FICHE PROJET EUROPEEN			
ACRONYME : PLAIR 2.018			
NOM COMPLET DU PROJET	Plateforme d'Indexation Régionale 2.018		
NUMERO DE CONVENTION	D15-22180		
DATE DE DEBUT	01/09/2015		
DATE DE FIN	30/06/2019		
COORDINATEURS	Thierry PAQUET		
• *Etablissement(s)*	• *Laboratoire(s)*	• *Responsable(s)*	• *Partenaire(s)*
	LITIS		
	IRIHS		
	DYSOLA		
	IDIT CUREJ		
	GRHIS		
	IDEE		
	ERIAC		
	CEREDI		
CONTACT			
SITE INTERNET DU LABORATOIRE ET PROJET			

DESCRIPTION DU PROJET

RESUME

PlaIR 2.018 est organisé en quatre grandes tâches, qui se proposent de développer des compléments de briques logicielles à partir des contributions des projets co-financés, et de développer des démonstrateurs à partir des briques disponibles. Nous décrivons dans les paragraphes qui suivent les orientations générales de ces différentes tâches.

WP1. Scan On Demand

Cet axe contribuera au développement de l'atelier numérique du programme Digital Paris Normandie, en prenant appui sur les acquis de PlaIR, dont notamment la plateforme de valorisation d'archives de journaux (PIVAJ). Le programme de travail consistera à consolider les briques logicielles acquises en visant deux extensions principales:

1- Généraliser la chaîne de dématérialisation de journaux à d'autres corpus, pour répondre aux besoins des Humanités Numériques qui, dans le cadre de l'Equipex Biblissima, s'intéressent à des corpus littéraires du patrimoine écrit du moyen âge et de la renaissance. 2- Intégrer un module de reconnaissance de caractères performant (ICR), multi-scripts (imprimé - manuscrit) et paramétrable par les connaissances linguistiques du domaine (lexiques et modèle de langage). Ces deux extensions bénéficieront des acquis les plus récents de l'équipe réalisés dans le cadre des projets, MAURDOR achevé en 2014, et DOD (Document On Demand) qui s'achève en 2015.

Sur le plan des démonstrateurs, l'accent sera mis sur deux applications importantes qui permettront de réaliser un continuum dans la chaîne de traitement documentaire au sein du consortium Digital Paris Normandie. Ce continuum peut être décrit schématiquement selon les phases de traitements suivantes : - Numérisation - Transcription automatique - Correction et annotation des textes transcripts - Edition. La première phase concerne l'acquisition et le traitement des images. Elle est prise en charge par l'équipe image du GREYC. La dernière phase concerne l'édition numérique des documents. Elle est prise en charge par les

équipes de la MRSH de Caen ainsi que par l'équipex Biblissima. Les deux phases intermédiaires sont prises en charge par le LITIS.

Un démonstrateur, s'appuyant sur l'architecture PIVAJ, exploitera les résultats des traitements automatiques pour proposer aux utilisateurs du consortium Digital Paris Normandie un environnement de travail leur permettant d'éditer les transcriptions produites automatiquement, de les corriger et de les annoter. L'exportation des textes et annotations dans un format TEI assurera le continuum de la chaîne documentaire souhaité vers les travaux d'édition.

WP2. Gestion et exploitation des Systèmes d'Organisation des Connaissances en Santé

Plusieurs briques logicielles vont être développées ou consolidées à partir des Systèmes d'Organisation des Connaissances (SOC) intégrés au Serveur 3M (Multi-terminologique, Multi-discipline, Multi-lingue) conçu et développé lors des travaux de thèse de Julien Grosjean (PlaIR). Ces SOC peuvent être exploités dans de nombreux contextes : indexation manuelle ou automatique, représentation et enrichissement de la connaissance, enseignement, recherche d'information, etc. Dans le domaine de la Santé en particulier, ces problématiques sont des enjeux majeurs et offrent des perspectives importantes dans la gestion de la connaissance, des contenus et même de la prise en charge des patients, en particulier avec le développement d'une brique «recherche d'information au sein d'un dossier patient informatisé». Ces technologies peuvent également être portées vers d'autres domaines car elles appartiennent plus généralement à celui des Sciences de l'Information.

L'architecture du projet PlaIR 2.018 sera partiellement modifiée par rapport à celle de PlaIR 2.0. En effet, pour des soucis d'efficacité d'un côté et du respect du Web sémantique de l'autre, la technologie noSQL sera implémentée dans le système. Cette solution permettra de s'affranchir de la "dépendance" aux outils d'Oracle choisis depuis 15 ans, et surtout de la possibilité de pousser les outils directement dans les hôpitaux, sans passer par les services Web via l'Internet qui posent naturellement des problèmes de confidentialité pour les données personnelles de santé. Ainsi, une grande partie de l'architecture logicielle doit être refondue et permettre la manipulation d'objets directement dans la couche de données sous forme de cache (mémoire vive - RAM). Ce type de structure est certes assez coûteux en termes de matériels mais constitue un nouveau paradigme dans le domaine des données massives (Big Data).

WP3. Accès personnalisé et édition collaborative d'un corpus de documents

Cette tâche s'intéresse à la prise en compte des utilisateurs d'une plateforme d'accès à un corpus numérique et de leurs modes d'utilisation de manière à enrichir le corpus de documents et les mécanismes de navigation et d'édition dans une approche Web 2.0.

Dans cette vision, les utilisateurs ne sont pas seulement consommateurs mais aussi producteurs de contenu. Dans les corpus considérés, l'édition collaborative permet aux utilisateurs dans certains cas de modifier les documents du corpus numérique (par exemple pour corriger des transcriptions automatiques par OCR) ou d'adjoindre des annotations à une partie d'un document, relativement à une recherche en cours, qui pourraient être exploitées dans de futurs processus de recherche. L'intérêt de cette démarche collaborative est d'enrichir et d'améliorer le corpus pendant son exploitation, mais, si elle n'est pas contrôlée, elle introduit aussi un risque de détérioration du corpus par l'ajout de bruit, d'informations fausses ou par vandalisme.

Nous nous intéresserons dans cette tâche de PlaIR 2.018 au développement d'outils de gestion de la confiance dans le but d'évaluer la qualité

	et la fiabilité des contributions des utilisateurs. Plusieurs problèmes sont à résoudre pour mettre en œuvre un processus de gestion de la confiance dans une plateforme collaborative : la définition de métriques d'estimation de la qualité d'une production associées à des algorithmes de calcul automatique ou semi-automatique, l'exploitation de ces mesures dans un processus de validation des contributions, la généralisation de la qualité des contributions en une confiance dans un contributeur. Les modèles de confiance développés dans le domaine des systèmes multi-agents (tq REGRET, FIRE, LIAR, EigenTrust, …) ont déjà prouvé leur adéquation à des systèmes collaboratifs (eg Wikipedia) et nous proposons d'adopter cette approche pour développer un modèle de confiance dans PlaIR 2.018. **WP4. Humanités numériques** Dans la continuité des projets FEDER LEONUM-SHS et INTEREG IVA DocExplore, l'objectif de cet axe est de renforcer, de développer et de rendre plus visibles les projets scientifiques stratégiques numériques en sciences humaines et sociales. Il constitue l'un des volets de la structuration des SHS en Haute Normandie et au-delà le projet s'inscrit dans le cadre de la future MRSH Normandie. En effet, ce projet vise à mutualiser un certain nombre de travaux des laboratoires autour des éditions de textes littéraires et musicaux, l'étude du patrimoine littéraire notamment Normand, l'interprétation des œuvres et notamment l'étude des phénomènes de réécriture, de réception et de transformation des textes. Les avantages d'une documentation numérique ou d'une base de données en ligne concernent à la fois : les possibilités de mise à jour immédiate et donc d'enrichissement des textes ou des bases de données, l'ajout immédiat de textes inédits, à leur place (dans une édition papier, il faut attendre une réédition, souvent tardive, et les nouveaux textes se trouvent rejetées en annexe, pour éviter de refaire une mise en page complète), la correction en temps réel avec l'apparition de lettres inédites ou la découverte d'un élément nouveau qui permet souvent de redater des lettres déjà connues ; des progrès dans la mise en œuvre et la consultation en raison de la nature même du support et l'accès à des informations pour des publics variés qu'ils soient chercheurs, enseignants-chercheurs, décideurs publics et privés, grand public.
OBJECTIFS	
IMPACTS ATTENDUS ET FINALITE	
RESULTATS	**LIVRABLES ATTENDUS SUITE A LA REALISATION DE L'OPERATION** WP 1.1 : Analyse intelligente de mise en page (2016) Livrable : briques logicielles d'analyses de structure de documents (notamment pour le WP 1.3) WP 1.2 : ICR multi-script et multi-lingue (2017) Livrable : briques logicielles de reconnaissance de l'écriture (notamment pour le WP 1.3) WP 1.3 : Démonstrateur Environnement de travail Digital Paris Normandie (2017-18, 18 mois) Livrable : démonstrateur web instancié sur quelques corpus sélectionnés au sein du consortium Digital Paris Normandie WP2 - SOC de santé (2016-2018, 36 mois) WP2.1 : Architecture de données et Logicielle basée sur la technologie noSQL (juin 2016) Livrables : - Choix d'une solution de base de données noSQL basé sur un état de l'art des solutions noSQL libres et des tests de développement et

	performances. - Rapport de migration des données et service web d'accès au cache noSQL. WP 2.2 : Migration et maintenance du Serveur 3M (décembre 2017) Livrables : - Adaptation du Serveur 3M à la solution noSQL choisie en WP 2.1 - Rapports annuels de mises à jour du Serveur 3M en fonction des nouveaux SOC, SOC mis à jour et enrichissements par les experts de l'équipe. WP 2.3 : Migration et améliorations de l'Extracteur de Concepts Multi-Terminologiques (ECMT) (2018) Livrable : Adaptation de l'ECMT à la solution noSQL choisie en DXX.1 et création d'un service web dédié. WP 3.1 : Edition d'annotations à un corpus numérique (2016) livrable : Module d'édition d'annotation dans PlaIDIT WP 3.2 : Modèle de confiance pour l'édition (2017) livrable : Service de calcul semi-automatique de la qualité d'une annotation WP 3.3 : Mise en œuvre de PlaDIT a l'IDIT (2018) livrable : Adaptation de la recherche d'information dans PlaIDIT aux annotations WP4 - Humanités Numériques (2016-2018, 36 mois) WP 4.1 : Recherche d'objets graphiques dans des images de documents (2016). Livrable : Un logiciel Web destiné à la valorisation de collections numérisées riches en images. WP 4.2 : Outils numériques pour l'interprétation des œuvres (2017) Livrable : Démonstrateur web. WP 4.3 : Usages de Scan on demand par les Humanités Numériques (2018) Livrable : - Des Edition numériques augmentées des corpus sélectionnés, accessibles sur le web - Retour d'expérience de Scan on demand.
MODALITES DE FINANCEMENT	
BUDGET TOTAL	1 695 642,49€
• *Niveau de soutien FEDER / FSE / FAEDER*	841 247,93€
• *Niveau de soutien région*	
• *Niveau de soutien Etat*	
• *Autofinancement*	
• *Autre*	11 000€
NOMBRE D'ALLOCATIONS DOCTORANTS	0
NOMBRE D'ALLOCATIONS ET POST-DOCTORANTS	0
L'Europe s'engage en Normandie avec le Fonds Européen de Développement Régional	