Un pipeline de traitement d'informations issues du web pour anticiper les dangers en santé des plantes. Méthodologie développée par la Plateforme ESV dans le cadre de la " Veille Sanitaire Internationale".

Sandy Duperier, Marie Grosdidier, Jean-Baptiste Louvet, Isabelle Pieretti, Anne Quillévéré-Hamard

Un pipeline de traitement d'informations issues du web pour anticiper les dangers en santé des plantes

Méthodologie développée par la Plateforme ESV dans le cadre de la « Veille Sanitaire Internationale »

Sandy DUPERIER[1,4]
Marie GROSDIDIER[1,4]
Jean-Baptiste LOUVET[1,4]
Isabelle PIERETTI[2,4]
Anne QUILLÉVÉRÉ-HAMARD[3,4]

Correspondance
sandy.duperier@inrae.fr

Résumé.

Dans un contexte de changement climatique et d'intensification des échanges internationaux, les situations sanitaires des végétaux peuvent évoluer rapidement, avec un risque sanitaire accru. Afin d'informer au mieux les gestionnaires des risques sanitaires, la Plateforme d'Épidémiosurveillance en Santé Végétale (ESV) a mis en place une Veille Sanitaire Internationale (VSI) qui condense les éléments d'intérêt dans des bulletins publics. Cette veille relaie des informations médiatiques et scientifiques permettant d'identifier et/ou de suivre les évolutions de foyers épidémiques dans une zone géographique donnée, ainsi que des avancées scientifiques qui présentent un intérêt opérationnel. Pour répondre à ces objectifs, la Plateforme ESV a développé sa propre méthode de veille, pour laquelle la collecte d'informations est basée notamment sur le *Web scraping*. La chaîne de traitement des articles est semi-automatique, hybridant codes informatiques, système d'information et apports humains, permettant la diffusion de ces derniers à travers des bulletins. À ce jour, les articles diffusés par la VSI de la Plateforme ESV représentent environ 1 % de la totalité des articles collectés. Ce taux de pertinence est en lien avec les méthodes utilisées. Dans le futur, un des enjeux de l'évolution de la chaîne de traitement sera d'améliorer les processus de collecte et de tri des articles sans dénaturer la qualité des informations récupérées puis diffusées. Pour ce faire, la VSI souhaite s'appuyer sur le développement d'outils informatiques opérationnels, notamment issus de travaux de recherche.

Mots-clés

Veille sanitaire, *web scraping*, média-scanning, santé végétale, organismes nuisibles, épidémiosurveillance.

1. INRAE, Unité BioSP, Domaine Saint-Paul, Site Agroparc, 84914 Avignon Cedex 9, France.
2. CIRAD, PHIM Plant Health Institute, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France.
3. Université de Lyon - ANSES, unité d'Épidémiologie et Appui à la Surveillance (EAS), 69364 Lyon Cedex 7, France.
4. Plateforme d'Épidémiosurveillance en Santé Végétale (Plateforme ESV).

A pipeline for processing web data for anticipating plant health risks

Methodology developed by the ESV Platform in the framework of "Global Health Monitoring"

Sandy DUPERIER[1,4]
Marie GROSDIDIER[1,4]
Jean-Baptiste LOUVET[1,4]
Isabelle PIERETTI[2,4]
Anne QUILLÉVÉRÉ-HAMARD[3,4]

Correspondence
sandy.duperier@inrae.fr

Abstract.

The context of climate change and the intensification of international trade leads to increasing risks for plant health as well as rapid changes in the latter. To improve the information available to those responsible for managing health risks, the French Epidemiological Plant Health Surveillance Plateform (named Plateforme ESV in french) has set up a global health monitoring system (named VSI in french) that synthetises items of interest in the public press. This monitoring relays links media and scientific information allowing to identify and/or monitor outbreaks progress in a given geographical area, as well as scientific progress of operational interest. To obtain these goals, the ESV Platform has developed its own monitoring method in which the collecting part is mainly based on web scraping. To process articles, the pipeline is semi-automatic, combining computer codes, information systems and human contributions, leading to their dissemination in newsletters. To date, the articles disseminated by the VSI of the ESV Platform make up about 1% of all the articles collected. This rate of pertinence is linked to the methods used.
In the future, one of the challenges of the pipeline will be to improve article collection and sorting process without degrading information quality collected and then disseminated. To do this, the VSI wants to focus on developing operational computing tools, in particular resulting from research activities.

Keywords

Health monitoring, web scraping, media-scanning, plant health, pests, épidemiological monitoring.

1. INRAE, Unité BioSP, Domaine Saint-Paul, Site Agroparc, 84914 Avignon Cedex 9, France.
2. CIRAD, PHIM Plant Health Institute, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France.
3. Université de Lyon - ANSES, unité d'Épidémiologie et Appui à la Surveillance (EAS), 69364 Lyon Cedex 7, France.
4. Plateforme d'Épidémiosurveillance en Santé Végétale (Plateforme ESV).

Introduction

La maîtrise des dangers sanitaires constitue aujourd'hui un enjeu majeur dans un contexte de changement climatique, d'augmentation des flux internationaux de voyageurs et de marchandises, de modification des habitudes de consommation, de l'usage des sols et des pratiques agricoles (Lannou *et al.*, 2023). Afin d'accroître la performance de sa politique de sécurité sanitaire, la France s'est dotée de trois plateformes d'épidémiosurveillance dans les domaines de la santé animale (2011), de la sécurité de la chaîne alimentaire (2018) et de la santé végétale (2018) (Amar et Dupuy, 2020). Ces plateformes ont un rôle d'appui et de conseil auprès des services compétents de l'État et, à leur demande, aux autres gestionnaires de dispositifs de surveillance, qu'ils soient publics ou privés. Chacune des trois plateformes d'épidémiosurveillance a développé son propre dispositif de veille sanitaire, basé sur des sources différences et des outils spécifiques. Ainsi, la Plateforme d'Épidémiosurveillance en Santé Animale (ESA) utilise le système d'information sanitaire de la Commission européenne (*Animal Disease Information System* – ADIS) et de l'Organisation Mondiale de la Santé Animale (*World Animal Health Information System* – WAHIS) pour les informations officielles. Les informations non officielles sont issues des médias *via* l'outil PADI-web (*Platform for Automated extraction of Disease Information from the web*), basé sur la collecte d'articles de *Google News* grâce à des flux RSS (Valentin *et al.*, 2020) et des réseaux de professionnels (ex : ProMed, réseaux d'experts thématiques). La Plateforme d'Épidémiosurveillance de la Chaîne Alimentaire (SCA) réalise sa veille avec l'outil MedISys[1] (Alomar *et al.*, 2016) qui est également utilisé par l'EFSA (*European Food Safety Authority*) et qui est basé sur le système EMM (*Europe Media Monitor*). Ce système est fondé sur la collecte de flux RSS (*Really Simple Syndication*) et des méthodes de *web scraping* (de l'anglais « *scraping* » « gratter/racler ») qui consiste à récupérer de façon automatique des informations textuelles sur le web. La suite de l'article précise des éléments de contexte puis décrit la méthodologie développée par la Plateforme ESV pour réaliser sa veille sanitaire.

La veille réalisée par la Plateforme d'Épidémiosurveillance en Santé Végétale (ESV) a pour objectifs la mise à jour des connaissances, pour les gestionnaires du risque et les acteurs de la surveillance, mais aussi l'anticipation de problématiques sanitaires qui pourraient avoir un impact en France (dans l'hexagone et les DROM). Ainsi, une équipe est chargée de réaliser une Veille Sanitaire Internationale (VSI) puis diffuse les informations publiquement. La Plateforme ESV a développé sa propre méthode de veille basée notamment sur le *web scraping*, en utilisant des outils génériques ciblant des canaux médiatiques généralistes et des canaux spécifiques (journaux agronomiques, journaux scientifiques et sites institutionnels). La veille est réalisée

La veille sanitaire

La veille permet de collecter, d'analyser et d'interpréter des informations dans un champ d'expertise donné, mais aussi de détecter des signaux dits faibles, pour anticiper les évolutions. Le concept de « signaux faibles », introduit par Igor H. Ansoff (Ansoff, 1975; Antoniou, 2006), fait référence à une information d'alerte précoce, de faible intensité, pouvant être annonciatrice d'une tendance ou d'un événement important. Pour être efficace, y compris pour traiter des signaux faibles, la veille doit être continue et itérative. La veille sanitaire cible les signaux pouvant représenter un risque pour la santé publique dans une perspective d'anticipation, d'alerte et d'action précoce. D'après Eilstein *et al.* (2012), la veille sanitaire inclue notamment des objets sanitaires « indirects » comme des signaux bibliographiques ou médiatiques. En général, la collecte de ces informations repose sur des sources en accès libre ou restreint telles que des notifications de réseaux sociaux, des listes de diffusion, des alertes utilisant des mots-clés ou encore des flux RSS (*Really Simple Syndication*) pour récupérer des contenus officiels et non officiels mis à jour sur le web. Dans le domaine de la santé des plantes, l'association de diverses sources d'informations a fait l'objet d'une étude rétrospective récente dans laquelle trois agrégateurs de données, y compris des données issues des réseaux sociaux, ont été utilisés pour identifier des informations clés pour deux espèces de ravageurs (Tateosian et al., 2023). Les résultats de ces travaux suggèrent que les informations en ligne et les réseaux sociaux sont des mines d'informations précieuses et complémentaires aux informations officielles pour la détection de signaux faibles.

1 MedISys = Système de surveillance des médias

de deux manières : (i) ciblée, sur des organismes de quarantaine prioritaires (OQP) définis par l'Union européenne (règlement UE 2016/2031), et (ii) généraliste, notamment pour détecter des signaux faibles pouvant amener à signaler des émergences d'agents pathogènes de ravageurs ou de nouvelles maladies. En termes réglementaires, les espèces jugées nocives, en raison de leurs effets néfastes pour les végétaux et les filières économiques qui en dépendent, sont nommées « organismes nuisibles » (ON) (NIMP 11, 2001). À ce jour, douze ON font l'objet d'une veille ciblée régulière par la VSI (liste complète : Annexe 1). Pour mener la veille de manière exhaustive et fiable, les veilleuses s'appuient sur un groupe d'une cinquantaine d'experts internationaux en santé des plantes, constitué de scientifiques, d'experts réglementaires et de gestionnaires du risque, pouvant être sollicités de façon ponctuelle sur des sujets en lien avec leur domaine d'expertise. Ce groupe peut transmettre à l'équipe de la VSI des informations issues de leurs réseaux régionaux et/ou internationaux. Ce travail collaboratif permet de diffuser des informations pertinentes au lectorat, principalement sous forme de bulletins hebdomadaires et mensuels. Ces bulletins peuvent être consultés sur le site de la Plateforme ESV et sont diffusés *via* une liste de diffusion[2].

Cet article présente les trois grandes étapes de la chaîne de traitement (ou *pipeline/workflow*) : la collecte des données, la classification et la sélection des articles pertinents par un comité éditorial (CE), puis la valorisation et la diffusion des informations (Figures 1 et 2), ainsi qu'un bilan associé.

Collecte des données

La Plateforme ESV a développé des scripts informatiques qui permettent de collecter les informations de manière automatique en utilisant du *web scraping*[3] et l'appel à une API[4] d'interrogation de moteurs de recherche. Les informations cibles sont requêtées à l'aide de mots-clés relatifs aux différents ON ciblés, contenant le nom de l'organisme ciblé (nom latin et/ou nom vernaculaire). Les mots-clés sont également choisis selon la source web ciblée.

Une partie de ces sources sont spécifiques à la santé végétale, typiquement des pages dédiées sur les sites officiels de gouvernements tels que ceux de la France, des États-Unis et du Royaume-Uni. D'autres sont des agrégateurs de sources tels que MedISys, PubMed ou Google Search (voir la liste complète : Annexe 2). Un certain nombre de mots-clés génériques (ex. : « santé des plantes ») permettent de réaliser une veille généraliste pour détecter des signaux faibles ou des informations sur des sujets transversaux en santé végétale. La période d'interrogation des pages web est définie dans la requête afin de cibler uniquement les informations parues la semaine précédant le jour du *web scraping* qui s'exécute tous les lundis.

Pour la bonne compréhension des explications, le mot « article » utilisé dans les parties suivantes fait référence au contenu d'une page web associée à une URL[5] correspondant à une source médiatique ou scientifique.

Récupération des URL grâce au *web scraping*

Comme énoncé précédemment, le *web scraping* consiste à extraire de façon automatique des informations textuelles sur des pages web en ciblant certaines parties du code de la page (via des requêtes HTTP). L'intérêt de cette méthode est d'automatiser la collecte et le stockage d'informations sans les dénaturer. Le *web scraping* est réalisé en deux étapes : le téléchargement et/ou la lecture du contenu de la page web puis la recherche d'informations en ciblant la ou les parties d'intérêt dans la page grâce aux éléments qui permettent de structurer celle-ci (ex : balises HTML - *Hypertext Markup Language*).

La Plateforme ESV réalise ce *web scraping* avec le langage R en utilisant les *packages* rvest (qui permet d'extraire du contenu à partir des pages web) et xml2 (qui permet d'analyser des documents XML [*Extensible Markup Language*]). Néanmoins, d'autres langages informatiques peuvent être utilisés tels que Python. Avant de s'intéresser au *web scraping*, il est important de comprendre comment le code source ou code HTML d'une page web est structuré. La consultation du code source des pages est assez simple (voir-le-code-source). Pour obtenir plus d'informations sur le *web scraping* avec R, plusieurs documents sources sont disponibles en ligne (ex : web-scraping-with-r). Un exemple concret du processus de *web scraping* pour une page web spécifique est détaillé dans la partie suivante.

2 Lien d'inscription à la liste de diffusion : https://groupes.renater.fr/sympa/subscribe/esv_veille_newsletter?previous_action=info

3 *Web scraping* (de l'anglais « scraping » = gratter/racler) qui consiste à récupérer de façon automatique des informations textuelles présentes sur les pages web.

4 Une API (*Application Programming Interface*) est une interface logicielle qui permet de « connecter » un logiciel ou un service à un autre logiciel ou service afin d'échanger des données et des fonctionnalités (CNIL).

5 URL signifie *Uniform Resource Locator* (ou, en français, « localisateur uniforme de ressource »). Il s'agit de l'adresse d'une ressource donnée, unique sur le web (What_is_a_URL).

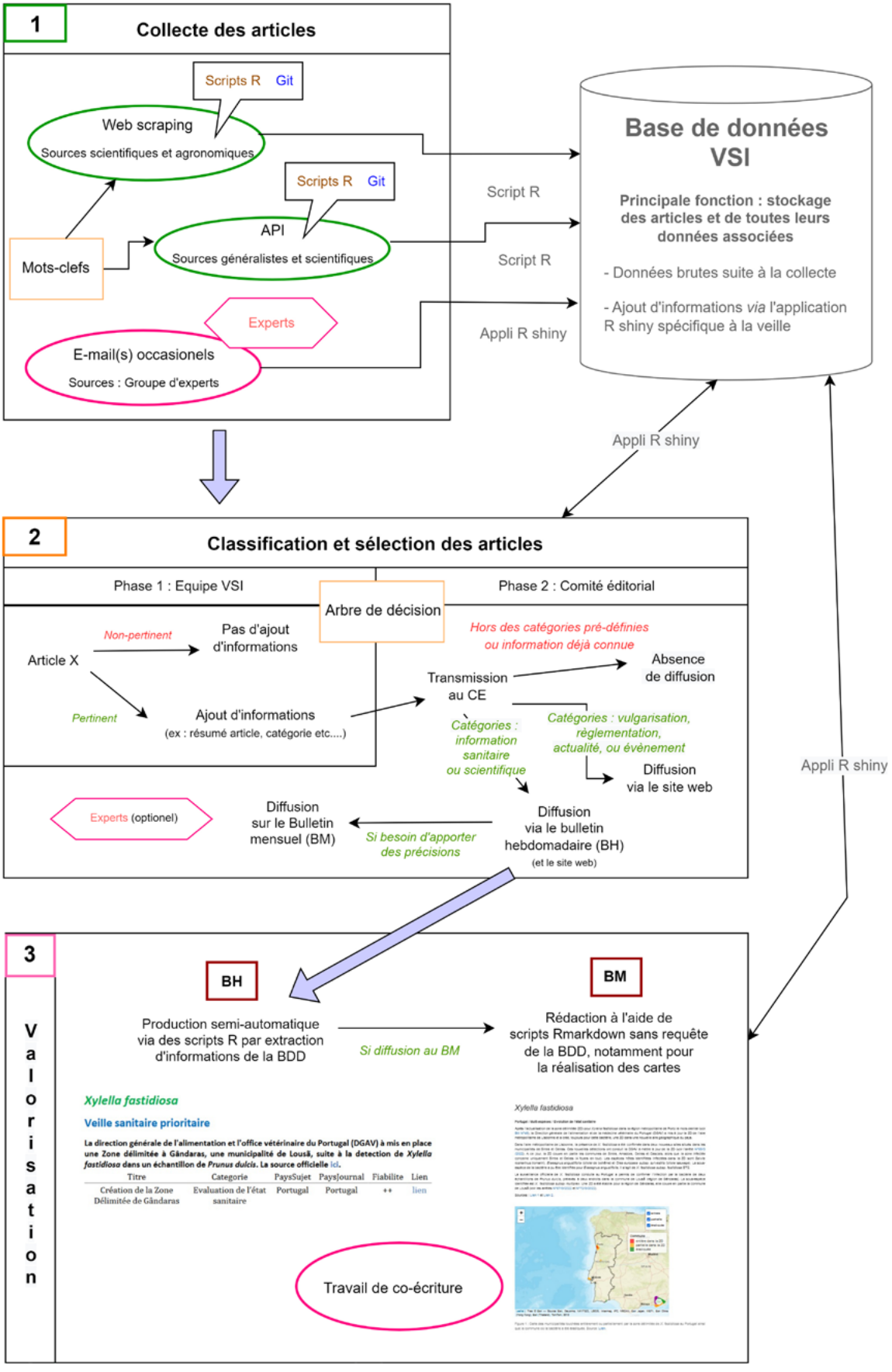

Figure 1. Schéma descriptif du processus, de la récupération des articles jusqu'à la production des bulletins. 1 : étape de collecte

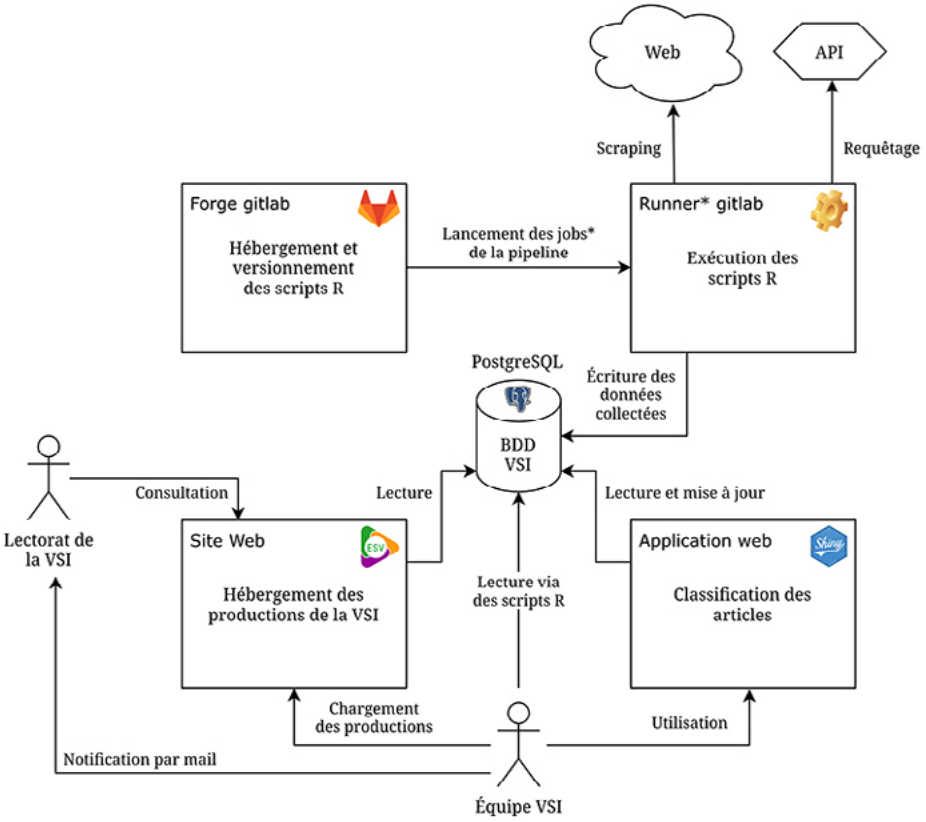

Figure 2. Schéma descriptif du processus de la récupération des articles jusqu'à la lecture des bulletins, du point de vue du système d'information.

Cas concret

Exemple d'un article scientifique sur le coléoptère ravageur *Popillia japonica*.

URL ciblée : https://www.frontiersin.org/articles/10.3389/finsc.2022.887659/full

Besoin : Collecter le titre, les auteurs, le résumé, le texte, la date de publication, le journal et le lien URL de l'article scientifique.

Étape 1 : Consultation du code source de la page web

Le scraping ou la récupération des informations des articles se base sur des balises présentes dans le code source de la page web. On y accède grâce à un clic droit sur la page web ouverte dans un navigateur et en choisissant « Inspecter » (variable selon les navigateurs, voir liens dans la partie « Récupération des URL grâce au *web scraping* »). Cette fonctionnalité permet notamment de visualiser le lien entre une zone de la page et le code source. De plus, avec un clic droit sur la partie d'intérêt du code source, on peut copier directement le « chemin CSS » ou « chemin Xpath ». Par exemple, pour récupérer le titre, le chemin CSS est : « body.journal-insect-science.section-invasive-insect-species div#main-content.page-container.sidemenu-collapsed div#article.boxed.white.no-padding div.container-fluid.main-container-xxl div#similar-articles.row div.new-wrapper main div.article-section div.article-container div.abstract-container div.JournalAbstract h1 ».

Étape 2 : Réalisation du script R

Avant de pouvoir chercher les éléments qui nous intéressent, il faut que la/les pages soient chargée(s) dans l'environnement R (utilisation du package «xml2»). Puis on récupère les éléments d'intérêt avec la syntaxe utilisée dans la page web en question (utilisation du package « rvest »). Les scripts permettant de répondre aux objectifs énoncés et les résultats obtenus sont détaillés ci-dessous (Figure 3, Tableau 1). Les scripts de l'exemple ci-dessous et d'un exemple supplémentaire sont disponibles dans la section Matériel supplémentaire (au format .R).

```
# Chargement des librairies
library(xml2)  # Lecture de page web
library(rvest) # Webscraping
library(dplyr) # R "moderne"
library(stringr) # Manipulation de chaines de caractères
library(lubridate) # Modification du format de dates

# URL ciblée
 searchURL <- "https://www.frontiersin.org/articles/10.3389/finsc.2022.887659/full"

# Lecture de la page web
 page <- xml2::read_html(searchURL)

# Sélection des informations d'intérêt grâce aux balises de la page web
 titre <- page %>%
      rvest::html_nodes(«body.journal-insect-science.section-invasive-insect-species  div#main-content.page-container.sidemenu-collapsed div#article.boxed.white.no-padding div.container-fluid.main-container-xxl div#similar-articles.row div.new-wrapper main div.article-section div.article-container div.abstract-container div.JournalAbstract h1») %>%
   rvest::html_text()

 auteurs <- page %>%
      rvest::html_nodes(«body.journal-insect-science.section-invasive-insect-species  div#main-content.page-container.sidemenu-collapsed div#article.boxed.white.no-padding div.container-fluid.main-container-xxl div#similar-articles.row div.new-wrapper main div.article-section div.article-container div.abstract-container div.JournalAbstract div.authors») %>%
   rvest::html_text()

 abstract <- page %>%
      rvest::html_nodes(«body.journal-insect-science.section-invasive-insect-species  div#main-content.page-container.sidemenu-collapsed div#article.boxed.white.no-padding div.container-fluid.main-container-xxl div#similar-articles.row div.new-wrapper main div.article-section div.article-container div.abstract-container div.JournalAbstract p») %>%
   rvest::html_text()

 text <- page %>%
      rvest::html_nodes(«body.journal-insect-science.section-invasive-insect-species  div#main-content.page-container.sidemenu-collapsed div#article.boxed.white.no-padding div.container-fluid.main-container-xxl div#similar-articles.row div.new-wrapper main div.article-section div.article-container div.abstract-container div.JournalFullText») %>%
   rvest::html_text()

 # (la date est traitée dans la partie suivante)
 journaldatelien <- page %>%
      rvest::html_nodes(«body.journal-insect-science.section-invasive-insect-species  div#main-content.page-container.sidemenu-collapsed div#article.boxed.white.no-padding  div.container-fluid.main-container-xxl div#similar-articles.row div.new-wrapper main div.article-section  div.article-container  div.abstract-container  div.article-header-container  div.header-bar-three-container  div.header-bar-three») %>%
   rvest::html_text()

 lien <- page %>%
      rvest::html_nodes(«body.journal-insect-science.section-invasive-insect-species  div#main-content.page-container.sidemenu-collapsed div#article.boxed.white.no-padding  div.container-fluid.main-container-xxl div#similar-articles.row div.new-wrapper main div.article-section  div.article-container  div.abstract-container  div.article-header-container  div.header-bar-three-container  div.header-bar-three a») %>%
   rvest::html_attr(«href»)

# Retrait des espaces des éléments extraits
 auteurs <- stringr::str_trim(auteurs)
 journaldatelien <- stringr::str_trim(journaldatelien)
 text <- stringr::str_trim(text)

# Remplacement des retours à la ligne par un espace
 text <- gsub('\n' ,' ', text)

# Séparation du journal et de la date (ne garde pas le lien « https »)
 journal <- unlist(stringr::str_split(journaldatelien,"([0-9][2} ..... [0-9][4])|(\r\n\r\n)")) #sépare sur une expression régulière ou \r\n\r\n
 journal <- paste0(journal[1:2], collapse = ' ')
 date <- stringr::str_extract(journaldatelien,»[0-9][2} ..... [0-9][4]»)
 date <- lubridate::dmy(date)

# Synthétise les données dans un tableau
 TabRecap <- data.frame(
   Lien = lien,
   Titre = titre,
   Auteurs = auteurs,
   DatePublication = date,
   Journal = journal,
   Abstract = abstract,
   Text = text)

# Export du tableau en format xlsx
# Hors «projet R», ajouter le chemin d'accès vers le fichier
 write.xlsx (TabRecap,  «CAS2_tabrecap.xlsx»,  row.names = FALSE)
```

Figure 3. Exemple de script R permettant de répondre aux objectifs décrits dans le « Cas concret ». Ce script permet de récupérer des informations sur une page web (*web scraping*)

Tableau 1 : Exemple de visualisation du résultat (tableau au format .xlsx) obtenu suite à l'application du script R décrit en figure 3 pour répondre aux objectifs décrits dans le « Cas concret »

Lien	Titre	Auteurs	Date Publication	Journal	Abstract	Text
https://doi.org/10.3389/finsc.2022.887659	Impact of Adult *Popillia japonica (Coleoptera: Scarabaeidae)* Foliar Feeding Injury on Fruit Yield and Quality of a Temperate, Cold-Hardy Wine Grape, 'Frontenac'	Dominique N. Ebbenga1*, Eric C. Burkness1, Matthew D. Clark2 and William D. Hutchison1	4/26/2022	Front. Insect Sci., Sec. Invasive Insect Species	*Popillia japonica* (Newman), is a highly polyphagous, invasive species, first recorded in the U.S. in 1916, and detected in Minnesota in the late 1960s. Historically, research on this pest in the Midwest U.S. has focused primarily on ornamental and turf crops, with little attention placed on adult feeding damage to fruit crops. [...]	Introduction Japanese beetle, *Popillia japonica* Newman (*Coleoptera: Scarabaeidae*), is an invasive insect, native to Japan (1). The species was first detected in the United States in New Jersey in 1916, and eventually found in Minnesota in 1968 (2). Since the arrival in the U.S., *P. japonica* have become a major pest in turfgrass, ornamental and horticultural settings (3, 4). [...]

Récupération des URL grâce à une interface de programmation d'application (API)

Dans le cadre du pipeline de la VSI, les recherches d'articles sur les moteurs de recherche Google Search et Google Scholar sont réalisées à l'aide de l'API Scale SERP. Cette API permet de récupérer dans un format structuré les résultats de recherches Google réalisées avec des mots-clés prédéfinis par l'équipe de la VSI.

Automatisation du processus de collecte et d'enregistrement dans la base de données VSI

L'automatisation de l'exécution des scripts tous les lundis à heure fixe permet un gain de temps et une plus grande régularité dans la récupération des informations. L'exécution automatique des scripts R de *web scraping* et de l'API est réalisée par l'instance GitLab du département MathNum de INRAE[6]. Afin de vérifier l'exécution des tâches, deux fichiers sont générés : le premier contient toutes les informations collectées (notamment les URL des articles), et le second recense les potentielles erreurs qui auraient pu survenir lors de l'exécution des scripts. À l'issue de cette étape, les informations obtenues sont enregistrées directement dans la base de données (BDD) VSI de la Plateforme ESV (BDD relationnelle PostgreSQL) grâce à l'exécution d'un script R utilisant des fonctions SQL[7]. Les informations sanitaires transmises par le réseau d'experts sont quant à elles intégrées manuellement dans la BDD VSI grâce à l'application web décrite ci-dessous.

Classification et sélection des articles

Phase de classification

Tous les lundis, deux à trois agents chargés de veille de la Plateforme ESV réalisent le jour même la classification des articles enregistrés dans la BDD VSI. Au cours de cette étape, les articles sont classés manuellement selon leur pertinence : (i) non pertinent, (ii) pertinent non remonté au comité éditorial (CE) (ex. : articles en doublon ou déjà vus les semaines précédentes), (iii) pertinent remonté au CE. Le choix de la pertinence de l'article est guidé par l'utilisation d'un arbre de décision basé sur différentes catégories auxquelles sont associés plusieurs critères (Tableau 2). Cet arbre de décision a été défini et validé par le CE, puis édité sous forme d'une carte mentale. Par exemple, un article portant sur le coût de la mise en place de mesures de gestion d'un ON est associé à la catégorie « Économie » et sera classé pertinent si l'ON en question est à la fois ciblé par la veille de la Plateforme ESV, qu'il est associé à

6 GitLab Community Edition (CE) est une plateforme de développement logiciel open source de gestion de version et de maintenance collaborative de code (https://about.gitlab.com/). L'instance GitLab du département MathNum est accessible à l'adresse : https://forgemia.inra.fr/

7 Le SQL (*Structured Query Language*) est un langage permettant de communiquer avec une base de données relationnelles (https://sql.sh/)

la « Surveillance » et/ou à la « Lutte » et que l'information est intéressante pour la France. Cette phase de tri manuel nécessite l'ouverture de l'article, sa traduction éventuelle, et une lecture permettant l'extraction des informations pertinentes (titre traduit, pays concerné par l'information, catégorie(s), fiabilité), et l'attribution d'un « sujet » (description très succincte des éléments d'intérêt traités par l'article). Toutes ces modifications sont réalisées grâce à une application web R shiny[8] directement connectée avec la BDD VSI. Les nouveaux éléments sont ainsi enregistrés dans la BDD VSI en temps réel.

Processus de sélection des articles à diffuser

Le CE est composé de sept personnes de quatre instituts partenaires de la Plateforme ESV (INRAE[9], Anses[10], Cirad[11] et DGAl[12]), avec des compétences scientifiques (biologie, phytopathologie, épidémiologie, entomologie) mais aussi des compétences réglementaires en santé des plantes. Ce dernier a pour mission d'analyser les articles préalablement classés afin de choisir de manière collégiale ceux qui seront résumés dans les bulletins produits par la VSI et/ou relayés sur le site internet de la Plateforme ESV. Pour réaliser ce travail, le CE peut s'appuyer sur la carte mentale (également utilisée lors de la phase précédente de tri) et sur le groupe d'experts de la VSI (experts spécialisés au niveau de thématiques et/ou de zones géographiques) pour obtenir des compléments d'informations. Le comité se réunit chaque semaine et prend part à la révision des résumés qui accompagnent les articles sélectionnés. Suite à ce travail, ces derniers sont modifiés dans la BDD VSI *via* l'application web (R-shiny). Un certain nombre d'informations accompagnant les articles sélectionnés sont *in fine* diffusées.

Tableau 2 : Précisions sur l'étape de classification des articles pertinents en fonction de catégories qui leur sont attribuées

Catégories de classification thématique	
Actualités	Réglementation
	Communication/Vulgarisation
Évènements	Communication/Vulgarisation
Sanitaire prioritaire	Notification de nouveaux cas
	Évaluation de l'état sanitaire
Sanitaire secondaire ou scientifique	Évaluation de l'état sanitaire
	Économie
	Prophylaxie
	Échelle de la population
	Échelle génétique et moléculaire
	Évaluation de l'état sanitaire
	Mesures de lutte
	Mesures de surveillance
	Méthodes d'analyse et de détection
	Méthodes pour améliorer la surveillance

Valorisation

Plusieurs types de productions sont réalisées dans le cadre des activités de veille de la Plateforme ESV. On y retrouve le bulletin hebdomadaire (BH) dans lequel sont relayés les articles (sanitaires et scientifiques) de la semaine précédente, validés par le CE, sous une forme condensée et structurée. Les liens vers un ou plusieurs articles pouvant faire référence au même sujet sont accompagnés d'un résumé en français (Figure 4A). La production des BH se fait de manière semi-automatisée incluant une extraction d'informations à partir de la BDD VSI à l'aide de scripts R. Les articles les plus pertinents du BH et dont le sujet est jugé intéressant à approfondir font l'objet de brèves plus conséquentes et vulgarisées pour être diffusées dans le bulletin mensuel (BM). Un des objectifs des BM est de synthétiser les informations tout en permettant de développer les aspects d'intérêt sanitaire et/ou opérationnels. Dans ce format mensuel, les informations sanitaires de différentes sources peuvent être illustrées avec des cartes interactives (Figure 4B). Ces dernières permettent de visualiser les évolutions de répartition géographique des détections, des éradications d'ON, ou encore des zones délimitées réglementées.

Les informations servant à réaliser les cartes sont extraites des textes collectés par les activités de veille décrites précédemment auxquelles s'ajoutent parfois des données issues d'investigations. La production des BM se fait à l'aide de scripts R-Markdown[13]. À une autre échelle de temps, l'ensemble des

8 Shiny est un package R permettant de développer des applications web de visualisation de données (https://shiny.rstudio.com/)

9 Institut national de recherche pour l'agriculture, l'alimentation et l'environnement

10 Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail

11 Organisme français de recherche agronomique et de coopération internationale pour le développement durable des régions tropicales et méditerranéennes

12 Direction générale de l'alimentation

13 Rmarkdown est un package R permettant d'exécuter du code et de générer des rapports prenant en charge des formats de sortie statiques et dynamiques (ex : graphs, cartes ; https://rmarkdown.rstudio.com/index.html)

Fusarium oxysporum f. sp. *cubense* Tropical race 4

A

Veille sanitaire prioritaire

L'état d'urgence a été déclaré le 19 janvier 2023 par le gouvernement Vénézuélien suite à l'émergence de Foc TR4 dans le pays.

Titre	Categorie	PaysSujet	PaysJournal	Fiabilite	Lien
Phytosanitary Emergency for presence of Foc R4T	Notifications de nouveaux cas	Venezuela	/	+++	lien
Le Venezuela confirme la présence de Fusarium Race 4 dans les plantations de trois États (foyers cités)	Notifications de nouveaux cas, Communication / vulgarisation	Venezuela	/	++	lien

Xylella fastidiosa

B

Portugal / Multi-espèces / Notification de nouveaux cas, évolution de la situation sanitaire

Depuis le début de l'épidémie au Portugal, 15 foyers ont été recensés entre janvier 2019 et décembre 2022 et à ce jour 77 espèces végétales ont été infectées, dont la vigne. Un des éléments les plus marquants a été la communication en décembre 2022 de la première détection dans l'UE de *Xylella fastidiosa* dans plusieurs espèces d'agrumes : le citronnier (*Citrus limon*), le pamplemousse (*C. paradisi*), la mandarine (*C.reticulata*) et l'orange douce (*C. sinensis*). Les analyses moléculaires ont permis d'identifier la sous-espèce *fastidiosa* de *X. fastidiosa*, rarement présente chez les agrumes. La détection de *Xylella* au Portugal avait entraîné la création de zones délimitées à Porto, à Lisbonne et dans quatre autres zones du pays (deux au nord et deux au centre du pays). Depuis notre dernier bilan sur la situation sanitaire de *Xylella* dans le pays (BM n°47), de nouvelles zones ont été délimitées à la fin du mois de décembre et en janvier, en particulier au nord et au centre du pays. En région nord, dans la municipalité d'Alijó, *Xylella* a été détectée sur pêcher (*Prunus persica*) et dans la commune de Mirandela sur un hibiscus de Syrie (*Hibiscus syriacus*) en pépinière. La sous-espèce *multiplex* a été identifiée sur un échantillon de romarin (*Salvia rosmarinus*) dans la commune de Trofa. En région centrale, la bactérie a été détectée dans des échantillons de lavande (*Lavandula stoechas*, *L. angustifolia*) dans la municipalité de Leiria, et sur olivier (*Olea europaea*) dans la municipalité de Tábua ; la sous-espèce *multiplex* a été caractérisée dans ces deux communes. Les mesures obligatoires de lutte et de surveillance ont été étendues aux nouvelles zones.

Sources : Lien 1, Lien 2, Lien 3, Lien 4.

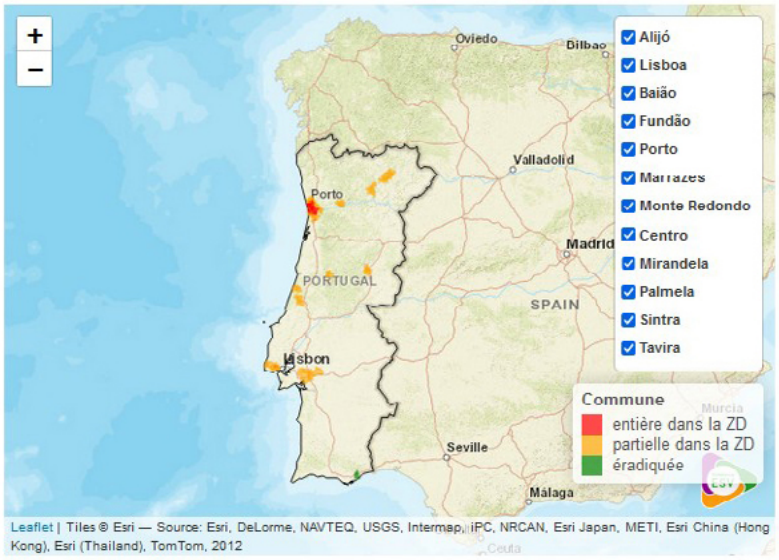

Figure 1 : Carte des municipalités touchées entièrement ou partiellement par la zone délimitée de *X. fastidiosa* au Portugal ainsi que la commune où la bactérie a été éradiquée. Sources : Lien 1, Lien 2,Lien 3, Lien 4.

Figure 4. Extraits du bulletin hebdomadaire (A) de la semaine 4 (2023, BH_s4) et du bulletin mensuel (B) n°48 (BM_48). Ces bulletins sont les productions principales de la VSI de la Plateforme ESV.

articles collectés permettent également d'alimenter des rapports rédigés par l'équipe de la VSI sur l'état d'avancement des connaissances de certains ON. Ces rapports dénommés « Point Sur » permettent de faire une synthèse des connaissances sanitaires, épidémiologiques et des actions de gestion mises en place pour limiter les maladies causées par ces ON. L'ensemble de ces productions, BH, BM et Point Sur, font l'objet d'un travail de relecture et de validation par les membres du comité éditorial puis sont diffusés sur le site internet de la Plateforme ESV sur une page dédiée à la VSI. La publication des bulletins sur le site s'accompagne d'une notification par mail à une liste de diffusion spécifique aux bulletins de la VSI de la Plateforme ESV. Les informations des catégories « actualités » ou « évènements » sont diffusées uniquement sur les pages spécifiques des ON ou sur les pages Actualités et Agenda du site web. Les informations présentes dans les BH sont également diffusées sur les pages spécifiques des ON lorsqu'elles existent.

Bilan du pipeline de la VSI

Quelques chiffres

En 2021 et 2022, respectivement 45 320 et 43 535 articles ont été collectés et stockés dans la BDD VSI. Cela équivaut à des valeurs médianes de 769 (2021) et 826 (2022) articles ouverts et classés par semaine. Ainsi, en 2021-2022, le temps correspondant à la phase 1 de classification (par l'équipe de la VSI) est estimé à environ 9 h de travail par semaine. La langue d'origine des articles est très diverse impliquant potentiellement une étape de traduction. En effet, sur les articles traités en 2021-2022, 113 langues différentes ont été détectées dont cinq majoritaires : l'anglais, l'italien, l'espagnol, le français et le portugais. Les langues décelées sont notamment liées à la langue parlée dans les pays où sont présents les ON suivis mais aussi à l'actualité sanitaire. Comme expliqué précédemment, la pertinence des articles collectés est évaluée à deux étapes : au moment de la classification (présélection), puis durant le CE pour la sélection finale. En 2021 et 2022, seuls 1,04 % des articles collectés ont été jugés pertinents pour être diffusés dans le BH (Tableau 3).

Ce taux de pertinence peut d'abord s'expliquer par une volonté de réaliser une veille couvrant un large spectre d'informations en utilisant des mots-clés peu restrictifs, permettant de capter les informations sanitaires majeures et les signaux faibles. D'autre part, il existe des contraintes inhérentes à la méthodologie de collecte qui récupère des informations sur le web « ouvert » et dépendent donc de l'indexation des pages, de leurs modalités de mise à jour, etc. De plus, pour que ces informations soient transmises au CE, elles doivent correspondre aux critères définis par l'arbre de décision, réduisant encore la taille du « filtre ».

Enfin, certains articles peuvent présenter une grande similarité, par exemple lorsqu'une information est reprise par plusieurs médias. Dans ce cas, tous les articles se voient attribuer un sujet, mais seuls quelques articles sont soumis au CE.

Enquête auprès du lectorat

Une enquête a été réalisée sur deux mois (décembre 2020 – janvier 2021) auprès du lectorat (*via* la liste de diffusion et le site web) afin de recueillir les avis et d'éventuelles suggestions d'amélioration concernant les bulletins. Fin 2020, le nombre d'abonnés à la liste de diffusion de la VSI était d'environ 300, et début 2022 il était de 406 abonnés. Les 53 réponses à cette enquête ont montré que les informations relayées dans les bulletins présentent une dimension opérationnelle importante permettant aux acteurs de la surveillance de mettre à jour leurs connaissances, communiquer avec leurs collaborateurs, et former leurs équipes. À titre d'exemple, la régularité des informations transmises a permis aux lecteurs d'être rapidement informés de la première détection de *Bactrocera dorsalis* à Hyères dans le Var

Tableau 3 : Proportions d'articles qui ont été jugés comme pertinents à différents niveaux au moment de la classification puis à l'issue du comité éditorial. Entre parenthèses, le nombre d'articles collectés sur l'année.

	Pourcentage d'articles en 2021 (et nombre total d'articles)	Pourcentage d'articles en 2022 (et nombre total d'articles)
Attribution d'un sujet (résumé de l'article)	10,76 (4 876)	10,64 (4 632)
Soumission au comité éditorial	4,52 (2 050)	3,52 (1 532)
Diffusion dans les bulletins hebdomadaires et sur le site internet	1,04 (473)	1,04 (452)

pendant l'été 2021. D'autre part, les informations publiées dans les bulletins sur *Xylella fastidiosa* dans les Pouilles en Italie ont contribué à sensibiliser les pépiniéristes d'oliviers français à cette problématique et au risque d'approvisionnement en Italie. Cette enquête a aussi fait ressortir un certain nombre de points à améliorer comme la mise en place d'une application de visualisation qui permettrait un suivi spatio-temporel des informations sanitaires rapportées ou encore le besoin de suivi d'un nombre plus important d'ON.

Conclusion

La méthodologie développée dans cet article constitue à notre connaissance, le premier outil de veille semi-automatique (hybridant codes informatiques, système d'information et apports humains) développé en santé végétale pour la France. Cet outil a l'avantage d'être adapté aux besoins de la Plateforme ESV et d'être évolutif car il permet l'ajout de nouvelles fonctionnalités et la mise à jour des requêtes en fonction du contexte sanitaire. De plus, les productions sont le fruit d'un travail de corédaction, de traduction, de synthèse, de vulgarisation et de cartographie qui permettent de rendre les informations provenant de plusieurs pays accessibles aux différents acteurs de la santé végétale en France. Le fonctionnement collaboratif permet également d'augmenter la qualité des informations diffusées. Cependant, la phase de classification des articles collectés est particulièrement chronophage et implique des tâches répétitives.

Afin d'augmenter le périmètre couvert par la veille, il semble incontournable d'automatiser une partie des activités. Les réflexions et les travaux autour de ces questions d'automatisation et le développement de prototypes sont intégrées dans deux projets de recherche dans lesquels la Plateforme ESV est impliquée : TIERS-ESV et BEYOND. Ces projets qui réunissent des biologistes, des épidémiologistes, des mathématiciens et des informaticiens devraient, à terme, permettre le développement de process robustes et automatisés. Le traitement naturel du langage[14] est une approche intégrée à ces projets, qui pourrait, à l'avenir, contribuer de manière significative à améliorer la surveillance des dangers sanitaires (Morris *et al*. 2022). L'ajout de nouvelles fonctionnalités pourrait également faciliter le suivi spatio-temporel des informations sanitaires et sa visualisation. Le pipeline de la VSI de la Plateforme ESV est en cours d'adaptation pour rendre possible l'implémentation d'outils issus des technologies et des projets cités.

Cette chaîne de traitement développée pour la santé végétale pourrait aussi, à terme, être adaptée à d'autres types de veille à l'échelle nationale (Plateformes) ou européenne (EFSA). ■

Remerciements

Nous remercions Carlène Trevennec (INRAE), Viviane Henaux (ANSES) et Samuel Soubeyrand (INRAE) pour leur relecture et leurs conseils.

14 Le traitement naturel du langage est un sous-domaine de l'intelligence artificielle qui a pour objectif de développer des programmes pour traiter et interpréter des énoncés oraux ou écrits (Morris *et al.*, 2022). Ces méthodes sont utilisées par exemple dans la traduction automatique de textes d'une langue à une autre.

Annexe 1 : Liste des organismes nuisibles ciblés dans la VSI au 13/10/22. Noms non exhaustifs basés sur les informations de l'OEPP[1] et du NCBI[2].

Nom scientifique validé ou candidat	Nom vernaculaire	Maladie(s)	Culture(s) touchée(s)	Précisions
Xylella fastidiosa	*Xylella*	Maladie de Pierce ; Syndrome de déclin rapide de l'olivier	Multiespèces	Bactérie
Bursaphelenchus xylophilus	Nématode du pin	Flétrissement du pin	Pins	Nématode
Grapevine flavescence dorée phytoplasma	Flavescence dorée	Jaunisse de la vigne	Vigne	Phytoplasme
Candidatus Liberibacter asiaticus	Bactérie du verdissement des citrus	Huanglongbing (HLB)	Agrumes	Bactérie
Candidatus Liberibacter africanus	Bactérie du verdissement des citrus	HLB	Agrumes	Bactérie
Candidatus Liberibacter americanus	Bactérie du verdissement des citrus	HLB	Agrumes	Bactérie
Spodoptera frugiperda	Légionnaire d'automne		Multiespèces	Lépidoptère
Tomato brown rugose fruit virus (ToBRFV)	Virus des fruits bruns et rugueux de la tomate		Tomates, Poivrons et Piment	Virus
Bactrocera dorsalis	Mouche orientale des fruits		Cultures fruitières et légumières	Insecte (Diptère)
Popillia japonica	Scarabée japonais		Multiespèces	Insecte (Coléoptère)
Fusarium oxysporum f. sp. cubense Tropical race 4 (FocTR4)		Fusariose du bananier	Culture de bananes	Champignon ascomycète
Bretziella fagacearum		Flétrissement américain du chêne	Chênes	Champignon ascomycète

1 European and Mediterranean Plant Protection Organization : https://gd.eppo.int/

2 National Center of Biotechnology : https://www.ncbi.nlm.nih.gov/

Annexe 2 : Liste des sources interrogées par le *web scraping*

	Source	Lien source	Type de veille	Statut source
Aggrégateurs de sources	EMM	https://emm.newsbrief.eu/	Événementielle	Officielle et non officielle
	MedISys	https://medisys.newsbrief.eu/medisys/homeedition/fr/home.html	Événementielle	Officielle et non officielle
	Google Search	https://www.google.fr	Événementielle	Officielle et non officielle
	Gouvernement UK	https://www.gov.uk/government/publications/plant-health-news	Événementielle	Officielle
	PestAlert	https://www.pestalerts.org/official-pest-reports https://www.pestalerts.org/emerging-pest-alert	Événementielle	Officielle
	PubMed	https://pubmed.ncbi.nlm.nih.gov/	Scientifique	Savante
	Google Scholar	https://scholar.google.com/	Scientifique	Savante
Source « directe »	USDA	https://www.aphis.usda.gov/aphis/ourfocus/planthealth	Événementielle	Officielle
	EPPO	https://www.eppo.int/	Événementielle	Officielle
	Boagri	https://info.agriculture.gouv.fr/gedei/site/bo-agri	Événementielle	Officielle
	Alimagri	https://agriculture.gouv.fr/sante-et-protection-des-vegetaux	Événementielle	Officielle
	Newsletter de Bruno Peiffer (Agent chargé de la veille à la DGAl)	Newsletter reçue par mail	Événementielle et scientifique	Officielle et non officielle
	Nature	https://www.nature.com/	Scientifique	Savante

Références

Amar, H. et Dupuy, C. (2020). Les plateformes d'épidémiosurveillance : un concept novateur au service de l'efficience des dispositifs de surveillance. Bulletin de l'Académie Vétérinaire de France, 173, pp. 200-5. Disponible sur : https://doi.org/10.3406/bavf.2020.70904.

Ansoff, H. I. (1975). Managing Strategic Surprise by Response to Weak Signals. California Management Review, 18(2), pp. 21-33. Disponible sur : https:// doi.org/10.2307/41164635.

Antoniou, P. H. et Sullivan, P. E. (dir.). (2006). The Igor Ansoff Anthology. Charleston (SC) : BookSurge Publishing.

Alomar, O., Batlle, A., Brunetti, J. M., García, R., Gil, R., Granollers, T., et al. (2016). Development and testing of the media monitoring tool MedISys for the monitoring, early identification and reporting of existing and emerging plant health threats. EFSA supporting publication, 13(12), 81 pp. Disponible sur : https://doi.org/10.2903/sp.efsa.2016.EN-1118.

Eilstein, D., Salines, G. et Desenclos, J. C. (2012). Veille sanitaire : outils, fonctions, processus. Revue d'épidémiologie et de Santé Publique, 60(5), p. 401-11. Disponible sur : https://doi.org/10.1016/j.respe.2012.03.005.

Lannou, C., Rasplus, J.-Y., Soubeyrand, S., Gautier, M., Rossi, J.-P. (dir.). (2023). Crises sanitaires en agriculture : Les espèces invasives sous surveillance. Versailles : Editions Quæ.

Morris, C. E., Géniaux, G., Nédellec, C., Sauvion, N. et Soubeyrand, S. (2022). One Health concepts and challenges for surveillance, forecasting, and mitigation of plant disease beyond the traditional scope of crop production. Plant Pathology, 71, pp. 86-97. Disponible sur : https://doi.org/10.1111/ppa.13446.

Tateosian, L. G., Saffer, A., Walden-Schreiner, C. et Shukunobe, M. (2023). Plant pest invasions, as seen through news and social media. Computers, Environment and Urban Systems, 100, p. 101922. Disponible sur : https://doi.org/10.1016/j.compenvurbsys.2022.101922.

Valentin, S., Arsevska, E., Falala, S., De Goër, J., Lancelot, R., Mercier, A., et al. (2020). PADI-web: A multilingual event-based surveillance system for monitoring animal infectious diseases. Computers and Electronics in Agriculture, 169, 105163. Disponible sur : https://doi.org/10.1016/j.compag.2019.105163.