIALISEES COMPLEMENTAIRES EN HAUTE-NORMANDIE DANS LE CADRE DU RESEAU STIC-SHS

L'offre de formation en relation avec le projet de réseau interrégional concerne 11 masters existants (auxquels s'ajoutent deux projets déposés dans le cadre des prochains contrats d'établissements) et 4 écoles doctorales.

Parmi les formations de Masters dispensées en Haute-Normandie, certains se distinguent tout particulièrement dans des thèmes en lien avec la problématique du document numérique.

Ainsi, le **Master Professionnel "Systèmes d'Acquisition et de Traitement de l'Information"** (SATI) de l'Université de Rouen a pour objectif de former des étudiants aux problématiques et aux solutions à mettre en oeuvre dans le domaine du Traitement Automatique de l'Information. La spécificité de cette spécialité réside dans son orientation "Analyse et Conception de Système de Traitement".

Egalement, le **Master Recherche spécialité Génie Informatique** proposé par l'Université de Rouen a pour finalité de former des étudiants d'une filière traitement de l'information aux problématiques de la recherche dans le domaine. La spécificité de cette formation réside dans son orientation " Système de Traitement " et vise à dispenser les connaissances fondamentales pour appréhender des applications dans le domaine du traitement de l'information.

Ces deux formations scientifiques abordent les aspects de numérisation du document, d'indexation et de recherche d'information.

Citons également le **Master Professionnel en sciences humaines et sociales, mention géographie : Information Spatiale et Territoire spécialité "Traitement de l'Information Géographique pour l'Aménagement et le Développement"** (TRIAD), proposé par l'Université de Rouen qui aborde des problématiques d'information et de document géographique d'aide à la décision. Il forme aux métiers liés au traitement de l'information géographique dans les domaines de l'aménagement et du développement (responsable de services SIG, Gestionnaire de systèmes de gestion de bases de données). Les secteurs d'activités professionnels sont particulièrement diversifiés : tourisme, développement rural, gestion des risques environnementaux, urbanisme, politique de la ville, marketing, géomarketing, logistique, édition, multimédia…

Sont en outre en projet à la fois un **Master Professionnel "Ingénierie de l'Information Electronique"**, compétence complémentaire destinée notamment aux étudiants en formations Linguistique et Lettres et un **Master "Systèmes de Traitement des Informations Multimédia"** (STIM).