Thèse 2020 - *Cirad – Inrae – Projet H2020 MOOD – Montpellier - France*

Proposition et mise en œuvre de méthodes génériques pour la veille épidémiologique fondée sur l'intégration de données textuelles hétérogènes

CONTEXTE ET PROBLEMATIQUE

La veille en santé animale a pour objectif l'alerte précoce vis-à-vis de dangers sanitaires connus ou émergents. Elle repose sur le recueil, le suivi et l'analyse quotidienne d'informations issues de sources officielles, telles que l'Organisation mondiale de la santé animale (OIE), et de sources non-officielles telles que les médias ou les réseaux sociaux (Hartley et al. (2010)). Plusieurs systèmes de biosurveillance, tels que MedISys (Mantero et al. (2011)), GPHIN (Blench (2008)) ou HealthMap (Freifeld et al. (2008)), sont ainsi dédiés à l'acquisition et à la diffusion de données issues de sources informelles. Par ailleurs, ces approches reposent sur une modération humaine à une ou plusieurs étapes de leur processus. Certains systèmes collectent les données à partir de sources officielles et non officielles (HealthMap, EWRS et GOARN) tandis que d'autres les collectent principalement via un réseau d'experts et d'abonnés (ProMED). Les utilisateurs des systèmes EWRS et ProMED mènent également une recherche manuelle sur le Web et d'autres systèmes pour trouver des informations sanitaires complémentaires (Barboza, 2014 ; Yu *et al.*, 2007). Le système IBIS utilise l'approche collaborative (« crowd-sourcing »). IBIS permet également d'analyser le contenu de chaque article et de contribuer à l'évaluation de termes automatiquement extraits et annotés : les maladies, les espèces touchées, les signes cliniques ainsi que le lieu d'évènement (Lyon, Mooney *et al.*, 2013 ; Lyon, Grossel *et al.*, 2013).

Dans ce contexte, les chercheurs des unités TETIS et ASTRE ont proposé et développé PADI-Web (Platform for Automated extraction of Disease Information from the web), un outil de biosurveillance des médias digitaux pour la détection de foyers de maladies animales (Arsevska *et al.*, 2018 ; Valentin *et al.*, 2020). PADI-web est intégré dans la thématique de Veille sanitaire internationale, au sein de la plateforme d'Epidémiosurveillance en santé animale (plateforme ESA). Depuis sa première version dédiée à la veille de sources en anglais, PADI-web a été enrichi en 2019 d'un nouveau classifieur reposant sur des méthodes d'apprentissage automatique et intègre les documents multilingues.

RESUME DU TRAVAIL PROPOSE

Outre les améliorations méthodologiques à mettre en œuvre pour enrichir la plateforme actuelle (prise en compte l'hétérogénéité des différentes sources mobilisées dans un contexte multilingue), le point le plus important du travail proposé dans cette thèse est de développer des méthodes génériques pour la veille épidémiologique fondée sur l'intégration de données textuelles hétérogènes. Ceci permettra de proposer et mettre en place un système appelé **PADI-Web *One Health*.** Pour répondre à cet enjeu, trois problématiques seront plus particulièrement étudiées.

→ Proposition d'un cadre généralisé qui intègre les différents types de veille

Depuis 2014, le système PADI-Web s'intéresse à la veille épidémiologique liée à la santé animale. L'objectif de cette thèse est d'étendre ces approches dans un cadre générique qui intègre la veille en santé végétale et alimentaire. En effet, outre certaines informations, en particulier *spatio-temporelles* et les problématiques associées (extraction et désambiguïsation d'informations spatiales dans les textes) qui sont par nature tout à fait génériques, la généricité de certains concepts *thématiques* (par exemple, les symptômes) devront être étudiées tout en prenant en compte la spécificité liée à chaque domaine.

→ Identification d'événements épidémiologiques fins dans les données multi-sources

La tâche proposée consiste à identifier les informations issues de donnés non structurées multilingues (dépêches, articles scientifiques, etc.) et de qualifier ces informations extraites (« confiance » à établir sur la base de la qualité des données, des sources et des approches automatiques utilisées). Une attention particulière sera portée à l'identification de signaux faibles. Les méthodes proposées combineront des approches d'apprentissage supervisées, des systèmes à base de règles et des méthodes de plongements lexicaux (*word embedding)*.

→ Fusion d'informations épidémiologiques issues de données hétérogènes

La dernière contribution attendue consistera à combiner les informations issues des organismes officiels (par exemple l'OIE) aux données non officielles obtenues par fouille de textes afin de proposer une méthode générique, robuste et complète.

PROJET

La thèse proposée s'inscrit dans le cadre du projet H2020 MOOD « *Monitoring Outbreak events for Disease surveillance in a data science context* » (https://mood-h2020.eu/). Ce projet, qui fédère 25 partenaires issus de 10 pays, a pour objectif d'améliorer la détection, la surveillance et l'évaluation des maladies infectieuses émergentes en Europe en utilisant les techniques d'exploration et d'analyse de données massives provenant de sources multiples. Il est porté par le Cirad (UMR ASTRE) avec une participation importante de l'UMR TETIS dans les WP2 et WP3.

PUBLICATIONS DE L'EQUIPE ASSOCIEES A CE SUJET

Valentin S, Arsevska E., Mercier A., Falala S., Rabatel J., Lancelot R., Roche M. PADI-web: an event-based surveillance system for detecting, classifying and processing online news. In *Post-Proceedings of 8th Language & Technology Conference, LTC 2017, Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence* - Springer, 2020, *to appear*

Valentin S, Arsevska E, Falala S, de Goër J, Lancelot R, Mercier A, Rabatel J, Roche M. PADI-web: a multilingual web-based biosurveillance system for the monitoring of animal infectious diseases. *Computer and Electronics for Agriculture*, Elsevier, 169: 105163, 2020

Arsevska E, Valentin S, Rabatel J, de Herve JG, Falala S, Lancelot R, Roche M. Web Monitoring of Emerging Animal Infectious Diseases Integrated in the French Epidemic Intelligence System in Animal Health. *PLOS One*; 13: e0199960, 2018.
Selected as one of the "Best paper" of IMIA Yearbook of Medical Informatics 2019.

Drury B., Roche M. A survey of the applications of text mining for agriculture. *Computers and Electronics in Agriculture*, 163 104864. 13 p. 2019.

Lossio-Ventura J.-A., Bian J., Jonquet C., Roche M., Teisseire M. 2018. A novel framework for biomedical entity sense induction. *Journal of Biomedical Informatics*, 84 : 31-41, 2018

CANDIDATURES

Le ou la doctorant(e) sera accueilli(e) à l'UMR TETIS. Le candidat retenu devra être de formation initiale en informatique, en biostatistique ou en épidémiologie mais avec des compétences solides dans les sciences informatiques.

Dossier de candidature (* : éléments obligatoires) à envoyer **avant le 26 juin 2020** :

- CV détaillé *
- lettre de motivation *
- relevés de notes (avec classement) *
- contacts pour recommandation *
- lettres de recommandation
- rapport du dernier stage réalisé

Les candidatures sont à envoyer par mail à :

- Mathieu Roche (Cirad, UMR TETIS) – mathieu.roche@cirad.fr
- Maguelonne Teisseire (Inrae, UMR TETIS) – maguelonne.teisseire@inrae.fr
- Renaud Lancelot (Cirad, UMR ASTRE) – renaud.lancelot@cirad.fr