

Proposition de contrat doctoral 2017

(Cirad – Montpellier - France)

INTITULE DU SUJET DE THESE

Extraction et agrégation d'informations issues de données multi-sources pour la veille internationale des maladies animales infectieuses

CONTEXTE ET PROBLEMATIQUE

La veille en santé animale, et notamment la détection précoce d'émergences au niveau mondial d'agents pathogènes, est l'un des moyens permettant de prévenir ou d'anticiper l'introduction en France de dangers sanitaires. Pour cela, dans le cadre de la Veille Sanitaire Internationale (VSI) de la Plateforme nationale d'épidémiosurveillance en santé animale (PNESA : <http://www.plateforme-esa.fr>), les chercheurs des unités TETIS et ASTRE développent depuis 2014 une plateforme dédiée à la veille automatique allant du recueil des données textuelles (dépêches) jusqu'à la restitution synthétique des informations extraites.

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, Grossel et al., 2013](#)).

Afin de trouver des informations pertinentes sur le Web, la plupart des systèmes de veille utilisent des combinaisons de termes et d'expressions en plusieurs langues. Les systèmes Argus et IBIS utilisent des termes de noms de maladies, d'agents pathogènes et de leurs variants ([Nelson, Brownstein et al., 2010](#) ; [Nelson, Yang et al., 2012](#)). Les systèmes MediSys et HealthMap utilisent en plus, des termes qui décrivent des signes cliniques et des mots clés qui caractérisent des foyers. Les termes sont proposés par des experts (systèmes GPHIN, MediSys) ([Keller, Freifeld et al., 2009](#) ; [Mantero et al., 2011](#)), proviennent d'un dictionnaire des pathogènes (système HealthMap) ([Brownstein et al., 2008](#)) ou d'une ontologie médicale, comme dans le projet BioCaster ([Collier, Doan et al., 2008](#)).

Travaux réalisés par les équipes des unités de Recherche ASTRE & TETIS

Le recueil des dépêches issues des médias électroniques s'appuie sur des requêtes à partir de mots-clés de maladies, d'hôtes et de signes cliniques pour collecter des articles de Google News. Ces mots-clés ont été définis par des experts (méthode Delphi) et/ou par des approches de fouille de textes. De nouvelles pondérations selon les sources de données ont été proposées ([Arsevska et al., 2016](#)). Les dépêches collectées sur la base des mots-clés sélectionnés par les experts sont prétraitées et normalisées avant d'être stockées dans une base de données. L'extraction d'information dans les dépêches collectées identifie les éléments clés (noms de maladies, lieux, dates, nombres et espèces d'animaux touchées). Elle repose sur des dictionnaires dédiés et des règles préalablement construites par un processus de fouille de données (extraction de motifs séquentiels et méthodes d'apprentissage automatique). Les résultats sur un corpus de 352 dépêches en anglais montrent une identification pertinente (*accuracy*) des informations spatiales d'environ 80% et de plus de 85% pour les autres types d'informations. Un des verrous scientifiques est l'ambiguïté des entités spatiales candidates. En effet, une localisation mentionnée dans un texte peut faire référence à plusieurs coordonnées spatiales (par exemple, le toponyme *Saint-*

Louis peut être associé au *Sénégal*, la *Réunion* et aux *Etats-Unis*). Lever automatiquement de telles ambiguïtés est donc crucial pour consolider le système actuel.

RESUME DU TRAVAIL PROPOSE

Outre les améliorations méthodologiques à mettre en œuvre (désambiguïsation des entités spatiales, traitement de l'aspect multilingue dans la chaîne complète, classification automatique), ce projet de thèse se concentrera sur la problématique de l'identification, la combinaison et la qualification des informations et indicateurs (en particulier les signaux faibles) issus des données multi-sources (dépêches officielles, textes libres, etc.) pour l'identification de l'émergence de maladies. Ce travail consistera à construire un modèle général prenant en compte les éléments de différentes sources afin d'améliorer les prédictions.

La première étape consiste à identifier les informations issues de données 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 syndromiques (par exemple, baisse de la production laitière). La seconde étape consistera à combiner ces informations intégrées à celles issues des organismes officiels (par exemple OIE) afin de proposer une méthode générique, robuste et complète. Les approches à proposer, implanter et expérimenter durant cette thèse pourront consister à concevoir des méthodes pour (i) agréger des modèles statistiques¹ (modèles épidémiologiques construits à partir de données issues de différentes sources et/ou obtenues par différentes approches) ; (ii) intégrer des pondérations spécifiques, en particulier pour les informations spatio-temporelles, selon leur provenance pour construire un modèle générique.

Les contributions scientifiques seront intégrées à la plateforme PADI-Web (Platform for Automated Extraction of Animal Disease Information from the Web) développée dans le cadre de la collaboration pluridisciplinaire TETIS-ASTRE avec des applications étudiées sur 5 maladies (peste porcine africaine, l'influenza aviaire, la fièvre catarrhale ovine, la fièvre aphteuse et la maladie de Schmallenberg).

REFERENCES :

Barboza, P. (2014). « Evaluation des systèmes d'intelligence épidémiologique appliquées à la détection précoce des maladies infectieuses au niveau mondial ». français. Thèse de doct. Paris : Université Pierre et Marie Curie - Paris VI.

Brownstein, J., C. Freifeld, B. Reis et K. Mandl (2008). Surveillance sans frontières : Internet-based emerging infectious disease intelligence and the HealthMap project.

Collier, N., S. Doan, A. Kawazoe, R. M. Goodwin, M. Conway, Y. Tateno, Q. H. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, M. Shigematsu et K. Taniguchi (2008). « BioCaster : Detecting public health rumors with a Web-based text mining system ». en. In : Bioinformatics 24.24, p. 2940–2941.

Keller, M., M. Blench, H. Tolentino, C. C. Freifeld, K. D. Mandl, A. Mawudeku, G. Eysenbach et J. S. Brownstein (2009). « Use of unstructured event-based reports for global infectious disease surveillance ». In : Emerging Infectious Diseases 15.5, p. 689-695.

Lyon, A., G. Grossel, M. Burgman et M. Nunn (2013). « Using internet intelligence to manage biosecurity risks : A case study for aquatic animal health ». In : Diversity and Distributions 19.5-6, p. 640–650.

Mantero, J., E. Centre, D. Prevention, J. Belyaeva, E. Commission et J. P. Linge (2011). How to maximise event-based surveillance web- systems : the example of ECDC / JRC collaboration to improve the performance of MedISys. English. Rapp. tech. Luxembourg : Publications Office, p. 1–22.

¹ La pertinence des modèles statistiques à utiliser (linéaires, non linéaires, non paramétriques, bayésiens, etc.) pourra être étudiée dans le cadre de ce projet de thèse.

Nelson, N., J. Brownstein et D. Hartley (2010). « Event-based biosurveillance of respiratory disease in Mexico, 2007-2009 : connection to the 2009 influenza A(H1N1) pandemic? » In : Euro surveillance : bulletin europ en sur les maladies transmissibles. European communicable disease bulletin 15.30, p. 19626.

Nelson, N., L. Yang, A. R. A. Reilly, J. E. Hardin et D. M. Hartley (2012). « Event-based internet biosurveillance : relation to epidemiological observation ». In : Emerg Themes, p. 4.

Yu, H. et D. Kaufman (2007). « A cognitive evaluation of four online search engines for answering definitional questions posed by physicians. » In : Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. T. 12. World Scientific, p. 328-339.

REFERENCES DE L'EQUIPE :

Publications :

Arsevska, E., Roche, M., Hendrikx, P., Chavernac, D., Falala, S., Lancelot, R., Dufour, B., 2016. Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web. *Computers and Electronics in Agriculture*, Elsevier, 123, 104-115.

Arsevska, E., Roche, M., Hendrikx, P., Chavernac, D., Falala, S., Lancelot, R., Dufour, B., 2016. Identification of associations between clinical signs and hosts to monitor the web for detection of animal disease outbreaks. *Int. J. Journal of Agricultural and Environmental Information Systems*, IGI, 7(3): 1-20.

Arsevska, E., Roche, M., Lancelot, R., Hendrikx, P., Dufour, B., Falala, S., Chavernac, D., 2016. Monitoring Disease Outbreak Events on the Web Using Text mining Approach and Domain Expert Knowledge. *In Proceedings of LREC 2016 (Int. Conf. on Language Resources and Evaluation)*, Slovenia, pp.3407-3411

Arsevska, E., Roche, M., Lancelot, R., Hendrikx, P., Dufour, B., 2014. Exploiting Textual Source Information for Epidemiosurveillance, in: B. S. Clos et Al. (Ed.). *In Proceedings of MTSR 2014: 8th Metadata and Semantics Research Conference*, Springer, pp. 359-361.

Lien détaillant le système PADI-Web :

<http://www.cirad.fr/nos-recherches/resultats-de-recherche/2016/veille-sanitaire-sur-le-web-un-outil-pour-prevenir-la-propagation-des-maladies-animales>

CANDIDATURES

Le ou la doctorant(e) sera accueilli(e) à l'UMR ASTRE. 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 17 juin 2017 :

- 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 :

Renaud Lancelot (Cirad, UMR ASTRE) – renaud.lancelot@cirad.fr

Mathieu Roche (Cirad, UMR TETIS) – mathieu.roche@cirad.fr

Thierry Lefrançois (Cirad, UMR ASTRE) – thierry.lefrancois@cirad.fr