

OFFRE DE STAGE MASTER 1^{ère} ou 2^{ème} - OHM-Littoral Méditerranéen – LabEx DRIIHM, RATTACHÉ A LA MESHS de Lille (USR 3185 CNRS)

INTITULÉ DU STAGE : EXTRACTION ET ANALYSE D'INFORMATION A PARTIR DES RESEAUX SOCIAUX : APPLICATION AUX DONNEES DU PARC NATIONAL DES CALANQUES.

CONTEXTE SCIENTIFIQUE : Ce stage de Master 1 ou 2 s'inscrit dans le projet de recherche intitulé « **Les réseaux numériques : un tournant pour l'analyse des relations homme-milieu ? Application au parc national des calanques** » **RENUROHM**. Ce projet a été sélectionné dans le cadre de l'Appel à Projets de Recherche 2016 du LabEx DRIIHM (Dispositif de Recherche Interdisciplinaire sur les Interactions Hommes-Milieus), au titre de l'Observatoire Hommes-Milieus Littoral Méditerranéen. L'Observatoire Hommes-Milieus "Littoral méditerranéen" est un dispositif de recherche interdisciplinaire du CNRS (INEE et INSHS). Son projet scientifique est l'étude de l'urbanisation et de l'anthropisation côtière en Méditerranée. Dans un contexte de changement des modes de gestion du littoral (GIZC), il s'intéresse à quatre systèmes socio-écologiques littoraux inégalement soumis à l'artificialisation et aux fréquentations touristiques et récréatives : le littoral marseillais, le Golfe d'Aigues-Mortes et les rivages corses de Balagne et du sud Bastia. Depuis une quinzaine d'années, on constate un développement des **humanités numériques** provoquant l'émergence d'un nouveau champ de recherche interdisciplinaire, la création de nouveaux corpus de données numérisées et la reconfiguration des pratiques ou des objets de recherche sous l'impact des contenus numériques diffusés sur le web et les réseaux sociaux numériques. La mobilisation des corpus de données numériques rassemblées sur le web et les réseaux sociaux peut représenter une nouvelle approche de recherche des relations entre l'environnement et la société. Des recherches récentes ont analysé le processus de création du Parc national des calanques (Deldrève et Deboudt, 2012 ; Deboudt et Deldrève, 2015). Dans le cadre du projet RENUROHM, un premier travail est en cours pour réaliser la cartographie des acteurs en lien avec le Parc national des calanques afin d'identifier les acteurs qui s'expriment sur des blogs, sites web ainsi que sur les réseaux sociaux, à propos de différents sujets en lien avec ce territoire. Nous nous appuyons pour cela sur une méthodologie pour la cartographie semi-automatisée des acteurs de domaine (Berthelot et al., 2016). Ce premier travail permettra de produire une première liste d'acteurs qui sera utilisée pendant le stage pour réaliser des tâches de fouille textuelle (Text Mining).

OBJECTIFS : TRAVAUX ATTENDUS : A partir d'un corpus déjà constitué (sites web et tweets), il s'agit de réaliser une analyse semi-automatique des contenus des tweets afin d'identifier les entités nommées (organisation, personnes et lieux) et les thématiques abordées dans le texte des tweets. Plus précisément, le stage consistera à intégrer et enrichir des chaînes de traitement automatique pour :

- Extraire des entités nommées à partir de listes préalablement fournies d'acteurs et de lieux et en se basant sur la méthode décrite dans (Zenasni et al., 2016) qui permet l'extraction de nouvelles formes de lieux dans les messages courts ;
- Extraire des thématiques en se basant sur des approches fouilles de textes (Pak et al. 2014) ;
- Evaluer les résultats obtenus ;
- Réaliser différentes analyses qualitatives et quantitatives sur les résultats obtenus et notamment répondre aux questions suivantes : Quels sont les acteurs qui s'expriment sur ces lieux/sujets ? Quelles sont les relations entre ces acteurs ? Quelles sont les évolutions observées selon différentes temporalités ?
- Participer au travail de valorisation des résultats en enrichissant le site Web du projet (en cours de construction).

INDEMNISATION, DURÉE ET LIEUX DE TRAVAIL :

- Gratification : **554,40 euros par mois** ;

- Ce stage d'une durée de 4 mois (**1^{er} avril 2017 – 31 juillet 2017**) se déroulera de manière partagée à l'Université Lille 3 et à l'Université Lille 1, dans les locaux du laboratoire Geriico (Lille 3 ;

<http://geriico.recherche.univ-lille3.fr/>) et du laboratoire TVES (Lille 1 ; <http://tves.univ-lille1.fr/>)

PROFIL DU CANDIDAT :

- Master 1^{ère} ou 2^{ème} année en cours ;
- Formation, compétences et qualités requises : Linguistique ; Traitement Automatique des Langues (TAL); Fouille de textes ; la maîtrise d'un langage de programmation est indispensable (java) ; la maîtrise du logiciel R serait un plus. Capacité à travailler en équipe et à distance. Une connaissance préalable de l'outre mer ou du Brésil sera appréciée.

CONTACTS ET CALENDRIER : Envoyer un CV détaillé, par mail (1 fichier PDF), **avant le lundi 6 mars 2017** à Eric Kergosien (eric.kergosien@univ-lille3.fr) et Amel Fraisse (amel.fraisse@univ-lille3.fr).

Bibliographie :

- Arsène S., 2013, *Vers une recomposition des pouvoirs : Internet et réseaux sociaux*, CERISCOPE Puissance, ;
- Deldrève V., Deboudt P. (dir.), 2012, *Le parc national des calanques : construction territoriale, concertation et usages*, QUAE, 231 p. ;
- Deboudt P., Deldrève V., 2015, Inégalités et concertation « encadrée » : le projet du parc national des calanques, in L. Mermet et D. Salles (dir.), *Environnement et transition écologique*, De Boeck éd., coll. Ouvertures Sociologiques, p. 151-166.
- Berthelot M.-A., Severo M., Kergosien E., 2016, , Cartographier les acteurs d'un territoire : une approche appliquée au patrimoine industriel textile du Nord-Pas-de-Calais, In 3^{ème} colloque international du CIST (CIST 2016), pp.6, Grenoble.
- Zenasni S., Kergosien E., Roche M., Teisseire M., 2016, Extracting new Spatial Entities and Relations from Short Messages, In the 8th International ACM Conference on Management of Digital EcoSystems (MEDES'2015), pp. 8, Hendaye (France).
- **Alexander Pak and Patrick Paroubek and Amel Fraisse and Gil Francopoulo (2014)**. *Normalization of Term Weighting Scheme for Sentiment Analysis*. Book Chapter, Human Language technology Challenges for Computer Science and Linguistics. Series: Lecture Notes in Artificial Intelligence, Springer, Vol. 8387. ISBN 978-3-319-08957-7. *Vetulani, Zygmunt, Mariani, Joseph (Eds.). May 27, 2014.*
- **Amel Fraisse and Patrick Paroubek (2014)**. Twitter as a Comparable Corpus to build Multilingual Affective Lexicons. In proceedings of the 7th International Workshop on Building and Using Comparable Corpora at LREC 2014 (BUCC 2014), pages 17-21. May 26-31, 2014. Reykjavik, Iceland.