Psylve - A Text-to-Ontology Information Extraction Framework for the Occurrence Distribution of Plant Pathogen Vectors

Elisa Lubrini

UNIVERSITÉ DE LORRAINE

MASTER'S THESIS

Psylve - A Text-to-Ontology Information Extraction Framework for the Occurrence Distribution of Plant Pathogen Vectors

Author:
Elisa LUBRINI

Supervisors:
Claire NÉDELLEC
Nicolas SAUVION
Reviewer:
Marianne CLAUSEL

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science in Natural Language Processing

in the

Plant Health Institute of Montpellier (PHIM)
Institut du Digital, Management et Cognition (IDMC)

March - August 2022

Declaration of Authorship

I, Elisa LUBRINI, declare that this thesis titled, "Psylve - A Text-to-Ontology Information Extraction Framework for the Occurrence Distribution of Plant Pathogen Vectors" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Date: 20/08/2022

UNIVERSITÉ DE LORRAINE

Abstract

Master of Science in Natural Language Processing

Psylve - A Text-to-Ontology Information Extraction Framework for the Occurrence Distribution of Plant Pathogen Vectors

by Elisa LUBRINI

Diseases due to insect-borne plant pathogens have a large negative effect on the world's agricultural industry. An effective way to anticipate disease outbreaks can be to infer risk maps of vector introduction and spread from known occurrence data. However, compiling this type of data manually is time consuming and laborious, especially due to the recent spike in publicly available data. To address this issue, this work describes attempts at facilitating researchers' workflows by using approaches to automate the extraction of vector related information from literature.

To carry out this automation, we developed PsylVe, a solution initially targeted at psyllid vectors that encompasses document recollection, Natural Language Processing (NLP) and Knowledge Representation (KR) techniques. PsylVe includes a working NLP pipeline, and a fully documented methodology. The NLP pipeline is based on the adaptation of an existing pipeline, Omnicrobe, on microbial biodiversity that bears many similarities with epidemic events.

We conducted a quantitative (precision, recall, and F1-score) and qualitative (six qualitative criteria for text mining pipeline evaluations) evaluation of results obtained with PsylVe and compared them to a manually compiled dataset of observations on *Cacopsylla pruni* responsible for the spread of a pathogenic bacterium in fruit tree orchards in Europe. From the outset, we designed the PsylVe Framework to be transferable to other plant disease vectors, as well as human and animal diseases. We have also designed an application for the extraction of texts from PDF documents and an original formal ontology that enables the representation of the data and of the knowledge on vector-borne diseases. Various projects in the MaIAGE department of INRAE have already started integrating the PsylVe framework in their workflow and concrete plans to develop it further were made in order to expand its usage to new biological domains.

Acknowledgements

I would like to take this opportunity to express my deepest gratitude to the outstanding people whose support and dedication made the journey of this Master's Degree so valuable and enriching.

First and foremost, I would like to thank my thesis reviewer **Marianne Clausel** for her encouraging support as well as the coordinators of the MSc Degree in Natural Language Processing **Maxime Amblard** and **Miguel Couceiro** for making this journey possible in the first place and going the extra mile to make the overall Master's experience as meaningful as possible.

I would also like to thank the aforementioned people in their quality of teachers for their extreme dedication to the respective subjects and the contagious passion they put into their work, which has been a source of inspiration during my studies.

Within the context of my internship, I would like to thank my supervisors: **Nicolas Sauvion**, an unending source of galvanising enthusiasm and remarkable domain expertise, and **Claire Nédellec**, whose distinct management skills and breadth of expertise made this interdisciplinary work possible.

Both tenaciously supported me during every phase of my internship and considerably contributed to the quality of my work.

Additionally, I would like to acknowledge the support of **Robert Bossy** for his technical expertise and dedication, and **Catherine Faron Zucker** for her help and availability for collaboration for future projects.

A special thank you goes to **Anna Mosolova**, my project partner during this whole degree, whose tireless ambition and industriousness greatly contributed to our academic achievements as a team.

Last, but not even remotely close to being the least, of superlative value was the support of: **William Soto**, **Eduardo Vallejo**, and **Luis Vasquez**, the combination of whose vast expertise and didactic aptness skyrocketed my technical abilities and my grasp of computer science topics, during these past two years, and **Shane Kaszefski** and **Dimitra Niaouri**, for sharing their immense knowledge of linguistic topics.

It is to all of their continuous support and collaboration that I owe a huge amount of both personal and academic achievements. I will be forever grateful to them for regularly engaging in inspiring, intellectually stimulating discussions and for their immense moral support as fellow students and great friends.

I would also like to thank all the other fellow students, teachers, friends and family members who, with immense patience and constancy, supported me in this intense but rewarding journey of mine.

Contents

List of Figures

List of Tables

Glossary

abiotic	Abiotic variables include all elements playing a role in the distribution of diseases that do not possess genetic material (as opposed to bacteria and viruses). Examples of abiotic variables are elements such as wind, water and heat.
AlvisNLP	AlvisNLP is a software for the assembling of NLP pipelines developed at MaIAGE, INRAE.
anaphora	Anaphora is a linguistic phenomenon where a same entity is referred to multiple times, possibly via a different wordform (Liddy, 1990).
Bash	Bash (Bourne Again Shell) is a Unix shell and command language. (Loshin, 2021).
BEYOND	BEYOND (2021- 2026) is a project funded by ANR.
biodiversity	Biodiversity, a contraction of the term "biological diversity" refers to the biotic variation that extents from all genes to ecosystems, (Wilson et al., 1988).
concept lattice	In graph theory, a concept lattice consists of a partially ordered set in which every pair of elements has a unique supremum (also called a least upper bound or join) and a unique infimum (also called a greatest lower bound or meet). They can be used to perform both inference and induction. (Belohlávek, 2008).
Deep Learning	Deep Learning is a part of the broader family of Machine Learning. Its methodology is based on deep artificial neural network architectures with representation learning. It is also a critical component of data science, which also covers statistics and predictive modelling. The aim of Deep Learning is to imitate the way that specific types of knowledge are acquired by humans (Burns and Brush, 2021).

disease surveillance	Disease surveillance is an information-based activity involving the collection, analysis and interpretation of large volumes of data originating from a variety of sources. The information collated is then used in a number of ways to evaluate the effectiveness of control and preventative health measures (HPSC, 2019).
ecological preference	Ecological preference refers to the associations of species with habitats and with the food they consume (Underwood, Chapman, and Crowe, 2004)
epidemiology	Epidemiology is the scientific, methodological, and data-driven study of the causes of health outcomes and illnesses in defined populations (Disease Control and Prevention, 2021).
F1-score	F1-score is a measurement whose primary application is to compare the performance of two classifiers. It is calculated by taking the harmonic mean of a classifier precision and recall (Team, 2022).
FCA	Formal Concept Analysis) (FCA) is a method to derive partial hierarchies among elements and the respective attributes (Ganter and Wille, 2012). An extension of FCA is constituted by *pattern structures* (Belfodil, 2019), a tool used for the database analysis in Section 4.1.2, useful in the identification of subgroups (Lumpe and Schmidt, 2015).
First Order Logic	First order logic is a subset of formal logic in mathematics. As opposed to propositional logic, first order logic covers predicates and quantification to define axioms and formulas. Each of its statements or phrases is broken down into a subject and a predicate. (Contributor, 2005).
geographical distribution	Geographical distribution refers to the natural arrangement and apportionment of the various forms of animals and plants in the different regions and localities of the earth (Merriam-Webster, 2022a).
GIS Fruits	GIS Fruits is a scientific interest group in the French fruit industry, involved in research, development, training and professional organization, with the aim of implementing a long term common strategy for research and agricultural innovation. It finances the internship whose work is described in the current thesis.
host plant	In the context of psyllid habitats, a host plant is a plant on which the insect completes its immature to adult life cycle. (Burckhardt et al., 2014).

Machine Learning	Machine learning (ML) is a type of AI method that allows software applications to become more accurate at predicting outcomes by "learning" how to predict from input data (Mitchell et al., 1990).
MaIAGE	MaIAGE stands for Mathématiques et Informatique Appliquées du Génome à l'Environnement (Applied Mathematics and Informatics from Genome to the Environment). It is a laboratory that brings together mathematicians, computer scientists, bioinformaticians, and biologists to work on biology, agronomy, and ecological challenges.
natural language	Natural language refers to any language that, as opposed to artificial ones, has evolved naturally among humans via use without deliberate planning or premeditation.
normalisation	Text normalisation refers to the process of reducing wordforms in a text to their base form (Baggia et al., 2010), for example by obtaining the root of a verb across conjugations or, in some languages, that of a modifier without its gender agreement.
Omnicrobe	Omnicrobe is an application that includes an NLP workflow and an open-access database of microbial habitats and phenotypes that employs extensive text mining and data fusion techniques (Dérozier et al., 2022a).
ontology	Ontologies are a formal device to better represent natural language and organise knowledge expressed by it (Alatrish, Tošić, and Milenković, 2014).
partially ordered set	A partially ordered set (or *poset*) is a set taken together with a partial order on it. Formally, a partially ordered set is defined as an ordered pair $P = (X, <=)$, where X is called the ground set of P and $<=$ is the partial order of P.
pathogen	Pathogen is a specific causative agent (such as a bacterium or virus) of disease (Merriam-Webster, 2022b).
pest	As a general term, pest refers to an organism that is undesirable due to its destructive nature. In biological terms it often refers to crops and animals. Insect pests in agricultural systems are a major source of crop production and storage damage globally (Dangles et al., 2009).

phylogenetic tree	In evolutionary biology, phylogenetic tree refers to a tree showing phylogenetic relationships of a set of taxa.
phytoplasma	Phytoplasma are bacteria causing diseases in plants with consequences ranging from yellowing of the leaves, to death.
pipeline	A pipeline in computer science, is a group of data processing components connected in sequence, each of which feeds data into the next component.
polysemy	Polysemy is the linguistic phenomenon where a token corresponds to an 'ambiguous word' bearing more than one possible meaning, usually determined by context.
precision	Precision is a measure of a machine learning model performance. It measures the quality of a positive prediction provided by the model. It is calculated by dividing the number of true positives by the total number of positive predictions, or else the number of true positives plus the number of false positives (Arora, Kanjilal, and Varshney, 2016).
psyllid	Psyllids or jumping plant-lice constitute the superfamily Psylloidea Latreille, 1807 of the hemipterous Sternorrhyncha Duméril, 1806 with worldwide about 4000 described species (Burckhardt, Ouvrard, and Percy, 2021).
PsylVe	PsylVe (March-August 2022) is an internship project hosted by INRAE. Within the context of this thesis, the name can refer to the PsylVe Framework, the tool developed, the ontology, or the overall project.
recall	Recall is a measure of a machine learning model performance. It measures the number of relevant elements retrieved by the model. In order to be calculated the number of true positive elements is divided by the number of the relevant elements (Arora, Kanjilal, and Varshney, 2016).
regex	Regular expressions are rules based on regular grammars that allow to recognize if a given input matches an expected pattern. A classical example is the use of rules to recognize string patterns.
shelter plant	In the context of psyllid habitats, a shelter plant is a plant on which the insect overwinters and occasionally feeds (Burckhardt et al., 2014).
species complex	A species complex is typically considered as a group of close, but distinct species (Brown, Frohlich, and Rosell, 1995).

state of the art	The term "state of the art" refers to the most recent/up-to-date version of a given technology, normally the best performing at any given time.
structured data	Structured data is a standardized format for presenting information and categorising it. HTML tables, HTML lists, and back-end Deep Web databases are all examples of structured data on the Web (Cafarella, Halevy, and Madhavan, 2011).
supervised Machine Learning	Supervised Machine Learning is the process of providing labelled data to the Machine Learning algorithm.
taxon	A taxon is any of the groups composing a taxonomy, independently from the rank (species, family, class, etc.). Examples of taxons used in this thesis are *C. pruni* - a species complex of psyllids species - and *Ca.* P. prunorum a species corresponding to a pathogen.
taxonomic nomenclature	Taxonomic nomenclature is a formal method of naming that is used to identify taxonomic groupings. Accurate and suitable application of nomenclature is often used to minimize ambiguities and errors (Rivera et al., 2014).
taxonomy	The study targeting the definition of groups of biological organisms based on shared characteristics, their naming and hierarchical structuring. It includes the development of phylogenetic trees which group related organisms and show the genetic relationships between them.
text mining	Text mining refers to the process of examining collections of documents to discover new information or help answer specific research questions. In the field of NLP, it includes tasks such as information retrieval, information extraction and question answering.
token	A token is the smallest unit which a corpus is made up of. It normally refers to punctuation and any unit of text separated by either spaces or punctuation itself, although some common tokenisation methods aim more generally at separating morphemes, whether concatenated or not. (Graën et al., 2018; Webster and Kit, 1992).
turtle	Turtle is a syntax of a language that allows non-XML serialization of RDF models, characterised by compactness and human readability.

unstructured data	Unstructured data refers to data that are not kept in a structured database format (Eberendu, 2016). Unstructured data often consists of bitmap images/objects, free text, email, and other non-database data formats (Feldman, Sanger, et al., 2007).
vector	A vector is any organism that acts as a carrier of an infectious agent from one species to another (Wilson et al., 2017).
Western Palaearctic	Western Palaearctic corresponds to the Western region of one of the eight biogeographic realms dividing the Earth's surface.
XML	Extensible Markup Language is a markup language and file format that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable.

Abbreviations

C. pruni	*Cacopsylla (Thamnopsylla) pruni* (Scopoli, 1753)
Ca. P. prunorum	'*Candidatus* Phytoplasma prunorum'
AI	Artificial Intelligence
ANR	Agence nationale de la recherche
AP	Apple Proliferation
ESFY	European stone fruit yellows
FCA	Formal Concept Analysis)
HTML	HyperText Markup Language
IE	Information Extraction
INRAE	National Research Institute for Agriculture, Food and Environment
IPCC	The Intergovernmental Panel on Climate Change
KB	Knowledge Base
KR	Knowledge Representation
MaIAGE	Mathématiques et informatique appliquée du génome à l'environnement
MIREOT	Minimum Information to Reference an External Ontology Terms
NCBI	National Center for Biotechnology Information
NER	Named Entity Recognition
NLP	Natural Language Processing
OCR	Optical Character Recognition
OWL	Web Ontology Language
PD	Pear Decline
PDF	Portable Document Format
PESV	Plateforme d'Epidémiosurveillance en Santé Végétale
PHIM	Plant Health Institute of Montpellier
PIA	Programme d'Investissements d'avenir
POS	Part Of Speech
RDF	Resource Description Framework

RE	Relation Extraction
SKOS	Simple Knowledge Organization System
SPARQL	SPARQL Protocol and RDF Query Language (recursive acronym)
XML	Extensible Markup Language

Chapter 1

Introduction

1.1 Research Context

Vector insects give rise to serious sanitary crises all over the world. Such an issue is particularly relevant nowadays as global changes, such as intensification of international exchanges and climate change, are favouring the rise of new diseases (Jarausch et al., 2019). Modelling the geographical distribution of vector insects is an increasingly popular approach in the field of plant pathogen risk forecasting (see Figure 1.1 for a visual representation of occurrence data). However, the output quality of forecasting models heavily relies on the quality of the input occurrence data, which is particularly labour intensive and time consuming to obtain.

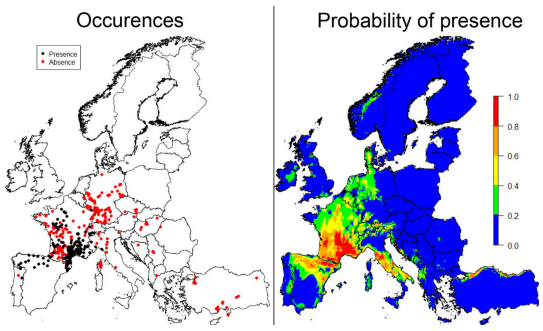


FIGURE 1.1: Theoretical geographical distribution map obtained for the psyllid *Cacopsylla (Thamnopsylla) pruni* (Scopoli, 1753), species A, from occurrence (presence/absence) data (Sauvion et al., 2021).

An example of such data was collected by the INRAE Montpellier team, which published a database of occurrence data of a vector insect, carrier of bacteria causing a disease of fruit plants in Europe (Sauvion et al., 2021). More precisely, the observed vectors are two psyllids of the *C. pruni* species complex (see Section 2.1 for more details on the biological aspects). Most of such occurrence data was manually extracted and structured from scientific documents.

The launch of an ANR project BEYOND in 2021 was the occasion for the Montpellier team to meet with researchers specialized in text mining from the MaIAGE unit of *INRAE Jouy-en-Josas*. This encounter gave rise to an idea for a NLP Master's internship proposal aimed at exploring possible solutions for the integration of text mining techniques in the automation of the occurrence data collection process. and then comparing their advantages (e.g. speed/ease of access to information) and their disadvantages/limitations in relation to traditional manual bibliographic research.

The internship constitutes a interdisciplinary work, co-supervised by a biologist-entomologist, Nicolas Sauvion (INRAE-PHIM, Montpellier) and NLP-specialist, Claire Nédellec (INRAE-MaIAGE, Jouy-en-Josas), in close collaboration with Robert Bossy (INRAE-MaIAGE, Jouy-en-Josas).

1.2 Objectives

The objective of the internship was originally set to the adaptation of an existing NLP pipeline for the automatisation of the extraction of *C. pruni* occurrence data, namely Information Extraction (IE), starting from a variety of documents (see Section 3.3 for a description of the document database) and comparing its performances to the manual approach. It was conceived as a very first step in the exploration of the specificities of the tasks. However, more ambitious results were considered thanks to the intern's keen interest in the biological domain and a swift progress during the initial stages. The original plans were thus outdone by the creation of a longer term project, named PsylVe, whose initial stage is described in the current thesis. The project was subdivided in three main stages, each characterised by an increasingly wide scope with respects to the one preceding it, as outlined in Table 2.1. The project established an Information Extraction (IE) framework optimised for reuse by projects in the disease surveillance domain, whose implementation choices were supported by extensive testing. Exceeding the expectations of the initial objectives, such framework included a thorough analysis of the results of the IE pipeline including an outline of working solutions for future improvements and a Knowledge Representation of the biological issue, allowing to set the bases for a reasonment model. All aspects of the framework were addressed and solid bases were set for future development.

FIGURE 1.2: PsylVe logo

1.3 Contents of this document

The current thesis presents the proposed solution to the automation of the occurrence data collection described above. Such solution took the form of a project planned to span a longer period of time than the internship itself (see Figure 1.3 for a project timeline). Such project was named PsylVe, inspired by two main concepts motivating it: *psyllid* insects and their role as *vector* of pathogens.

The following chapters will be organised as follows. Chapter 2 will offer some background to the problems both from the perspectives of the subjects involved in it, namely Biology, Natural Language Processing (NLP) and Knowledge Representation (KR); Chapter 3 will provide an overview of works and tools related to the project; Chapter 4 will dive into the methodological aspects of the project, emphasising the modularity of the overall framework and the ease of replication of each

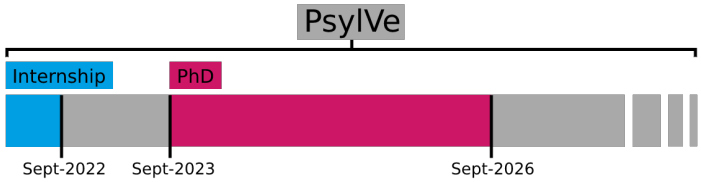


FIGURE 1.3: Expected timeline of the PsylVe project

step, even in the case of integration into a different system; Chapter 5 will present a few experiments to better illustrate the outcome of the various pipeline modules and evaluate the overall outcomes; finally, Chapter 6 will draw the conclusions and describe the plan for future development of the project.

Given the high level of interdisciplinarity of this project and, consequently, the heterogeneity of the audience to which this document is addressed, a didactic effort was made with the assumption that each of the introductory topics proposed will be unfamiliar to at least one reader. Therefore all explanations, including those related to the Master's field of study, were approached from a beginner's standpoint. And all the field-specific terminology, abbreviations, and notation were gathered into a glossary which can be consulted at the beginning of the document, with the objective of making this thesis as accessible as possible independently from the reader's background.

Regarding the terminology used to refer to PsylVe specific concepts, they will be referenced as follows.

The Internship - The employment framework encompassing the work carried out by the author of the current thesis within a 6 month time period, from March to August, 2022.

The PsylVe project - The project whose scope encompasses both the current internship and future employment frameworks based on what outlined in the current thesis (see Figure 1.3 for a visual representation).

The PsylVe Framework - A computational framework for occurrence data extraction (currently optimised for psyllid occurrence data, but applicable to other domains), including a working pipeline, fully documented methodology and directions for further development as new pipelines are created.

Framework module - In the context of the PsylVe Framework, a "module" is either each of the main tasks composing it, with its relative subtasks (see Figure 1.4). At the time of writing, the Framework, is constituted of five modules: namely *database assembling and analysis*, *data preprocessing*, *knowledge base development*, *information extraction*, and *evaluation*.

The PsylVe Pipeline - The pipeline developed for the PsylVe project, based on the PsylVe Framework. Pipelines based on the PsylVe Framework can use the PsylVe Pipeline as reference for a fully working example.

Pipeline module - In the context of the PsylVe Pipeline, a "module" is either each of the main tasks composing it. At the time of writing, the Pipeline, is constituted of five modules: namely *data analysis*, *text extraction*, *ontology development*, *information extraction*, and *evaluation*.

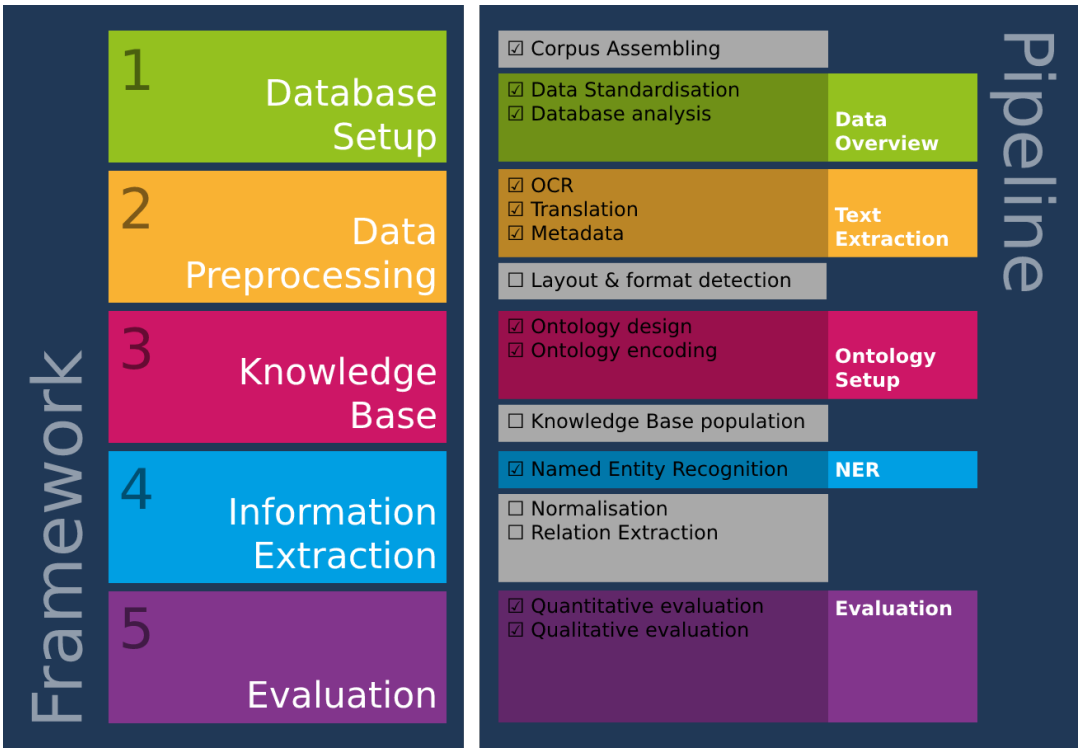


FIGURE 1.4: Corresponding modules in the *PsylVe Framework* and *Pipeline*. Greyed out tasks were not part of the Internship.

The *C. pruni* database - The database published by Sauvion et al. (2021) including: (1) unstructured data, in the form of documents recording the occurrence data of *C. pruni*, and (2) structured data in tabular form, both about manually extracted occurrences and document metadata. See Figure 3.1 for a visual representation of its composition.

The *C. pruni* document dataset - The set of documents contained in the *C. pruni* database.

The *C. pruni* document metadata - The tabular structured metadata of the *C. pruni* document database.

The *C. pruni* occurrence dataset - The occurrence data, recorded in tabular form as it appears in the *C. pruni* database[1]. Occurrence data is the recorded presence or absence of a certain taxon in a certain geographical place on a certain date.

[1]Occurrence data downloadable from `https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.15454/VC9UR5`

Chapter 2

Background

This chapter serves as a purposely rudimentary introduction to: (1) some relevant biological aspects (Section 2.1), with a focus on disease surveillance and taxonomy, (2) Natural Language Processing (NLP) (Section 2.2), and (3) Knowledge Representation (KR) (Section 2.3). As mentioned in Chapter 1, each section is targeted to readers who are unfamiliar with the corresponding research domain and is therefore conceived to be simple and approachable in its explanations.

Additionally, this chapter will allow to better understand the research context in which the PsylVe project was developed, both from the point of view of the biological issues addressed and the more technical IE and KR challenges encountered. At the end of each of the following sections, a subsection is dedicated to linking the corresponding research context with the PsylVe methodology and objectives (Section 2.1.4, Section 2.2.3, and Section 2.3.2).

2.1 Biological Aspects

This section will provide a brief overview of disease surveillance, some basic features of taxonomy and nomenclature, and the relationship between psyllid vectors and plant pathogens involved in the scope of this project, while considering the related diseases. Familiarising oneself with this topic, at the very least on a surface level, allows a better understanding of the implemented NLP strategies and the KR choices, both detailed in Chapter 4.

2.1.1 Vectors in disease surveillance

Vector insects can play a critical role in sanitary crises, drastically influencing the spread of some dangerous diseases. One of the most widely recognised examples in the human health domain are mosquitoes as vectors of viruses causing dengue fever, yellow fever, and Chikungunya. In plant health, virus- and bacteria-carrying insects cause agricultural losses year after year (Lefèvre et al., 2022).

Global changes such as intensification of international exchanges and climate change, enabled the rise of new diseases (Steffek et al., 2012). A solution to efficiently fight these diseases while minimising the damage caused to the environment would be to foresee them or detect them as quickly as possible in order to put into place a variety of prophylactic measures.

In the field of epidemiology, disease surveillance is a practice consisting in monitoring the progress of diseases with the ultimate goal of curbing any negative consequences associated with their outbreak and spread (Ristaino et al., 2021; Trivellone et al., 2022). In particular, plant diseases can have an enormous economic impact on the agricultural sector (Savary, Willocquet, and Pethybridge, 2019), making their

surveillance, forecasting and mitigation of an extensively researched issue (Morris et al., 2022). It has been observed that some of the most destructive plant diseases known to humankind have been caused by vector-born pathogens (Marie-Jeanne et al., 2020; Perilla-Henao and Casteel, 2016; Gilbertson et al., 2015), among which, those caused by psyllids (Jarausch et al., 2019).

In this context, the ANR project BEYOND[1], launched in February 2021, conceives a multidisciplinary setting for plant disease surveillance where risk forecasting is achieved through the processing of multiple types of data: information from documents of various nature[2], data from plant-pathogen risk-forecasting networks for the detection of abiotic variables (such as wind and water), data on interconnection between regions via natural elements and artificial transportation systems, historical and forecast meteorological trends, etc.

One of the most popular strategies for plant pathogen risk forecasting is that of recording occurrences of pathogen vectors. Nonetheless, the quality of the input data remains one of the most critical factors in the output quality of forecasting models. As mentioned by (Sauvion et al., 2021), such input data includes occurrence data, historical data, insect-specific ecological preferences and relationships with host plant. Occurrence data in particular is the one whose extraction the PsylVe framework is aiming to automatise.

Occurrence data collection

Unfortunately, collecting occurrence data via classical approaches (e.g. manually gathering data from online and physical sources) is, more often than not, extremely time consuming (Sauvion et al., 2021), in spite of their value for the objective of epidemiological risk assessment[3].

In 2021, the *INRAE Montpellier* team published a database of occurrence data of psyllids vectors of the species complex, namely *Cacopsylla (Thamnopsylla) pruni* (Scopoli, 1753) (*C. pruni*), carrying of a phytoplasma '*Candidatus* Phytoplasma prunorum' (*Ca.* P. prunorum) responsible for the European stone fruit yellows (ESFY) on cultivated *Prunus* (especially *Prunus armeniaca* L. - the common apricot tree - and *Prunus salicina* Lindl. - the Japanese plum tree). This database, published on the INRAE data repository (Sauvion et al., 2021), is the result of collection efforts that spanned over multiple years, including field inspections and both online and offline[4] information searches. The compiled data allowed to model the current and predicted geographical distribution of psyllid vectors in at the European scale in the form of maps while taking into account variables related to climate change (IPCC). See Figure 1.1 for an example of occurrence data usage for geographical mapping visualisations..

PsylVe, was co-supervised by Nicolas Sauvion (INRAE-PHIM, Montpellier) and Claire Nédellec (INRAE-MaIAGE, Jouy-en-Josas), in close collaboration with Robert Bossy (INRAE-MaIAGE, Jouy-en-Josas) (see Figure 2.1 for a full organigram). The internship was funded by the consortium GIS Fruits within the ANR BEYOND project and the governmental initiative "*cultiver et de protéger autrement*". The French

[1] `https://www6.inrae.fr/beyond/`

[2] Documents include, for example, recent articles available in digital form and scans of old books.

[3] An example of valuable historical data having a considerable impact on risk-assessment is that of a mosquito-borne West Nile Virus as described by Suarez and Tsutsui (2004).

[4] Not all documents are available in digital form. Institutions such as museums made physical documents available for these purposes.

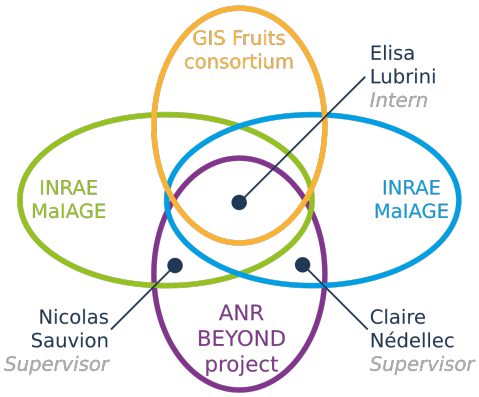


FIGURE 2.1: The placement of the author and supervisors of this thesis with respect to the hosting organisations: INRAE-PHIM, INRAE-MaIAGE, BEYOND and GIS Fruits

priority research program "*cultiver et de protéger autrement*", included in the 3rd future PIA (PIA3) promotes the use of sustainable agricultural techniques. Its objective is to identify alternatives to the use of phytosanitary products by mobilizing the levers of agroecology, biocontrol, genetics and prophylaxis. The call for projects was launched in June 2019, with the BEYOND project being one of the successful candidates. The objective of the BEYOND framework is to integrate an interdisciplinary approach to the data collection process and compare it to the classic, manual approach (e.g. in terms of speed/easiness of access to data) in order to identify relative pros and cons of each method. In particular, the BEYOND project envisions as the ultimate objective the creation of an extremely versatile framework for the development of such a tool, in order to allow researchers specialising on the forecasting of various diseases, beyond those covered by PsylVe, to have access to computational assistance.

2.1.2 Taxonomy and Nomenclature

As stressed above, a critical part of text analysis is the recognition and normalisation of the names of organisms that are involved in an interaction, i.e. pests, plants and vectors. Taxonomys have been since Linné (1746) the way to share a common and formal representation of the organisms. Such resources are extensively integrated in the PsylVe project. More precisely, in biology, *taxonomy* is the study targeting the definition of groups of biological organisms based on shared characteristics, their naming and hierarchical structuring.

In this context, *nomenclature* refers to the system of names assigned to taxa. Names belonging to this system are case and format sensitive, as a capital first letter and italic formatting are used to denote specific ranking within the taxonomy(Lapage et al., 1992).

Regarding the hierarchical structure of a taxonomy, large taxonomically structured ontologies can be modelled following an partially ordered set structure (Kaiser and Schmidt, 2011; Kaiser, Schmidt, and Joslyn, 2006), making it possible to derive a concept lattice useful for representations as formal ontologies.

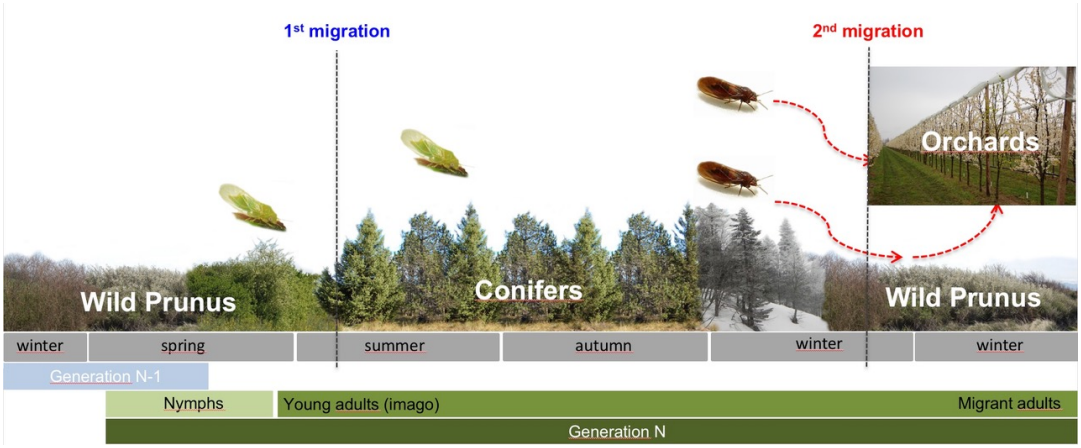


FIGURE 2.2: The biological life cycle of psyllid *C. pruni*
illustrating in particular the alternation between host plants (Prunus sp.) and shelter plants (conifers).

Given that many pivotal documents in the database are particularly old, it is important to mention that the designed way to refer to a specific taxon often changes over time, as new species are discovered and phylogenetic tree representations are updated. In Chapter 3, some external resources, including taxonomical ones, will be presented. Two of the most relevant resources for the first stage of the PsylVe project are the NCBI Taxonomy Database (see Section 3.2.2) and the Psyl'list database (see Section 3.3)

2.1.3 Psyllids as vectors of phytoplasma

As stressed above, a central piece of the occurrence descriptions to be extracted from documents are vector mentions. More precisely, organisms that have the ability to inoculate pathogens in other organisms (e.g. mosquitoes, ticks, aphids, nematodes, fungi). For plant pathogens, in particular, movement from one plant to another is essential for their spread and long-term survival, and insects play a key role as vectors (Herrbach et al., 2013). Both organisms, the pathogen and the vector, are classified as pests, as both can cause the insurgence of a disease in the targeted host plant.

The scope of the PsylVe project points at jumping plant lices, also called psyllids, of the genus *Cacopsylla*. The psyllids (Psylloidea family) are small piercing sap-sucking insects of the order Hemiptera, phylogenetically related to aphids, mealy bugs and cicadas (Ouvrard, Chalise, and Percy, 2015). The life cycle of numerous species of *Cacopsylla* relies on plants of the *Rosaceae* family, in particular fruit trees of the genera *Prunus* (e.g. apricot tree), *Pyrus* (e.g. pear tree), and *Malus* (e.g. apple tree).

Several *Cacopsylla* species have been described as vector of phytoplasma on such fruit trees, with a close specificity of the insect-plant and host-pathogen interactions (i.e. each species of psyllid transmits a single species of phytoplasma to trees belonging to one single genus). For example, both psyllid species of the *C. pruni* species complex are known to only transmit the *Ca.* P. prunorum phytoplasma to *Prunus*, causing in it the European stone fruit yellows (ESFY) disease. Other species are known to transmit phytoplasma responsible for Apple Proliferation (AP) and Pear Decline (PD) respectively on apple and pear trees (Assunta et al., 2019).

Bacteria such as phytoplasma and their vectors are native to Europe, where they are widely present in orchards, as well as in wild habitats, with its spread in the latter impeding vector eradication and, therefore, complete elimination of the related fruit tree diseases (Steffek et al., 2012).

Psyllid vectors of phytoplasma do not reside on their preferred host plant for the duration of a whole year, but rather migrate, once reached adulthood, usually in early or late summer (depending on the species) to *shelter plants* (also called *overwintering* plants, as mentioned by Burckhardt et al. (2014)), usually conifers. The following year, usually at the end of winter, psyllids migrate again to their host plants to reproduce. (Thébaud et al., 2009).

This double migration at varying distances between lowland and mid-mountain regions plays an essential role in the spread of phytoplasmas on large spatial scales (Marie-Jeanne et al., 2020). This example is a good illustration of the complexity of plant-vector-phytopathogen interactions. In epidemiology, it is necessary to decipher this complexity as well as possible in order to hope to implement relevant control measures. Relying on examples of pathosystems already well described in the literature can be very useful to save time in understanding these biological interactions.

2.1.4 *PsylVe*: The biological challenges

Although the PsylVe framework is envisioned to be easily replicable within different epidemiological domains and for different taxa, during the course of the Internship the scope was set to the species complex *C. pruni*. The host specificity of psyllids allows to gradually expand the scope of the PsylVe project, containing the work of the current thesis within the smallest possible scope: the triplet (1) *Prunus* as host plant, (2) *C. pruni* as vector, (3) *Ca.* P. prunorum as pathogen. In the following stages of the PsylVe project, the scope will be progressively extended to neighbouring taxa in the phylogenetic tree (see Table 2.1 for a full description of the scope of the PsylVe project).

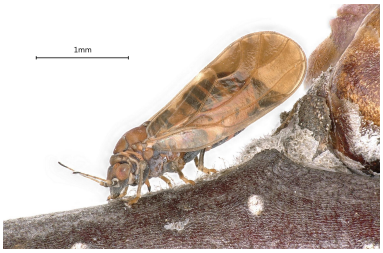


FIGURE 2.3: Female of the psyllid *C. pruni* feeding on a blackthorn (*Prunus spinosa*) branch. Photo credit: Sauvion N., INRAE

In light of the considerations above, occurrence data treated within the PsylVe project will be geographically limited to the Western Palaearctic, meaning that terms designing locations outside of this scope will not need treatment. Additionally, the migration habits of psyllids causes the occurrence data to encompass more species than merely those designated as host plants, which will affect the number of taxa to be considered alongside their respective terminology.

With regards to the challenges related to taxonomic nomenclature, the overarching biological scope spanning over entomology, botanics and microbiology means that the terminology will follow standards associated with the respective domains. Additionally, such standards are in constant development and will present a challenge in the extraction of taxa from documents published in such a wide range of historical periods.

	Stage 1 (internship scope)	**Stage 2** (PhD scope)	**Stage 3** (BEYOND scope)
Host Plant	*Prunus*	*Rosaceae*	*Rosaceae* + *Rutaceae* + *Solanaceae*
Vector	*C. pruni*	*Cacopsylla* spp.	psyllids
Pathogen	*Ca.* P. prunorum	phytoplasma of group X	phytoplasma+ liberibacter
Disease	ESFY	All diseases caused by phytoplasma transmitted by psyllids on *Rosaceae*	All diseases caused by bacteria transmitted by psyllids
Geography	Western Palaearctic		

TABLE 2.1: Scope of the corpus on which the pipeline is optimised

2.2 Text mining

Text mining refers to the computational treatment of natural language data for the purpose of automatising the extraction of structured data. As an example, the database published by Sauvion et al. (2021) is divided in two parts: the unstructured data, composed of the documents (books, articles, etc.), namely the corpus, from now on referred to as the *C. pruni* document database, mentioning the presence or absence of *C. pruni* in a certain place at a certain time, and the structured data, namely tabular data that was manually compiled by filling slots such as observed taxon, location, and time corresponding to what mentioned in each of the documents .

The task of automatising the extraction of specific predefined information from unstructured data has been known for thirty years as Information Extraction (IE) and it is usually based on NLP techniques. It is important to note that, before an IE pipeline can be used, the input data usually needs to be preprocessed in order to ease its treatment via computational methods.

In the following subsections, a brief introduction to document preprocessing and NLP methods for IE will be followed by the description of specific subtasks relevant to the PsylVe project.

2.2.1 Data Preprocessing

In order for a IE pipeline to be able to process input data, the data must respect some standards and requirements. Below, some relevant subtasks are discussed.

Standardisation of structured data When data is stored in the form of a table, it is important for certain standards to be respected in order to facilitate machine readability. Unfortunately, in many research domains, especially those who historically experienced lack of computational automation and a prevalence of manual methods, such standards are not well known nor followed.

As an example (see Section 4.1 for a complete list of the applied standards), when treating tabular data, column names will be treated as variables and should therefore follow variable naming conventions, such as absence of spacing or special characters. It is however common, in the above mentioned research domains, for

researchers to store the data in a format carrying more resemblance to natural language, using spaces and punctuation as it normally appears in unstructured data. Ideally, the C. pruni occurrence database schema would follow the Darwin Core standard Group, 2009, whose objective is to simplify the sharing of biological information in an extensible format (Baskauf and Webb, 2016).

Optical Character Recognition When developing natural language pipelines, the expected corpus in natural language is expected to be in a machine readable format. However, older documents are only available as images of physical documents (such as scans of book pages), sometimes embedded in a PDF, but rarely being accompanied by an accurate machine readable transcription of their content.

Optical Character Recognition (OCR) techniques make use of computational tools for the recognition of written characters and allow to obtain a machine readable version of a document. Recently, such tools have evolved to allow recognition of various document layouts, such as multi-column articles, and have begun to include more and more special characters, facilitating, by consequence, the recognition of letters specific to languages other than English. Additionally, when processing a corpus with mixed formats, the identification of the actual given document format is necessary in order to trigger the relevant processing tool.

Translation Having a machine readable input is sometimes not enough, as some NLP tools set specific input requirements, such as restrictions on accepted languages. Most NLP tools are optimised to work with English and some of them do not allow treatment of other languages at all. Terminological resources also pose similar constraint given their prevalent availability in English.

For this reason, when treating documents in different languages, their content might sometimes need translation depending on which tools are to be integrated in the NLP pipeline. Luckily, some of the most used language models for translation are easily and freely accessible and can be used in integration within a pipeline for easier automation of the task.

2.2.2 NLP and Information Extraction

Natural Language Processing (NLP) refers to the set of computational techniques for the treatment of natural language data. Having emerged in the 1950s as an intersection of Artificial Intelligence and linguistics, it started gradually expanding and encompassing elements from different areas of research, providing useful tools to various fields, among which are biology and medicine. More specifically, information retrieval, named entity recognition, sentiment analysis and question answering are some examples of NLP tasks applied in the aforementioned fields.

The initial stages of information retrieval were heavily influenced by the Message Understanding Conferences during the 1990's (Grishman and Sundheim, 1996), soon after which it started to be applied to the biomedical field (Spyns, 1996). This continued during the 2000's (Meystre et al., 2008), and more recently, the importance of information retrieval became clear after the outburst of the COVID-19 pandemic (Chen et al., 2021).

Similarly in the study of Ristaino et al. (2021), a new set of tools, including disease monitoring and enhanced detection technologies such as pathogen sensors, predictive modeling, and data analytics, was proposed in order to avert future outbreaks of plant disease pandemics associated to global human pandemics. Additionally,

in the study of Molik et al. (2021) NLP tasks and metabarcoding techniques were used to reveal pathogen-environment associations. As an example, in their study, the investigation of ecological niches of rare pathogens was facilitated by the use of text-mining analysis techniques.

NLP techniques can be used for tasks such as Information Extraction (IE) where the treatment of natural language allows for the detection of required data within a text (Nédellec, Nazarenko, and Bossy, 2009). Information Extraction (IE) refers to the research field that aims at the development of tools and scientific methodology facilitating the access to document content via computational methods The development of this field of research emerged in an attempt to improve the managing of knowledge in an era of rapid growth of textual information. Initial IE systems created were rather superficial in retrieving and treating concepts, focusing mostly on targeted text exploration and being far from in-depth semantic analysis tools .These constraints of the initial IE systems highlighted the necessity for new approaches, formally based on text analysis and ontological knowledge. Nowadays, IE systems are advanced tools for identifying relevant information from text documents organising them in structured forms (Zhou et al., 2005), with the aim to fill predefined form slots or templates with the extracted information.For achieving these tasks, a variety of linguistic processes are involved among which are text segmentation, anaphora resolution, polysemy detection. With respect to the overall process followed during the information extraction, preprocessing and extraction rules (e.g., regular expressions and other patterns) facilitate the identification of subsections and therefore the interpretation of the targeted text. An in depth review of these topics can be found in Nédellec, Nazarenko, and Bossy (2009).

The main subtasks of an IE pipeline are usually enchained in the following order: (1) Recognition of relevant entity mentions in the inputted text; (2) Normalisation of the mentions by means of an external reference; (3) Extraction of relations among such mentions.

Named Entity Recognition

Named Entity Recognition (NER), also known as *(named) entity identification*, *entity chunking*, and *entity extraction*, refers to the process of recognizing and extracting named entities such as persons, locations, and organisations from natural language texts (Mansouri, Affendey, and Mamat, 2008). The importance of NER is prominent in a variety of NLP tasks among which is information extraction. Its objective is the detection of terms associated with predefined entities in natural language text. Specifically, given the entity *C. pruni*, all the different terms used to refer to it (such as taxonomic nomenclature variations) should be detected and linked to the corresponding entity.

A number of different approaches could be used to tackle such problem, among which are the creation of rules[5] and the projection of a lexical resource onto the inputted text. Specifically, knowing that every one of the different notations of referring to *C. pruni* ends with the token "pruni" and that the probability of another entity being referred to with the same string, would mean that creating a rule for detecting all entities ending with the token "pruni" could be a reasonable option. On the other hand, if we were to expand such a rule to many taxa, using an external lexical reference could prove to be extremely useful. For example, resources such as the NCBI Taxonomy Database contain naming variations for a number of taxa and can be easily projected onto the text.

[5]Rules usually make use of regular expressions to automate the recognition of certain patterns.

Normalisation

An important subtask of IE is *entity normalisation*, often referred to as *entity disambiguation*, *entity grounding*, or *entity linking* (Ferré et al., 2020). It refers to the task of associating entity mentions found in a text with categories associated to a predefined vocabulary. It is often considered as a classification task, with the entity mentions being classified to the corresponding categories (Ferré et al., 2020). Traditionally, this classification process is based on the similarity of the entity mentions and the concepts associated with the predefined vocabulary.

Normalisation by Machine Learning and Deep Learning is a very active research domain. Some recent approaches called joint learning achieve entity detection and entity normalisation in a single step. State of the art tools achieve better performances than rule-based methods but are still complex to implement due to their computational costs. Due to the high amount of data to be processed via the PsylVe pipeline, rule based methods were prioritised.

Relation Extraction

Relation Extraction (RE) is a crucial step towards the creation of advanced natural language understanding applications (Bach and Badaskar, 2007). It refers to the process of determining connections between entities or events (Nadkarni, Ohno-Machado, and Chapman, 2011). Such connections or other relations can be expressed as "causes" and "treats" (Nadkarni, Ohno-Machado, and Chapman, 2011) in the domain of biology and medicine among others. Two main steps can be identified within the process of RE. The first one is determining which candidate entities or events mentioned in the text are actually linked and the second is the classification of the detected relations into predetermined classes (Nguyen and Grishman, 2015). These steps tend to be grouped in a single step by recent Deep Learning approaches. As noted by Nguyen and Grishman (2015) the two prevalent methods of the past decade have been the *feature-based method* and the *kernel-based method*. The former makes use of a sequence of features that correspond to labelled examples in an n-dimensional space according to the number of features (Moncecchi, Minel, and Wonsever, 2010). Despite the benefits of this method, the main issue arising is that data in natural language sentences are not always easily represented via feature vectors but rather tend to better fit into graph representations (Moncecchi, Minel, and Wonsever, 2010). Such cases lead to high dimensional vectors that make feature extraction a rather species complex task, which is often associated with computational power challenges (Moncecchi, Minel, and Wonsever, 2010). The kernel method, on the other hand, was a more suitable method as it allows the implicit calculation of dot-products in high dimensionality spaces, without the need for an explicit representation of each vector (Moncecchi, Minel, and Wonsever, 2010). The most recent and successful methods are based on Deep Learning [6]. The PsylVe project planned the use of RE-BERT developed by (Tang et al., to appear, 2022) and the textitBibliome group at MaIAGE in the next few weeks. The tool that I have selected for the first version of my pipeline is based on rules and trigger words.

[6] Pre-trained transformer models such as BERT have achieved state of the art results for most recent NLP tasks, including RE approaches. Conventionally, an RE task is treated as a special kind of text classification problem. A common pipeline is to first initialise a BERT model with pre-trained weights, and then fine-tune the model on the dataset of interest (Devlin et al., 2019).

2.2.3 *PsylVe*: a text mining framework

The content of existing databases, combined with the potential of NLP techniques (and the structurality offered by KR methods; see Section 2.3), open up very promising prospects for better knowledge of the potential circulation of pathogens and, thus, for better short-term anticipation of prophylactic decisions. This is the ultimate objective of the ANR BEYOND project

The objective of PsylVe is to build a database as exhaustive as possible of occurrence data of psyllids, focusing on their role as vectors of phytoplasmas infecting fruit trees. This will be carried out using a text mining approach and integration of structured data in the pipeline. The work of the Internship focuses on occurrence data of one species of psyllid: *C. pruni*, but is a first step in building solid foundations for a larger, more general framework. The work will be part of the more general framework of extraction of broader information such as habitats of plant pathogens, host plants, their phenotypes, habitat conditions or diseases developed in the BEYOND project by the MaIAGE unit and PESV platform. In 2007, the MaIAGE unit at INRAE developed the AlvisNLP software (Ba and Bossy, 2016) to facilitate NLP pipeline implementations. . AlvisNLP offers a library of supervised Machine Learning methods based on neural architectures and rule-based approaches exploiting linguistic, lexical, terminological and conceptual information (thesauri, taxonomic nomenclatures, ontologies) (Aubertot et al., 2015) for the purpose of building NLP pipelines. The work carried out for the current thesis consisted in proposing a computational solution to improve the reaction speed in case of disease outbreak or expansion by easing the access to psyllids occurrence data by disease surveillance platforms. Additionally, the method used aims to be flexible enough to be used as a framework for related disease surveillance tasks. Many pipelines for various applications in the biomedical domain have been developed using it. Among them, the Omnicrobe pipeline (see Section 3.1) extracts entities and relations about organisms and habitats in the microbiology field.

The framework was designed to be composed of five stages: database assembling and overview, document preprocessing, ontology development, information extraction Information Extraction, and pipeline evaluation. Refer to Section 2.3.1 for a presentation of the ontology development and its technical challenges. The remaining four modules of the PsylVe framework were organised as follows:

Dataset composition analysis

Visualising the composition of a database allows one to better understand its patterns and identify potential characteristics to be exploited. As mentioned above, the language(s) in which each document was redacted and year of publication (and consequently, likelihood for a machine readable version of the text to be available) constitute critical information for the pipeline input. In the case of the *C. pruni* document database, such data was provided by the published manually compiled metadata. See Section 4.1 for a more complete overview of the methodology used.

Document preprocessing

Given the unavailability of accurate embedded machine readable text for a large portion of the *C. pruni* document database, a pipeline for text extraction was implemented to ensure availability of documents independently of whether the source was physical or digital. Additionally, a translation module was added in order to make the extracted text available for processing by the chosen NLP tools. See Section 4.2 for a more detailed presentation of this module features.

Information extraction

The IE pipeline was based on an existing pipeline, also developed using AlvisNLP, that of Omnicrobe(Dérozier et al., 2022b) (See Section 3.1 for an introduction to the pipeline). The first step was to evaluate the current results produced by Omnicrobe on the provided set of documents in order to identify possible issues preventing accurate IE results. After analysing the results, a list of issues to be addressed was drafted, together with proposed solutions and their limitations. Some of such issues were addressed within the work of the current thesis, while others were outlined, and solutions proposed, with the objective of building a solid starting point following work. The prioritisation of tasks relied on the best compromise between resources and quality of the results.

Evaluation

A qualitative and quantitative evaluation of the pipeline results was carried out, highlighting the improvements compared to the existing AlvisNLP-based Omnicrobe pipeline.

2.3 Knowledge Representation

In order to be useful, the information extracted from documents, such as those to be processed by the PsylVe pipeline, must be represented in a computer readable format that allows automatic deduction of new information via reasoning. Knowledge Representation (KR) and Reasoning refers to the sub-field of Artificial Intelligence (AI) that facilitates the creation of computer systems which are able to provide an accurate reasoning on a machine interpretable conceptual representation of the world (Stephan, Pascal, Andreas, et al., 2007). Such interpretation of the world is intended to resemble the interpretation provided by human reasoning (Stephan, Pascal, Andreas, et al., 2007).

2.3.1 Knowledge bases and Ontologies

In KR, a Knowledge Base (KB) is a representation of knowledge on a certain topic, serving as a description of concepts and relationships that facilitate the sharing of such knowledge (Gruber, 2018). The concepts, normally referred to as *classes*, are associated with *properties* describing some of their characteristics (Noy and McGuinness, 2001). Classes being the main focus of most ontologies often have *subclasses* which represent more specific concepts than those of their *superclass* and can be linked to one another via *hierarchical* (or *taxonomical*) or *non hierarchical relations* (Noy and McGuinness, 2001). In most description languages, relations can be either binary, or unary, with the latter also known as *attribute*.

As an example, in the current version of the PsylVe ontology, `GenomeBearingEntity` is a superclass of `Organism`, meaning they are related via at least one relation, that is hierarchical. `HasHostPlant` is an example of ontological, non hierarchical, binary relation, and `has_participant` is an example of attribute.

For the creation of an ontology the definition of its domain and scope is of a particular importance (Noy and McGuinness, 2001). Considering reusing existing ontologies is also important as many are already available on the Web and in the literature for this purpose (Noy and McGuinness, 2001). For the creation of an ontology the following steps are being followed as presented in Noy and McGuinness (2001): First, all classes of the ontology should be defined. For this step the enumeration of important terms in the ontology would facilitate the definition of the ontology domain and scope. Next, a taxonomic hierarchy should be applied for the arrangement of the classes and subclasses. This hierarchy can be implemented either using a top-down approach, a bottom-up one or combination of both. Finally, properties and their relative restrictions should be defined for each class.

All instances of classes in an ontology may compose a knowledge base (Noy and McGuinness, 2001). Knowledge-based systems operate with computational models focusing on specific domains of interest (Stephan, Pascal, Andreas, et al., 2007). The symbols used in those models correspond to domain artefacts of the physical world like relationships, events, or objects (Stephan, Pascal, Andreas, et al., 2007), making a KB the union of an ontology and its corresponding instances. Common languages for ontology and KBs representations are RDF-based, such as the popular OWL or SKOS. In this project, RDF graphs will be written using the turtle syntax. Within the PsylVe project, entities and relations are not sufficient to detect novelties and contradictions; and inference rules need to be implemented. Inference rules allow reasoning. An example within the PsylVe ontology is that any `Psyllid` which is a `CarrierOf` of `Phytoplasma` will always belong to the class `Vector`.

In addition to the creation of the KB it is also common practice to define commands to query it. A common way of retrieving information from KBs is based on the SPARQL query language.

For this project, the ontology will be a subset of the KBs. Specifically, the ontology will only deal with the knowledge regarding the behaviour of classes, subclasses, and relationships, and not the specific instances of members of these classes and relationships. The ontology together with the instances constitute the KB.

2.3.2 *PsylVe*: a framework for structured data

As stressed in Chapter 1, the initial program of the Internship was extended by adding, among other features, a significant module on the representation of biological knowledge in the form of an ontology. The structure of the KB was planned in order for occurrences extracted from the text to be instances of the ontology classes and relations. This expressive framework will allow the detection of contradictions between new occurrences and existing knowledge and the discovery of new knowledge. Building this representation relies on Knowledge Representation (KR) tools and methods.

Structured data from external sources was and will be integrated, prioritising the resources mentioned in Chapter 3. Defining the domain scope, main classes and relations was possible thanks to continuous exchanges with the designed domain expert, Nicolas Sauvion, and the KR specialist, Catherine Faron Zucker (see Section 4.5 for the presentation of the methodology and plans for future development)

2.4 Pipeline evaluation

For the evaluation of the pipeline the qualitative criteria for text mining pipeline evaluation presented in Spinakis and Peristera (2004) were combined with common quantitative scores for evaluation as mentioned as follows.

2.4.1 Qualitative evaluation

For the evaluation of the pipeline the following criteria for text mining pipeline evaluation were taken into consideration as presented in Spinakis and Peristera (2004). For an analysis on how the PsylVe pipeline meets the criteria, see Chapter 5.

Criterion 1: Data retrieval and result evaluation This criterion focuses on the quantitative aspect of the result evaluation. Therefore, the scores mentioned in Section 2.4.2 will be used. See other criteria for the qualitative evaluation.

Criterion 2: Integration with other sources Integrating a text mining pipeline with external resources allows for more and possible better results.

Criterion 3: Output format flexibility Allowing users to enable different outputs improves the integration of the pipeline with other workflows.

Criterion 4: Availability of result statistics and analysis Providing a pipeline to help the user visualise statistics and carry out an analysis of the results is vital for a better understanding of the mechanisms and results of the pipeline.

Criterion 5: Combining linguistic and statistical methods Text mining is an inherently interdisciplinary field that leverages the union of linguistic research and computational power to automatise tasks that would be otherwise labour intensive. Therefore, exploiting tools and methods from both domains is key to optimal results.

Criterion 6: Online tool availability Online availability of a tool widens its accessibility and potential userbase.

Criterion 7: Result visualisation quality Visualising results in an easy-to-understand and informative way allows for an overall better user experience and better integration in the user's workflow.

2.4.2 Quantitative evaluation

In IE, *precision* and *recall* are performance metrics that apply to data retrieved from a corpus. Sometimes, in order to give a comprehensive score of a pipeline, the harmonic mean of the two is calculated, with the result being usually referred to as *F1-score*.

Precision The precision score evaluates the ratio of correct predictions over the complete set of predictions. *False positives*, i.e. wrongly detecting strings as entity mentions, negatively impact the precision. The precision indicates how correct are entities predicted by the pipeline.

$$precision = \frac{true\_positives}{true\_positives + false\_positives} \tag{2.1}$$

Recall The recall score evaluates the ratio of correct predictions over the complete set of entities that should have been predicted. *False negatives*, i.e. missing an entity mention, negatively impacts the recall. The recall indicates how comprehensive is the set of entities predicted by the pipeline.

$$recall = \frac{true\_positives}{true\_positives + false\_negatives} \tag{2.2}$$

F1-score The F1-score is the harmonic mean of precision and recall and is usually used to provide an overall score for a pipeline predictions.

$$F_1 = \frac{precision \cdot recall}{precision + recall} \tag{2.3}$$

Chapter 3

Related Works

This chapter focuses on the presentation of related works, with a focus on external resources that were either integrated in the PsylVe pipeline or planned to be integrated in future developments of the pipeline. Resources were divided into two types: other pipelines (see Section 3.1) and structured data, i.e. ontologies (see Section 3.2) and databases (see Section 3.3).

3.1 Pipelines

Omnicrobe Omnicrobe is a comprehensive text mining and data fusion approach-based open-access database of microbial habitats and phenotypes, whose purpose is to link and share data (Dérozier et al., 2022a). The Omnicrobe database schema is made up of elements of biological relevance that are connected together by particular relations as presented by Dérozier et al. (2022a). Microorganisms, habitats, phenotypes, and uses are the four sorts of entities defined. They are connected by three sorts of relations: (1) the `lives_in` relation, which connects a microorganism to its habitat; (2) the `exhibits` relation, which connects a microorganism to its phenotype; and (3) the `studied_for` relation, which connects a microorganism to its use. This formal schema organizes the information in the Omnicrobe database, defines the categories of data to be retrieved from the data sources, and leads the extraction process.

3.2 Ontologies

Ontologies were chosen according to the criteria listed by Malone et al., 2016 and using the Minimum Information to Reference an External Ontology Terms (MIREOT) method (Courtot et al., 2009). In the following subsections, organised by domain of interest, a paragraph was dedicated to briefly present each of them and their relevance to the PsylVe project.

3.2.1 Plant health

OntoBiotope OntoBiotope is an publicly available ontology developed and maintained by MaIAGE and INRAE Nédellec et al. (2018). The concepts mainly cover microorganisms and their habitats and phenotypes. The ontology was developed as part of the Omnicrobe pipeline (see Section 3.1) similarly to how the PsylVe ontology is being developed to be integrated in the PsylVe pipeline.

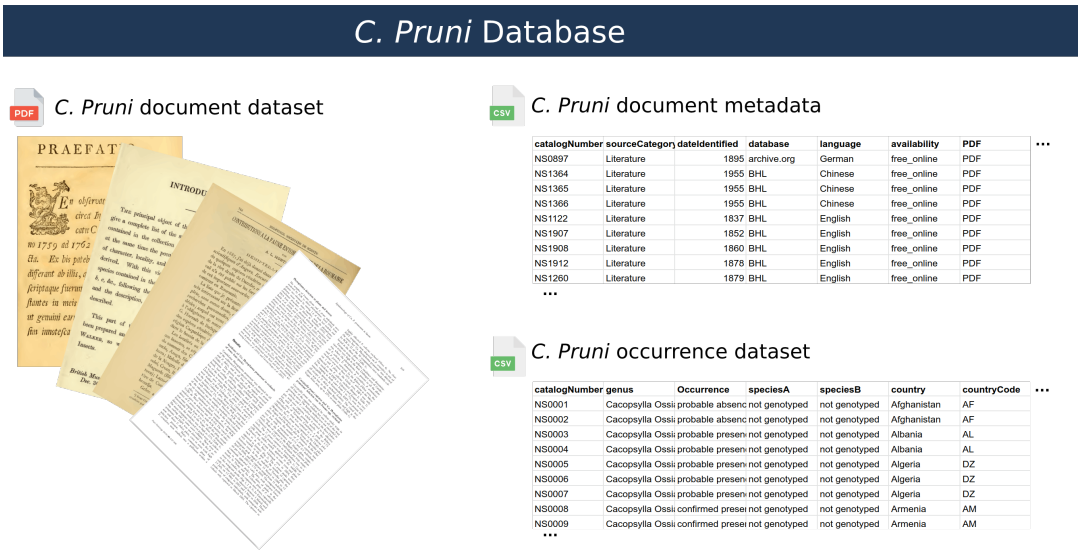


FIGURE 3.1: The three datasets composing the *C. pruni* database (Sauvion et al., 2021)

Plant Health Threats Ontology *Plant Health Threats Ontology* 2016 Another useful ontology in the plant health domain is the Plant Health Threats Ontology. Integrating EPPO (see Section) classes, it has some useful elements helping define pivotal aspects of plant diseases (e.g. a `symptomExpression` class to represent plant disease symptoms).

3.2.2 Taxonomy

NCBI The National Center for Biotechnology Information (NCBI) Taxonomy Database is one of the most popular ontologies in the biological domain and contains a curated categorization and nomenclature for all taxa found in public sequencing databases (Schulz, Stenzhorn, and Boeker, 2008).

TAXREF TAXREF is the national taxonomic reference system for the fauna, flora and fungi of France, developed and distributed by the Muséum national d'Histoire naturelle (MNHN) as part of the implementation of the National Natural Heritage Inventory Information System. (SINP)(Gargominy, 2022)

LSPN LSPN is a membership platform dedicated to facilitating personal learning, development, and progress in the life sciences through networking and information exchange. Being more than a simple content distribution platform it comprises of one of the most prominent networks for life science professionals (*LSPN connect* 2022).

3.2.3 Agricultural sectors

the agricultural sector is one of the most important sectors in the European economy (https://data.europa.eu/en/datastories/open-data-agricultural-sector). However agricultural community is made up of a wide range of actors, and thus lexical heterogeneity exists between these different set of actors. To resolve interoperability issues, recent generation of information system related to agricultural domain have benefited from the use of ontologies (Roussey et al., 2010).

EPPO EPPO Global is a database maintained and regularly updated by the Secretariat of the European and Mediterranean Plant Protection Organization. Its aim is to give pest-specific information created and gathered by EPPO. Specifically, information for over 90 000 species of agricultural, forestry, and plant protection interest is available along with scientific names, synonyms, common names in several languages, taxonomic positions, and EPPO codes for each species (*Eppo Global Database* 2022).

SORE Service Oriented Requirements Engineering (SORE) shares some activities with traditional Requirements Engineering but its focus on the identification of services and workflows used to modeling applications – developed and running in a Service Oriented Architecture framework – and on their reuse (Shahzadi and Tahir, 2016).

FCU (French Crop Usage) This ontology represents a thesaurus of cultivated plant organised by agriculture usages in France such as human food, industry, cattle feed (Darnala et al., 2022).

3.2.4 Time and Location

Time and location are two very important entity types when recording the circumstances of an event and it is common to reuse external resources for the standardisation of their recording.

TimeML This ontology is one of the most complete annotation schemes for the annotation of temporal aspects of ontological events. Thanks to its wide userbase, it has been recently proposed to become an ISO standard. The ontology allows for definition of events, temporal expressions related to them, and any links between the two (Del Gratta Riccardo and Ruimy, 2008).

geonames GeoNames is a geospatial dataset providing geographical data and metadata of around 7 million unique placenames from all over the world collected from several sources that are available for download free of charge (www.geonames.org) (Maltese and Farazi, 2013).

3.3 Databases

BHL The Biodiversity Heritage Library improves research methodology by collaboratively making biodiversity literature openly available to the world as part of a global biodiversity community.

Psyl'list Psyl'list is an online database, hosted by the Muséum National d'Histoire Naturel de Paris MNHN, dedicated to psyllids of world wildlife, whose main contributor is David Ouvrard (ANSES-Montpellier). This tools aims to organize taxonomic data available, but very scattered in literature, on this Hemiptera superfamily. Additional information is also easily accessible about geographical distribution, host plants, etc. This information, in particular the valid names of psyllid species, is continuously updated.

***C. pruni* occurrence dataset** The *C. pruni* occurrence dataset (Sauvion et al., 2021) contains 1975 lines and 33 columns and follows Darwin Core standard. Each line represents the mention of an occurrence in a document. The duplicates of the same observation are not represented. Conversely, there is one line per distinct occurrence. Beyond occurrences and document reference, the database includes very relevant metadata for the automatisation about the language, the source and the availability of the document.

Chapter 4

Methodology and results

This chapter outlines the methodology used to develop the PsylVe framework and the results obtained by each module, while Chapter 5 will focus on their evaluation. The purpose is to provide an easily replicable methodology in order for it to be applied to similar data for the same purposes of aiding manual occurrence detection from texts.

The PsylVe framework is divided into four processing modules, with an additional fifth module for evaluation that is explored in Chapter 5. The modules discussed in this chapter are: data overview (Section 4.1), data preprocessing (Section 4.2), Knowledge Base (Section 4.3), and information extraction (Section 4.5).

4.1 Database assembling and analysis

The creation of the document corpus falls outside of the scope of the work carried out by the author since Nicolas Sauvion, supervisor of this dissertation previously achieved the database assembling. Nonetheless, standardisation of the *C. pruni* occurrence database was needed for it to be processed. An analysis of the corpus composition was carried out using the provided metadata in order to highlight the peculiarities of the corpus and suggest relevant directions.

4.1.1 Standardisation of the data

The *C. pruni* document metadata includes 1975 lines and 13 columns. As opposed to the *C. pruni* occurrence dataset, no formatting/naming standards were originally assigned to the *C. pruni* document dataset and metadata. Thus, it was not easily processable by computational tools. The *C. pruni* document database was renamed and the corresponding *C. pruni* observation database was reformatted in order to respect the following standards.

Since the following list comprises standards that were already respected, in order for the methodology to be replicable in future works, whether any changes had to be done or not will be mentioned, for each of the standard, along with examples.

- Use the first row as column headers. *(no changes)*
- Use the first column as row names. *(no changes)*
- Column names must be unique. *(no changes)*
- CellRow names should be unique. ***(applied)***

  Example:
  Modified the document metadata table to make each cellrow refer to a single document.

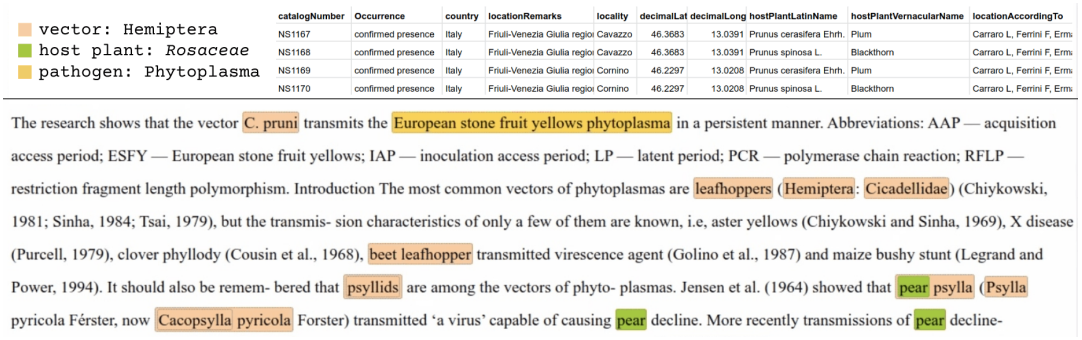


FIGURE 4.1: An example of occurrences processed by the Omnicrobe pipeline.

- Avoid names with blank spaces. ***(applied)***

  Example:
  Column names in document metadata contained space characters; underscore characters were used for substitution.

- Avoid names with special symbols. ***(applied)***

  Example:
  On top of space characters, document names contained special characters such as commas and colons. Such characters were removed.

- Start variable names with letters. *(no changes)*

- Avoid empty rows. *(no changes)*

- Avoid empty cells: use NAs ("not available"). ***(applied)***

  Example:
  Blank cells in document metadata were substituted with NAs.

4.1.2 Data composition analysis

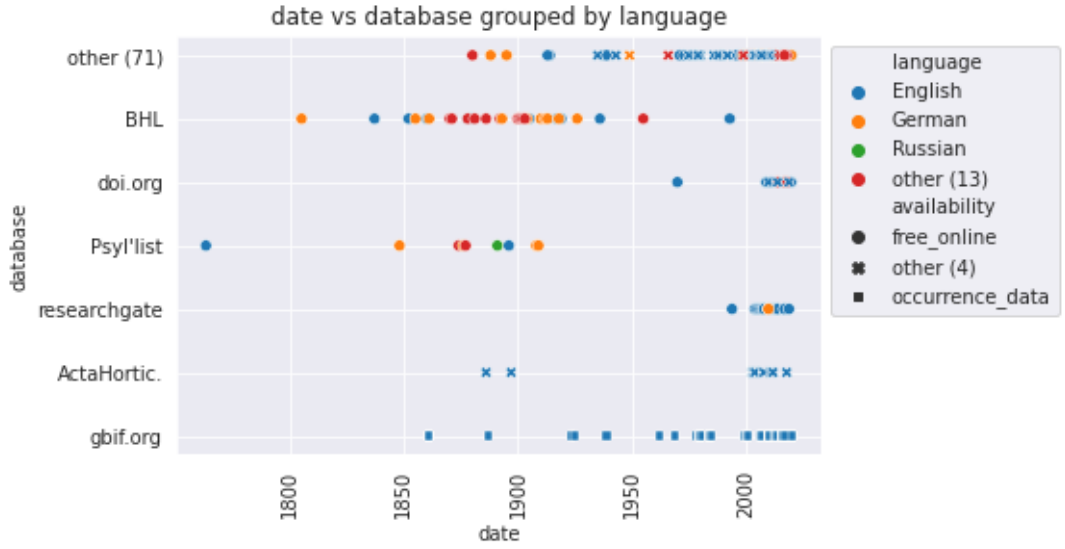
The analysis was carried out on the *C. pruni* document metadata, in order to predict possible challenges (e.g. text extraction issues, nomenclature irregularities) in the treatment of the corresponding documents.

Methodology

Some initial plots were automatically generated by looping over the most common associations[1]. This allowed a human-readable representation of the most interesting feature correlations based on which to design more insightful visualisations, examples of which will be illustrated in this section. A full list of figures, both in their first automatically produced version and the manual one, can be accessed via the corresponding module on the GitHub repository[2].

[1] More specifically, the x axis was set to the only ordinal variable of the dataset, while the y axis and hue grouping changed at each iteration over the produced subgroups.

[2] `https://github.com/e-lubrini/psylve/tree/internship`



l

FIGURE 4.2: Distribution of documents in the largest database, plotted against publication dates and grouped by language and accessibility.

Database composition analysis results

The following two figures (Figure 4.2 and Figure 4.3) were designed starting from some of the detected correlations. In both plots we can immediately notice how distributions are generally negatively skewed along the x (publishing date) axis. This is due to increasing document availability as online publishing was popularised.

In Figure 4.2, we can see a scatter plot of the distribution of documents in the six most popular databases and a group of the remaining ones plotted against their year of publication. When looking at the average publication date and earliest published document in each database, two of them stand out as valuable resources for relatively old documents: the BHL and Psyl'list (see Section 3.3 for a presentation of such databases). Additional attributes were plotted via marker colour and shape: document language and availability, respectively.

Again, BHL and Psyl'list stand out from the rest, in this case for the variety of languages included, when compared to the prevalence of English documents retrieved from other sources. When focusing on the colours we can notice a direct correlation between publishing date and prevalence of English data, which is easily explainable by the relatively recent establishment of English as a lingua franca in the research domain.

Regarding public availability of resources, we can observe a high resource to availability correlation, meaning that whether a document is available or not (or the structured occurrence data is not accompanied by a document, such as the case of the GBIF database) largely depends on where such document is published.

In Figure 4.3, accessibility was plotted against publishing date and grouped by whether the document is in English, in a violin plot. Two interesting accessibility levels are the case when only occurrence data is published (`occurrence_data`) and that of only a physical version of the document being available (`paper_version_only`). Such distributions stand out as being extremely polarising on the language dimension, with occurrence data being exclusively published in English and the paper

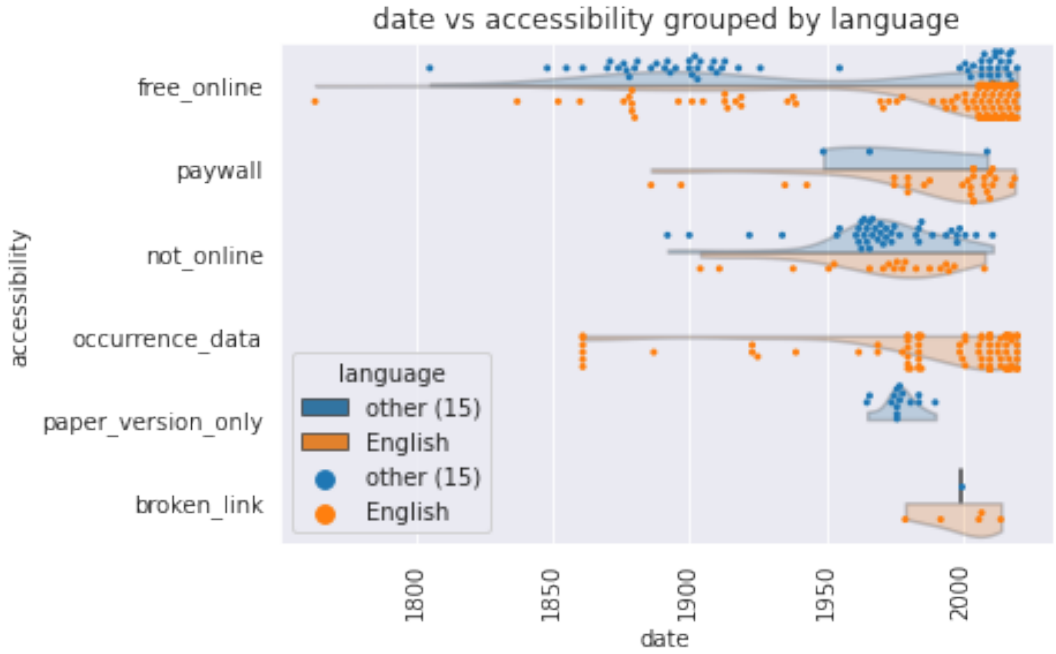


FIGURE 4.3: Caption

version of documents in the dataset being exclusively in languages other than English. This reflects how the use of English in the field has risen together with the use of structured data, such as occurrence data, both trends converging towards an improvement of internationally standardised scientific communication.

One last remark can be made on the case of accessibility via paywall. The distribution of non-English documents only accessible via paywalls presents a sharp cut[3]) at a 70 year distance mark from the year in which the database was assembled. This is thought to be due to European[4] copyright laws often install a copyright for collective works of the duration of 70 years after publication or the author's death, as it is the case of copyright law in France[5]. Scattered markers were superimposed on the violin plot as the relativity of the distribution was suspected to contribute to the representation being misleading. Three markers, corresponding to three non-English documents were found on the accessibility via paywall distribution. It is therefore important to mention that given the small sample size, assumptions based on it could be inaccurate and not representative of larger portions of data.

Consequences on the pipeline

Given the above mentioned remarks, some considerations are due before proceeding to the treatment of the data.

[3]Distribution cuts were set as described in the parameter table available at `https://github.com/e-lubrini/psylve/blob/internship/src/parameters.md`.

[4]The scope of the dataset was set to European occurrence data.

[5]`https://wipolex.wipo.int/en/text/363403`

Given the redaction language of physical documents with no machine readable text associated with them, it seems that OCR tools will be particularly important for languages other than English, meaning that recognition of foreign letters such as characters with graphic accents might be crucial. Additionally, given the variety of publication years and sources, OCR compatibility with a variety of document graphic layouts might be needed.

As shown in Figure 4.3, languages other than English amount to a total of 15 and, as seem to be quite common especially in documents from historical databases, meaning that integration of translation in the pipeline might unlock access to valuable historical data.

Finally, such diversity in language of publication might drastically affect the NER task, given, respectively: (1) the potential variety of language-specific vernacular names, that is, terms to design taxa that do not follow the taxonomic nomenclature standards, and (2) historical changes in taxonomic nomenclature relatively to specific taxa.

4.2 Text extraction

The text extraction module corresponds to a Bash application that was developed with the objective of extracting text from the corpus of documents. Since the application was expected to be (and currently is) used by similar parallel projects covering different biological domains, thorough documentation on how to execute thepipelinewith similar data was redacted. Documentation of the text extraction module is available on the PsylVe GitHub repository[6].

4.2.1 The structure of the application

The text extraction application allows input of two different types: images and PDF. After the detection of the input format, the application covers 7 different tasks, plus an evaluation of the overall module: Image to PDF conversion, PDF to image conversion, OCR text extraction, embedded text extraction, language recognition, translation to English, spellcheck, evaluation. The following subsections will explore the reasoning behind each stage of the pipeline.

Image to PDF conversion

The thematic domain of application concerned by the scope of this project is characterised by a high degree of specialisation. Furthermore, there is a great amount of relevant information held in documents currently unavailable online. The image to PDF conversion step of the pipeline allows for pictures of physical documents to be input and transformed to PDF formathis to be processed alongside the rest of the document database.

PDF to image conversion

This step was created to standardise the input data before passing it to the OCR tools. Depending on the chosen tool, the input might need to be in PDF or image format. This step ensures that if the chosen tool only allows image input, all the documents will be converted to sets of images.

[6] `https://github.com/e-lubrini/psylve`

OCR text extraction

For the extraction of text via OCR, a total of five tools were integrated in thepipeline (see Table 4.3 for a full list of tools used). The user can choose the preferred tool by editing the configuration file and additional tools can be easily integrated by following the instructions in the documentation. The tools differ along various parameters, such as special character recognition and multi-column content alignment. In order to evaluate the best tool for any one specific database of documents, an evaluation pipeline was implemented. Based on the results, the user can make a more informed decision on which tool to use for the OCR extraction and select it from the configuration file.

Evaluation of the task An evaluation pipeline was provided in the form of a notebook in order for the user to identify the OCR tool that best performs on a specific dataset. The performance of some OCR tools could be affected depending on various characteristics of the input documents, such as language, presence of special characters, page layout, structural composition. For this reason, the best performing tool for the documents in the PsylVe database might not be the best for other databases. Evaluating the pipeline with different parameters and tools is therefore encouraged.

Embedded text extraction

Some documents are embedded with additional XML information that can sometimes be useful depending on the domain of application. In the case of biology as a domain of application, the formatting is sometimes important when categorising information. For an example of the role of formatting in nomenclature, see section Section 2.1.2.

Language recognition and translation to English

The data output by the text extraction module described in this section is to be input to an information extraction module assembled using AlvisNLP. Since the Omnicrobe pipeline only works with English, the extracted texts that are in different languages need to be translated to English. For this purpose a language recognition enchained with translation was implemented.

Spellcheck

The tool comparison pipeline described in Section 4.2.1 uses a spell checker tool to evaluate the quality of the tool output, computing a score based on the amount of correctly spelled words, since the number of OCR errors should be correlated to the number of spelling errors.

4.2.2 Future usage

The text extraction application is currently being integrated to be a module in the AlvisNLP (Ba and Bossy, 2016) tool to facilitate its usage in similar pipelines developed at the MaIAGE, firstly the future pipelines to be conceived within the BEYOND project.

4.3 KB Development

This section will describe the development of the PsylVe ontology, which will allow reasonment on the occurrences extracted from texts, treated as instances of the hereby defined classes. The ontology will not only include the annotation scheme for the IE extraction from text, but also a comprehensive biological framework of knowledge linked to the already existing resources, such as the NCBI Taxonomic Database, and planned to be linked to additional resources such as the Ontobiotope ontology and the Psyl'list database (see Chapter 3 for a full presentation of the referenced external resources). In the future, the model will allow entities corresponding to the ontology detected by the IE module (Section 4.5) to populate a KB and, according to preset model restraint, allow the expansion of the ontology.

The development of the ontology consisted in three main phases: initial drafting, refinement, and encoding.

4.3.1 Initial drafting

For this phase the process followed was the one described by Noy and McGuinness, 2001, whose steps and results can be seen in Table 4.2.

The first step towards drafting the ontology consisted in defining a domain and scope with the help of competency questions.

Example: A competency question that helped define the scope was: *"Will a certain plant in a certain area be likely to host a certain pathogen vector?"*. By consequence the scope was set to include the transmission of diseases by vectors, as described in Table 4.2.

External resources were also considered. In particular classes from the NCBI Taxonomy Database were integrated in the ontology.

Example: the "organism" class in the PsylVe ontology corresponds to the "cellular organism" in the NCBI Taxonomy Database.

In order to begin defining elements of the ontologies, a list of important terms was redacted, which were subsequently assigned to the different roles in the ontology, that is classes, properties and relations.

Example: The term "vector", a pivotal term within the domain of the ontology, was initially assigned the role of class.

4.3.2 Refinement

The second phase consisted in refining the ontology via a continuous exchange with a domain expert[7] and a knowledge representation expert[8] to improve the structure and content of the ontology.

[7] Nicolas Sauvion (entomologist)
[8] Prof. Catherine Faron Zucker

Step 1. Determine the domain and scope of the ontology	*Objective:* Ontology-based text mining for geographical distribution of pathogen vectors. *Competency questions* • Will a certain production line be affected by a certain disease this year? • Will a certain plant in a certain area be likely to host a certain pathogen vector? • Can a certain organism host a certain pathogen? *Scope:* The scope of classes includes: the roles played in the pathogen transmission interaction between organisms, diseases resulting from vected pathogens, and human activities affected. For the scope of instances, see Table 2.1.
Step 2. Consider reusing existing ontologies	See Section 3.2 for a presentation of external ontologies.
Step 3. Enumerate important terms in the ontology	*Terms:* vector, organism, pathogen, disease, human activity, location, time.
Step 4. Define the classes and the class hierarchy	*Classes:* InanimateEntity, Disease, Host, HostPlant, ShelterPlant, FoodPlant, Guest.
Step 5. Define the properties of classes—slots	*Properties:* ill, organic, living, nonLiving, isPest, mouthPieces. *Relations:* Transmits, Carries.
Step 6. Define the facets of the slots	*Transmits*: cardinality = 1; domain = Insect; co-domain = Pathogen. *Carries*: cardinality = na; domain = Insect; co-domain = Pathogen.
Step 7. Create instances	Given the definition of instances (see Section 2.3.1) used in the PsylVe project, i.e. occurrence data, instances were considered to be out of scope for the current development of the KB

TABLE 4.2: Steps in the first draft of the ontology.

The domain expert provided a comprehensive summary of the domain knowledge in natural language, while the knowledge representation expert provided a plan for its representation via logical formalisms.

The result of this initial phase was a two-column table containing natural language sentences in the left column and their First Order Logic representation on the right. This was followed by regular interactions, alternating between specialists, as mentioned above, in an attempt to improve the representations in preparation for the encoding of the ontology (see Section 4.3.3).

Example: Although initially, a relation `VectorOf` had been defined, it became apparent that the predicate implied the relation to be binary, while, in reality, transmission relationships are ternary, involving a vector, a host plant, and a pathogen. As expressed above, the purpose of this phase was to prepare the representations to be encoded in a logic description language, making ternary relationships impossible to be defined as relations. Therefore, such relationships were defined as classes, to be instantiated and linked via secondary relations to each of the participants. For example, the relation `VectorOf` was substituted with the class `PathogenTransmissionRelation`, whose instances are linked to subclasses of `VectingCarrier` (the vector), `Plant` (the infected plant), and `Pathogen` (the pathogen being transmitted).

More concretely, in the case of the transmission relationship in the scope of the Internship, the class `PathogenTransmissionRelation` has to be instantiated and linked to the respective NCBI classes representing the taxa *C. pruni*, *Prunus*, and *Ca.* P. prunorum.

$$
\begin{aligned}
&\mathit{MemberOf}(r, \texttt{PathogenTransmissionRelation}) \\
&\qquad \iff \exists v, hp, p \; (\\
&\qquad\quad \mathit{MemberOf}(v, \texttt{VectingCarrier}) \\
&\qquad\quad \wedge \texttt{has\_participant}(r, v) \\
&\qquad\quad \wedge \mathit{MemberOf}(hp, \texttt{Plant}) \\
&\qquad\quad \wedge \texttt{has\_participant}(r, hp) \\
&\qquad\quad \wedge \mathit{MemberOf}(p, \texttt{Pathogen}) \\
&\qquad\quad \wedge \texttt{has\_participant}(r, p) \\
&\qquad\quad)
\end{aligned}
$$

This should be read as *r is an instance of a transmission relation if and only if it has an associated vectorcarrier, an associated plant and an associated pathogen*.

4.3.3 Encoding

The final step was the encoding of the ontology using turtle syntax. This specific syntax was chosen thanks to its human readability and compact representation style. The result of this step is a turtle file[9] that can be found on the project repository. It includes 45 `owl:Class` and 9 `owl:ObjectProperty` .

[9] `https://github.com/e-lubrini/psylve/blob/main/src/ontology/ontology.ttl`

4.4 Information Extraction from text

The objective of the pipeline is to extract observation data contained in various documents. In the context of vector geographical distribution mapping, observations are conceptually constituted by a quaternary relation between four entities: a vector, its host plant, and the date and location of the observation and practically correspond to classes and properties as described above. Some preliminary experiments were carried out to have an overview of possible issues to be targeted in the implementation of NER methods. In this chapter such preliminary experiments will be addressed in Section 4.4.1, together with a discussion of the results, followed by the methodology used for the NER in Section 4.5.1.

Since the relationship extraction was left for future development, the results of the NER subtask are the final results of the last processing module and, thus, the final results before the evaluation. The results of the IE module will be discussed in the evaluation.

4.4.1 Preliminary experiments

According to the original Internship objectives, the PsylVe IE module was planned to be inspired on the Omnicrobe IE pipeline. In order to identify problems to be addressed in the current and future development of the IE module, four representative and short texts extracted from the *C. pruni* document database were processed using the Omnicrobe IE pipeline.

The extracted texts resulting from the corresponding pipeline module were processed with the Omnicrobe NER pipeline and the resulting NER annotations were manually evaluated and discussed with the two supervisors, specialists of biology and NLP, respectively (see **??**)[10] (Tamanini, 1955; Gjonov, Cassar, and Mifsud, 2020; Horwood and Fitch, 1919; Özgen[1], Gözüaçık, and Burckhardt, 2012).

Methodology

To visualise the annotations, we used html format: the results of the four documents being processed were three HTML files containing an annotated version of their contents, with the annotations reflecting the entities that were detected by the Omnicrobe pipeline, namely microorganisms, taxa, and hosts. Six types of annotation errors were defined to be checked: three related to annotation boundaries, i.e. too wide, too narrow, and skewed, two related to detection, i.e. false positive and false negative, and one related to labelling, i.e. wrong label.

The HTML files were manually analysed in order to detect these annotation errors: In the following stage of the analysis, a cause of such annotation errors was identified within the pipeline. A total of 9 causes were identified: case sensitivity, discontinued entity coordination, encoding, foreign term, morphological variation, outdated term, pattern recognition, POS tagging, polysemy, pragmatics, and text extraction.

[10] Given the labour extensive nature of manual annotation tasks, only the results of a limited number of documents were annotated (see Section 4.4.1). More precisely, manual efforts were optimised by choosing the smallest document sample that would be linguistically representative of the whole dataset. Finally a total number of four documents was selected taking into account different periods of publication, presence or not of embedded text, and variety of languages.

Results and discussion

After the manual annotation[11] of the Omnicrobe pipeline results on the selected subset of the preprocessed *C. pruni* document database, a number of issues were identified.

Case sensitivity

Example Undetected string: "HEDGES"

Issue Some tokens, especially those composing vernacular names, which are not subject to taxonomic nomenclature formatting standards, are occasionally capitalised, often due to document formatting choices.

Solutions

– Disabling case sensitivity when a whole multi-character token is capitalised.

pros - Since taxonomic nomenclature does not apply capitalisation on whole tokens to encode information (as opposed as italicisation, which can be used to encode ranking information), case sensitivity of rules can be deactivated in cases of a whole tokens being capitalised, as it signals taxonomic nomenclature formatting standards are not being respected.

cons - The main downside is constituted by polysemy, as capitalisation could be used either for style formatting reasons or to point to a specific meaning, such as in the case of acronyms.

Discontinued entity via coordination

Example Undetected string: "Urtica dioica" in "Urtica urens and dioica"

Issue Wordforms belonging to the same entity can be discontinued, for example, via coordination, where a first wordform belonging to two separate entities is not repeated after the coordination, as it remains implied.

Solutions

– When an entity α composed of a sequence of n tokens precedes a coordination, all combinations of length n composed of the first $n - x$ tokens of α and x tokens following the coordination, for all $0 < x < n$ will be tested with the NER algorithm.[12]

pros - Both recall and precision of this method should approximate to one with coordinated entities of the same length.

[11] Full annotationfor full annotations) available at `https://github.com/e-lubrini/psylve/tree/internship/src/ner/data/preliminary_experiment.csv`

[12] In the provided examples, the entity α "Urtica urens" was detected, preceding a coordination. Given that the entity is of length $n = 2$, the only possible $0 < x < n$ is $x = 1$ so the first $n - x = 1$ token(s) of α - "Urtica" - will be joined with the $x = 1$ token(s) after the coordination - "diotica" - and "Urtica diotica", of length n, will be checked with the NER algorithm to decide whether it corresponds to one of the requested entities.

cons - Checking all coordinations preceded by an entity could be time consuming. Additionally patterns such as attributes separated by Additionally, complexity will increase with the number of tokens in consideration. State of the art algorithms will take a-t least $O(3^{n/3})$ steps, where n is the number of tokens (Wang et al., 2021).

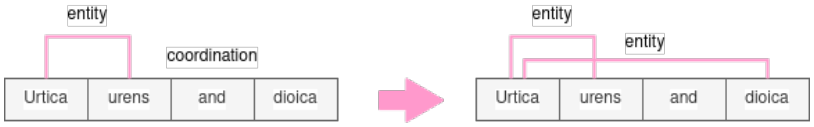


FIGURE 4.4: Detection of discontinued entity via coordination

Encoding

Example Undetected string: "Lady's Fingers"

Issue The visual similarity of some characters, in particular special characters, can cause errors by the OCR tools.

Solutions

- Detection of characters' common properties during NER[13].

  pros - Flexibility in the choice of treatment depending on the task to be performed and, possibly, other variables, such as Unicode character groups (e.g. treating some Unicode groups but not others) or document metadata (e.g. whether the text was extracted by OCR or was already embedded).

  cons - This solution would need the implementation of a new module in the AlvisNLP software.

- Normalisation of characters according to common Unicode properties during the text extraction module.

  pros - Avoiding possibly having to address the issue multiple times in several following tasks, by solving the problem at the root.

  cons - Reduced flexibility in the choice of treatment during following pipeline tasks, since some data will be lost in normalisation.

Foreign vernacular name

Example Undetected string: "psillidi"

Issue Vernacular names sometimes differ from language to language, therefore, translated documents in languages other than English might still include foreign terms as vernacular names. The problem has a greater impact when the taxonomic nomenclature is not used, as the identification of the taxon only relies on the normalisation of such foreign vernacular terms.

[13] In Unicode, the encoding system used by AlvisNLP, character properties are properties that sets of characters have in common, which value can be used to define character groups. For example, all characters with the property `White_Space=yes` are grouped into the set of 'whitespace' characters. This can be useful to process all white spaces independently of the specific character used.

Solutions

- Exploit a multilingual database.

  pros - Can be systematically applied to most well-resourced languages.

  cons - Any two languages could use similar vernacular names for different taxon or might not have an equivalent in English.

- Normalise wordforms via language-specific morphological rules.

  pros - Can be applied in absence of structured data in a given language.

  cons - Any two languages could use similar vernacular names for different taxon or might not have an equivalent in English, therefore morphological rules might not apply. Additionally, building rules for each language of interest might be time consuming and involve extensive linguistic research.

Morphological variation linked to POS

Example Undetected string: "auchenorrhynchan"

Issue Sometimes, taxa present morphological variations, usually to allow it to take on a different POS within the sentence[14].

Solutions

- Create morphological rules.

  pros - If rules are exhaustive, they can solve the need for additional data on morphological variations.

  cons - Recall could be low if morphological rules are not exhaustive. Additionally, building rules for each language of interest might be time consuming and involve extensive linguistic research.

- Relax restrictions. For example allow an entity to contain the desired term while allowing the boundaries of the entity to span larger than the bonds of the recognised characters.

  pros - This solution is easy and quick to implement.

  cons - Could generate false positives, such as in the case of some short taxon names that can be often contained in common English words.

polysemy

Example Mislabelled token: "flesh".

Issue Some tokens can point to a different meaning depending on the context[15]

Solutions

- Polysemy is a common problem to which Machine Learning solutions seem to hold the state of the art.

[14] In the provided example, "auchenorrhynchan" is composed of the morphemes "Auchenorrhyncha" and the suffix "-an" which is generally used to turn a noun into an adjective and/or denote individuals belonging to a group (e.g. Europe versus European)

[15] In the provided example, flesh was mislabelled as corresponding to the concept of the flesh of an animal, while in the context it referred to the plesh of a fruit.

pros - Current state of the art.

cons - Machine Learning approaches are often considered to be resource intensive and their quality depends on availability of a well performing model or good quality input.

– Word distance and similarities between entities.

pros - Less resource intensive and variety in the approaches available to calculate similarity.

cons - Less performing than state of the art methods.

POS tagging

Example Undetected string: "stalk" in "flower-stalk"

Issue When a token in a noun phrase is used as attribute and has an attribute for the following token, if the first is considered an entity, the second should be part of it too.

Solutions

– A rule can be implemented using POS tagging patterns as trigger.

pros - Relatively simple to implement and it gives flexibility to address various POS patterns.

cons - Higly dependent on the exhaustivity of the implemented rules. Because of this extensive linguistic research might need to be carried out, or *ad-hoc* rules could be implemented based on manual annotation of the issues in the treatment of the data.

Text extraction issue

Example Undetected string: "Payllopsis fraxini" because of incorrect spelling (correct spelling: "Psyllopsis fraxini")

Issue The text extraction fails to convert the image text to the right set of characters.

Solutions

– Introduce a taxon autocorrector to optimize the occurrence dataset taxa.

pros - It can correct tokens that are not normally included in other dictionaries.

cons - If it is not very precise it will result it will overestimate the amount of occurrences in a giving text, resulting on higher numbers of false positive detection.

Unknown term

Example Undetected string: Bullace

Issue Some terms are not included in the exploited resources therefore the Named Entity Recognition algorithm is unable to recognize them.

Solutions

– Integrate further lexical resources into the NER algorithm.

  pros - If we find lexical resources that cover many of the not included entities we can increase the coverage of the system.

  cons - Regardless of how many resources are included there is still a possibility that not all the required entities are included, therefore becoming less efficient without necessarily improving the coverage.

– Manually add synonyms encountered in the text.

  pros - It allows domain experts to continuously add knowledge, making it possible to keep updating the knowledge base as needed.

  cons - Manually adding synonyms is time consuming and the process is not efficient.

Pattern recognition

Hearst's patterns such as enumeration, first described by HearstHearst, 1992.

Examples

Undetected string: "Blackthorn" in "Blackberry, Blackthorn, Blackthorn-May, Buckthorn"

Undetected string: "Blackthorn Chats" in "Blackthorn Chats are the young shoots"

Issue Failures in other methods can be stemmed via syntactic structure analyses.

Solutions Some syntactic patterns commonly containing entities can be detected via POS[16] and punctuation[17] detection.

pros - For the most common patterns the recall could drastically improve, especially if compared to the relative low effort required to build such rules.

cons - Depending on the rule, false positives could be common.

4.5 Information Extraction from text

The objective of the pipeline is to extract observation data contained in various documents. In the context of vector geographical distribution mapping, observations are conceptually constituted by a quaternary relation between a vector, its host plant, and the date and location of the observation.

Within the scope of the Internship, only NER was carried out. Normalisation and RE were left for future work (see Section 6.2).

[16] As shown in one of the provided examples, a noun phrase followed by a copula verb and a noun phrase containing an entity represents a common pattern in which the first noun phrase should be also recognised as an entity.

[17] In one of the provided examples, a list of terms is detected, although not all relevant terms were detected as entities. A rule could be set to detect enumeration of entities in order to recognise all elements.

4.5.1 NER

In this section, the strategies applied for the extraction of each of the four targeted entities will be briefly discussed. Four different NER pipeline versions were implemented. A different approach was used in each of the four versions, for the recognition of taxonomic entities, i.e. those corresponding to the vector(*C. pruni*) and the host plant (host plant), while the same strategy was implemented across versions for the recognition of location and date respectively. Following is a list of the strategies with their relative pros and cons. Full results and discussion will be provided in Chapter 5.

Vector Given the scope of the current thesis (see Table 2.1) the vectors to be recognised were set to the species complex *C. pruni*.

V1 Psyl'list projection of *C. pruni* - The names and historical synonyms of the desired vectors (i.e. *C. pruni*) were extracted from the Psyl'list database and projected onto the extracted texts.

> **pros -** As opposed to the NCBI dataset, Psyl'list has richer historical records on the evolution of nomenclature around psyllid taxa, which means that more word forms could potentially be recognised as belonging to a the desired entity, namely the taxon corresponding to *C. pruni*.
>
> **cons -** It is a very specific dataset with a a relatively small size compared to the NCBI.

V2 NCBI taxon identification of *C. pruni* - An AlvisNLP module was used to extract word forms recognised by the NCBI Taxonomy Database (database presented in Section 3.2).

> **pros -** The module has been already successfully implemented in the Omnicrobe pipeline. The amount of taxa included in the database allows for flexibility regarding the choices of vectors to be included.
>
> **cons -** Since the current scope only spans over a precise species complex, the range of taxa included in the NCBI database is not being used to its full potential. Additionally, the NCBI database might not be the best resource in terms of recording historical nomenclature changes for each taxon and, given the width of the publication date range of the inputted documents, it might need to be complemented with external resources.

V3 NCBI taxon identification of psyllid - The same strategy that was used in version 2, but with the scope extended from *C. pruni* to psyllid.

> **pros -** This extending of the scope allows a larger number of vector entities to be extracted from the dataset. We expect this version to have the highest recall of all versions.
>
> **cons -** Since the scope is wider, other species in the same taxon may be captured as vector entities, which is expected to decrease the overall precision of the results.

V4 'pruni' string match - A string match was used corresponding to the species name *C. pruni*.

pros - Since only the species of the vectorentities are extracted from a given dataset, it is expected that the precision may be higher than version 3.

cons - While precision is expected to be higher due to the specificity of this particular string match, it is also expected to decrease recall in instances where the species name is not included.

Host Plant Given the scope of the current thesis (see Table 2.1) the host plant to be recognised were set to the *Prunus* genus.

V1 Psyl'list projection of *Prunus* - The names and historical synonyms of the desired host plants (i.e. *Prunus*) were extracted from the Psyl'list database and projected onto the extracted texts.

pros - As opposed to the NCBI dataset, Psyl'list has richer historical records on the evolution of nomenclature around psyllid taxa, which means that more word forms could potentially be recognised as belonging to a the desired entity, namely the taxon corresponding to *Prunus*.

cons - It is a very specific dataset with a a relatively small size compared to the NCBI.

V2 NCBI taxon identification of *Prunus* - An AlvisNLP module was used to extract word forms recognised by the NCBI Taxonomy Database (database presented in Section 3.2).

pros - The module has been already successfully implemented in the Omnicrobe pipeline. The amount of taxa included in the database allows for flexibility regarding the choices of host plants to be included.

cons - Since the current scope only spans over a precise species complex, the range of taxa included in the NCBI database is not being used to its full potential. Additionally, the NCBI database might not be the best resource in terms of recording historical nomenclature changes for each taxon and, given the width of the publication date range of the inputted documents, it might need to be complemented with external resources.

V3 NCBI taxon identification of psyllid - The same strategy that was used in version 2, but with the scope extended from *Prunus* to *Rosaceae*.

pros - This extending of the scope allows a larger number of host plants entities to be extracted from the dataset. We expect this version to have the highest recall of all versions.

cons - Since the scope is wider, other species in the same taxon may be captured as host plants entities, which is expected to decrease the overall precision of the results.

V4 'Prunus' string match - A string match was used corresponding to the species name *Prunus*.

pros - Since only the species of the host plants entities are extracted from a given dataset, it is expected that the precision may be higher than version 3.

cons - While precision is expected to be higher due to the specificity of this particular string match, it is also expected to decrease recall in instances where the species name is not included.

Date In concordance with the format of the manually extracted occurrence data, only the year is needed.

A RE detecting four consecutive digits starting with 1 or 2 was instantiated.

pros - A simple RE is easy and quick to implement. Chances of false positives are relatively low, at least in the given application domain[18].

cons - Whenever four consecutive digits starting with 1 or 2 are detected, they will be automatically classified as dates, since no other discriminant was set.

Location According to the available manually annotated dataset, the scope of the location detection was set to country names.

Within the scope of the current thesis' evaluation, only country names were needed. A list of countries extracted from Wikidata[19] was projected onto the text.

pros - The solution was easy and quick to implement, on top of being precise and easily customisable.

cons - No cons were detected for the purpose of identifying the country; however if one wanted to detect a more precise location, a more comprehensive resource should be used.

4.5.2 Evaluation

The evaluation of the IE module of the PsylVepipelinecorresponds to the evaluation of the whole pipeline, since the results outputted by this module are the final result expected by the pipeline. See Chapter 5 for the evaluation of the pipeline.

[18] four-digit codes and numbers are not commonly found in the database documents, except when referring to observation or publication years.

[19] `https://www.wikidata.org/wiki/Wikidata:WikiProject_Countries`

Module	Task	Main Tools and Resources
Data Composition Analysis	FCA	Python packages: `pysubgroup`
	Visualisation	Python packages: `matplotlib` `seaborn`
Text Extraction	Image to PDF	Python packages: `fpdf.Fpdf`
	PDF to image	Python packages: `fitz` `PIL.Image`
	OCR text extraction	Python packages: `pytesseract` `Pypdf4` `pdfminer.six` `Pypdf4` `TIKA` `pdfreader`
	Embedded text extraction	Bash software: `GROBID`
	Language Recognition	Python package: `fasttext` Model: `lid.176.ftz`
	Translation to English	Python packages: `deep_translator`
	Spellcheck	Python packages: `enchant`
Ontology Development		Description frameworks: `owl` `rdf` `rdfs`
Information Extraction	Projection	AlvisNLP modules: `TabularProjector`
	Segmentation	AlvisNLP modules: `WoSMig` SeSMig
	Pattern matching	AlvisNLP modules: `PatternMatcher` `RegExp`
	Result export	AlvisNLP modules: `TabularExport` `QuickHTML`

TABLE 4.3: NLP tools used for each task

Chapter 5

Evaluation

The pipeline developed during the Internship produced valuable results and is already being integrated in workflows to be exploited in various other projects at MaIAGE. This pipeline represents a prototype to be used as the basis for more ambitious developments. The evaluation of the proposed framework on our pipeline produced satisfying results that show it is useful and functional; however, this results should not be considered representative of the overall quality of the framework since they only apply to a specific task of our pipeline.

For this reason, both quantitative and qualitative criteria were taken into account in the evaluation module. The criteria used for the two evaluations are those mentioned in Section 2.4.

5.1 Quantitative evaluation

As mentioned in Section 2.4.2, three main scores will be used for the evaluation of the PsylVe pipeline, each of which has its own implications on the related workflow and can be prioritised over the others depending on the context.

In the case of the current pipeline, it is important to remember that the results that are being evaluated are the output of a Named Entity Recognition pipeline, while the reference data against which the results are being evaluated is not composed of all entities found in the text, but rather a subset of entities that take part in specific relations. Because the RE functionality has not yet been implemented, the objective of this Internship was to maximise the recall, so that as many relevant candidates as possible can occur in the RE phase.

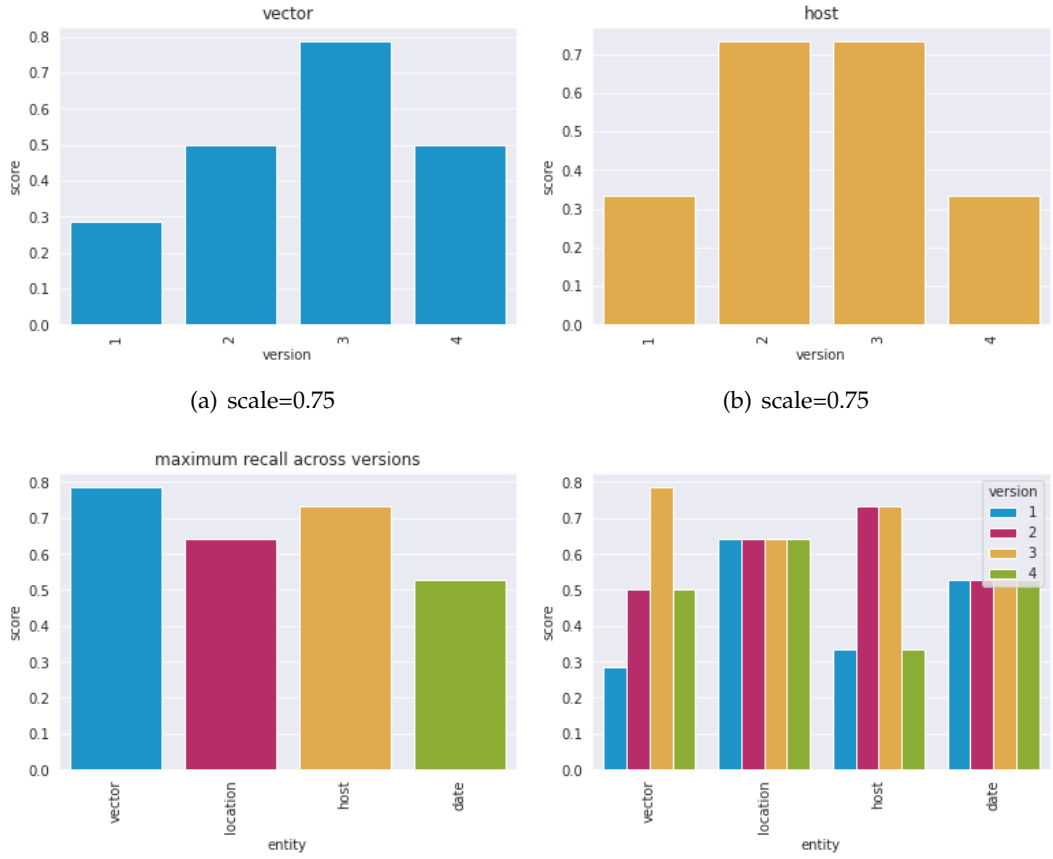
	Entity			
	Vector	Host	Location	Date
NER V1	*C. pruni* in Psyl'list	Host plants of *C. pruni* in Psyl'list	Wikipedia list	regex
NER V2	*C. pruni* in NCBI	*Prunus* in NCBI	Wikipedia list	regex
NER V3	psyllid in NCBI	*Rosaceae* in NCBI	Wikipedia list	regex
NER V4	"pruni" string match	"Prunus" string match	Wikipedia list	regex

TABLE 5.1: Versions of the NER pipelines according to approach combination

5.1.1 Recall

In Figure 5.1, the recall scores were plotted against the entities to be extracted and the NER pipeline versions, as defined in **??**. It is important to point out that *C. pruni* occurrence dataset contains some inexact data. On top of that, the multiple layers of automation reduces the overall quality of the results.

Figure 5.1 (a), shows the recall plotted against the pipeline versions in the extraction of vector entities as described in Table 5.1. In the first version, the list of *C. pruni* published by Psyl'list shows a low recall because of the strong constrains imposed by the exact nomenclature. As expected, widening the scope from the species *C. pruni* to all psyllid entities drastically increased the recall. In Figure 5.1 (b), the recall was plotted against the pipeline versions when extracting host plant entities. Figure 5.1 (c) shows the maximum recall score which was obtained by the third version of the pipeline. Figure 5.1 (d) shows the recall scores for all versions of the pipeline grouped by the entity types.



(a) scale=0.75 (b) scale=0.75

(c) scale=0.75 (d) scale=0.75

FIGURE 5.1: (a) Recall for vector entities across pipeline versions (b) Recall for host entities across pipeline versions (c) Maximum recall for each entity type across pipeline versions (d) Recall scores for all versions grouped by entities

5.1.2 Precision

Figure 5.2 shows the precision scores of all versions of the PsylVe Pipeline grouped by entity type.

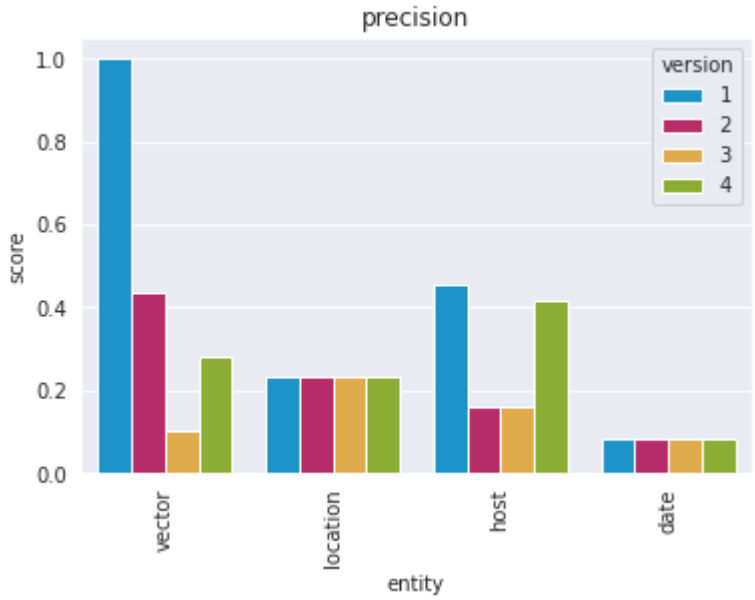


FIGURE 5.2: Precision scores for all versions grouped by entities

The precision score is negatively affected in the NER stage before relationship extraction. Since the relationship extraction module has yet to be implemented all entities are being extracted regardless of whether or not they are related to entities present in the dataset. This includes, for example, entities that do not form part of an observation.

5.1.3 F1-score

The bias on the precision score described on Section 5.1.2 also affects the F1 score as can be sen on Figure 5.3.

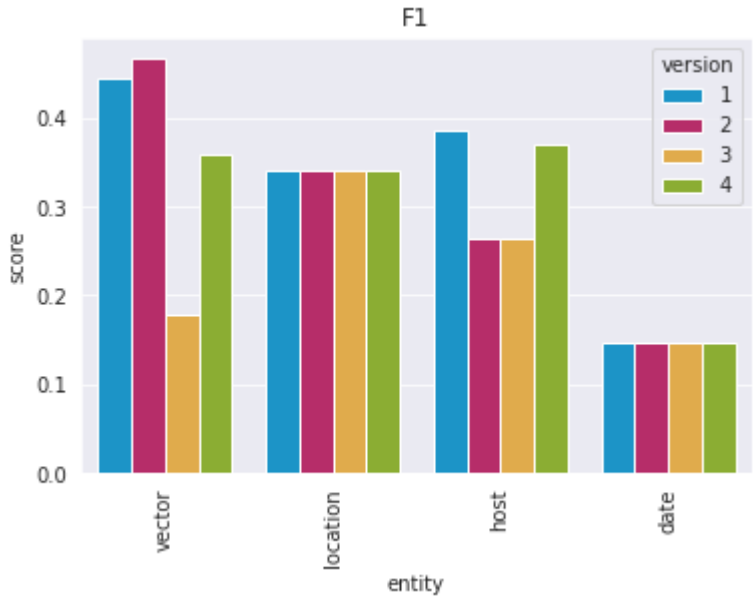


FIGURE 5.3: F1 scores for all versions grouped by entities

5.2 Qualitative evaluation

5.2.1 Evaluation criteria

For this stage of the text mining pipeline evaluation the following criteria were taken into consideration as presented in Spinakis and Peristera (2004).

Criterion 1: Data retrieval and result evaluation This criterion focuses on the quantitative aspect of the result evaluation. Therefore, the scores mentioned in Section 2.4.2 were used. See other criteria for the qualitative evaluation.

Criterion 2: Integration with other sources Integrating a text mining pipeline with external resources allows for more and possible better results. The PsylVe pipeline was integrated with a number of resources, which were presented in Chapter 3. To recap, both features of an external pipeline (see Section 3.1) and structured data (see Section 3.2 and Section 3.3) were integrated. This allows for compatibility with a similar pipeline (Omnicrobe[1]) and compatibility with knowledge bases that are widely used in the plant health domain. Additionally, the database integration allowed for rigorous methodical access to peer-reviewed resources.

Criterion 3: Output format flexibility Allowing users to enable different outputs improves the integration of the pipeline with other workflows. A number of output formats are allowed within the PsylVe pipeline. The ones used for this thesis' experiments were `json` and `csv` for the quantitative evaluation and `html` for the qualitative evaluation.

Criterion 4: Availability of result statistics and analysis In this project, the availability of result statistics and analysis was done.

As mentioned in section /ref the availability of statistics and analysis tools allow the user to have a better overview of the results. Tools for evaluating the PsylVe Pipeline are provided. These tools include:

steps for evaluation: design experiments analyse the results propose solutions - attempt to anticipate the pros and cons of each solution.

Criterion 5: Combining linguistic and statistic methods In order to exploit the potential of both linguistic knowledge and statistical tools, both can be integrated into the PsylVe Pipeline with ease. This is due to its highly customisable nature.

An example of a linguistic tool that was implemented in this version of the pipeline is NER. Due to a combination of time constraints of the Internship and the size of the dataset to be processed, the implementation of statistical tools is left for future iterations of the PsylVe Pipeline.

Criterion 6: Online tool availability Online availability of a tool widens its accessibility and potential userbase. The source code for the PsylVe pipeline is easily accessible via the PsylVe GitHub repository[2]. Additionally, detailed documentation was included in order to increase the ease of implementation in different domains.

[1] `https://omnicrobe.migale.inrae.fr/`

[2] `https://github.com/e-lubrini/psylve`

FIGURE 5.4: An example of HTML output of the pipeline.

Criterion 7: Result visualisation quality Visualising results in an easy-to-understand and informative way allows for an overall better user experience and better integration in the user's workflow. The PsylVe tool allows for an HTML output that visually points to the requested information within the inputted text (see Figure 5.4).

Chapter 6

Conclusion

The original goal of this Internship was to create an add-on for the Omnicrobe Pipeline which automatically extracted *C. pruni* occurrence data. Over the course of the Internship, the scope has expanded far beyond a simple add-on into a full-fledged framework (as seen in Figure 1.4) with a proof-of-concept pipeline.

6.1 Results

The quantitative results were deemed satisfactory considering: (1) the amount of consecutive levels of automation and (2) amount of features implemented relative to the resources assigned to the Internship[1].

In terms of qualitative results, the most obvious measure of success relates to how well the framework contributes towards the objectives of the BEYOND project. Since the framework is designed to be modular and easy to implement, it allows biologists to save time and resources when analysing large datasets. Due to this, the framework has already begun to be used in various projects in the MaIAGE department of INRAE. However, some limitations of the current version of the framework remain.

Due to the numerous consecutive layers of automation, both precision and recall are not perfect, therefore the workflow of geographical distribution mapping should be completed by manual intervention. The final section discusses future steps which could be taken to improve the performance of pipelines created using this framework.

6.2 Future Work

In this section, some possible future steps for each of the framework modules will be outlined.

6.2.1 Database Setup

For what concerns the setup of the database, some improvements could still be made to the tabular data, such as implementation of multiple tables within a SQL database, in order to avoid information repetition and ensure proper indexing. Additionally, automatic tools for metadata extraction that are currently being used in the data preprocessing module could be moved earlier in the pipeline, in order to add valuable information in the data analysis subtask.

[1]The implementation of more elaborate approaches (e.g. implementing current state of the art for RE) and expensive tools (e.g. commercial OCR tools) was not feasible due to time and budget constraints.

6.2.2 Data Preprocessing

Improvements of the OCR subtask could drastically improve results. As mentioned in Section 4.5, the implementation of a taxon autocorrector would compensate for the faults of the implemented OCR tools. Integration of other OCR tools could also be considered.

6.2.3 Knowledge Base

The PsylVe ontology will be continuously improved by adding and refining its elements and integrated into the IE pipeline. More ambitious developments would include a system to cyclically populate a KB, propose and integrate ontology expansions based on the newly acquired data, and use such expanded ontology to detect more data.

6.2.4 Information Extraction

The Internship focused on the NER part of the IE, therefore normalisation and RE have yet to be implemented and will be integrated in the following stage of the project. Regarding the NER subtask, state of the art Machine Learning approaches will be implemented.

6.2.5 Evaluation

Additional evaluation workflows could be implemented to evaluate each module. Regarding the specific pipeline targeting *C. pruni* occurrences, the improvement and correction of the mistakes in the *C. pruni* occurrence dataset, which was used for the evaluation, in predicted to play an important role in the improvement of the quantitative results.

Bibliography

Alatrish, Emhimed Salem, Dušan Tošić, and Nikola Milenković (2014). "Building ontologies for different natural languages". In: *Computer Science and Information Systems* 11.2, pp. 623–644. DOI: 10.2298/CSIS130429023A.

Arora, Monika, Uma Kanjilal, and Dinesh Varshney (2016). "Evaluation of information retrieval: precision and recall". In: *International Journal of Indian Culture and Business Management* 12.2, pp. 224–236. DOI: 10.1504/IJICBM.2016.074482.

Assunta, Bertaccini, G Weintraub Phyllis, Pratap Rao Govind, and Mori Nicola (2019). *Phytoplasmas: Plant pathogenic bacteria-II: Transmission and management of phytoplasma-associated diseases*. Springer Singapore. DOI: 10.1007/978-981-13-2832-9.

Aubertot, Jean-Noël, Jean-Marc Barbier, Alain Carpentier, Jean-Joël Gril, Laurence Guichard, Philippe Lucas, Serge Savary, I. Savini, and Marc (éditeurs) Voltz (2015). "Pesticides, agriculture et environnement : Réduire l'utilisation des pesticides et en limiter les impacts environnementaux". In: *Expertise scientifique collective, synthèse du rapport, INRA et Cemagref (France)*, p. 64. URL: https://www.inrae.fr/sites/default/files/pdf/synthese-expertise-68-pages.pdf.

Ba, Mouhamadou and Robert Bossy (2016). "Interoperability of corpus processing workflow engines: the case of. AlvisNLP/ML in OpenMinTeD". In: *Meeting of working Group Medicago sativa*. Portoroz, Slovenia, np. DOI: https://hal.archives-ouvertes.fr/hal-01455853/document.

Bach, Nguyen and Sameer Badaskar (2007). "A review of relation extraction". In: *Report. Language Technologies Institute, Carnegie Mellon University*. URL: https://nguyenbh.github.io/publication/bach-badaskar-2007.

Baggia, Paolo, Paul Bagshaw, Michael Bodell, Zhi Huang De, Lou Xiaoyan, Scott McGlashan, Jianhua Tao, Yan Jun, Hu Fang, Yongguo Kang, Helen Meng, Wang Xia, Xia Hairong, and Zhiyong Wu (2010). In: *Speech synthesis markup language (SSML) version 1.1*. URL: https://www.w3.org/TR/speech-synthesis11/.

Baskauf, Steven J and Campbell O Webb (2016). "Darwin-SW: Darwin Core-based terms for expressing biodiversity data as RDF". In: *Semantic Web* 7.6, pp. 629–643.

Belfodil, Adnene (Oct. 2019). "Exceptional model mining for behavioral data analysis". Theses. Université de Lyon. URL: https://hal.archives-ouvertes.fr/tel-02335097.

Belohlávek, Radim (2008). "Relational Data, Formal Concept Analysis, and Graded Attributes". In: *Handbook of Research on Fuzzy Information Processing in Databases*.

Brown, Judith K., D.R. Frohlich, and R.C. Rosell (1995). "The sweetpotato or silverleaf whiteflies: biotypes of Bemisia tabaci or a species complex?" In: *Annual review of entomology* 40.1, pp. 511–534. DOI: 10.1146/annurev.en.40.010195.002455.

Burckhardt, Daniel, David Ouvrard, and Diana M Percy (2021). "An updated classification of the jumping plant-lice (Hemiptera: Psylloidea) integrating molecular and morphological evidence". In: *European Journal of Taxonomy* 736, pp. 137–182. DOI: 10.5852/ejt.2021.736.1257.

Burckhardt, Daniel, David Ouvrard, Dalva Queiroz, and Diana Percy (2014). "Psyllid host-plants (Hemiptera: Psylloidea): resolving a semantic problem". In: *Florida entomologist* 97.1, pp. 242–246. DOI: 10.1653/024.097.0132.

Burns, Ed and Kate Brush (Mar. 2021). *What is deep learning and how does it work?* URL: https://www.techtarget.com/searchenterpriseai/definition/deep-learning-deep-neural-network.

Cafarella, Michael J, Alon Halevy, and Jayant Madhavan (2011). "Structured data on the web". In: *Communications of the ACM* 54.2, pp. 72–79. DOI: 10.1145/1897816.1897839.

Chen, Qingyu, Robert Leaman, Alexis Allot, Ling Luo, Chih-Hsuan Wei, Shankai Yan, and Zhiyong Lu (2021). "Artificial intelligence in action: addressing the COVID-19 pandemic with natural language processing". In: *Annual review of biomedical data science* 4, pp. 313–339. DOI: 10.1146/annurev-biodatasci-021821-061045.

Contributor, TechTarget (Sept. 2005). *What is first-order logic? - definition from whatis.com*. URL: https://www.techtarget.com/whatis/definition/first-order-logic.

Courtot, Mélanie, Frank Gibson, Allyson L Lister, James Malone, Daniel Schober, R Brinkman, and Alan Ruttenberg (2009). "MIREOT: the minimum information to reference an external ontology term". In: *Nature Preceding*. DOI: 10.1038/npre.2009.3576.1.

Dangles, Olivier, Verónica Mesías, Verónica Crespo-Perez, and Jean-François Silvain (2009). "Crop damage increases with pest species diversity: evidence from potato tuber moths in the tropical Andes". In: *Journal of Applied Ecology* 46.5, pp. 1115–1121. DOI: 10.1111/j.1365-2664.2009.01703.x.

Darnala, Baptiste, Florence Amardeilh, Catherine Roussey, Konstantin Todorov, and Clement Jonquet (2022). "Ontological Representation of Cultivated Plants: Linking Botanical and Agricultural Usages". In: *1st Workshop on Modular Knowledge @ESWC 2022, Hersonissos, Greece*. URL: https://hal.archives-ouvertes.fr/hal-03679652.

Del Gratta Riccardo, Caselli Tommaso and Nilda Ruimy (2008). "TIMEML: An ontological mapping onto UIMA Type Systems". In: *ICGL 2008, The First International Conference on Global Interoperability for Language Resources, Hong Kong*. URL: https://publications.cnr.it/doc/84724.

Dérozier, Sandra, Robert Bossy, Louise Deléger, Mouhamadou Ba, Estelle Chaix, Olivier Harlé, Valentin Loux, Hélène Falentin, and Claire Nédellec (2022a). "Omnicrobe, an open-access database of microbial habitats and phenotypes using a comprehensive text mining and data fusion approach". In: *bioRxiv*. DOI: 10.1101/2022.07.21.500958.

— (2022b). "Omnicrobe, an open-access database of microbial habitats and phenotypes using a comprehensive text mining and data fusion approach". In: *bioRxiv*. DOI: 10.1101/2022.07.21.500958. URL: https://biorxiv.org/cgi/content/short/2022.07.21.500958v1.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Disease Control, Centers for and Prevention (Dec. 2021). *Lesson 1: Introduction to Epidemiology*. en-us. URL: https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section1.html.

Eberendu, Adanma Cecilia (2016). "Unstructured Data: an overview of the data of Big Data". In: *International Journal of Computer Trends and Technology* 38.1, pp. 46–50. DOI: 10.14445/22312803/IJCTT-V38P109.

Eppo Global Database (2022). URL: https://gd.eppo.int/.

Feldman, Ronen, James Sanger, et al. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press. DOI: 10.5860/CHOICE.44-5684.

Ferré, Arnaud, Louise Deléger, Robert Bossy, Pierre Zweigenbaum, and Claire Nédellec (2020). "C-Norm: a neural approach to few-shot entity normalization". In: *BMC bioinformatics* 21:579. DOI: 10.1186/s12859-020-03886-8.

Ganter, Bernhard and Rudolf Wille (2012). *Formal concept analysis: mathematical foundations*. Springer-Verlag Berlin and Heidelberg GmbH & Co. K, p. 284. DOI: 10.1007/978-3-642-59830-2.

Gargominy, Olivier (2022). "TAXREF. Version 4.8". In: *GBIF, UMS PatriNat (OFB-CNRS-MNHN). Paris*. 2022-08-17. DOI: 10.15468/vqueam.

Gilbertson, Robert L, Ozgur Batuman, Craig G Webster, Scott Adkins, et al. (2015). "Role of the insect supervectors Bemisia tabaci and Frankliniella occidentalis in the emergence and global spread of plant viruses". In: *Annu. Rev. Virol* 2.1, pp. 67–93. DOI: 10.1146/annurev-virology-031413-085410.

Gjonov, Ilia, Thomas Cassar, and David Mifsud (2020). "New records of Hemiptera from the Maltese Islands". In: DOI: 10.17387/BULLENTSOCMALTA.2020.16.

Graën, Johannes, Mara Bertamini, Martin Volk, Mark Cieliebak, Don Tuggener, and Fernando Benites (2018). "Cutter–a universal multilingual tokenizer". In: *CEUR Workshop Proceedings*. 2226. CEUR-WS, pp. 75–81. DOI: 10.5167/UZH-157243.

Grishman, Ralph and Beth Sundheim (1996). "Message Understanding Conference-6: A Brief History". In: *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. URL: https://aclanthology.org/C96-1079.

Group, Darwin Core Task (2009). *Darwin Core*. URL: http://www.tdwg.org/standards/450.

Gruber, Tom (2018). *Ontology.* URL: https://tomgruber.org/writing/ontology-in-encyclopedia-of-dbs.pdf.

Hearst, Marti A (1992). "Automatic acquisition of hyponyms from large text corpora". In: *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Herrbach, Etienne, Nicolas Sauvion, Elisabeth Boudon-Padieu, Jean-Michel Lett, Bernard Reynaud, and René Sforza (2013). "Une relation trophique originale: la vection entomophile d'agents pathogènes". In: *Interactions insectes-plantes, edited by N. Sauvion, P.-A. Calatayud, D. Thièry and F. Marion-Poll. IRD Editions Editions QUAE 2013*, pp. 511–548. URL: https://hal.archives-ouvertes.fr/hal-01927478.

Horwood, A. R. and J. N. Fitch (1919). *A new British flora: British wild flowers in their natural haunts*. Gresham. DOI: 10.5962/bhl.title.18045.

HPSC (Mar. 19, 2019). *What is disease surveillance?* URL: https://www.hpsc.ie/abouthpsc/whatisdiseasesurveillance/.

Jarausch, Barbara, Rosemarie Tedeschi, Nicolas Sauvion, Jürgen Gross, and Wolfgang Jarausch (2019). "Psyllid vectors". In: *Phytoplasmas: plant pathogenic Bacteria-II*. Springer, pp. 53–78. DOI: 10.1007/978-981-13-2832-9_3.

Kaiser, Tim B and Stefan E Schmidt (2011). "Some remarks on the relation between annotated ordered sets and pattern structures". In: *International Conference on Pattern Recognition and Machine Intelligence*. Springer, pp. 43–48. DOI: 10.1007/978-3-642-21786-9_9.

Kaiser, Tim B, Stefan E Schmidt, and Cliff A Joslyn (2006). "Concept lattice representations of annotated taxonomies". In: *International Conference on Concept Lattices and Their Applications*. Springer, pp. 214–225. DOI: 10.1007/978-3-540-78921-5_14.

Lapage, SP, PHA Sneath, EF Lessel, VBD Skerman, HPR Seeliger, and WA Clark (1992). "International code of nomenclature of bacteria: bacteriological code, 1990 revision". In.

Lefèvre, Thierry, Nicolas Sauvion, Rodrigo PP Almeida, Florence Fournet, and Haoues Alout (2022). "The ecological significance of arthropod vectors of plant, animal, and human pathogens". In: *Trends in Parasitology*.

Liddy, Elizabeth D. (1990). "Anaphora in natural language processing and information retrieval". In: *Information processing & management* 26.1, pp. 39–52. DOI: 10.1016/0306-4573(90)90008-P.

Linné, Carl von (1746). *Caroli Linnaei... Fauna svecica, sistens animalia Sveciae regni: Quadrupedia, Aves, Amphibia, Pisces, Insecta, Vermes, distributa per classes & ordines, genera & species, cum differentiis specierum, synonymis autorum, nominibus incolarum, locis habitationum, descriptionibus insectorum*. sumtu & literis L. Salvii.

Loshin, Peter (Dec. 2021). *What is bash? (Bourne again shell)*. URL: https://www.techtarget.com/searchdatacenter/definition/bash-Bourne-Again-Shell.

LSPN connect (2022). URL: https://www.lspnconnect.com/.

Lumpe, Lars and Stefan E Schmidt (2015). "Pattern Structures and Their Morphisms." In: Sadok Ben Yahia, Jan Konecny (Eds.), CLA, pp. 171–179. URL: http://ceur-ws.org/Vol-1466/proceedings-cla2015.pdf#page=183.

Malone, James, Robert Stevens, Simon Jupp, Tom Hancocks, Helen Parkinson, and Cath Brooksbank (2016). "Ten simple rules for selecting a bio-ontology". In: *PLoS computational biology* 12.2, e1004743. DOI: 10.1371/journal.pcbi.1004743.

Maltese, Vincenzo and Feroz Farazi (2013). "A semantic schema for GeoNames". In: *Technical Report # DISI-13-004*. URL: http://eprints.biblio.unitn.it/4088/1/techRep004.pdf.

Mansouri, Alireza, Lilly Suriani Affendey, and Ali Mamat (2008). "Named entity recognition approaches". In: *International Journal of Computer Science and Network Security* 8.2, pp. 339–344.

Marie-Jeanne, Véronique, François Bonnot, Gaël Thébaud, Jean Peccoud, Gérard Labonne, and Nicolas Sauvion (2020). "Multi-scale spatial genetic structure of the vector-borne pathogen 'Candidatus phytoplasma prunorum' in orchards and in wild habitats". In: *Scientific Reports* 10.1, p. 5002. DOI: 10.1038/s41598-020-61908-0. URL: https://hal.inrae.fr/hal-02548441.

Merriam-Webster, Incorporated (2022a). *Geographical Distribution*. URL: https://www.merriam-webster.com/dictionary/geographical\%20distribution.

— (2022b). *Pathogen*. en. URL: https://www.merriam-webster.com/dictionary/pathogen.

Meystre, Stéphane M, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle (2008). "Extracting information from textual documents in the electronic health record: a review of recent research". In: *Yearbook of medical informatics* 17.01, pp. 128–144.

Mitchell, Tom, Bruce Buchanan, Gerald DeJong, Thomas Dietterich, Paul Rosenbloom, and Alex Waibel (1990). "Machine learning". In: *Annual review of computer science* 4.1, pp. 417–433.

Molik, David C, DeAndre Tomlinson, Shane Davitt, Eric L Morgan, Matthew Sisk, Benjamin Roche, Natalie Meyers, and Michael E Pfrender (2021). "Combining natural language processing and metabarcoding to reveal pathogen-environment associations". In: *PLoS neglected tropical diseases* 15.4, e0008755. DOI: 10.1371/journal.pntd.0008755.

Moncecchi, Guillermo, Jean-Luc Minel, and Dina Wonsever (2010). "A survey of kernel methods for relation extraction". In: *Workshop on NLP and Web-based Technologies. Bibliography 33*.

Morris, Cindy E, Ghislain Geniaux, Claire Nédellec, Nicolas Sauvion, and Samuel Soubeyrand (2022). "One Health concepts and challenges for surveillance, forecasting, and mitigation of plant disease beyond the traditional scope of crop production". In: *Plant Pathology* 71.1, pp. 86–97. DOI: 10.1111/ppa.13446.

Nadkarni, Prakash M, Lucila Ohno-Machado, and Wendy W Chapman (2011). "Natural language processing: an introduction". In: *Journal of the American Medical Informatics Association* 18.5, pp. 544–551. DOI: 10.1136/amiajnl-2011-000464.

Nédellec, Claire, Robert Bossy, Estelle Chaix, and Louise Deléger (2018). "Text-mining and ontologies: new approaches to knowledge discovery of microbial diversity". In: *arXiv preprint arXiv:1805.04107*. DOI: 10.48550/arXiv.1805.04107.

Nédellec, Claire, Adeline Nazarenko, and Robert Bossy (2009). "Information extraction". In: *Staab, S., Studer, R. (eds) Handbook on ontologies*. International Handbooks on Information Systems. Springer,Berlin, Heidelberg, pp. 663–685. DOI: 10.1007/978-3-540-92673-3_30.

Nguyen, Thien Huu and Ralph Grishman (2015). "Relation extraction: Perspective from convolutional neural networks". In: *Proceedings of the 1st workshop on vector space modeling for natural language processing*, pp. 39–48. DOI: 10.3115/v1/W15-1506.

Noy, Natalya F and Deborah L McGuinness (2001). "Ontology development 101: A guide to creating your first ontology". In: URL: https://corais.org/sites/default/files/ontology_development_101_aguide_to_creating_your_first_ontology.pdf.

Ouvrard, David, Pragya Chalise, and Diana M Percy (2015). "Host-plant leaps versus host-plant shuffle: a global survey reveals contrasting patterns in an oligophagous insect group (Hemiptera, Psylloidea)". In: *Systematics and Biodiversity* 13.5, pp. 434–454. DOI: 10.1080/14772000.2015.1046969.

Özgen[1], İnanç, Celalettin Gözüaçık, and Daniel Burckhardt (2012). "Contribution to the knowledge of Psylloidea (Hemiptera) of Southeastern Anatolia, Turkey". In: DOI: 10.13140/RG.2.1.4887.6400.

Perilla-Henao, Laura M and Clare L Casteel (2016). "Vector-borne bacterial plant pathogens: interactions with hemipteran insects and plants". In: *Frontiers in Plant Science* 7, p. 1163. DOI: 10.3389/fpls.2016.01163.

Plant Health Threats Ontology (2016). URL: http://akswnc7.informatik.uni-leipzig.de/dstreitmatter/archivo/rhizomik.net/ontologies--PlantHealthThreats--owl--ttl/2020.06.10-211502/ontologies--PlantHealthThreats--owl--ttl_type=generatedDocu.html.

Ristaino, Jean B, Pamela K Anderson, Daniel P Bebber, Kate A Brauman, Nik J Cunniffe, Nina V Fedoroff, Cambria Finegold, Karen A Garrett, Christopher A Gilligan, Christopher M Jones, et al. (2021). "The persistent threat of emerging plant disease pandemics to global food security". In: *Proceedings of the National Academy of Sciences* 118.23, e2022239118. DOI: 10.1073/pnas.2022239118.

Rivera, Diego, Robert Allkin, Concepción Obón, Francisco Alcaraz, Rob Verpoorte, and Michael Heinrich (2014). "What is in a name? The need for accurate scientific nomenclature for plants". In: *Journal of Ethnopharmacology* 152.3, pp. 393–402. DOI: 10.1016/j.jep.2013.12.022.

Roussey, Catherine, Vincent Soulignac, Jean-Claude Champomier, Vincent Abt, and Jean-Pierre Chanet (2010). "Ontologies in agriculture". In: *AgEng 2010, International Conference on Agricultural Engineering*. Cemagref, p–p.

Sauvion, Nicolas, Jean Peccoud, Christine N Meynard, and David Ouvrard (2021). "Occurrence data for the two cryptic species of Cacopsylla pruni (Hemiptera: Psylloidea)". In: *Biodiversity Data Journal* 9. DOI: 10.3897/BDJ.9.e68860.

Savary, S, L Willocquet, and SJ Pethybridge (2019). "The global burden of pathogens and pests on major food crops". In: *The global burden of pathogens and pests on major food crops. Nat. Ecol. Evol* 3, pp. 430–439. DOI: 10.1038/s41559-018-0793-y.

Schulz, Stefan, Holger Stenzhorn, and Martin Boeker (2008). "The ontology of biological taxa". In: *Bioinformatics* 24.13, pp. i313–i321. DOI: 10.1093/bioinformatics/btn158.

Shahzadi, Sidra and Sidra Tahir (2016). "Ontological Framework for Alignment of Web Services with Requirements in Service Oriented Requirement Engineering (SORE)". In: *International Journal of Software Engineering and Its Applications* 10.11, pp. 255–270. DOI: 10.14257/ijseia.2016.10.11.21.

Spinakis, Antonis and Paraskevi Peristera (2004). "Text mining tools: Evaluation methods and criteria". In: *Text Mining and its Applications*. Springer, pp. 131–149. DOI: 10.1007/978-3-540-45219-5_10.

Spyns, Peter (1996). "Natural language processing in medicine: an overview". In: *Methods of information in medicine* 35.04/05, pp. 285–301.

Steffek, R, S Follak, N Sauvion, G Labonne, and A MacLeod (2012). "Distribution of 'Candidatus Phytoplasma prunorum'and its vector Cacopsylla pruni in European fruit-growing areas: a review". In: *EPPO bulletin* 42.2, pp. 191–202. DOI: 10.2903/sp.efsa.2012.EN-319.

Stephan, Grimm, Hitzler Pascal, Abecker Andreas, et al. (2007). "Knowledge representation and ontologies". In: *Semantic web services*. Springer, pp. 51–105. DOI: 10.1007/978-1-84882-448-5_14.

Suarez, Andrew V and Neil D Tsutsui (2004). "The value of museum collections for research and society". In: *BioScience* 54.1, pp. 66–74.

Tamanini, Livio (1955). *Alcuni nuovi reperti di psillidi italiani e francesi*. Bollettino della Societa Entomologica Italiana 85: 10-11. ISBN: 0373-3491.

Tang, Anfu, Louise Delèger, Robert Bossy, Pierre Zweigenbaum, and Claire Nédellec (to appear, 2022). "Do syntactic trees enhance domain-specific BERT models for relation extraction?" In: *Database journal*.

Team, Educative Answers (2022). *What is the F1-score?* URL: https://www.educative.io/answers/what-is-the-f1-score.

Thébaud, Gaël, Michel Yvon, Rémi Alary, Nicolas Sauvion, and Gérard Labonne (2009). "Efficient transmission of 'Candidatus Phytoplasma prunorum'is delayed by eight months due to a long latency in its host-alternating vector". In: *Phytopathology* 99.3, pp. 265–273. DOI: 10.1094/PHYTO-99-3-0265.

Trivellone, Valeria, Eric P Hoberg, Walter A Boeger, and Daniel R Brooks (2022). "Food security and emerging infectious disease: risk assessment and risk management". In: *Royal Society Open Science* 9.2, p. 211687. DOI: 10.1098/rsos.211687.

Underwood, A.J., M.G. Chapman, and T.P. Crowe (2004). "Identifying and understanding ecological preferences for habitat or prey". In: *Journal of Experimental Marine Biology and Ecology* 300.1. VOLUME 300 Special Issue, pp. 161–187. ISSN: 0022-0981. DOI: https://doi.org/10.1016/j.jembe.2003.12.006. URL: https://www.sciencedirect.com/science/article/pii/S002209810400005X.

Wang, Yucheng, Bowen Yu, Hongsong Zhu, Tingwen Liu, Nan Yu, and Limin Sun (2021). "Discontinuous Named Entity Recognition as Maximal Clique Discovery". In: *ACL*.

Webster, Jonathan J and Chunyu Kit (1992). "Tokenization as the initial phase in NLP". In: *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*. DOI: 10.3115/992424.992434.

Wilson, Anthony James, Eric René Morgan, Mark Booth, Rachel Norman, Sarah Elizabeth Perkins, Heidi Christine Hauffe, Nicole Mideo, Janis Antonovics, Hamish McCallum, and Andy Fenton (2017). "What is a vector?" In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1719, p. 20160085. DOI: 10.1098/rstb.2016.0085.

Wilson, Edward O et al. (1988). "Biodiversity". In: *The National Academies Press*. URL: https://nap.nationalacademies.org/catalog/989/biodiversity.

Zhou, GuoDong, Jian Su, Jie Zhang, and Min Zhang (2005). "Exploring various knowledge in relation extraction". In: *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)*, pp. 427–434. DOI: 10.3115/1219840.1219893.