PROPOSITION DE STAGE 2021-2022

Le GIS Fruits souhaite soutenir des stages étudiants de 6 mois, niveau Master 2 sur le thème des fruits et offre pour cela de financer des bourses de stages réalisés dans des labos INRAE. Le sujet proposé doit :

i) s'inscrire dans les axes thématiques du GIS,
ii) être construit en partenariat entre au moins 3 membres du GIS*,
iii) le stagiaire doit être encadré par un maître de stage INRAE.

** Les trois partenaires proposant le stage ne doivent pas appartenir à la même unité.*

>Axes thématiques du GIS : http://www.gis-fruits.org/Le-GIS-Fruits/Axes-thematiques
>Partenaires du GIS : http://www.gis-fruits.org/Le-GIS-Fruits/Membres-fondateurs

Organismes partenaires : (1) INRAE **(2)** *CTIFL* **(3)** *CEP Innovation*
Dont l'école membre du GIS le cas échéant : Institut Agro
Lieux du stage : INRAE Montpellier, PHIM (Plant Health Institute of Plant)
Durée : 6 mois
Dates : 1er semestre 2022
Niveau : Stage de fin d'études BAC + 5 (Option Ingénieur, ou Master 2)
Profil du stage : Recherche appliquée

Date de diffusion : 27-10-2021

INTITULE DU STAGE : Application du text mining à la recherche de données d'occurrences des psylles vecteurs de phytoplasmes des arbres fruitiers.

Contexte et problématique :

Les maladies causées par les phytoplasmes aux arbres fruitiers ont des impacts économiques importants en Europe [A1]. Ces bactéries s'attaquent à différentes rosacées (*Prunus*, pommiers et poiriers), elles sont transmises par des insectes vecteurs psylles du genre *Cacopsylla*, et sont à l'origine de trois maladies : *l'European stone fruit yellows* (ESFY), *l'Apple Proliferation* (AP) et le *Pear Decline* (PD) [A2] Ces bactéries et leurs vecteurs sont originaires d'Europe où ils sont largement présents dans les vergers, ainsi que dans les habitats sauvages, ce qui empêche l'éradication des vecteurs et, par conséquent, l'endiguement des maladies. Les psylles vecteurs sont contrôlés principalement par des insecticides, mais l'évolution des pratiques agricoles (ex. réduction de l'utilisation des pesticides dans le cadre du plan EcoPhyto en France) et les réglementations européennes (ex. pathogènes retirés de la liste des organismes de quarantaine) pourraient être, voire sont déjà, la source de nouvelles émergences. Malgré les efforts de la recherche pour mieux comprendre la biologie et l'écologie des psylles vecteurs (ou potentiels vecteurs) de phytoplasmes (Action COST FA0807 2013, [A3-A8], la présence de ces insectes dans certaines parties de l'Europe, et même dans d'autres parties du monde touchées par ces maladies, reste incertaine [A9]. Or, les occurrences représentant l'étendue et la variabilité dans l'aire de répartition actuelle d'une espèce donnée sont essentielles pour caractériser et cartographier sa distribution potentielle dans le cadre de scénarios d'introduction accidentelle ou de changement climatique. Réduire cette incertitude sur la distribution géographique des vecteurs permettrait de mieux évaluer les risques posés par les phytoplasmes des arbres fruitiers et d'aider à la prise de décisions pour gérer ces risques à différentes échelles spatiales [A3].

Les modèles de distribution des espèces (SDM) sont devenus le principal outil de prédiction pour atteindre cet objectif. Les SDM ont prouvé leur utilité, entre autres, en biologie des invasions et en biologie de la conservation. En pathologie végétale, les SDM sont également de plus en plus utilisés pour prédire les distributions potentielles des phytopathogènes vectoriels. Cependant, la fiabilité de ces modèles dépend fortement de la qualité des données d'occurrence qui sont utilisées en entrée. Or, obtenir des données de haute

qualité pour cartographier correctement la distribution d'une espèce est tout sauf une sinécure. Jusqu'à récemment, la démarche pouvait se résumer à un travail laborieux de 'fouille de texte' manuelle s'apparentant à une chasse au trésor avec ses difficultés (ex. accès aux références très anciennes), ses pièges (ex. erreurs de traduction) et ses énigmes typiques (ex. synonymie des noms d'espèces, confusion dans les noms de localités). Il nous aura ainsi fallu plusieurs années pour rassembler les données d'occurrences des deux espèces de psylles du complexe d'espèce *Cacopsylla pruni*. Nous avons récemment publié notre démarche, la base de données des 1975 occurrences rassemblées et les cartes générées [A10, A11].

Le web a démultiplié dans des proportions gigantesques la disponibilité de documents publiés et stockés numériquement y compris des documents très anciens (ex. manuscrit de Scopoli 1763 décrivant pour la première fois *C. pruni*), rendant la perspective d'une exploration systématique manuelle inatteignable.

Le text mining, ou plus précisément l'extraction automatique d'informations a pour objectif d'extraire et de structurer les informations contenues dans ce type de documents grâce à la mise en œuvre de techniques statistiques ou de machine learning, de traitement automatique de la langue et d'ingénierie de la connaissance. La grande diversité du vocabulaire utilisé a fait de la normalisation automatique des mentions du texte en fonction d'un référentiel, une étape majeure de cette extraction [A12, A14]. Un exemple en est le rattachement des mentions textuelles 'European stone fruit yellows', 'mycoplasma-like organism' ou 'European prunus phytoplasmas' à la référence taxonomique '*Candidatus* Phytoplasma prunorum'. Ainsi, la base Florilege agrège et normalise les informations obtenues automatiquement par extraction d'information à partir de plus de 3 millions de documents sur les microbes et leurs habitats - dont les phytoplasmes et leurs vecteurs (http://migale.jouy.inrae.fr/florilege/#&searchByTaxon=Candidatus%20Phytoplasma), [A13, A15]. En épidémiosurveillance, cette approche commence à être utilisée notamment par la Plateforme d'Epidémiosurveillance en Santé Végétale comme outil de veille (https://plateforme-esv.fr/).

Les résultats du text mining, combinés avec d'autres informations (ex. analyses des flux longs distances : air, échanges commerciaux, etc), dans des modèles prédictifs ouvre des perspectives très prometteuses pour une meilleure anticipation à court terme des décisions prophylactiques et pour une meilleure connaissance du potentiel de circulation des agents pathogènes. C'est l'objectif du projet ANR BEYOND (2021-2026, https://www6.inrae.fr/beyond/) dans lequel s'inscrit ce stage M2.

Objectifs généraux du stage / Résultats attendus :

Obj.: constituer une base de données aussi exhaustive que possible des données d'occurrences des psylles vecteurs de phytoplasme des arbres fruitiers par une approche text mining.

Le projet de Master porte sur l'extraction automatique et la modélisation de connaissances à partir de données textuelles pour obtenir des occurrences datées et géolocalisées. Le stage sera ciblé sur les données d'occurrences de trois espèces de psylles vecteurs.

Le travail s'inscrira dans le cadre plus général de l'extraction d'informations plus larges tels que les habitats de phytopathogènes (vectés ou non), les plantes hôtes, leurs phénotypes, les conditions d'habitabilité ou les maladies développé dans le projet BEYOND par l'unité MAIAGE et PESV. Un pipeline général de text mining inspiré de celui de Florilege et basé sur la plateforme AlvisNLP est en cours de développement dans l'unité MaIAGE (équipe Bibliome et plateforme bioinformatique Migale) (https://github.com/Bibliome/alvisnlp) [A16]. AlvisNLP exploite des méthodes d'apprentissage automatique supervisé basées sur des architectures neuronales et des approches à base de règles exploitant des informations linguistiques, lexicales, terminologiques et conceptuelles (thésaurus, nomenclatures, ontologies) [A17]. Le travail consistera à intervenir sur ce pipeline pour en améliorer la qualité sur la question spécifique des psylles.

Il s'agira dans un premier temps d'évaluer les résultats actuels produits par la plateforme en fonction des résultats attendus, puis d'identifier les étapes à consolider et enfin d'intervenir sur les étapes présentant le meilleur compromis investissement / qualité. Plusieurs dimensions complémentaires sont envisagées, (1) l'enrichissement et potentiellement la restructuration des thesaurus utilisés pour décrire les psylles, les espèces hôtes et les phytoplasmes, (2) l'exploitation de nouvelles sources documentaires, et (3) l'ajout de règles améliorant le traitement des ambiguïtés.

Trois résultats principaux sont attendus : (i) en nous appuyant le cas d'école 'Cacopsylla pruni' , le stagiaire mettra en œuvre du text ming et pourra ensuite comparer les avantages de cette approche (ex. rapidité/facilitation d'accès à l'information) et ses inconvénients par rapport à une recherche bibliographique manuelle classique ; (ii) fort de ce retour d'expérience croisée, le stagiaire constituera une base de données d'occurrences des psylles vecteurs des phytoplasmes responsables de l'*Apple Proliferation* et du *Pear Decline*. Ces résultats seront rendus librement accessibles au travers une base de donnée déposée dans le DataServe INRAE [A11], avant d'être publiés comme nous l'avons fait pour *C. pruni* [A10]. A terme, ces informations serviront à modéliser les aires de distributions actuelles, potentielles et futures des vecteurs, et à cartographier

les zones à risque pour les pommiers et les poiriers. Ce travail est en cours de publication pour les Prunus ; (iii) les développements du stagiaire en text mining et son analyse approfondie des limitations et avantages de l'approche automatisée ouvrira des perspectives de généralisation à d'autres pathosystèmes (projet BEYOND).

Ce travail nécessitera une collaboration étroite entre un entomologiste spécialiste des insectes vecteurs d'une part (Nicolas Sauvion, unité PHIM, https://umr-phim.cirad.fr) et l'unité MaIAGE, https://maiage.inrae.fr/) dont une spécialiste de la recherche en extraction d'information orientée vers la connaissance (Claire Nédellec) et Robert Bossy, le responsable de la plateforme AlvisNLP qui apportera une formation et plus généralement l'accompagnement à son usage. Le stagiaire recevra également le soutien ponctuel de MaIAGE, Louise Deléger (CR) pour la partie text mining du pipeline et Mouhamadou Ba (IR) pour sa mise en œuvre informatique.
Le stage sera principalement localisé à PHIM et des visites ponctuelles à MaIAGE particulièrement en début de stage, seront programmées pour former le stagiaire. Des réunions régulières en visio complèteront le dispositif.

Publications de l'équipe d'accueil et/ou relative au sujet (et/ou au projet dans lequel s'insère le stage) :

[A1] Hadidi A, Barba M, Candresse T, Jelkmann W (2011) Virus and virus-like diseases of pome and stone fruits. The American Phytopathological Society Press, St Paul, Minnesota. [ISBN 978-0-89054-396-2] https://doi.org/10.1094/9780890545010

[A2] Jarausch B, Tedeschi R, **Sauvion N**, Gross J, Jarausch W (2019) Chapter 3: Psyllid vectors. In: Bertaccini PW, Rao GP, Mori N (Eds) Transmission and management of phytoplasma associated diseases. II. Springer, Singapore. [ISBN 978-981-13-2831-2]. https://doi.org/10.1007/978-981-13-2832-9_11

[A3] MacLeod A. (…), **Sauvion N.,** et al (2012) Pest risk assessment for the European Community plant health: a comparative approach with case studies. *Supporting Publications 2012*:EN-319 [1053 pp.] Available online: https://efsa.onlinelibrary.wiley.com/doi/pdf/10.2903/sp.efsa.2012.EN-319

[A4] Thébaud G., Yvon M., Alary R., **Sauvion N.** & Labonne G. (2009) Efficient transmission of '*Candidatus* Phytoplasma prunorum' is delayed by eight months due to a long latency in its host-alternating vector. *Phytopathology* 99: 265-273. https://apsjournals.apsnet.org/doi/10.1094/PHYTO-99-3-0265

[A5] **Sauvion N.,** Lachenaud O. Mondor-Genson, G., Rasplus J-Y. & Labonne, G., (2009). Nine polymorphic microsatellite loci from the psyllid *Cacopsylla pruni* (Scopoli), the vector of European stone fruits yellows. *Molecular Ecology Resources* 9: 1196-1197. https://doi.org/10.1111/j.1755-0998.2009.02604.x

[A6] Peccoud J., Labonne G., **Sauvion N.** (2013) Molecular tests to assign individuals within the *Cacospylla pruni* complex. *PLoS ONE* 8: e72454. https://hal.archives-ouvertes.fr/hal-01922692

[A7] Peccoud J., Pleydell D.R.J., **Sauvion N.** (2018) A framework for estimating the effects of sequential reproductive barriers: implementation using Bayesian models with field data from cryptic species. *Evolution* 72-11: 2503-2512 doi:10.1111/evo.13595 https://onlinelibrary.wiley.com/doi/pdf/10.1111/evo.13595

[A8] 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 https://doi.org/10.1038/s41598-020-61908-0

[A9] Steffek, R., Follak, S., **Sauvion, N.,** Labonne, G., MacLeod, A. (2012) Distribution of '*Candidatus* Phytoplasma prunorum' and its vector *Cacopsylla pruni* in European fruit growing areas: a review. *EPPO Bulletin* 42: 191-202. https://doi.org/10.1111/epp.2567

[A10] **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 https://hal.archives-ouvertes.fr/hal-03230951v2

[A11] **Sauvion N.** (2020) Compilation of occurrence data for two psyllid species of the *Cacopsylla pruni* complex (Hemiptera: Psylloidea). https://doi.org/10.15454/VC9UR5, Portail Data INRAE, V6

[A12] Ferré, A., Zweigenbaum, P., & **Nédellec, C.** (2017). Representation of complex terms in a vector space structured by an ontology for a normalization task. In Proceedings of the BioNLP 2017 Workshop, Association for Computational Linguistics, Vancouver, 8 pages, Canada 2017. https://www.aclweb.org/anthology/W17-2312.pdf

[A13] Chaix, E., Deléger, L., Bossy, R., & **Nédellec, C.** (2019). Text mining tools for extracting information about microbial biodiversity in food. *Food microbiology*, 81, 63-75. https://doi.org/10.1016/j.fm.2018.04.011

[A14] Ferré, A., Deléger, L., Bossy, R., Zweigenbaum, P., **Nédellec, C.** (2020) C-Norm: a neural approach to few-shot entity normalization. *BMC Bioinformatics* 21, 579 https://doi.org/10.1186/s12859-020-03886-8

[A15] Sandra Dérozier, Louise Deléger, Estelle Chaix, Reda Mekdad, Mouhamadou Ba, **Robert Bossy**, Delphine Sicard, Valentin Loux, Hélène Falentin, **Claire Nédellec**. Florilege, a database gathering microbial habitats, phenotypes and uses. *Jobim 2020*, 30 juin 2020 https://hal.archives-ouvertes.fr/hal-01827946/file/Florilege_JOBIM2018.pdf

[A16] Mouhamadou Ba and **Robert Bossy**. Interoperability of corpus processing workflow engines: the case of. AlvisNLP/ML in OpenMinTeD. LREC Workshop on Cross-Platform Text Mining and Natural Language, Portoroz May 2016. https://interop2016.github.io/pdf/INTEROP-4.pdf

[A17] Zorana Ratkovic, Wiktoria Golik, Pierre Warnier: Event extraction of bacteria biotopes: a knowledge-intensive NLP-based approach. *BMC Bioinformatics* 13(S-11): S8 (2012). https://doi.org/10.1186/1471-2105-13-S11-S8

ACTIVITES DOMINANTES CONFIEES AU STAGIAIRE :

- Enrichissement d'ontologies spécifiques à la question posée de l'extraction d'informations épidémiologiques
- Contribution à la définition d'un corpus documentaire
- Mise en œuvre d'un pipeline de text mining dédié au projet
- Constitution d'un base de données normalisée en vue de son dépôt dans un DataServe.

PROFIL REQUIS :

- Dernière année de Formation Supérieure BAC + 5
- Compétences en Science de la Vie avec au moins l'une des spécialités suivantes : écologie microbienne, entomologie, santé des plantes, systématique
- Compétences en informatique : systèmes d'information, programmation dans un langage parmi Java, Python, ou R.
- Une expérience en traitement automatique de la langue, en apprentissage automatique, en représentation des connaissances serait un plus.
- Un fort intérêt pour le travail pluridisciplinaire et l'analyse de l'écrit sont souhaités
- Aptitude au travail en équipe
- Aisance à communiquer oralement.
- Langues : bonne maitrise de la lecture de l'anglais (lu).
- Permis de conduire (le cas échéant) : pas nécessaire

INDEMNISATION (SUR BUDGET INRAE-GIS FRUITS) :

Selon la règlementation en vigueur pour 2022 (environ 628 €/mois)

AVANTAGES PROPOSES (le cas échéant) :

- logement : le labo d'accueil aidera à trouver un logement
- restauration : restauration d'entreprise sur place

CONTACT MAITRE DE STAGE INRAE :

Nicolas SAUVION, Ingénieur de Recherche
nicolas.sauvion@inrae.fr
https://cv.archives-ouvertes.fr/nicolas-sauvion

PHIM (Plant Health Institute of Montpellier)
CIRAD TA A-54 K,
Campus International de Baillarguet
34398 Montpellier cedex 5
Tél. : 06 31 50 63 89
https://umr-phim.cirad.fr/recherche/comprendre-les-epidemies-dans-les-champs-prism/equipe-forisk