Projet de Thèse AVIVA

***A**ssistance à la **v**eille sanitaire et scientifique des **i**nsectes **v**ecteurs en santé des plantes par l'**a**cquisition de connaissances fiables à partir de documents.*

Contacts
Claire Nédellec (claire.nedellec@inrae.fr), Nicolas Sauvion (Nicolas.sauvion@inrae.fr), Catherine Faron (faron@i3s.unice.fr).

Rattachement administratif

Ecole Doctorale Sciences et Technologies de l'Information et de la Communication
Spécialité Informatique
Unité de recherche MaIAGE – Mathématiques et Informatiques appliquées du Génome à l'Environnement
Equipe Bibliome : aquisition et formalisation de connaissance à partir de texte.
Pôle B. Données, Connaissances, apprentissage et interactions
Établissement de préparation de la thèse Université Paris-Saclay – GS Informatique et Sciences du numérique
Référent Faculté des Sciences d'Orsay
Année universitaire de 1ère inscription en doctorat
2023
Date de début de la thèse
01/10/2023
Date limite de candidature
30/06/2023 à 23h59
Thèse interdisciplinaire
Oui

Projet Doctoral

Direction de thèse
NEDELLEC Claire - 36282
01 34 65 28 78
claire.nedellec@inrae.fr
HDR: oui

Modalités d'encadrement, de suivi de la formation et d'avancement des recherches du doctorant
Claire NEDELLEC (MaIAGE-INRAE) [DR-HDR]
Nicolas SAUVION (PHIM-INRAE) [IR-HDR]
Catherine FARON ((Wimmics, Inria et I3S, Univ. Côte d'Azur) [PR- HDR]

Robert Bossy (IR, MaiAGE) et Marie Grosdidier (IR Plateforme ESP, BioSP-INRAE) contribueront à l'encadrement.
L'étudiant.e partagera son temps entre MaIAGE (Jouy-en-Josas) et Wimmics (Sophia Antipolis) sur des durées longues (1 à 2 ans) et fera des visites régulières à PHIM (Montpellier).

Titre

Assistance à la veille sanitaire et scientifique des insectes vecteurs en santé des plantes par l'acquisition de connaissances fiables à partir de documents.

Mots-clés

Traitement Automatique de la Langue, Extraction d'Information, Qualité des Données, Représentation des Connaissances, Raisonnement, Risque biologique, Santé des Cultures, Insectes Vecteurs, Interaction Biologique

EN

Title

Assistance in the sanitary and scientific monitoring of insect vectors for plant health by the acquisition of reliable knowledge from documents.

Keywords
Natural Language Processing, Information Extraction, Data Quality, Knowledge Representation, Reasoning, Biological Risk, Crop Health, Insect Vectors, Biological Interaction

Résumé
En santé du végétal, des insectes porteurs de virus ou de bactéries causent chaque année d'importants dégâts à l'agriculture malgré la lutte dont ils sont l'objet. La veille scientifique et sanitaire porte sur un nombre croissant de documents et d'objets pour comprendre et anticiper ces phénomènes. Elle manque cruellement d'outils pour extraire, structurer et évaluer la pertinence des informations publiées au regard de la connaissance scientifique sur les interactions entre espèces et leur distribution géographique. La pertinence dépend de la nouveauté et de la validité de l'information. Celles-ci sont mesurables à partir d'indices linguistiques (ex. first report), biologiques (ex. *in vitro*), bibliographiques (ex. réseau de citation), et par leur cohérence avec la connaissance scientifique du domaine. L'adaptation de méthodes de traitement automatique de la langue (TAL) (extraction d'information relationnelle, normalisation par des ontologies) permettra d'extraire les informations et de les lier à la connaissance des pathosystèmes dans une représentation formelle cohérente. Le mécanisme d'évaluation de la pertinence combinera (1) le degré de confiance dans la prédiction de l'algorithme de TAL, (2) les indices de fiabilité de l'information des documents et (3) l'évolutivité de la connaissance du pathosystème pour produire une mesure explicable utilisant le raisonnement en logique possibiliste. Le projet est centré sur huit bactéries réglementées d'intérêt économique majeurs. Ces modèles biologiques ont été aussi choisis pour leurs propriétés favorables à la généralisation des méthodes dans les domaines mobilisés, santé des plantes, veille et intelligence artificielle.

Abstract
Insects carrying viruses or bacteria species cause each year important damage to agriculture despite their control. The scientific and sanitary watch is based on a growing number of documents and objects in order to understand better and anticipate these phenomena. Watch cruelly lacks tools to extract, structure, and evaluate the relevance of published information with respect to scientific knowledge on the interactions between species and their geographical distribution. Relevance depends on the novelty and validity of the information. These can be measured by linguistic (e.g. a first report), biological (e.g. *in vitro* mention), and bibliographic (e.g. citation network) clues and by their coherence with the scientific knowledge of the field.

The adaptation of Natural Language Processing (NLP) methods (relation information extraction, normalization by ontologies) will allow the extraction of the information and link it to the knowledge of the pathosystems in a coherent formal representation. The relevance evaluation mechanism will combine (1) the degree of confidence in the prediction of the NLP algorithm, (2) the reliability indicators of the document information, and (3) the evolution of the knowledge of the pathosystem in order to produce an explainable measure using possibilistic logic reasoning. The project focuses on eight regulated bacteria of major economic interest. These biological models were chosen for their properties that are conducive to the generalization of methods in the fields of plant health, surveillance, and artificial intelligence.

Thématique
La quantité de connaissances sur les vecteurs de maladies d'intérêt majeur pour l'agriculture s'accroît de manière substantielle à un rythme souvent très rapide. Les interactions vecteur-plante-pathogène peuvent être plus ou moins spécifiques et parfois très complexes. Bien les décrire est essentiel à la compréhension des cycles biologiques des épidémies, et *in fine* à la conception de méthodes de lutte efficaces. La majeure partie de l'information est d'abord décrite sous forme documentaire (Barboza et al., 2014). Maintenir une expertise scientifique et sanitaire à jour dans ce domaine très évolutif requiert une veille continue de sources d'information de plus en plus abondantes et dispersées dans des publications scientifiques primaires, ou non, des revues, des textes réglementaires, des *news*, *etc*. Or, la collecte et la formalisation de ces informations restent essentiellement manuelles et laborieuses qu'elles soient réalisées par les scientifiques dans leur domaine de spécialité ou par les spécialistes de la veille sanitaire alors qu'elles sont très utiles notamment pour caractériser des niches écologiques, dresser des cartes d'occurrences, inférer des cartes de risque, ou nourrir des modèles épidémiologiques (Morris et al., 2022).
Notre projet vise à développer des outils pour assister la veille scientifique ou sanitaire à la fois dans la collecte, la formalisation, le partage et la mise à jour de connaissances fiables dans une ambition d'amélioration de la compréhension, de la surveillance et du contrôle des maladies à transmission vectorielle.

Domaine

L'approche retenue consiste à automatiser la mise à jour une base de connaissance initiale avec de nouvelles observations et connaissances extraites de documents au fur et à mesure de leur publication. Cette base de connaissances sera principalement alimentée par l'extraction automatique d'informations à partir de textes, réalisée par des méthodes de traitement automatique de la langue (TAL).

L'exploitation des méthodes de TAL appliquées à la veille scientifique et sanitaire est confrontée à trois difficultés, la qualité variable des prédictions automatiques du TAL et l'abondance et le manque de pertinence de certaines sources. Bien que les progrès récents en TAL soient très considérables, l'horizon de 100% d'informations extraites et interprétées avec succès est loin d'être atteint. De plus, ces outils produisent un très grand nombre d'informations redondantes et plus ou moins pertinentes, qui croît avec l'afflux de documents. Ces informations sont traitées aujourd'hui manuellement pour répondre à l'exigence de grande qualité de l'activité de veille sanitaire.

Nous proposons de développer des méthodes originales d'assistance à la veille par une sélection intelligente et explicable des informations extraites des textes basées sur l'évaluation de leur fiabilité et de leur nouveauté pour en faciliter le tri, l'interprétation et la prise de décision. Il est nécessaire à la fois de rendre compte de l'évolution de la connaissance [ex. rôle très nouveau de l'aulne dans la propagation de la Flavescence dorée de la vigne (Malembic et al., 2020)], tout en disqualifiant les connaissances mal extraites, non confirmées ou fausses [ex. capacité d'une cicadelle à transmettre un phytoplasme du pommier au laboratoire (Hegab & El-Zohairy 1986) jamais confirmée depuis en verger (Fischnaller et al. 2020)]. Le projet de thèse répond à cette double contrainte, particulière à l'écologie. L'approche proposée pour la conception d'un système d'aide à la veille scientifique et sanitaire se situe à l'interface des domaines de la biologie, du TAL, de la représentation des connaissances et raisonnement automatique et de la veille scientifique et technique.

Objectifs

Nous proposons un projet en santé des plantes pour l'extraction et la représentation d'observations et de connaissances relationnelles entre espèces vecteurs, plantes hôtes et agents pathogènes à partir de documents. La structuration de la base de connaissance sera standardisée grâce à des taxinomies et à des ontologies. L'utilisation de référentiels sémantiques partagés permettra leur exploitation conjointe avec d'autres données (observations, données météo, etc.) selon les objectifs du projet ANR BEYOND (Morris et al., 2022) et en cohérence avec les principes FAIR (Wilkinson et al., 2016). Cette partie du projet repose sur l'adaptation aux pathosystèmes à vecteurs de méthodes d'Intelligence Artificielle existantes.

Des approches originales de la mesure de la pertinence et la fiabilité de l'information sont envisagées, telle que l'exploitation d'un faisceau d'éléments complémentaires obtenus par différentes stratégies : des indices extraits du texte d'ordres linguistique et biologique, des indices bibliographiques, et enfin le raisonnement pour estimer la cohérence des informations extraites avec la connaissance biologique préexistante. Nous pensons qu'un nouvel élément important de l'appréciation de la cohérence d'une base de connaissance sur les pathosystèmes en évolution continue est la « propension au changement » et l'acceptabilité de nouvelles connaissances telles qu'elles peuvent être établies par des experts. Pour cela, nous proposons d'ajouter à la représentation des connaissances, une caractérisation de l'évolutivité attendue des interactions biologiques connues. Par exemple, « attendons-nous dans un futur proche une extension rapide des découvertes de nouveaux hôtes de *Xylella fastidiosa*, et dans quelle famille sont-ils plus probablement attendus ? ».

Le choix des modèles biologiques, restreint au groupe des hémiptères vecteurs des huit bactéries phytopathogènes réglementées à l'échelle de l'Europe est guidé par les forts enjeux économiques des maladies ciblées et par leur intérêt pour la recherche en informatique telles que la diversité documentaire et linguistique, et le nombre très variable d'interactions biologiques connues et à découvrir. La question biologique sur la spécificité des couples vecteurs-hôtes servira de fil rouge à l'étude. Nous ambitionnons dès l'initiation du projet une généricité des méthodes pour la veille sur les arthropodes vecteurs (moustiques, tiques, pucerons, ...) dans leur ensemble et de nouvelles approches sur le traitement de la validité et de la nouveauté de connaissances scientifiques d'origine documentaire.

Contexte

Les méthodes d'extraction d'information (TAL) en biologie extraient et agrègent indifféremment les informations, valides ou non, qui représentent des observations, et des connaissances plus générales, avec un degré de précision plus ou moins élevé selon des approches dites d'*ontology learning*, ou *knowledge base*

population (Ji and Grishman, 2011). Nous réutiliserons ces méthodes en distinguant les informations en fonction de leur nature et de leur qualité.

Anomalie ou nouveauté ? Les approches de détection d'anomalie calculent la plausibilité par la détection d'*outliers* par rapport à une distribution mesurée sur un grand nombre de données, par exemple sur le web (Keller et Lapata, 2003), ce qui est inapproprié pour notre objectif de découverte de *nouvelles connaissances*. Dans notre cadre, ces approches sont insuffisantes : le degré de nouveauté dépend certes de la similarité avec les observations passées, mais surtout avec le consensus scientifique.

Véracité. La recherche sur la détection d'informations douteuses ou fausses dans des textes (« fact checking ») s'est concentrée sur les médias et réseaux sociaux, plus récemment dans le domaine médical (ex. COVID-19, allégation de bienfaits de médicaments) en recherchant des preuves dans d'autres phrases (Waden et al., 2020) ou d'autres textes (*wikipédia* pour Fan et al., 2020), mais pas dans des bases externes. Le contexte de la veille scientifique est différent de celui du *fact checking*, il n'y a pas de la part des auteurs d'intention de désinformation ou de l'ignorance, mais un *continuum* entre hypothèse, possible mauvaise interprétation de résultats (ex. conditions expérimentales limitées), non reproductibilité (résultats non confirmés), jusqu'à la remise en cause et l'obsolescence.

Détection. Concrètement, la publication apporte aux veilleurs des indicateurs d'ordre linguistique sur l'intention de l'auteur. L'extraction automatisée de ces éléments linguistiques, documentaires et biologiques pose de nouvelles questions de recherche que nous souhaitons traiter ici. Par exemple, une connaissance nouvelle est marquée par exemple par *first report* ou *new host*, une référence à une connaissance publiée est indiquée par une ou plusieurs citations dont le contexte indique l'intention et l'orientation négative ou positive (Abu-Jbara et al., 2013). L'emplacement de l'information dans les sections d'une publication, le nombre de citations de la publication, la fiabilité du support (journal, rapport officiel, *news*) sont autant d'éléments pris en compte (Aksnes et al., 2019). Dans notre cadre, le protocole expérimental (ex. *in vitro vs in vivo*) et la méthode d'identification de l'espèce sont déterminants.

Evolutivité de la connaissance. Nous proposons une représentation formelle du degré de stabilité de la connaissance, c'est-à-dire de qualifier *a priori* manuellement les relations biologiques des pathosystèmes comme évolutives ou non, et de l'utiliser pour évaluer la vraisemblance de la nouvelle information grâce à un raisonnement basé sur des logiques non classiques. La combinaison des indicateurs issus de la publication et de la validité de la connaissance permettra de définir une mesure intégrée de vraisemblance. Le classement des informations extraites des textes selon ces mesures de vraisemblance et de nouveauté facilitera la décision du veilleur qui décidera d'ajouter ou non l'information à la base de connaissance comme nouvelle occurrence ou nouvelle connaissance, selon sa nature, associée à un degré de plausibilité.

Méthode

Nous proposons de façon originale d'exploiter le contenu de la publication, son contexte historique et bibliographique et les connaissances du domaine pour inférer un degré de vraisemblance de l'information extraite de la publication et automatiser la démarche de veille manuelle.

Pour calculer un degré de nouveauté explicable de l'information extraite par rapport à la connaissance et aux observations, nous proposons une nouvelle approche : représenter explicitement les relations entre connaissances du domaine (ontologie, référentiel sémantique) et information sur les observations d'agents pathogènes, leurs vecteurs et leurs hôtes avec leur localisation et leur date. Concrètement, les informations d'observations structurées en événements représenteront des instances de la connaissance représentées par une ontologie, c'est-à-dire, les taxa (vecteurs, agents pathogènes et hôtes), les informations temporelles ou phénologiques et les informations spatiales seront liés aux référentiels appropriés, suivant en cela *Core Plant Plant Health Threat Ontology* (Alomar et al., 2016). La représentation des connaissances est donc un élément clé du dispositif : les langages du web sémantique seront utilisés de manière classique pour représenter les connaissances, RDF pour les occurrences, OWL (*Web Ontology Language*) pour les connaissances ontologiques, et SKOS (*Simple Organization System*) pour les terminologies. Les connaissances biologiques pourront être représentées par des définitions de classes en OWL, des contraintes en SHACL (*Shapes Constraint Language*) et des règles complémentaires en SWRL (*Semantic Web Rule Language*), suivant en cela le consensus du domaine du web sémantique. De façon complémentaire, les travaux décrits dans (Tettamanzi et al., 2017) sur l'évaluation de la vraisemblance d'axiomes OWL reposant sur la théorie des possibilités et ceux décrits dans (Felin et al. 2023) sur la validation probabiliste de contraintes SHACL constituent une proposition pertinente et adaptable à notre projet pour le traitement de l'évolutivité de la connaissance biologique.

Les étapes du projet sont les suivantes, (1) l'étude bibliographique de l'analyse d'indices de fiabilité et de nouveauté dans les publications et de leur représentation, leur adaptation au domaine et identification de

nouveaux indices spécifiques au domaine. Les approches de TAL incluront les méthodes récentes de *deep learning* à base de modèles de langues pour prédire la pertinence biologique en fonction du contexte (BioBERT (Lee et al., 2020) ou SciBERT (Beltagy et al., 2019) et les méthodes plus récentes qui exploitent la connaissance sous forme de graphe (Hao et al., 2020). L'implémentation des méthodes de TAL choisies sera suivie de tests.
(2) La 2ème partie portera sur la représentation de la base de connaissance biologique et de son évolutivité appropriée au raisonnement en contexte certain et incertain et la définition d'un mécanisme de vérification de cohérence et d'estimation de la vraisemblance.
(3) La combinaison des facteurs de décision des méthodes de TAL et de la représentation des connaissances et évaluation des propositions sur la spécificité des interactions relations psylles-arbres fruitiers fera l'objet de la dernière partie. La pondération des différents critères prendra en compte le compromis temps de curation/risque d'information absente (Abdelrahman, 2020).

Références bibliographiques

Equipes d'accueil

Chaix E., Deléger L., Bossy R., & Nédellec C. (2019) Text mining tools for extracting information about microbial biodiversity in food. *Food Microbiology* 63-75 https://doi.org/10.1016/j.fm.2018.04.011

Felin, R., Faron, C. and Tettamanzi A.G.B. (2023). A Framework to Include and Exploit Probabilistic Information in SHACL Validation Reports. To appear in *Proc. of the 20th International European Semantic Web Conference (ESWC 2023)*.

Ferré A., Bossy R., Ba M., Deléger L., Lavergne T., Zweigenbaum P., & Nédellec C. (2020) Handling Entity Normalization with no Annotated Corpus: Weakly Supervised Methods Based on Distributional Representation and Ontological Information. *Proc. of* LREC-2020, 1959–1966

Marie-Jeanne V., Bonnot F., Thébaud G., Peccoud J., Labonne G., & Sauvion N. (2020) Multi-scale spatial genetic structure of the vector-borne pathogen 'Candidatus phytoplasma prunorum' in orchards and in wild habitats. *Scientific Reports* 10, 5002

Morris C.E., Geniaux G., Nédellec C., Sauvion N., & Soubeyrand S. (2022) One Health concepts and challenges for surveillance, forecasting and mitigation of plant disease beyond the traditional scope of crop production. *Plant Pathology*, 71, 86-97

Sauvion N., Peccoud J., Meynard C., Ouvrard D. (2021) Occurrence data for the two Cacopsylla pruni cryptic species (Hemiptera: Psylloidea) *Biodiversity Data Journal* 9, pp.e68860. ff10.3897/BDJ.9.e68860

Tettamanzi A.G., Faron-Zucker C., & Gandon F. (2017) Possibilistic testing of OWL axioms against RDF data. *International Journal of Approximate Reasoning*, 91, 114-130.

Références

Abu-Jbara, A., Ezra, J., & Radev, D. (2013). Purpose and polarity of citation: Towards NLP-based bibliometrics. In *Proceedings of the 2013 Conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 596-606).

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *Sage Open*, 9(1), 2158244019829575

Alomar, O., Batlle, A., Brunetti, J. M., García, R., Gil, R., Granollers, T., ... & Virgili-Gomà, J. (2016). Development and testing of the media monitoring tool med is YS for the monitoring, early identification and reporting of existing and emerging plant health threats. *EFSA Supporting Publications*, 13(12), 1118E

Arsevska E., Valentin S., Rabatel J., De Goër de Hervé J., Falala S., Lancelot R., Roche M. (2018). Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System. *PloS One*, 13 (8) e0199960.

Barboza P, Vaillant L, Le Strat Y, Hartley DM, Nelson NP, Mawudeku A, et al. (2014) Factors Influencing Performance of Internet-Based Biosurveillance Systems Used in Epidemic Intelligence for Early Detection of Infectious Diseases Outbreaks. *PLoS ONE* 9(3): e90536. https://doi.org/10.1371/journal.pone.0090536

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv*:1903.10676.

Dagan I., Dolan B., Magnini B., and Roth D. (2009). Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii.

Doan S., Ngo Q.-H., Kawazoe A., Collier N. (2008) Global Health Monitor. A Web-based System for Detecting and Mapping Infectious Diseases, *Proc. of the International Joint Conference on Natural Language Processing* (IJCNLP), pages 951-956 https://www.aclweb.org/anthology/I08-2140.pdf

European Food Safety Authority (EFSA), Delbianco A, Gibin D, Pasinato L, Boscia D, Morelli M. (2021). Update of the *Xylella* spp. host plant database - systematic literature search up to 31 December 2021. *EFSA J.* 2022 Jun 15;. doi: 10.2903/j.efsa.2022.7356. PMID: 35734284; PMC9198695.

Fan A., Piktus A., Petroni F., Wenzek, G. Saeidi M., Vlachos A., Bordes A., and Riedel S. (2020). Generating fact checking briefs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (EMNLP), pages 7147–7161, Online. Association for Computational Linguistics.

Färber, M., & Rettinger, A. (2013). A semantic wiki for novelty search on documents. In *Proc. of the 13th Dutch-Belgian Workshop on Information Retrieval* (pp. 60-61).

Ghosal, T., Saikh, T., Biswas, T., Ekbal, A., & Bhattacharyya, P. (2022). Novelty Detection: A Perspective from Natural Language Processing. *Computational Linguistics*, 48(1), 77-117.

Guo Z., Schlichtkrull M., Vlachos A. (2022) A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*; 10 178–206. doi: https://doi.org/10.1162/tacl_a_00454
Hao, B., Zhu, H., & Paschalidis, I. C. (2020). Enhancing clinical BERT embedding using a biomedical knowledge base. In *28th International Conference on Computational Linguistics* (COLING 2020).
Huttunen, S. (2020). *Information Extraction and linguistic characteristics of texts: exploring scenarios and text types*. PhD dissertation, University of Helsinki, Finland.
Ji, H., & Grishman, R. (2011, June). Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 1148-1158).
Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234-1240.
Malembic-Maher, S., Desqué, D., Khalil, D., Salar, P., Bergey, B., Danet, J. L., ... & Foissac, X. (2020). When a Palearctic bacterium meets a Nearctic insect vector: Genetic and ecological insights into the emergence of the grapevine *Flavescence dorée* epidemics in Europe. *PLoS pathogens*, *16*(3), e1007967.
Sarrouti, M., Abacha, A. B., M'rabet, Y., & Demner-Fushman, D. (2021). Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics*: EMNLP 2021 (pp. 3499-3512).
Ouvrard, D. (2022) Psyl'list - The World Psylloidea Database. http://www.hemiptera-databases.com/psyllist - searched on 26 September 2022 doi:10.5519/0029634
Rees, E., Ng, V., Gachon, P., Mawudeku, A., McKenney, D., Pedlar, J., ... & Knox, J. (2019). Early detection and prediction of infectious disease outbreaks. *CCDR*, 45(5).
Abdelrahman, N. S. (2020). *Text Mining for Precision Medicine: Natural Language Processing, Machine Learning and Information Extraction for Knowledge Discovery in the Health Domain* (Doctoral dissertation, Utrecht University).
Sauvion N (2020) Compilation of occurrence data for two psyllid species of the *Cacopsylla pruni* complex (Hemiptera: Psylloidea). 10.15454/VC9UR5, Portail Data INRAE.
Steinberger R, Fuart F, Pouliquen B, Van Der Goot E. (2008) MedISys: A Multilingual Media Monitoring Tool for Medical Intelligence and Early Warning. In Conference Proceedings: Global Risk Forum GRF Davos *Proceedings of the International Disaster and Risk Conference*. Davos (Switzerland). p. 612-614. JRC45523
Thorne J. and Vlachos A. (2018). Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
Vlachos A. and Riedel S. (2015). Identification and verification of simple claims about statistical properties. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2596–2601, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1312
Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., & Hajishirzi, H. (2020). Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (EMNLP), pages 7534–7550, Online. Association for Computational Linguistics.
Wang, S., Durrett, G., & Erk, K. (2018). Modeling semantic plausibility by injecting world knowledge. *arXiv preprint arXiv*:1804.00619.
Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018. https://doi.org/10.1038/sdata.2016.18

Conditions scientifiques matérielles et financières

Pour constituer la base de connaissance de connaissance initiale nous utiliserons les bases de données publiques et en ligne qui décrivent les vecteurs, hôtes, maladies et distribution géographique, pour chacune des huit bactéries réglementées, dont l' *EPPO Global Database*, ISC (*Invasive Species Compendium*) de CABI (*Centre for Agricultural Bioscience International*) et Psyl'list (Ouvrard, 2022). Des listes publiques des communes pour certaines espèces suivies, par exemple Xylella pourront être intégrées à cette base pour l'évaluation de la nouveauté spatiale. Les corpus documentaires sont obtenus à partir des requêtes de veille scientifique et sanitaire des partenaires du projet et celles publiées par l'EFSA par exemple (Delbianco et al., 2022). La plateforme de développement de TAL AlvisNLP de MaIAGE (Dérozier et al., 2022 ; Chaix et al., 2019) et son *workflow* ESV (épidémiologie en santé du végétal) seront utilisés pour extraire les textes des documents, identifier les entités, les normaliser (Ferré et al., 2021) et extraire les relations entre microbes, vecteurs, hôtes, localisations et dates (Tang et al., 2022). L'étudiant.e aura accès aux moyens de calcul et de stockage de l'unité MaIAGE et aux GPU de lab.IA nécessaires aux approches de *deep learning* utilisées pour l'extraction d'information.

Collaborations envisagées

MaIAGE apporte principalement la compétence en TAL appliquée à la santé des plantes, PHIM la compétence en épidémiologie végétale centrée sur les insectes vecteurs et Wimmics apporte la compétence en représentation des connaissances et raisonnement. La plateforme ESV apporte la compétence en veille en épidémiologie végétale.

Les partenaires ont une expérience récente mais significative de collaboration interdisciplinaire dans les projets suivants dont cette thèse est le prolongement : (1) ANR PPR BEYOND *Building epidemiological surveillance & prophylaxis with observations near & distant* (2021-2024) sur l'extraction d'information sur la biologie des psylles des arbres fruitiers (MaIAGE/PHIM/PESV) (2) TIERS-ESV *Traitement de l'Information et Expertise des Risques Sanitaires pour l'Epidémiosurveillance en Santé Végétal* (2020-2022). (MaIAGE/PESV) (3) ANR D2KAB

(2019-2024). *Data to Knowledge in Agriculture and Biodiversity* sur l'intégration de connaissance et de données de cultures, basée sur des ontologies. (MaIAGE/Wimmics).

Objectifs de valorisation des travaux de recherche du doctorant : diffusion, publication et confidentialité, droit à la propriété intellectuelle,...:

En fonction des résultats et du degré d'originalité des méthodes, les conférences de TAL (*Coling*, *EMNLP*, *LREC*, *BioNLP*, *ACL*) et de représentation des connaissances (*EKAW*, *ESWC*, *ISWC*) seront pertinentes. Les journaux tels que *BMC Bioinformatics*, *Plos One* ou *Plant pathology* seront appropriés pour les travaux en biologie.
Les logiciels et données seront publiées sous des licences libres (Apache, CC-BY). Les résultats seront valorisés par la plateforme ESV, partenaire du projet.

Financement du projet doctoral

Le projet ne dispose pas de financement complet au 21/04/2023. 1/2 financement INRAE au 21/04/2023. Plusieurs demandes sont en cours pour financement en 2023.

Candidat

Profil et compétences recherchées

L'étudiant.e possèdera une formation approfondie en IA, TAL et représentation des connaissances acquise en Master. Une expérience de travail significative ou une formation en biologie sera un plus. Il/elle possèdera des compétences solides en développement informatique.

Profile and skills required

The student will have an extensive background in AI, NLP and knowledge representation acquired in a Master's degree. Significant work experience or training in biology will be a plus. He/she will have advanced skills in computer development.

Niveau de français A1
Niveau d'anglais B2