Élaboration d'une méthode semi-automatique pour l'identification et le traitement des signaux d'émergence pour la veille internationale sur les maladies animales infectieuses

Elena Arsevska

NNT : 2017SACLS008

Thèse de doctorat de l'Université Paris-Saclay préparée à l'Université Paris-Sud

Ecole doctorale n°570
Santé publique
Spécialité de doctorat : Epidémiologie

par

Mme. Elena Arsevska

Élaboration d'une méthode semi-automatique pour l'identification et le traitement des signaux d'émergence pour la veille internationale sur les maladies animales infectieuses

Thèse présentée et soutenue à Paris, le 31 Janvier 2017.

Composition du Jury :

Mme.	Juliette Dibie	Professeur AgroParisTech, France	(Rapporteur)
M.	Philippe Sabatier	Professeur Université Joseph Fourier, VetAgro Sup, France	(Rapporteur)
M.	Didier Calavas	Chercheur Anses, France	(Examinateur)
Mme.	Agnès Waret-Szkuta	Maître de Conférences Ecole véterinare nationale de Toulouse, France	(Examinateur)
M.	Loïc Josseran	Professeur Université de Versailles Saint-Quentin-en-Yvelines, France	(Président)
Mme.	Barbara Dufour	Professeur Ecole véterinaire d'Alfort, France	(Directeur de thèse)
M.	Renaud Lancelot	Chercheur Cirad, France	(Invité, Encadrant)
M.	Mathieu Roche	Chercheur Cirad, France	(Invité, Co-encadrant)

Посветено на моето семејство

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor Renaud Lancelot. It has been an honour to be his Ph.D. student. He passed me, both consciously and unconsciously, the love towards epidemiology. He also showed me what it means to use the research to solve various animal health challenges on the field, especially in the countries in development. I could not have imagined having a better advisor and mentor for my thesis.

I am also thankful to my second advisor, Dr. Mathieu Roche for the excellent example he has provided as a computer scientist. He helped me discover new methods of data analysis. His enthusiasm for research was contagious and motivational for me, even during tough times in the Ph.D. pursuit. I have no doubts that this is the area I would like to explore in my future research.

I express my highest gratitude to my thesis director, Prof. Barbara Dufour for her continuous support, presence and motivation during my thesis. Her original ideas and immense knowledge in the domain of epidemiology made my Ph.D. experience productive and stimulating. Her guidance helped me in all the time of research and writing of this thesis.

Besides my advisors, I would like to thank the rest of my thesis committee : Prof. Juliette Dibie, Prof. Philippe Sabatier, Prof. Loïc Josseran, Dr. Agnès Waret-Szkuta and Dr. Didier Calavas for their insightful comments, but also for the questions which will incent me to widen my research from various perspectives.

I am grateful to the French veterinary services, the Directorate General for Food (DGAL) for financing this work. I thank Dr. Thierry Lefrançois, for the continuous support as a Director of our Unit for Control of exotic and emerging animal diseases (CMAEE) in Cirad, Montpellier. With the financial and moral support, I was privileged to attend numerous trainings, conferences and courses and never to miss anything during my Ph.D.

I thank all the members of the French epidemic intelligence team for international monitoring of animal health (french : *Veille sanitaire internationale*). Over the past three years, we have shown that our network is successful because of our full devotion in the protection of human and animal health. I thank all the experts and stakeholders that actively contributed to the work and improvement of our network.

The realisation of the tool PADI-web would not have been possible without the tremendous contribution of Julien, Jocelyn, Sylvain, Mathieu and the students Clément, Baptiste, Max and Thomas.

I thank Marion for correcting the French in the manuscript. I thank her for the long hours spent in decoding my French.

My time in Montpellier was made enjoyable in large part due to the many friends and colleagues that became a part of my life. I am grateful for the unforgettable moments spent

with my very dear colleagues and friends Pachka, Tiffany, Alizé, Ahmadou, Andrea, Sylvie, Cécile, Berenice, Esmaile, Denise, Vito, Antonio, Francesca, and Amaury and Moises with their families. I am most thankful to Caroline with whom, besides sharing the office, we shared wonderful moments as friends. Special thanks to Sylvain who was an excellent working partner and a best friend during the past three years. I also thank all the other colleagues from Cirad, CMAEE for the collaboration and joyful times spent together.

I also thank all my friends in Macedonia for their continuous support and encouragement. I would not have been here without the support of the colleagues from the Food and Veterinary Agency of Macedonia – thank you tremendously for believing in me.

Finally, I thank Paulo for being an excellent partner, friend and a colleague. We have shared all the ups and downs that the thesis brings – and we made it. Thanks Dell for loving us so much.

Lastly, I would like to thank my family Aleksovi and Kralevski for all their love and encouragement. For my mother who raised me with love towards science and supported me in all my pursuits. I love you all very much.

Thank you. You are all invited to visit me wherever the future takes me.

Table des matières

Liste des figures

Liste des tableaux

Liste des équations

Travaux issus de la thèse

Articles entrant dans le cadre de la thèse avec comité de lecture :

— Arsevska, E., Roche, M., Hendrikx, P., Chavernac, D., Falala, S., Lancelot, R., Dufour, B., 2016. Identification of terms for detecting early signals of emerging infectious disease outbreaks on the Web. Computers and Electronics in Agriculture. 123, 104-115. (Annexe A.1).

— Arsevska, E., Roche, M., Hendrikx, P., Chavernac, D., Falala, S., Lancelot, R., Dufour, B., 2016. Identification of associations between clinical signs and hosts to monitor the Web for detection of animal disease outbreaks. International Journal of Agricultural and Environmental Information Systems. 7 :3, 1-20. (Annexe A.2).

— Arsevska, E., Rabatel, J., Goer, J., Falala, S., Roche, M., Lancelot, R., Dufour, B. Web monitoring of exotic animal infectious diseases integrated in the French epidemic intelligence system. PlosOne. En cours de rédaction.

Actes de conférences internationales avec comité de lecture :

— Arsevska, E., Roche, M., Lancelot, R., Hendrikx, P., Dufour, B., Falala, S., Chavernac, D., 2016. Monitoring Disease Outbreak Events on the Web Using Text mining Approach and Domain Expert Knowledge., in : N. Calzolari et Al. (Ed.). In Proceedings of LREC 2016 (International Conference on Language Resources and Evaluation), Portorož, Slovenia, pp. 3407-3411.

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

Communications orales :

— Arsevska, E., Mercier, A., Lefrancois, T., Lancelot, R., Peiffer, B., Bronner, A., Bournez, L., Cauchard, J., Etore, F., Hendrikx, P., Calavas, D. International monitoring of animal health threats : the case of the French epidemic intelligence team. Presenté à la AITVM-STVM conference, "Tropical Animal Diseases and Veterinary Public Health : Joining Forces to Meet Future Global Challenges", 2016 Berlin, Germany.

— Falala, S., De Goër De Herve, J., Arsevska, E., Roche, M., Rabatel, J., Chavernac, D., Hendrikx, P., Lefrancois, T., Dufour, B., Lancelot, R. Système de veille sanitaire pour analyser l'émergence et la propagation de maladies animales. Presenté à la conférence Ingénierie des Connaissances (IC 2016) et l'atelier d'intégration de sources/masses de données hétérogènes et ontologies, dans le domaine des sciences du Vivant et de l'Environnement (IN-OVIVE 4ème edition), 2016, Montpellier, France.

— Arsevska, E., Roche, M., Lancelot, R., Hendrikx, P., Dufour, B., Falala, S., Chavernac, D. Disease outbreak articles as a source of queries for detection of signals of

disease emergence on the Web. Presenté à la 14th Conference of the International Society for Veterinary Epidemiology and Economics (ISVEE 2014), 2015, Merida, Mexico.

— Arsevska, E., Peiffer, B., Perrin, Jean-B., Marcé, C., Hendrikx, P., Etore, F., Collignon, C., Lancelot, R., Lefrançois, T., Calavas, D. La veille sanitaire internationale en santé animale en France. Presentée à la Journée de la Plateforme Esa, 2015, Maisons – Alfort, France.

Autre publication en relation avec la thèse :

— Arsevska, E., Bronner, A., Calavas, D., Cauchard, J., Caufour, P., Falala, S., Hamon, M., Hendrikx, P., Lancelot, R., Mercier, A., Séverine, R., Tisseuil C. Dermatose nodulaire contagieuse des bovins : état des connaissances et situation épidémiologique dans les Balkans au 31 juillet 2016. Bull. Epid. Santé Anim. Alim. 2016, (74) : 25-29.

— Arsevska, E., Lancelot, R., El Mamy B., Cêtre-Sossah C. Situation épidémiologique de la fièvre de la Vallée du Rift en Afrique de l'Ouest et du Nord. Bull. Epid. Santé Anim. Alim. 2016, (74) : 25-29.

— Mercier, A., Arsevska, E., Lancelot, R., Diallo, A., Libeau, G. Situation épidémiologique de la peste des petits ruminants (PPR) en Europe de l'Est et au Moyen-Orient. Bull. Epid. Santé Anim. Alim. 2016, (74) : 30.

— Le Potier, MF., Arsevska, E., Marcé, C. Persistance de la peste porcine africaine en Europe de l'Est. Bull. Epid. Santé Anim. Alim. 2015, (70) : 28-29.

— Arsevska E., Calavas D., Dominguez M., Hendrikx P., Lancelot R., Lefrançois T., Peiffer B., Perrin J.B. Des laboratoires de référence à la veille sanitaire internationale en France. Bull. Epid. Santé Anim. Alim. 2015, (66) : 16-18.

— Arsevska, E., Balenghien, T., Garros, C., Lancelot, R., Zientara, S., Sailleau, C., Bréard, E. Fièvre catarrhale ovine en Europe en 2014 : épizootie dans les Balkans, progression de la circulation en Italie et en Espagne. Bull. Epid. Santé Anim. Alim. 2014, (69) : 20-24.

— Arsevska, E., Dominguez, M., Peiffer, B., Perrin, Jean-B., Marcé, C., Hendrikx, P., Etore, F., Collignon, C., Lancelot, R., Lefrançois, T., Calavas, D., Développement d'une veille sanitaire internationale en santé animale dans le cadre de la Plateforme ESA. Bull. Epid. Santé Anim. Alim. 2014, (63) : 30-31.

— Arsevska, E., Calavas, D., de Sales Lima, FE., Faye, B., Hendrikx, P., Lancelot, R., Lefrancois, T., Libeau, G. 2014. Coronavirus du Syndrome respiratoire du Moyen-Orient (MERS-CoV) : quel réservoir animal ? Bull. Epid. Santé Anim. Alim. 2014, (56) : 15-17.

— Arsevska, E., Calavas, D., Dominguez, M., Guis, H., Hendrikx, P., Lancelot, R., Peiffer, B., Perrin, J.B. 2014. Fièvre catarrhale ovine en Sardaigne – un point épidémiologique pour les années 2012 et 2013. Bull. Epid. Santé Anim. Alim. 2014, (56) : 13-14.

— Arsevska, E., Calavas, D., Dominguez, M., Hendrikx, P., Lancelot, R., Lefrancois,

T., Le Potier, MF., Peiffer, B., Perrin, JB. Peste porcine africaine en Sardaigne en 2014 – de l'enzootie à l'épizootie ? Bull. Epid. Santé Anim. Alim. 2014, (56) : 11-12.

Collaboration à des projets :

- Animation du dispositif (novembre 2013 - mars 2016) de Veille sanitaire internationale en santé animale en France. Plateforme national d'épidémiosurveillance en santé animale.
- Projet de licence. Acquisition de données textuelles liées à l'épidémiologie. Baptiste Belot. IUT de l'Université de Montpellier. Année 2014 – 2015.
- Projet de licence.Text Mining for Monitoring Disease Emergence from the Information Published on the Web. Clément Hemeury. IUT de l'Université de Montpellier. Année 2014 - 2015.
- Projet de post-doctorat. Information extraction from non-structured texts from the Web. Julien Rabatel. Année 2015-2016.
- Projet de licence. Interface Web pour la Veille sanitaire internationale en santé animale en France. Max Devaud et Thomas Filiol. IUT de l'Université de Clermont – Fernard. Développement d'applications intranet/ internet. Année 2015 - 2016.

Liste des abréviations

ADNS	Système de Déclaration des Maladies Animales
ANSES	Agence Nationale de la Sécurité Sanitaire
AUC	Aire sous la Courbe
BTV	Virus de la Bluetongue
CA VSI	Cellule d'Animation de la Veille Sanitaire Internationale
CDC	Centre pour le Contrôle et la Prévention des Maladies
CE	Commission Européenne
CMI	Information Mutuelle Cubique
DGAL	Directorat Général de l'Alimentation
EAT	Extraction Automatique de Termes
ECDC	Centre Européen de Prévention et Contrôle de maladies
EI	Extraction d'Information
EMPRES	Système de Prévention et de Réponse Rapide Contre les Maladies
EN	Entités Nominales
ESA	Épidémiosurveillance en Santé Animale
EWRS	Système d'Alerte Précoce et Rapide
FA	Fièvre Aphteuse
FAO	Organisation des Nations Unies pour l'Alimentation et l'Agriculture
FCO	Fièvre Catarrhale Ovine
GPHIN	Réseau Mondial d'Information en Santé Publique
IA	Influenza Aviaire
IAHP	Influenza Aviaire Hautement Pathogène
IBIS	Système International de Biosurveillance
ICD	Classification Internationale des Maladies
IDF	Fréquence Inverse de Documents
IE	Intelligence Épidémiologique
JRC	Centre de Recherche Commun
MedISys	Système d'Information Médicale
MeSH	Vedettes-matières Médicales
MI	Information Mutuelle
NB	Naïve Bayes
OIE	Organisation Mondiale de la Santé Animale
OMS	Organisation Mondiale de la Santé
ONG	Organisation Non Gouvernementale
PADI-web	Plateforme pour l'extraction automatique d'information sanitaire sur le Web
PMI-IR	Information Mutuelle Ponctuelle et
RI	Recherche d'Information
PPA	Peste Porcine Africaine
PROMED	Programme pour la Veille des Maladies Émergentes
RI	Recherche d'Information

ROC	Fonction d'Efficacité du Récepteur
RSI	Règlement Sanitaire International
SBE	Surveillance Fondée sur des Évènements
SBI	Surveillance Fondée sur des Indicateurs
SBV	Maladie de Schmallenberg
SNOMED	Nomenclature Systématique en Médecine Clinique
SVM	Machines à Vecteurs de Support
TALN	Traitement Automatique de Langage Naturel
TESSY	Système de Surveillance Européen
TF	Fréquence du Terme
TF-IDF	Fréquence du Terme-Fréquence Inverse de Document
UE	Union Européenne
VPP	Valeur Prédictive Positive
VSI	Veille Sanitaire Internationale
WAHID	Base de Données du Système Mondial d'Information Sanitaire
WAHIS	Système Mondial d'Information Sanitaire

Chapitre 1

Introduction

1.1 Contexte

Tandis que la mondialisation a généré de nombreux bénéfices pour la société, elle a également soulevé de nouveaux défis, particulièrement en ce qui concerne la santé humaine, la santé animale et la protection environnementale (Sherman, 2010). Les agents pathogènes de maladies infectieuses ignorent les frontières et traversent de continent en continent, en se servant des nouveaux contacts entre les biens, les humains et les animaux (Wood, 2007).

Durant les dernières décennies, les systèmes traditionnels de surveillance, fondés sur la surveillance des maladies connues, se sont révélés insuffisants pour détecter précocement de nouvelles émergences. Il est donc nécessaire de développer des méthodes originales et robustes de détection précoce des signaux des maladies exotiques et émergentes, notamment à travers l'analyse du contenu du Web. Ces méthodes assureraient la mise en place rapide de mesures de contrôle et de prévention (Amato-Gauci et al., 2008).

Grâce à l'utilisation du Web et des réseaux sociaux qui permettent de diffuser des informations relatives aux maladies infectieuses, une réelle prise de conscience s'est développée au cours des dernières décennies (Brownstein et al., 2009). Cette prolifération rapide d'information sous forme de texte non-structuré disponible dans une multitude de référentiels sur le Web a mis en exergue un nouveau défi : comment accéder et utiliser cette information issue du Web d'une manière efficace pour une veille épidémiologique ? Cela a entraîné une utilisation croissante des techniques d'analyse des données et de fouille de textes (« text mining ») afin de découvrir de nouvelles connaissances épidémiologiques contenues dans ces données textuelles non-structurées.

L'inscription d'une nouvelle thématique « Veille sanitaire internationale » (VSI) dans le programme d'activité de la Plateforme nationale d'épidémiosurveillance en santé animale (Plateforme ESA) a été décidée en 2013. La mission de la VSI est d'identifier, suivre et analyser les signaux de dangers sanitaires (en santé animale au sens large) menaçant le territoire français. L'analyse de ces signaux permettra de produire une information sanitaire claire dont le risque pourra être évalué par l'ANSES (Agence Nationale de la Sécurité Sanitaire) et géré par la DGAL (Direction Générale de l'Alimentation). La nature

des dangers sanitaires n'est pas prédéterminée, mais la priorité est donnée aux maladies exotiques et à leurs signaux non spécifiques (Arsevska et al., 2014c).

Au quotidien, la veille est essentiellement réalisée par la recherche, l'acquisition, l'analyse et la communication d'information épidémiologique par des analystes d'une Cellule d'Animation de la VSI en santé animale (CA VSI). Les données sont issues : *i)* de sources formelles (données structurées), comme la déclaration des maladies par des organismes internationaux, et *ii)* de sources informelles (données non-structurées), comme des média, les communications par des laboratoires de référence, des réseaux de surveillance ou des référents thématiques et géographiques qui collaborent avec la CA VSI.

La veille de la CA VSI qui est fondée sur des sources informelles présente plusieurs limites. La recherche manuelle sur le Web est fastidieuse et prend du temps. Cela s'applique également à la lecture individuelle des différents articles de média (articles), à la sélection d'un contenu pertinent ainsi que la préparation de rapports avec les dangers sanitaires repérés. Par exemple, pour faciliter la veille en santé animale, l'un des analystes de la CA VSI s'appuie sur une liste prioritaire contenant des centaines de sujets et de mots-clés. Ce derniers, lui permettent de créer des requêtes automatiques sur Google news et de préparer des synthèses quotidiennes résumant le titre de l'article acquis, la date de publication, la source et le lien vers le contenu (la page Web source). Ces synthèses sont ensuite envoyées par des courriels à un groupe d'utilisateurs restreint. Les listes des sujets et de mots-clés ne sont ni exhaustives, ni suffisamment souples et flexibles pour permettre leurs mises à jour régulières. Les synthèses quotidiennes ne permettent pas une interprétation épidémiologique claire et rapide de la situation sanitaire. De telles limites retardent la réactivité de la CA VSI et augmentent le risque éventuel d'introduction de dangers sanitaires en France.

1.2 Objectifs de la thèse

Le travail de cette thèse répond aux besoins de ce dispositif de VSI en France, se focalisant toutefois sur l'acquisition, la catégorisation et l'extraction de l'information sanitaire à partir des données non-structurées du Web (articles). L'objectif général de ce travail consiste à intégrer de nouvelles connaissances épidémiologiques et informatiques pour proposer une méthode originale de veille sur le Web pour des maladies animales infectieuses exotiques et émergentes au niveau international. Dans ce contexte, cette thèse propose de répondre aux questions suivantes :

1. Comment acquérir automatiquement les données sanitaires issus du Web ?
2. Comment extraire l'information épidémiologique à partir de ces données ?
3. Comment représenter l'information de manière synthétique et facile à analyser par des évaluateurs et gestionnaires du risque ?

1.3 Plan

Cette thèse est structurée en quatre chapitres. Ce premier chapitre introduit le contexte général à la thèse. Le chapitre 2 décrit l'état de l'art des différentes approches de surveillance : traditionnelles (fondées sur des indicateurs) et complémentaires (fondées sur des évènements), ainsi que leurs enjeux. Le chapitre 3 décrit notre proposition d'une approche de fouille de textes pour une veille sanitaire en santé animale à partir du Web. La démarche globale s'appuie sur les étapes d'acquisition de documents du Web, de leur classification automatique en différentes catégories, de l'extraction de l'information sanitaire à partir de documents pertinents, et la visualisation et l'analyse de l'information sanitaire. Le chapitre 4 aborde les conclusions et les perspectives de ce travail.

Chapitre 2

État de l'art des différentes approches de veille

2.1 Définitions et contexte

Telle que définie par (Paquet et al., 2006), l'intelligence épidémiologique (IE) comprend toutes les activités liées à la détection précoce de dangers sanitaires susceptibles de représenter un risque pour la santé. Ces activités comprennent l'évaluation du risque, nécessaire à la mise en place de mesures de prévention et contrôle appropriées. L'IE repose sur deux composantes :

— **les indicateurs** (SBI, « indicator based surveillance ») qui se réfèrent à des données structurées recueillies par le biais de systèmes de surveillance traditionnels (en routine), telles que des déclarations officielles des foyers par les autorités sanitaires à l'Organisation mondiale de la santé (OMS) ou à l'Organisation mondiale de la santé animale (OIE),
— **les événements** (SBE, « event based surveillance ») qui se réfèrent à des données non structurées collectées depuis des sources hétérogènes, comme le Web, les réseaux sociaux, des experts de terrain, etc.

L'Annexe B résume les principales caractéristiques de ce type de surveillance (veille) fondé sur des indicateurs et des événements. Les SBI et SBE ne sont pas nécessairement des systèmes de veille distincts, sans rapport entre eux. Ils ont une fonction d'alerte rapide, agissant de concert pour contribuer à la veille sanitaire internationale (Figure 2.1). Néanmoins, en considérant leur mise en œuvre opérationnelle, la définition de la procédure d'IE pour ces deux composantes est essentielle. Ces procédures seront successivement décrites dans la section qui suit.

Conceptuellement, la méthodologie de l'IE peut être détaillée en trois étapes : *i)* l'acquisition de données, *ii)* la vérification et l'analyse de ces données et *iii)* la communication de l'information sanitaire (Cakici et al., 2010, Paquet et al., 2006).

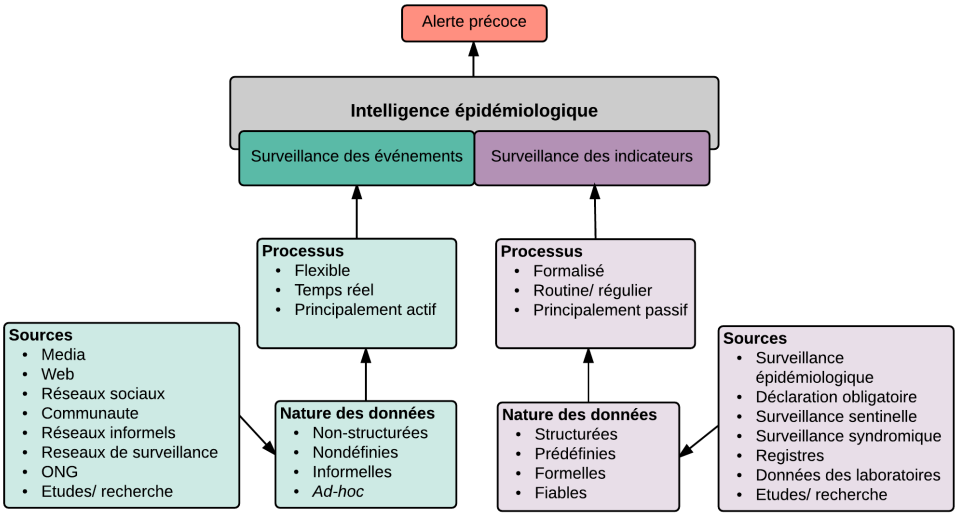


Figure 2.1 – Concept de l'intelligence épidémiologique et les composantes fondées sur les indicateurs et sur les événements (source : Barboza, 2014)

La **première étape, l'acquisition des données**, consiste à rassembler des listes de sources de données, à définir des stratégies de recueil de données (selon les différents types de sources), à formater et filtrer les données acquises et à définir et mettre en place des solutions de stockage. Par exemple, la collecte de données via les systèmes de surveillance traditionnels (basés sur les indicateurs, SBI) comprend le recueil systématique de variables qualitatives (nombres de cas, de foyers, de syndromes, etc.) à partir de systèmes de surveillance pré-établis tels que les résultats des programmes de surveillance nationaux rapportés à l'OIE ou l'OMS. Quant aux systèmes de surveillance basés sur les évènements (SBE), de par leur définition propre, ils se rapportent aux données non structurées captées à partir de sources d'informations de toute nature.

La **deuxième étape, l'analyse de données**, repose sur une grande variété de méthodes de fouille de textes et statistiques utilisées pour extraire et décrire des indicateurs épidémiologiques depuis les données collectées (lieu des foyers, date, espèces affectées et nombre de cas, etc.). Pour l'étude des événements, le processus consiste de plus à évaluer la pertinence des signaux détectés par le système (c'est-à-dire la vérification des signaux).

La **dernière étape, la communication d'informations**, comporte des méthodes de traduction nécessaires à la communication claire et synthétique des résultats d'analyse, en ciblant des utilisateurs qui ne sont pas nécessairement spécialistes. Les résultats peuvent être présentés sous plusieurs formes, notamment des informations sous forme textuelle partagés avec les utilisateurs par courriel, par des réseaux sociaux, ou encore par les pages Web dédiées. Ces messages font souvent référence à la maladie, au pays, à l'espèce sensible, à l'incidence ou à la prévalence, à la morbidité, à la mortalité, à la létalité, etc. Il peuvent aussi contenir des cartes de distribution de foyers ou des séries chronologiques avec le nombre de foyers.

2.2 Surveillance fondée sur des indicateurs

Bénéficiant du soutien de l'État, les origines de la SBI remontent au 19ème siècle, tout particulièrement en Europe et aux États-Unis. Ce ne sera cependant qu'au 20ème siècle que seront signées les premières conventions internationales pour lutter contre le choléra, la peste, la fièvre jaune et le grippe chez les humains ainsi que contre les pestes porcines/ bovines et la fièvre aphteuse (FA) chez les animaux (Choi, 2012 ; Sherman, 2007). La nécessité de coordonner la surveillance et de partager les informations sur la situation sanitaire mondiale a abouti à la création de deux organisations internationales : l'OIE en 1924 (Vallat et al., 2013) pour lutter contre les maladies infectieuses animales et l'OMS en 1948 pour les maladies transmissibles aux humains (zoonoses [1]). Une autre organisation internationale, l'Organisation des Nations Unies pour l'alimentation et l'agriculture (FAO), fut créée en 1946 dans le but d'améliorer la productivité agricole dans les pays en développement, notamment via l'amélioration de l'action vétérinaire (Tableau 2.1).

Après les années 1990, la technologie des systèmes d'information, comme l'accès au Web, l'utilisation de bases de données en ligne et le développement des systèmes d'information géographique, ont considérablement évolué, incitant les organisations internationales à mettre en place des systèmes de déclarations de maladies plus transparents via les plateformes Web appropriées.

En 1994, la FAO lança un système de prévention et de réponse rapide contre les maladies transfrontalières (Empres) se focalisant dans un premier temps sur les pestes porcines et l'influenza aviaire. Depuis 2004, Empres-i la plateforme Web de la FAO, ouverte à la communauté, permet de visualiser des données épidémiologiques sur plus de 34 maladies transfrontalières (Tableau 2.1) (Martin et al., 2007).

En 1996, l'OIE a mis en place un système de déclaration de maladies sécurisé en ligne. Ce système a été modernisé en 2006 grâce à une interface Web ouverte à la communauté (WAHIS : Système Mondial d'Information Sanitaire) (Ben Jebara, 2010). Cette interface permet l'accès à des informations épidémiologiques concernant plus de 100 maladies infectieuses des animaux terrestres et aquatiques, signalées dans plus de 180 pays (Tableau 2.1).

Au niveau européen, deux systèmes sont responsables de la collecte, la centralisation et le partage des données sanitaires pour les maladies infectieuses. Le système de déclaration de maladies animales (ADNS), créé en 1982, centralise et analyse les données sanitaires de 45 maladies animales exotiques susceptibles d'émerger en Europe afin de pouvoir alerter les pays européens en cas de risque d'introduction. De manière similaire, le système de déclaration en ligne TESSY, créé en 2004, permet le partage de l'information sanitaire au sujet de 52 maladies infectieuses humaines (Tableau 2.1) (Ammon et al., 2010 ; Zeller et al., 2013).

1. Maladies et infections dont les agents pathogènes se transmettent naturellement des animaux vertébrés à l'homme et vice-versa

Le sous-chapitre suivant présente les principales caractéristiques des systèmes de surveillance fondés sur des indicateurs, également synthétisés dans le Tableau 2.1. L'Annexe C.1 présente des exemples d'interface graphique pour ce type de systèmes (SBI).

Tableau 2.1 – Principales caractéristiques des systèmes fondés sur la surveillance des indicateurs (SBI)

Nom du système	WAHIS	EWARS	Empres-i	ADNS	TESSY
Année de constitution	1996	2000	2004	1998	2004
Couverture géographique	Mondial	Mondial	Mondial	Européen	Européen
Accès	Public	Restreint et public	Public	Restreint	Restreint et public
No. langues	3	1	1	1	1
Type de risque [a]	A	H	A, H	A	H
Si risque humain, no. maladies/ syndromes[a]	/	55 maladies H/ 7 syndromes	2 maladies H	/	52 maladies H
Si risque animal, no. maladies/ syndromes [a]	90 maladies AT, 18 maladies AA	/	34 maladies AT	32 maladies AT, 13 maladis AA	/
Sources de données [b]	Formelles	Formelles, Informelles	Formelles, Informelles	Formelles	Formelles
Acquisition de données [c]	A/ M	A/ M	A	A	A
Validation du contenu des événements	Oui/Humaine	Oui/Humaine	Non	Non	Non
Archive des événements	Oui	Oui	Oui	Oui	Oui
Carte géographique	Oui	Non	Oui	Oui	Oui
Graphiques	Oui	Non	Oui	Non	Oui
Statistique	Oui	Non	Non	Non	Oui
Communication aux utilisateurs, sauf interface	Email, RSS flux, Smartphone	RSS flux, Bulletin	RSS flux, Bulletin	Email	Email
Références	Ben Jebara, 2007	Formenty et al., 2006	Martin et al., 2007	European Council, 2016	Amato-Gauci et al., 2008

a. H=humaine, A=animale, AT = animaux terrestres, AA = animaux aquatiques

b. Source formelle = source officielle, information validée par les instances compétentes, Source informelle = tout autre source avec une information non-validée par les instances compétentes

c. M=Manuelle, A=Automatique

2.2.1 Acquisition

La plupart des systèmes internationaux fondés sur des indicateurs, collectent et centralisent des données structurées sur des foyers de maladies infectieuses. Ces données proviennent généralement de sources officielles et formelles, telles que les autorités sanitaires nationales (services vétérinaires, services de santé publique, etc.). Selon les réglementations internationales, les cas de maladies listées par l'OIE ou par l'OMS doivent être signalés auprès de ces organismes dans les 24 heures suivant leur détection. De même, doivent être signalés tout changement dans la distribution ou l'augmentation de l'incidence, de la virulence, de la morbidité ou de la mortalité liée à l'agent étiologique d'une maladie ; mais aussi la présence d'une maladie dans un hôte inhabituel ou encore la détection d'une maladie émergente nouvelle. Pour les maladies enzootiques, l'OIE exige la production d'un rapport synthétique biannuel par les instances nationales.

L'application Web d'Empres-i permet de collecter, stocker, analyser et communiquer des informations venant de sources officielles et non-officielles. Les données provenant de sources officielles, comme l'OIE et l'OMS, sont automatiquement acquises, extraites et sauvegardées dans la base de données. Cependant, les données en provenance des experts de la FAO sur le terrain, des laboratoires de référence, des autorités nationales et des articles de recherche (sources non-officielles, non-formelles) sont saisies manuellement par des analystes d'Empres-i (Vallat et al., 2013).

Les autorités sanitaires des états membres de l'Union Européenne (UE) contribuent au système TESSY pour transférer leurs données de systèmes de surveillance des maladies infectieuses et leurs données de systèmes sentinelles à intervalles réguliers. La fréquence des déclarations dépend de la gravité et de la contagiosité des infections : journalière (*e.g.*, légionellose), hebdomadaire (*e.g.*, grippe, fièvre du Nil occidental), mensuelle (*e.g.*, salmonellose) ou annuelle (*e.g.*, tuberculose) (Guglielmetti et al., 2006). Les données sont déposées en ligne par des utilisateurs assermentés par les autorités nationales compétentes. Le processus de dépôt des données en ligne est contrôlé par plusieurs étapes de validation automatique. Ainsi, les incohérences ou erreurs éventuelles sont détectées par le système qui avertit le fournisseur de données afin que celui-ci procède à une vérification (Ammon et al., 2010). Des règles similaires de contrôle automatique de déclaration en ligne s'appliquent au système ADNS.

2.2.2 Analyse

Les principaux systèmes internationaux fondés sur des indicateurs sont alimentés par des données structurées qui comprennent des informations géographiques, cliniques, épidémiologiques et de laboratoires décrivant les foyers de maladies et les cas.

Les données sont en grande partie analysées de manière automatique et descriptive. La plupart des interfaces Web des systèmes SBI permettent aux utilisateurs de représenter les foyers de maladies sur des cartes géographiques statiques, comme notamment le cas

de l'interface des systèmes ADNS et WAHIS. Le système WAHIS fournit également des tableaux synthétiques d'incidence par pays, ainsi que les mesures de surveillance et de contrôle et les capacités des laboratoires vétérinaires en matière de diagnostic. Cela permet aux utilisateurs de mieux évaluer et gérer les risques sanitaires.

Plus sophistiquées, les interfaces des systèmes Empres-i et TESSY permettent une représentation sous forme de graphiques de tendances et de cartes interactives illustrant des données agrégées par pays. De plus, pour faciliter la visualisation, Empres-i permet aux utilisateurs d'ajouter de l'information sous forme de couches, on peut ainsi représenter sur une même carte l'incidence de la maladie et la densité de bétail estimée par la FAO (Amato-Gauci et al., 2008 ; Ammon et al., 2010 ; Clements et al., 2002).

Les interfaces des systèmes TESSY, ADNS et l'Empres-i ont un avantage sur le système WAHIS ; ils offrent à l'utilisateur la possibilité de télécharger des données sur les foyers sous forme tabulaire et structurée. Les tableaux contiennent les pays avec la zone administrative, les coordonnées du foyer, la date d'observation des premiers signes cliniques, la date de la confirmation en laboratoire et la date de déclaration d'événement, les caractéristiques de l'agent pathogène, les espèces sensibles, les nombres de cas infectés, les nombres de morts, etc.

2.2.3 Communication

Les avancées récentes en matière de technologies de l'information ont transformé les systèmes SBI en outils de gestion garantissant la déclaration immédiate d'alertes pour des maladies. Les utilisateurs sont alertés par des e-mails, des flux RSS, des interfaces Web ou des applications sur des téléphones portables (WAHIS). Les alertes fournissent une information sanitaire plus ou moins complète selon le niveau de confidentialité : public (WAHIS, TESSY, GOARN) ou en accès restreint (ADNS, GOARN).

Les systèmes SBI produisent également des rapports épidémiologiques réguliers, comme : *a)* les bulletins épidémiologiques de l'OMS et d'ECDC (hebdomadaires, mensuelles, annuelles) attestant de la situation épidémiologique *i)* des maladies couvertes par le Règlement Sanitaire International (RSI) et *ii)* des maladies transmissibles importantes pour la santé publique, ou *b)* les bulletins épidémiologiques d'Empres-i sur les mesures de contrôle pour les maladies animales transfrontalières (Zeller et al., 2013).

2.2.4 Limites et enjeux

En raison de sa capacité à centraliser efficacement une multitude de données à différentes échelles spatiales et temporelles et à les traduire en informations claires et synthétiques, la SBI reste le pilier central de la surveillance épidémiologique des maladies infectieuses

connues. Néanmoins, elle s'est révélée peu efficace lorsqu'il s'agit de détecter l'émergence de maladies nouvelles (Zeller et al., 2013). Plusieurs approches cherchent à compléter les systèmes de surveillance traditionnels afin d'améliorer leur capacité à détecter de nouvelles menaces sanitaires. Plusieurs d'entre elles telles que : *i)* la surveillance sentinelle (médecins / vétérinaires volontaires inscrits au réseau) ou *ii)* la surveillance syndromique (signes cliniques) et *iii)* la « surveillance électronique » basée sur les données disponibles sur le Web, s'appuient sur la collecte systématique de données sanitaires (*e.g.*, mortalité, avortements, fièvre, etc.) qui sont ensuite compilées sous la forme de plusieurs indicateurs épidémiologiques (Amato-Gauci et al., 2008). L'utilisation de l'ensemble des sources d'information disponible demeure donc essentielle pour la détection précoce des épizooties et des risques sanitaires internationaux.

Dans le cadre de la santé publique, selon son importance (épidémiologique et spatiale), un événement sanitaire peut avoir un impact politique et économique majeur, comme l'épidémie du virus Ebola qui fait rage en Afrique de l'Ouest en 2013. La transparence des déclarations de foyers est donc cruciale. Dans le cadre de la santé animale, les conséquences des maladies ne suscitent généralement pas de telles préoccupations médiatiques (surtout au niveau international), à l'exception des zoonoses. Cependant, dans un pays jusque-là indemne, l'émergence de maladies animales infectieuses fortement contagieuses, telles que la peste porcine africaine (PPA), peut être dévastatrice pour l'économie nationale et la population animale. En effet, les populations animales ne sont pas nécessairement parées (absence de résistance naturelle ou de vaccin) pour lutter contre des maladies exotiques et peuvent être gravement affectées par l'introduction de ces pathogènes. Cela entraîne la mise en place, généralement coûteuses, de mesures de gestion de la maladie, par exemple les restrictions imposées aux déplacements ou des abattages massifs d'animaux (Amato-Gauci et al., 2008 ; Vallat et al., 2013).

Beaucoup de pays ne déclarent pas immédiatement les nouveaux foyers. Cela peut être dû au fait que dans beaucoup de pays, les services vétérinaires sont sous-dimensionnés avec peu de personnel et des capacités de laboratoire insuffisantes (Lightner, 2012). Par ailleurs, il peut être délicat pour certains laboratoires de déclarer des résultats positifs car ils peuvent être soumis à une obligation de confidentialité auprès du demandeur d'analyse (Ben Jebara, 2010 ; Vallat et al., 2013). Un exemple est l'introduction de la PPA pour la première fois dans les pays du Caucase (en Géorgie). Malgré l'observation des premiers signes cliniques de la PPA chez les porcs clandestins en avril 2007, la confirmation (et donc la déclaration) de cette maladie a eu lieu en juin 2007. Ce délai de confirmation du virus de la PPA a eu pour conséquence une épizootie majeure dans les pays de l'Europe de l'Est. Cette épizootie est le risque important d'introduction de la PPA dans les pays indemnes de l'Europe de l'Ouest (Sánchez-Vizcaíno et al., 2013).

Finalement, dans la plupart des cas, les données déclarées par les pays ne sont pas directement comparables d'un pays à l'autre car les systèmes de surveillance y sont différents. Dans le futur, il semble désormais nécessaire de se concentrer sur la pérennisation et

l'harmonisation des systèmes de surveillance, notamment en ce qui concerne la définition de cas, les stratégies d'échantillonnage, les tests diagnostiques pour confirmer des pathogènes, etc. (Ammon et al., 2010 ; Zeller et al., 2013).

2.3 Surveillance fondée sur des événements

La surveillance fondée sur des événements (veille ou SBE) est un domaine récent de l'IE. Son origine historique remonte aux années 1950, après la seconde guerre mondiale, avec la création du service d'intelligence épidémiologique du Centre pour le contrôle et la prévention des maladies aux États-Unis (CDC). Ce centre est considéré comme l'un des premiers systèmes d'alerte précoce contre des menaces sanitaires (système de biosurveillance) (Langmuir, 1980).

Depuis les années 1990, les systèmes de SBE utilisent des technologies de l'information pour rechercher en continu, sur le Web (pages Web des média, des autorités sanitaires, etc.) et les réseaux sociaux (Twitter, Facebook, blogs), des informations sanitaires pouvant servir à détecter des signaux de maladies infectieuses. C'est au milieu des années 1990 que les systèmes de SBE se sont développés, notamment en Amérique du Nord. Le système ProMED fut le premier. Développé en 1994, il a pour objectif l'échange des rapports concernant les foyers de maladies émergentes ou les événements sanitaires inhabituels par les experts du monde entier (Woodall, 2001 ; Yu et al., 2006).

En 1997, un prototype de système international de biosurveillance, GPHIN, a été développé dans le cadre du partenariat entre le gouvernement du Canada et l'OMS. L'objectif était de déterminer si l'utilisation de dépêches médiatiques du Web était envisageable et efficace pour identifier des événements inhabituels pour des maladies nouvelles et exotiques et par conséquent alerter rapidement les autorités compétentes afin de mettre en place des mesures préventives (Keller et al., 2009b ; Mykhalovskiy et al., 2006). L'efficacité du système GPHIN a été démontrée lors de l'épidémie de SRAS (« Syndrome Respiratoire Aigu Sévère ») dans le sud de la Chine en 2003 quand l'alerte a été donnée avec anticipation en s'appuyant sur des informations publiées dans des média électroniques chinois (Mykhalovskiy et al., 2006). L'émergence du SRAS et son extension mondiale (pandémique) a favorisé le développement de plusieurs systèmes de biosurveillance à l'échelle internationale (Bohigas et al., 2009).

En 2004, le Centre médical pour les maladies infectieuses de Georgetown aux États Unis, a développé le système Argus, un système de biosurveillance dont le but est d'identifier les dangers sanitaires potentiels pour les États-Unis.

La même année, à la demande de la Commission Européenne (CE), le Centre de recherche commun (JRC, « Joint research center ») a lancé MedISys, un agrégateur automatique de dépêches Avec une couverture Web internationale pour plus de cinq mille sujets différents (en santé animale, santé publique, ainsi que les menaces d'attaques chimiques,

nucléaires et bioterroristes) qui prend en compte plus de 50 langues cet agrégateur automatique est, à ce jour, l'un des plus complets qui existe (Alomar et al., 2015 ; Mantero et al., 2011 ; Steinberger et al., 2008).

Par ailleurs, en 2006, l'Université de Tokyo a proposé le projet BioCaster et l'Université de Harvard a lancé le projet HealthMap (Brownstein et al., 2008 ; Freifeld et al., 2008) – systèmes de biosurveillance fondés principalement sur la veille des flux RSS et de Twitter.

En 2011, l'Université de Melbourne a conçu le système IBIS (« International Biosecurity Intelligence System ») avec pour objectif d'identifier les menaces sanitaires pour l'Australie. Il s'agit du premier système de biosurveillance regroupant plus soixante agents pathogènes des maladies animales aquatiques et, depuis peu, plus de quarante maladies animales chez les animaux terrestres (Lyon et al., 2013b ; Lyon et al., 2013a).

Les systèmes SBE sont classés en trois catégories différentes selon leurs objectifs. Ces derniers peuvent être : *i)* de détecter de manière précoce de nouvelles émergences (HealthMap, BioCaster, GPHIN, MedISys et IBIS), *ii)* d'améliorer la communication (ProMED et EWRS) et *iii)* de compléter les systèmes déjà en place (HealthMap et IBIS qui facilitent la visualisation des rapports ProMED).

La plupart des systèmes SBE sont gérés par des Universités (Argus, BioCaster, IBIS et HealthMap), des organisations non gouvernementales, ONG (ProMED) ou des agences gouvernementales (EWRS, GOARN, GPHIN et MedISys). Les systèmes de biosurveillance gérés par des Universités ou des ONG ne sont pas soumis aux contraintes imposées par les gouvernements en ce qui concerne la diffusion d'informations. Par ailleurs, étant gérées par des professionnels, elles jouissent d'une grande crédibilité (Woodall, 2001).

Tous les systèmes fondés sur des événements traitent de nombreuses maladies infectieuses, principalement humaines et dans une moindre mesure, animales. Le système IBIS aborde la veille principalement pour des maladies d'animaux aquatiques et d'animaux terrestres. Les spécialistes de ProMed s'intéressant aux maladies nouvelles et à la propagation des épidémies à de nouvelles zones ou à de nouvelles populations, la liste des maladies couvertes par ce système de surveillance n'est pas fermée (Yu et al., 2006). MedISys rapporte aussi des informations concernant des signes cliniques neurologiques, respiratoires, gastro-entériques, éruptifs et hémorragiques chez les humains, via HealthMap qui, en plus, prend en considération des syndromes comme la mortalité et la fièvre chez les humains ou la mortalité chez les animaux.

Le sous-chapitre suivant présente les caractéristiques principales des systèmes fondés sur la SBE, également synthétisés dans le Tableau 2.2. L'Annexe C.2 présente des exemples d'interfaces de systèmes fondés sur des événements.

Tableau 2.2 – Principales caractéristiques des systèmes fondés sur la surveillance des événements (SBE)

Caractéristiques	ProMED	GPHIN	MedISys	Argus	BioCaster	HealthMap	IBIS
Année de constitution	1994	1997	2004	2004	2006	2006	2013
Couverture géographique	Mondial	Mondial	Mondial	Mondial	Mondial [2]	Mondial	Mondial[2]
Accès [3]	P	R	P, R	R	P	P	P
No. langues	5	9	50	40	13	7	1
Type de risque [4]	H, A, V	H, A, V, E	H, A, V, E	H, A, V, E	H, A, V, E	H, A, V, E	A, V
Si risque humain, no.maladies/ syndromes	Illimité	NA	140/ 5	130	182	50/ 5	/
Si risque animale [5] **no.maladies/ syndromes**	Illimité	NA	56 AT	NA	46 AT	18/ 1 AT	40 AT/ 60 AA
Modération [6]	H	SA	A	H	A	SA	SA
Sources [7]	F, IF	F, IF	F, IF	IF	IF	F, IF	IF
Acquisition de données [8]	M	A/ M	A	A/ M	A	A/ M	A/ M
Type de sources [9]	W, U	W	W	W	W	W, U, T	W, U, T
Dé-duplication des articles[6]	M	A et M	A	M et A	A	A	A
Recherches sur Web [10]	M	T, O, A	T, O, A	T, O	O	T, O, A	NA

2. Focus Asie du Sud-Est
3. P=Public, R=Restreint
4. H=humaine, A=animale, V=végétale, E=environnementale
5. AT = animaux terrestres, AA = animaux aquatiques
6. Modération = intervention humaine (analystes) dans le processus, A = Automatique, H = Humaine, SA = Semi-automatique
7. Source formelle (F)= source officielle, information validée par les autorités compétentes, Source informelle (IF)= source non-officielle, information non validée par les autorités compétentes
8. M=Manuelle, A=Automatique
9. W=Web, T=Twitter, U = Utilisateurs
10. Sélection de documents pertinentes selon une T= Terminologie, O=Ontologie, M=Manuelle, A=Autre

Caractéristiques	ProMED	GPHIN	MedISys	Argus	BioCaster	HealthMap	IBIS
Fouille de textes/ données [11]	NA	NB, SVM	NB	NB, SVM	NB	NB	NA
Évaluation du risque	Oui	Oui	Non	Oui	Non	Oui	Oui
Archive des événements	Oui	Oui	Oui	Oui	Non	Oui	NA
Carte géographique	Non	Non	Oui	Oui	Oui	Oui	Oui
Chronologie temporelle	Non	Non	Oui	Non	Oui	Oui	Non
Communication, sauf interface	Email	Email	Email, RSS	Email	Email, RSS, Twitter	Email, RSS, Twitter	Email
Références	Cowen et al., 2006 ; Woodall, 2001 ; Madoff, 2004	Mykhalovskiy et al., 2006 ; Hartley et al., 2010 ; Keller et al., 2009b	Hartley et al., 2010 ; Linge et al., 2009 ; Mantero et al., 2011	Collier et al., 2008 ; Hartley et al., 2010	Brownstein et al., 2008 ; Freifeld et al., 2008 ; Keller et al., 2009b	Lyon et al., 2013b ; Lyon et al., 2013a	Lyon et al., 2013b

11. NB = Naïve Bayes, SVM = Support Vector Machine

2.3.1 Acquisition

La plupart des systèmes SBE collectent des données de façon automatique. La collecte est principalement réalisée à partir de sources non officielles sur le Web, tels que les flux RSS, les agrégateurs de nouvelles ou les pages Web. Parmi les systèmes qui se concentrent uniquement sur ces sources non-officielles, les plus connus sont les systèmes Argus, GPHIN, IBIS, BioCaster et MedISys. D'autres systèmes, comme HealthMap, EWRS et GOARN, collectent à la fois des données issues de sources officielles et non-officielles.

Les analystes du système ProMED effectuent des recherches manuelles sur différentes pages Web afin de compléter les informations sanitaires recueilles principalement grâce à ses abonnés (on parle d'informations « de première main ») (Yu et al., 2006).

Le système IBIS utilise l'approche collaborative (« Crowd sourcing ») qui permet aux utilisateurs de partager des articles concernant des événements sanitaires ou d'analyser la pertinence de chaque article acquis. IBIS analyse également le contenu de chaque article et contribue à l'évaluation de termes automatiquement extraits : les maladies, les espèces touchées, les signes cliniques ainsi que le lieu d'événements (Lyon et al., 2013b ; Lyon et al., 2013a).

Afin d'extraire des informations pertinentes du Web, la plupart des systèmes fondés sur la SBE utilisent des combinaisons de termes et d'expressions de recherche en plusieurs langues. Les systèmes Argus et IBIS utilisent des termes de noms de maladies, d'agents pathogènes et leurs sérotypes (Nelson et al., 2010 ; Nelson et al., 2012). Les systèmes MedISys et HealthMap utilisent en plus, des termes qui décrivent des signes cliniques et des mots-clés qui caractérisent des foyers (« outbreak », « unknown disease », « case », etc.). Les termes sont proposés par des experts (systèmes GPHIN, MedISys) (Keller et al., 2009b ; Mantero et al., 2011) ou proviennent d'un dictionnaire de pathogènes de maladies (système HealthMap) (Brownstein et al., 2008) ou encore d'une ontologie médicale (projet BioCaster) (Collier et al., 2008).

Pour réduire la quantité de documents recueillis et pour améliorer l'extraction de l'information épidémiologique, les systèmes actuels utilisent des techniques de fouille de textes pour le filtrage et la classification des données. Ces techniques permettent notamment de supprimer les doublons et les documents non pertinents. Par exemple, les systèmes GPHIN et MedISys attribuent une valeur aux termes utilisés afin de les classer ; un seuil minimal est défini et seuls les documents dont la valeur attribuée est supérieure à ce seuil sont conservés (*e.g.*, le terme « dengue » apparaît au moins trois fois dans le texte ou le terme« dengue » est suivi par le terme « foyer ») (Mantero et al., 2011). Le système MedISys prend aussi en considération une liste de termes non pertinents. L'association d'un de ces termes avec un terme pertinent (*ex.*, les termes « dengue » et « concert ») a pour effet d'exclure automatiquement le document.

Enfin, les systèmes fondés sur la SBE utilisent des techniques automatiques de classification de documents fondées sur des algorithmes d'apprentissage automatique. Les algorithmes habituellement utilisés pour la classification automatique de documents en temps réel sont Naïve Bayes (NB) et les machines à vecteurs de support (SVM) (Keller et al., 2009b).

2.3.2 Analyse

L'analyse des données collectées par les systèmes fondés sur la SBE est délicate car, dans la plupart des cas, il s'agit de textes non structurés nécessitant d'être préalablement traités afin d'en extraire l'information désirée. La précision et l'authenticité de l'information, comme le lieu et la date de l'événement, le nom de la maladie (si mentionnée), les signes cliniques et le nombre de cas, constituent un autre défi du travail d'analyse.

Pour l'extraction de l'information sanitaire, la plupart des systèmes (BioCaster, MedISys, HealthMap, Argus et GPHIN) utilisent des méthodes de fouille de textes bien connues dans la littérature. Pour cette étape, le système IBIS dépend d'un outil commercial (AlchemyApi [12]). Notons que la plupart des systèmes intègrent des outils de traduction automatique pour ce placer dans un contexte multilingue.

Avant de les partager avec les utilisateurs ou avant la publication en ligne, des analystes professionnels (modérateurs humains), comme dans les systèmes Argus, GPHIN, GOARN, EWRS et ProMed-mail, HealthMap évaluent chaque article (Mykhalovskiy et al., 2006).

L'aspect collaboratif du système IBIS implique que sa communauté s'engage à réaliser deux tâches principales : *i)* l'estimation de la pertinence des articles collectés et *ii)* la vérification de la catégorie associée à ces articles (approche semi-automatique). Les systèmes BioCaster et MedISys postent automatiquement les articles qui ont été collectés et catégorisés.

2.3.3 Communication

Un des principaux moyens de diffuser l'information est sous la forme de pages Web comprenant une interface composée de cartes géographiques des lieux de foyers ou d'événements sanitaires (*e.g.*, BioCaster, IBIS ou HealthMap). HealthMap, propose aux utilisateurs la possibilité d'extraire des informations sous la forme de tableaux structurés (source, date de publication, pays, nom de la maladie, langue, lieu et coordonnées, intitulé de l'événement, nombre de cas et de morts).

12. http://www.alchemyapi.com/

Les abonnés à la version gratuite d'IBIS et de ProMED reçoivent des alertes quotidiennes par courriel comportant une interprétation des événements par des analystes professionnels (ProMED) ou un résumé de l'article acquis automatiquement et évalué comme pertinent par les collaborateurs d'IBIS. Depuis 2016, les abonnés à ProMED peuvent utiliser un lien menant à HealthMap pour visualiser géographiquement leurs articles avec des alertes.

Dû aux coûts de création et de maintenance, ou pour des raisons de confidentialité, les systèmes Argus, GPHIN, EWRS et GOARN diffusent leurs informations via des portails sécurisés ou à accès restreint ou payant (abonnement). Ces informations sont réservées à un réseau fermé d'organismes universitaires, d'Agences gouvernementales et de santé publique (Nelson et al., 2012 ; Nelson et al., 2010 ; Mykhalovskiy et al., 2006).

Enfin, deux niveaux d'accès sont offerts aux utilisateurs du système MedISys : un accès gratuit au grand public et un accès sécurisé pour les membres de la Commission Européenne (CE) et des organisations gouvernementales. Ceux-ci ont accès à plus de fonctionnalités, de catégories et de sources d'information. Le site public de MedISys sert d'agrégateur de dépêches, avec une liste d'articles classés par sujet, des cartes géographiques et des tableaux de fréquence par pays (Velasco et al., 2014).

2.3.4 Limites et enjeux

Afin de trouver des informations sur le Web, les principaux systèmes fondés sur la SBE utilisent un vocabulaire spécifique, tel que le nom des maladies et les termes liés aux signes cliniques regroupés en syndromes. Cependant, le mode d'identification des termes utilisés pour effectuer des recherches sur le Web n'est pas clair, en particulier en santé animale. Identifier un vocabulaire approprié pour les maladies infectieuses animales représente un défi en raison de l'existence de nombreux hôtes et d'un vocabulaire moins formel que chez l'Homme pour désigner les signes cliniques. En effet, un même agent pathogène peut affecter de nombreux hôtes, comme le virus de la fièvre aphteuse (bovins, ovins, caprins, porcins), avec des signes cliniques typiques (vésicules et des ulcères sur les muqueuses) ou non spécifiques (fièvre, faiblesse, diarrhée, etc.) (Santamaria et al., 2011 ; Smith-Akin et al., 2007).

L'acquisition et la catégorisation des articles sont souvent réalisées automatiquement. Par ailleurs, les articles ne sont pas toujours validés par des professionnels, ni interprétés par des épidémiologistes avant diffusion, ce qui peut remettre en cause leur fiabilité. De ce fait, la qualité de l'information est moindre que celle des informations provenant de sources officielles.

Par ailleurs, la fréquence des mises à jour n'étant pas standardisée, les articles et informations peuvent être répétés. Les systèmes fondés sur la SBE reçoivent quotidiennement un nombre considérable d'alertes, compliquant le travail d'interprétation des signaux. Les données sont souvent incomplètes et tardivement disponibles, voire obsolètes quand

elles parviennent. De plus, l'information n'est pas toujours efficacement communiquée relativement aux besoins des utilisateurs. Les interfaces sont chargées d'un nombre considérable d'événements avec des cartes géographiques ou des rapports qui sont essentiellement descriptifs et ne facilitent pas la décision. C'est la raison pour laquelle, les systèmes fondés sur la SBE ne sont pas utilisés comme des sources primaires mais comme des sources complémentaires d'informations.

Considérant ces enjeux, l'objectif principal de cette thèse est de fournir une méthode permettant de traiter les données non structurées collectées sur le Web. Le traitement des données non-structurées est une problématique ouverte nécessitant des approches originales de traitement automatique du langage naturel (TALN) combinant des méthodes linguistiques, statistiques et sémantiques. Plus précisément, dans les chapitres suivants nous décrivons le processus d'acquisition automatique des documents du Web, la classification des documents acquis et l'extraction de l'information sanitaire. Enfin, nous évaluons l'ensemble de ces étapes via un outil de veille sanitaire sur le Web, développé au cours de cette thèse, nommé PADI-web (« Platform for automated extraction of disease information from the Web »).

Chapitre 3

Processus de fouille de textes pour la veille sanitaire internationale

3.1 Approche proposée

L'approche de fouille de textes et les moyens de communication des résultats proposés dans ce manuscrit (Figure 3.1) sont appliqués au dispositif de la VSI en santé animale. L'approche s'articule autour de trois étapes principales : l'acquisition de documents publiés sur le Web, la catégorisation des documents et enfin l'extraction automatique de l'information sanitaire.

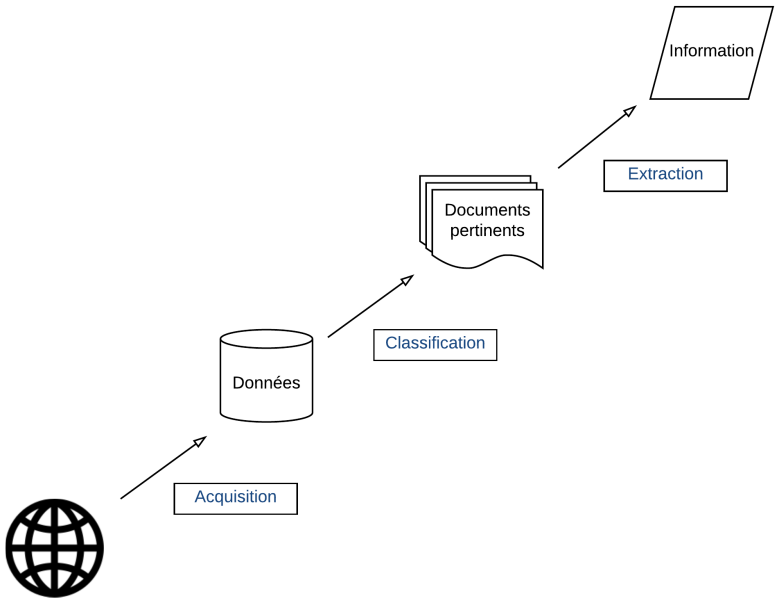


Figure 3.1 – Processus de fouille de textes pour la veille sanitaire sur le Web

Première étape : l'acquisition automatique des documents du Web (section 3.2). Cette étape est fondée sur des requêtes (recherche par mots-clés). Les termes permettant de constituer les requêtes sont extraits automatiquement à partir d'un corpus d'articles jugés pertinents pour la maladie ciblée.

Deuxième étape : la classification automatique de documents selon leur contenu (section 3.3). Deux catégories d'articles sont sélectionnés et classés. La première catégorie « nouveaux cas », est constituée de documents pertinents qui décrivent des événements sanitaires liés à l'apparition de maladies exotiques (*e.g.*, « The veterinary services confirmed yesterday an outbreak of African swine fever in a pig farm near Lusaka ») ou d'événements inhabituels dus à l'apparition de maladies émergentes (*e.g.*, « Since last monday, an increased death of wild bors has been observed by hunters in Alytus county »). La seconde catégorie est constituée d'articles non pertinents, elle regroupe elle-même deux sous-catégories. La sous-catégorie « bilan » regroupe les documents qui décrivent l'impact économique de l'apparition de foyers (*e.g.*, « Due to a recent outbreak of African swine fever in Russia, Poland has imposed a ban for introduction of live pigs and products thereof »). La sous-catégorie « général », regroupe les documents qui décrivent, de manière générale, des maladies (*e.g.*, « African swine fever is a highly contagious infectious disease in domestic and wild pigs, enzootic in sub-saharan Africa, eastern Europe and the island of Sardinia in Italy »). Étant donné à la quantité importante d'articles acquis, nous utilisons des techniques d'apprentissage automatique (« machine learning ») pour la classification des données.

Troisième étape : l'extraction automatique de l'information épidémiologique à partir de documents pertinents (section 3.4). L'objectif de cette étape est d'identifier et d'extraire, dans les documents pertinents, les informations concernant des événements sanitaires émergents : le nom de la maladie, si celui-ci est mentionné dans les données textuelles, la date et le lieu des foyers, les signes cliniques, les espèces affectées, etc.

Notre approche de fouille de textes est développée de manière générique. Toutefois, pour évaluer sa pertinence sur des cas spécifiques, quatre maladies modèles ont été sélectionnées en 2013, en collaboration avec la Cellule d'animation de la VSI (CA VSI) en France. Une priorité a été donnée aux maladies animales infectieuses considérées comme exotiques pour la France (ANSES, 2012 ; MAAF, 2013). Il s'agit des maladies ayant des conséquences importantes sur la santé animale et l'économie du pays et dont le risque d'introduction en France est particulièrement élevé. Les critères de choix supplémentaires de ces maladies modèles sont décrits ci-dessous :

— l'émergence au niveau européen : cas de la **peste porcine africaine** (PPA), une maladie épizootique en Europe de l'Est et aux pays Baltes ;
— l'échange commercial important entre la France et les pays d'Afrique et d'Asie où la **fièvre aphteuse** (FA) est enzootique ;
— la circulation dans les pays voisins : cas de la **fièvre catarrhale ovine** (FCO) avec des sérotypes du virus qui ne sont pas présents en France et circulent actuellement en Italie, en Espagne et dans l'Europe de l'Est ;
— une maladie nouvelle : cas de **Schmallenberg** (SBV), découverte en 2011 en Allemagne.

3.1.1 Maladies modèles

Les caractéristiques principales des quatre maladies modèles sont décrites dans la partie suivante ainsi qu'en Annexe D.

3.1.1.1 Peste porcine africaine

La peste porcine africaine (PPA) est une maladie virale, hautement contagieuse. Elle affecte les espèces de la famille des suidés (porcs, phacochères et sangliers), toutes les classes d'âge y sont sensibles. La maladie se caractérise par une forte fièvre, une perte d'appétit, des hémorragies au niveau de la peau et des organes internes. Les taux de mortalité engendrés par un foyer de PPR peuvent atteindre 100% dans les formes aiguë. En raison des lourdes conséquences associées à cette maladie, la détection d'un cas doit obligatoirement être signalée à l'OIE.

La PPA est enzootique dans les pays de l'Afrique sub-saharienne. Éradiquée de l'Europe de l'Ouest depuis la fin des années 1990, à l'exception d'une forme enzootique en Sardaigne, la PPA a été ré-introduite en Géorgie en juin 2007 (Figure 3.2). La souche introduite ayant été identifiée comme étant proche de celles connues en Afrique de l'Est et à Madagascar, l'hypothèse d'un déchargement de quartiers de porcs contaminés par un bateau de croisière est la plus probable pour expliquer l'introduction du virus sur le continent Eurasien. Au gré du commerce des porcs et de la viande de porc, le virus s'est rapidement propagé dans les populations domestiques et les populations de suidés sauvages (les sangliers). Tour à tour, l'Azerbaïdjan et la Russie ont été touchés au cours de l'année 2008, aussi bien au niveau de la faune sauvage que domestique. En 2009, malgré les interventions des autorités russes, le virus a été détecté sur des porcs aux portes de l'Europe : au Nord de la région de Saint-Pétersbourg près de la frontière entre l'Estonie et la Finlande. En 2012, l'Ukraine et la Biélorussie déclaraient leurs premiers foyers et en 2014, le virus fut introduit pour une première fois dans les pays Baltes (Lettonie, Lituanie, Estonie et la Pologne) (Le Potier et al., 2015).

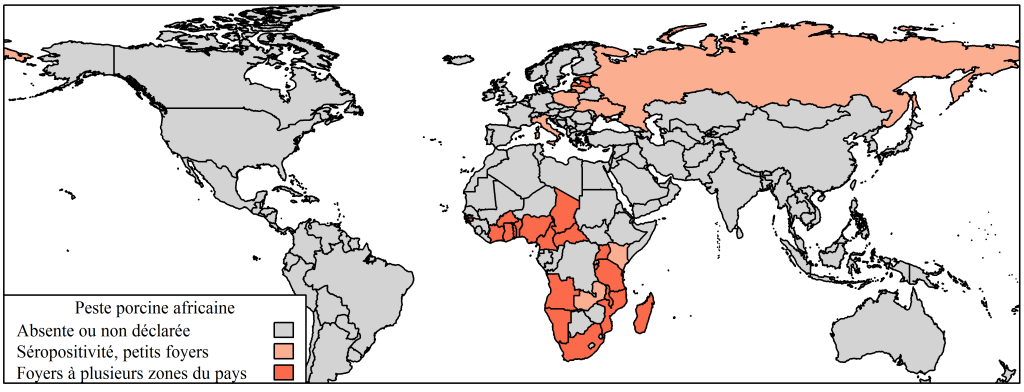


Figure 3.2 – Distribution de la peste porcine africaine au niveau mondial (2014-2016, au 1er mai 2016)

3.1.1.2 Fièvre aphteuse

La fièvre aphteuse (FA) est une maladie virale des artiodactyles. Hautement contagieuse, elle affecte principalement les bovins, les porcins et les caprins. Elle se caractérise par une hyperthermie et provoque des lésions nasales, buccales, pédales et mammaires qui débutent par des vésicules (Rodeia, 2008). Dans une population sensible, l'introduction du virus provoque des taux de morbidité proches de 100%. La maladie est rarement fatale chez les animaux adultes mais les taux de mortalité chez les jeunes sont souvent élevés. Cette maladie doit, elle aussi, obligatoirement être déclarée auprès de l'OIE.

La FA est enzootique dans des régions de l'Asie, de l'Afrique et du Moyen-Orient (Figure 3.3). En Europe, la maladie étant enzootique en Turquie, ce pays a été considéré comme une source des incursions sporadiques de FA en Europe, notamment en Grèce et en Bulgarie (Alexandrov et al., 2013 ; Saeed et al., 2015). Notamment en raison de la localisation des foyers en zones forestières, l'hypothèse la plus probable pour expliquer ces introductions est celle d'une contamination des élevages par des sangliers infectés venant de la Thrace de Turquie. Durant les dernières décennies, la Turquie a intensifié le contrôle de la FA, incluant la vaccination et le contrôle des mouvements des animaux. Cependant, le commerce intensif et les mouvements illégaux entre la Turquie et les pays voisins compliquent l'éradication du virus (Alexandrov et al., 2013 ; Rodeia, 2008).

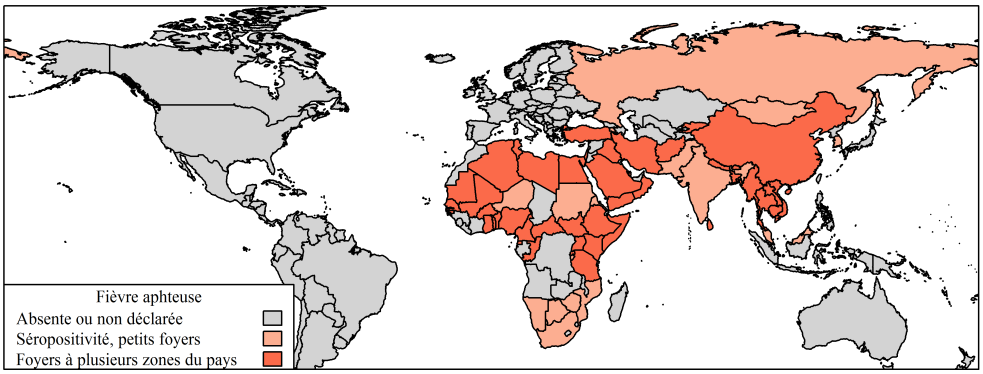


Figure 3.3 – Distribution de la fièvre aphteuse au niveau mondial (2014-2016, au 1er mai 2016)

3.1.1.3 Fièvre catarrhale ovine

La fièvre catarrhale ovine (FCO) ou maladie de la langue bleue (Bluetongue, BT) est une maladie virale, non contagieuse (à transmission vectorielle). Elle affecte les ruminants, essentiellement les ovins mais aussi les bovins et les caprins. L'infection se transmet par certains espèces d'insectes vecteurs du genre *Culicoides*. Les symptômes les plus graves touchent les ovins, où les taux de morbidité peuvent atteindre 100% et ceux de mortalité entre 2 et 30% (pouvant aller jusqu'à 70%). Cette maladie est, elle aussi, à déclaration obligatoire à l'OIE.

La distribution de la FCO est similaire à celle des vecteurs, elle est ainsi largement distribuée en Afrique, en Asie, en Australie, en Europe, en Amérique du Nord et sur plusieurs îles des zones tropicales et subtropicales (Figure 3.4). Le virus persiste dans les secteurs où le climat permet aux vecteurs de survivre à l'hiver (Wilson et al., 2009).

En Europe, la circulation active de différents sérotypes de la FCO (notamment BTV-1 et BTV-4) est régulièrement déclarée dans le sud d'Espagne et en Italie. Jusqu'à présent, il semblerait que le virus de la FCO n'ait pas passé les frontières françaises que ce soit depuis l'Espagne ou depuis l'Italie continentale. Cependant, des sauts d'introduction de l'infection liés à des mouvements d'animaux sont toujours possibles. Depuis 2014, il y a une ré-émergence du BTV-4 en Europe de l'Est (Arsevska et al., 2014a). La dernière grande épizootie dans cette région était en 2006 (sérotype BTV-8). Il n'existe pas *a priori* de barrière physique capable de limiter la progression du BTV-4 vers le Nord-Est de l'Europe. Les espèces de *Culicoides* de l'ensemble *Obsoletus* étant certainement présentes en abondance dans ces zones, le contrôle de la progression vers le Nord-Ouest dépend principalement de l'efficacité des mesures de lutte.

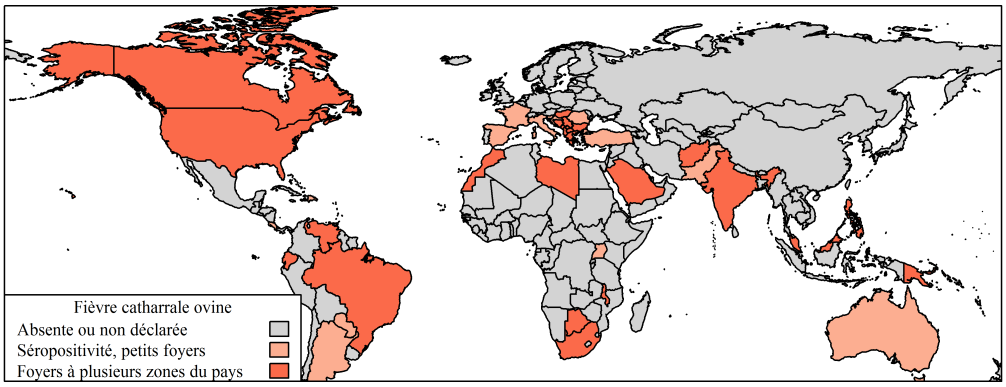


Figure 3.4 – Distribution de la fièvre catarrhale ovine au niveau mondial (2014-2016, au 1er mai 2016)

3.1.1.4 Schmallenberg

La maladie de Schmallenberg (SBV) est une maladie virale, non contagieuse (à transmission vectorielle). Elle affecte les ruminants, essentiellement les bovins mais aussi les ovins et les caprins. La SBV se manifeste, chez le bovin adulte, par une baisse de la production laitière, de la fièvre, une diarrhée et des avortements. Une affection congénitale est également décrite chez les agneaux, les veaux et les chevreaux, caractérisée par des malformations de type arthrogrypose/ hydranencéphalie. Il n'est pas obligatoire de déclarer cette maladie à l'OIE.

La première description de Schmallenberg eut lieu en août 2011, en Allemagne, suivie par des signalement dans de nombreux pays d'Europe (Belgique, France, Danemark, Luxembourg, Pays Bas, Grande Bretagne, Espagne et Italie) (Figure 3.5). Ces pays ont observé une fréquence élevée de diarrhée aqueuse associée à une hyperthermie transitoire,

une chute de production laitière significative (jusqu'à 50%), des avortements et malformations congénitales. La maladie a ré-émergé, en 2012, en France, en Allemagne et en Grande-Bretagne, ainsi que dans les pays nordiques et l'Europe centrale. En 2013, des séroconversions ont été signalées dans plusieurs pays d'Europe de l'Est (Conraths et al., 2013a ; Conraths et al., 2013b ; Wernike et al., 2013). Peu d'informations sont disponibles au sujet des foyers de SBV ayant eu lieu à partir de 2014.

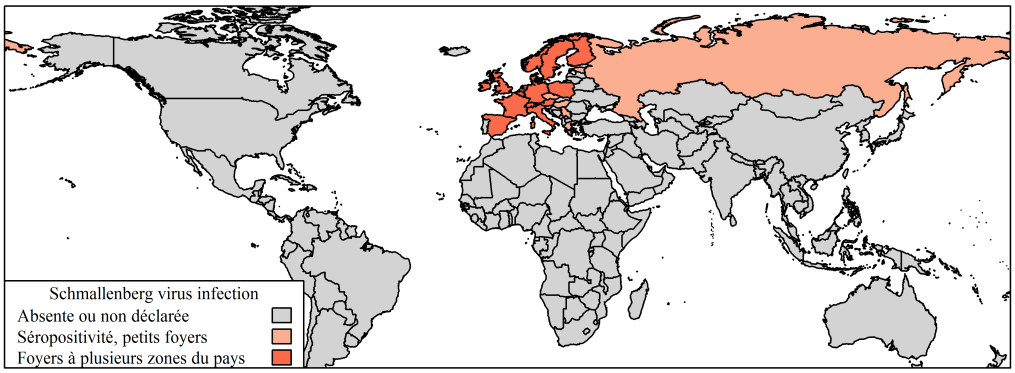


Figure 3.5 – Distribution de l'infection avec le virus de Schmallenberg au niveau mondial (2014-2016, au 1er mai 2016)

Les maladies modèles s'appuient toutes sur des caractéristiques du même type : informations spatiales, temporelles, signes cliniques et hôtes, ce que notre approche générique permet de traiter. Seulement leur instantiation change de manière très différente selon les maladies (lieux différents d'apparition des foyers, période temporelle variable, etc.). Ces différentes instantiations sont contenues dans les corpus décrits dans la sous-section qui suit (sous-section 3.1.2). Le principal défi de cette thèse était d'identifier ces caractéristiques de manière automatique dans les textes non structurés (sous-section 3.2.1), afin de fournir une analyse épidémiologique de la situation sanitaire associée aux maladies exotiques et émergentes au niveau international (section 3.5).

3.1.2 Corpus pour des maladies modèles

Pour la mise en place du processus de fouille de textes et les évaluations associées nous avons rassemblé un ensemble de données textuelles (documents) appelé « corpus ». Le premier corpus est constitué d'articles de média. Le second est composé de résumés scientifiques.

Nous avons collecté ce corpus manuellement en août et en septembre 2014 à partir du moteur de recherche Google et de la base de données de littérature biomédicale PubMed. Le moteur de recherche Google est exploité par plusieurs systèmes de biosurveillance du Web (Hartley et al., 2010 ; Keller et al., 2009b). La base de données PubMed est la plus grande base de données électronique en ce qui concerne la littérature biomédicale. PubMed est mise à jour régulièrement et contient plus de 24 millions d'articles scientifiques (Falagas et al., 2008).

Pour l'acquisition des documents, nous avons effectué des requêtes sur Google et PubMed en anglais, en utilisant la combinaison : « nom de la maladie » et « outbreak ». Par exemple, pour acquérir des données au sujet de la PPA, nous avons utilisé la requête : « african swine fever » AND « outbreak ».

Les articles de média ont été validés et catégorisés selon leur contenu comme :

— Catégorie **« nouveaux cas »** (pertinent) pour les articles décrivant l'apparition des foyers pour chacune des maladies modèles. Par exemple, pour la PPA, les articles contenant des informations cruciales à propos de la présence soupçonnée ou bien confirmée de la maladie ou de signes cliniques inexpliqués chez le suidé (date, lieu, espèces affectées, nombre de cas ou de foyers, etc.). Au total, 181 articles pour la PPA, 84 pour la FA, 92 pour la FCO et 148 pour le SBV constituent ce corpus.
— Catégorie **« bilan »** (non pertinent) pour les articles décrivant un bilan ou l'impact économique d'un foyer pour un pays ou alors quand l'information concernant des foyers est secondaire. Au total, ce corpus est fondée de 92 articles pour la PPA, 44 pour la FA, 19 pour la FCO et 23 pour le SBV.
— Catégorie **« général »** (non pertinent) pour les articles ne décrivant que des généralités au sujet des maladies modèles. Au total, 272 articles pour la PPA, 147 pour la FA, 109 pour la FCO et 181 pour le SBV constituent ce corpus.

Le contenu des résumés scientifiques a été validé et catégorisé selon leur contenu comme :

— Catégorie « **épidémiologie** » (pertinent) pour des résumés indexés avec le terme « épidémiologie » dans le thésaurus MeSH (« Medical Subject Heading ») de PubMed. Quand le résumé n'est pas indexé, il es évalué selon le contenu épidémiologique. Au total, ce corpus est fondée de 45 résumés pour la PPA, 143 pour la FA, 116 pour la FCO et 53 pour le SBV.
— Tous les autres résumés ont été considérés comme non pertinents. Au total, 73 résumés pour la PPA, 269 pour la FA, 373 pour la FCO et 48 pour le SBV composent ce corpus.

Tous les documents pertinents traitent de sujets en rapport avec *i)* les foyers de la PPA en Europe de l'Est ou Afrique sub-saharienne, *ii)* les foyers de la FA en Afrique du Nord ou *iii)* les foyers de la FCO dans les Balkans, sur une période donnée entre 2011 et 2014.

Dans les sections suivantes, nous présenterons les trois étapes principales de notre approche de fouille de textes pour la VSI, ainsi que les résultats des évaluations de ces étapes. Le corpus de documents pour des maladies modèles (sous-section 3.1.2) sera utilisé pour plusieurs de ces évaluations, telles que l'extraction de terminologie et la classification automatique des documents (sections 3.2.1 et 3.3).

L'une de nos contributions principales est de pouvoir traiter les données non structurées collectées à partir du Web. Le traitement des données non structurées est une problématique ouverte nécessitant des approches de traitement automatique du langage naturel (TALN) originales combinant des méthodes linguistiques, statistiques et sémantiques.

Plus précisément, dans les sections suivantes nous décrivons le processus d'acquisition automatique des documents du Web, la classification des documents collectés et l'extraction d'information sanitaire. Enfin, nous évaluons l'ensemble de ces étapes *via* l'outil PADI-web, développé dans le cadre de cette thèse.

3.2 Acquisition automatique des données

Cette partie du manuscrit présente la première étape de notre approche de fouille de textes pour la veille internationale en santé animale sur le Web (Figure 3.6) : l'acquisition automatique des données. Cette étape facilite la construction de la terminologie spécialisée pour des recherches automatiques sur le Web. Ceci permettra une veille du Web pour des maladies animales infectieuses exotiques et émergentes.

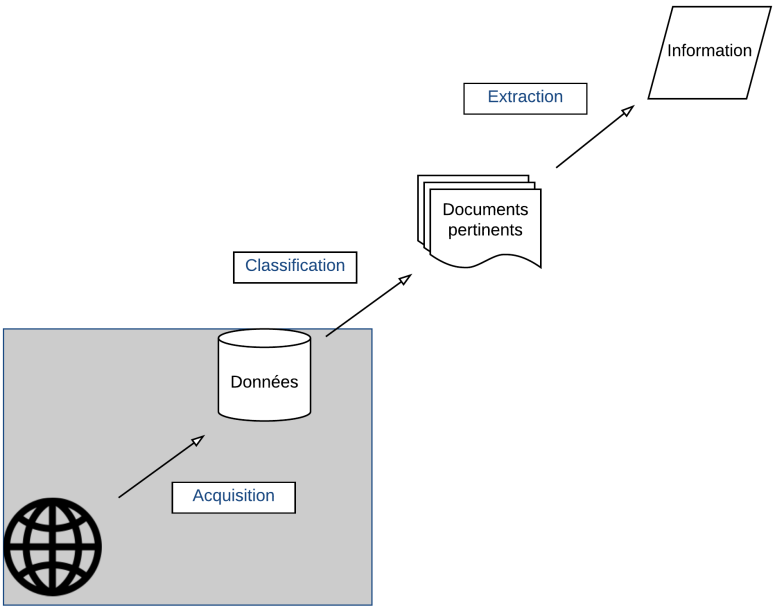


Figure 3.6 – Focus sur la première étape du processus de fouille de textes pour la veille sanitaire sur le Web

Afin de trouver des informations sur le Web, les systèmes fondés sur la SBE utilisent un vocabulaire spécifique, tel que le nom des maladies et les termes liés aux signes cliniques regroupés en syndromes. Par exemple, le système Argus utilise des termes de noms de maladies, d'agents pathogènes et leurs sérotypes (Nelson et al., 2010 ; Nelson et al., 2012). Les systèmes GPHIN, MedISys et HealthMap utilisent en plus, des termes qui caractérisent des signes cliniques. Les termes sont proposés par des experts (pour GPHIN et MedISys) (Keller et al., 2009b ; Mantero et al., 2011), sont obtenus d'un dictionnaire des pathogènes (pour HealthMap) (Brownstein et al., 2008) ou d'une ontologie médicale, comme dans le cadre du projet BioCaster (Collier et al., 2008).

Cependant, le mode d'identification des termes utilisé pour effectuer des recherches sur le Web n'est pas clair, en particulier en santé animale. Ainsi, le vocabulaire pour désigner des signes cliniques est moins formel que chez l'Homme. En effet, un même agent pathogène peut affecter de nombreux hôtes, comme le virus de la FA (bovins, ovins, caprins, porcins), avec des signes cliniques spécifiques (vésicules et des ulcères sur les muqueuses), mais également non spécifiques (fièvre, faiblesse, diarrhée, etc.) (Santamaria et al., 2011 ; Smith-Akin et al., 2007).

Comparé aux autres systèmes fondés sur la SBE, l'originalité de notre approche est l'élaboration du processus de construction de la terminologie pour des recherches automatiques sur le Web. Cette terminologie est obtenue à l'aide d'approches de fouille de textes (sous-section 3.2.1.1) appliquées sur un corpus de documents jugés pertinents (sous-section 3.1.2). La terminologie est complétée par des propositions d'experts du domaine (sous-section 3.2.1.4.2). Des approches de fouille de textes sont également utilisées pour identifier les meilleures associations entre différents termes (*e.g.,* hôtes et signes cliniques) (sous-section 3.2.2). La Figure 3.7 résume cette partie de notre approche.

Dans ce contexte, le travail présenté dans cette sous-section a donné lieu aux deux articles scientifiques publiés :

1. Nouvelle mesure pour le classement automatique de termes extraits avec une approche de fouille de textes à partir de documents pertinents (Article 1 de l'Annexe A.1, Arsevska et al., 2016d).

2. Nouvelles mesures pour l'association automatique entre les termes hôtes et termes relatifs aux signes cliniques pour assurer une veille syndromique (Article 2 de l'Annexe A.2, Arsevska et al., 2016c).

3.2.1 Extraction automatique des termes et nouvelle mesure de classement de termes

Les méthodes d'extraction automatique de termes (EAT) visent à extraire automatiquement des termes de spécialité à partir d'un corpus. Ces méthodes sont essentielles pour l'acquisition des connaissances d'un domaine, par exemple, pour la mise à jour de lexiques. Les méthodes d'EAT peuvent alors être appliquées aux documents sanitaires sur le Web (articles de média et résumés scientifiques). L'EAT à partir de ces documents est alors utile pour mieux comprendre les caractéristiques épidémiologiques des épizooties des maladies animales infectieuses (Figure 3.7).

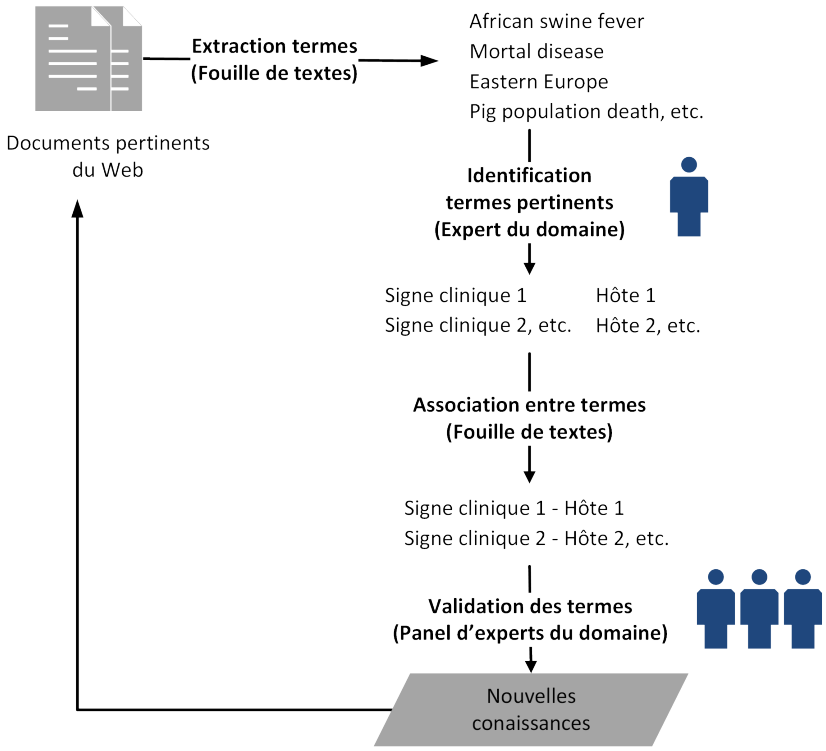


Figure 3.7 – Processus d'identification des termes pour la veille sanitaire sur le Web

3.2.1.1 Extraction des termes avec le processus de fouille de textes

Dans un tel contexte, les données textuelles issues d'articles de média et des résumés scientifiques recèlent des informations précieuses que des méthodes de fouille de textes peuvent mettre en exergue, en particulier dans le domaine de l'agriculture (Roche et al., 2015).

Les processus de fouille de textes sont souvent composés de deux phases successives. Dans un premier temps, ces méthodes consistent à extraire les descripteurs linguistiques les plus significatifs à partir de documents. Les descripteurs linguistiques peuvent être des mots (*e.g.*, « pig »), généralement plus simples à extraire, mais aussi des termes composés, plus difficiles à identifier (*e.g.*, « high mortality »). Ces derniers représentent le matériau de base afin d'associer une certaine sémantique aux documents. Par exemple, les termes « pig » et « high mortality » présents dans un document mettent en lumière une thématique liée à l'émergence des pestes porcines ou une autre pathologie importante chez les suidés.

La deuxième phase du processus consiste à exploiter ces termes pour, par exemple, classer automatiquement les documents dans des catégories (*e.g.*, documents qui décrivent des foyers de maladies, cf. sous-section 3.1.2). Cette classification repose sur le postulat suivant : si des documents possèdent de nombreux termes en commun alors ils peuvent être regroupés.

Afin d'identifier une terminologie adaptée pour des recherches automatiques sur le Web, nous avons extrait des termes candidats à partir de deux corpus pour chaque maladie modèle (sous-section 3.1.2). Le premier corpus est constitué d'articles de média (catégorie « nouveaux cas »). Le deuxième corpus est composé de résumés scientifiques (catégorie « épidémiologie »).

Pour l'extraction automatique des termes, nous avons utilisé et étendu certaines fonctionnalités de *BioTex*, un outil qui combine information statistique et linguistique pour sélectionner et classer les termes à partir des textes issus du domaine biomédical (Lossio-Ventura et al., 2014 ; Lossio-Ventura et al., 2016).

Les informations statistiques apportent une pondération des termes candidats extraits, telles que la fréquence des termes (*e.g.*, « sanglier » ou « mortalité ») qui sont présents dans plusieurs documents pour la PPA. Par ailleurs, pour effectuer une sélection des termes biomédicaux, *BioTex* utilise sur des mesures de discréminance et d'autres méthodes de pondérations qui calculent, par exemple, la dépendance des mots composant les termes complexes. Pour sélectionner les termes candidats, *BioTex* s'appuie, en entre autre, sur la mesure TF-IDF (« Term Frequency-Inverse Document Frequency »). Cette mesure donne un poids plus important aux termes spécifiques d'un document (Salton, 1983). Ainsi, pour attribuer un poids de TF-IDF, il est nécessaire, dans un premier temps, de calculer la fréquence d'un terme, TF (« Term Frequency »). Puis, la fréquence inverse de document, IDF (« Inverse Document Frequency ») mesure l'importance du terme dans l'ensemble du corpus. Elle consiste à calculer le logarithme de l'inverse de la proportion de documents du corpus qui contiennent le terme.

BioTex prend en compte deux facteurs pour extraire la terminologie. Tout d'abord, l'approche extrait des termes selon des patrons syntaxiques définis (nom-adjectif, adjectif-nom, nom-préposition-nom, etc.). Après un tel filtrage linguistique, un autre filtrage statistique est appliqué. Celui-ci mesure l'association entre les mots des termes composés (par exemple, « pig holding high mortality ») en utilisant une mesure appelée C-value (Frantzi et al., 2000) tout en intégrant la pondération TF-IDF. Le but de C-value est d'améliorer l'extraction des termes composés (expressions) particulièrement adaptés pour les domaines de spécialité (*e.g.*, agriculture, santé animale).

Notons que le logiciel *BioTex* propose deux types d'extraction (1200 termes extraits au maximum en utilisant la version en ligne) : *i)* termes composés uniquement, et *ii)* termes simples et composés. Ces termes sélectionnés sont ceux obtenant les meilleures pondérations statistiques selon les mesures précédemment présentées et détaillées ailleurs (Lossio-Ventura et al., 2016).

3.2.1.2 Nouvelle mesure de classement de termes

Une des étapes clés de notre approche a consisté à proposer une nouvelle mesure pour classer les termes extraits avec *BioTex* à partir des documents pertinents (sous-section

3.1.2). Le but de la nouvelle mesure de classement est d'aider les épidémiologistes dans la sélection des termes pertinents, qui peuvent ensuite être utilisés pour des recherches automatiques sur le Web.

Plus précisément, notre nouvelle mesure *w(t)* (formule (3.1)), prend en considération : *i)* le rang initialement assigné par *BioTex* à chaque terme, et *ii)* la qualité des sources du Web. Avec cette mesure, la valeur de *w(t)* prend en compte le rang des termes donné par *BioTex* et la qualité des sources (avec une pondération associée).

Dans la formule 3.1 propre à la mesure *w(t)*, *t* représente le terme, S_i la source Web, $S_i(t)$ est le rang de *BioTex* pour le terme *t* d'une source Web S_i et α_i est le poids attribué pour une source Web S_i.

$$w(t) = \sum \alpha_i \times \frac{1}{rank_{S_{i(t)}}} \text{ avec } \alpha_i \in [0,1] \text{ et } \sum \alpha_i = 1 \tag{3.1}$$

Par exemple, pour la PPA, à partir de 2400 termes extraits avec *BioTex*, nous avons retenu 135 termes comme des termes de référence, c'est-à-dire pertinents pour caractériser l'émergence de la PPA (Figure 3.8 et Annexe E.1). Au total, 77 termes ont été identifiés comme pertinents de la source S_1 (PubMed) et 58 termes de la source S_2 (Google). Afin de respecter cette proportion de pertinence pour la mesure *w(t)*, nous avons appliqué les poids suivants : $\alpha_1 = 0,57$ et $\alpha_2 = 0,43$. La valeur de ces poids qui peut dépendre des maladies est discutée plus loin dans cette section du manuscrit.

Ainsi, si un terme est présent dans les listes obtenues avec les deux sources, nous appliquons une pondération spécifique. Par exemple, le terme « wild boar » est automatiquement classé à la première position à partir de la source PubMed et en troisième position à partir des documents issus de la source Google. Pour ce terme *t*, nous obtenons une valeur de *w(t)* de 0,713, calculée comme suit : $((1/1) \times 0,57) + ((1/3) \times 0,43)$ (Tableau 3.1).

Ainsi, si un terme est présent dans une des listes obtenues avec les deux sources, nous appliquons une autre pondération spécifique. Par exemple, le terme « asfv introduction » est automatiquement classé en quatrième position à partir de la source PubMed et à un rang très éloigne avec les documents issus de Google (rang 1201). Pour ce terme *t* nous obtenons une valeur de *w(t)* de 0,143, calculée comme suit : $((1/4) \times 0,57) + ((1/1201) \times 0,43)$ (Tableau 3.1).

Tableau 3.1 – Reclassement des termes extraits avec *BioTex* selon la mesure *w(t)*. L'exemple concerne la peste porcine africaine

Rang ***BioTex***	**Terme extrait**	**Pertinence du terme**	**Source**	**Mesure** ***w(t)***
1	wild boar	pertinent - hôte	Google	0,713
3	wild boar	pertinent - hôte	PubMed	0,713
1	wild boars	pertinent - hôte	PubMed	0,620
3	wild boars	pertinent - hôte	Google	0,620
2	european union	non pertinent	PubMed	0,285
2	mr speaker	non pertinent	Google	0,215
4	asfv introduction	pertinent - maladie	PubMed	0,143

Nous avons appliqué le même principe pour les trois autres maladies modèles. Pour la FA, 60 termes ont été sélectionnés comme étant pertinents, dont 50 issus de PubMed ($\alpha_1 = 0,83$) et 10 de Google ($\alpha_2 = 0,17$). Pour la FCO, 115 termes ont été sélectionnés comme étant pertinents, dont 87 de PubMed ($\alpha_1 = 0,76$) et 22 de Google ($\alpha_2 = 0,24$). Pour le SBV, 141 termes ont été sélectionnés comme étant pertinents, dont 124 de PubMed ($\alpha_1 = 0,88$) et 17 de Google ($\alpha_2 = 0,12$). Notons que pour toutes les maladies, la proportion de termes pertinents issus de PubMed est plus élevée que pour les termes issus de Google.

Si nous appliquons une pondération générique pour toutes les maladies modèles, il s'avère que pour la mesure *w(t)* nous pouvons affecter les poids suivants : $\alpha_1 = 0,76$ pour les termes issus de la source PubMed (moyenne des valeurs de α_1 des sources S_1 pour les n maladies) et $\alpha_2 = 0,24$ pour les termes issus de la source Google (moyenne des valeurs de α_2 des sources S_2 pour les n maladies).

La Figure 3.8 présente le nombre et les principales caractéristiques des termes retenus comme des termes de référence pour les quatre maladies modèles. Tous les termes retenus (à partir des sources Google et PubMed) sont présentés en Annexe E de ce manuscrit.

Pour la PPA 58 termes (43%) décrivent le nom de la maladie et 18 termes (13%) la diagnose différentielle, 28 termes (21%) l'hôte, 26 termes (19%) des signes cliniques généraux, 4 termes (3%) des signes hémorragiques, 1 terme (1%) un signe pathologique.

Pour la FA, 38 termes (63%) décrivent le nom de la maladie, 15 termes (25%) l'hôte, 5 termes (8%) des signes cliniques cutanés/ muqueux et 2 termes (3%) des signes généraux.

Pour la FCO, 68 termes (59%) décrivent le nom de la maladie, 22 termes (19%) l'hôte, 14 termes (12%) des signes pathologiques, 6 termes (5%) des signes généraux, 3 termes (3%) des signes reproductifs et 2 termes (2%) des signes hémorragiques.

Pour le SBV, 60 termes (39%) décrivent le nom de la maladie, 48 termes (32%) l'hôte, 17 termes des signes cliniques périnatales (11%), 9 termes (6%) des signes reproductifs, 5 termes (3%) des signes généraux, 4 termes (3%) des signes nerveux et locomoteurs, 2 termes (1%) des signes digestifs et un terme (1%) un signe respiratoire.

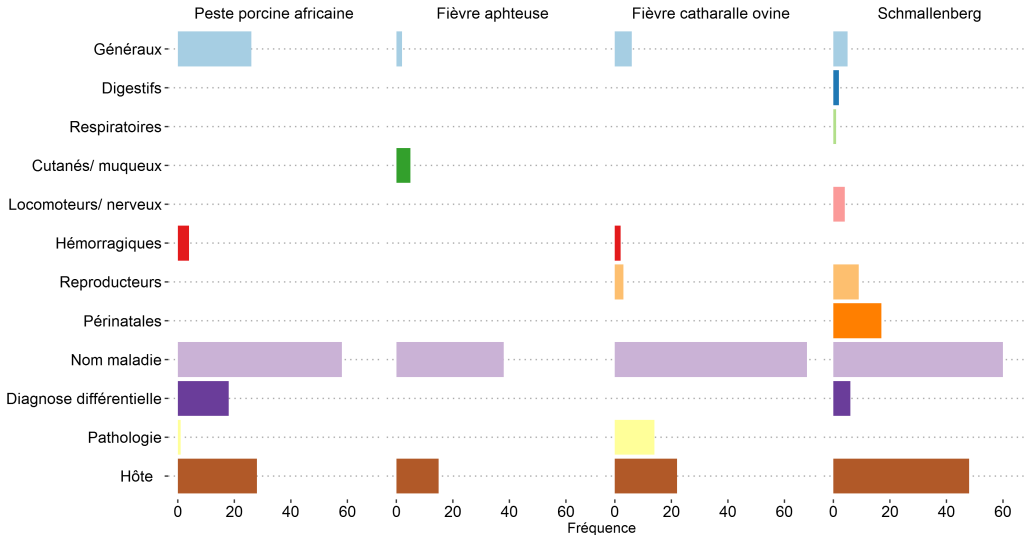


Figure 3.8 – Caractéristiques des termes jugés pertinents pour caractériser les maladies modèles. Les termes sont obtenus avec un processus de fouille de textes

3.2.1.3 Évaluation et protocole expérimental

Parmi les méthodes statistiques que nous appliquons pour évaluer la performance de la nouvelle mesure $w(t)$; le cœur de notre évaluation est fondé sur la contribution d'un panel d'experts du domaine à l'aide d'une méthode Delphi. Nous rassemblons des avis d'experts sur la pertinence des termes extraites avec le processus de fouille de textes. Nous avons pour but de mettre en évidence des convergences et des consensus sur les termes les mieux adaptés pour des recherches sur le Web afin de détecter des signaux de maladies infectieuses exotiques ou nouvelles.

La méthode Delphi apporte un éclairage des experts en vue d'une aide à la décision et d'une vérification de l'opportunité et de la faisabilité de notre approche de fouille de textes. Cette méthode nous permettra également d'identifier les différentes contraintes des approches de fouille de textes. Nous sollicitions un panel d'experts ayant une bonne connaissance pratique (les éleveurs, les chasseurs), scientifique (virologistes, épidémiologistes) ou administrative (gestionnaires de risques) pour nos maladies modèles, ayant une légitimité suffisante pour exprimer un avis représentatif du groupe d'acteurs auquel elle appartient. Finalement, ces différents acteurs, sont tous concernés par des nouvelles émergences ; leur rôle est complémentaire au processus de fouille de textes.

3.2.1.3.1 Performance de la nouvelle mesure de classement de termes. Pour évaluer la performance de la nouvelle mesure de classement $w(t)$, nous avons évalué sa capacité à classer les termes pertinents en début de liste.

Dans ce cadre nous avons effectué une analyse des courbes ROC (« Receiver Operating Characteristics »). Les évaluations ont été réalisées pour les 2400 termes (1200 termes par source) reclassées selon la nouvelle mesure $w(t)$ pour chacune des maladies modèles. Ces

analyses ont permis d'avoir une vue globale de la performance de la mesure $w(t)$ pour sélectionner en priorité les termes pertinents du point de vue épidémiologique.

La notion de courbe ROC est initialement issue du traitement du signal. Les courbes ROC sont couramment utilisées dans le domaine de la médecine pour évaluer la validité des tests diagnostiques. Les courbes ROC présentent en abscisse le taux de faux positifs (dans notre cas, taux de termes non pertinents) et en ordonnée le taux de vrais positifs (taux de termes pertinentes). L'intérêt principal des courbes ROC est le fait de ne pas tenir compte d'un éventuel déséquilibre entre le nombre de termes pertinents et non pertinents. Par ailleurs, l'aire sous la courbe ROC (AUC, « Area Under the Curve ») peut être vue comme la mesure globale de l'efficacité d'une mesure d'intérêt. Précisons que le critère relatif à l'aire sous la courbe est équivalent au test statistique de Wilcoxon-Mann-Whitney (voir les travaux d'Akobeng, 2007 ; Yan et al., 2003).

Dans le cas correspondant au classement des termes en utilisant des mesures statistiques, une courbe ROC idéale correspond au fait d'obtenir tous les termes pertinents en début de liste et tous les termes non pertinents en fin de liste. Cette situation correspond à une AUC de 1. La diagonale correspond aux performances d'un système aléatoire, progrès du taux de vrais positifs s'accompagnant d'une dégradation équivalente du taux de faux positifs. Une telle situation correspond à AUC = 0.5 (Hanley et al., 1982). Enfin, si les termes triés par intérêt décroissant sont tels que tous les termes pertinents sont précédés par les non pertinents, alors nous obtenons AUC = 0. Une mesure d'intérêt efficace pour ordonner les termes consiste donc à obtenir une aire sous la courbe (AUC) ROC la plus importante possible ce qui est strictement équivalent à minimiser la somme des rangs des exemples positifs (Akobeng, 2007).

3.2.1.3.2 Approche collaborative pour évaluer les termes extraits par un processus de fouille de textes. Conçue en 1950 (Dalkey et al., 1963), la méthode Delphi est l'une des plus anciennes méthodes formelles de sollicitation d'experts. Son objectif principal est d'atteindre le consensus du groupe à travers le partage itératif des réponses (Gustafson et al., 2013). Il s'agit d'un processus de quantification des avis d'experts afin de répondre aux questions scientifiques pour lesquelles les résultats obtenus avec la recherche traditionnelle ne sont pas suffisants ou ne sont pas encore disponibles. Par exemple, Economopoulou et al., 2014 ont utilisé l'avis d'experts pour hiérarchiser de maladies exotiques pour l'UE. Mantero et al., 2011 ont utilisé l'avis d'experts afin d'identifier les mots clés qui doivent être obligatoirement présents dans les articles collectés du Web pour le système MedISys. Furrer et al., 2015 ont sollicité des experts pour évaluer des termes décrivent des signes cliniques en santé animale extraits avec une méthode de fouille de textes.

Il est évident qu'une enquête Delphi ne repose pas sur un échantillon statistique représentatif de l'ensemble de la population. Il s'agit plutôt d'un mécanisme de prise de décision de groupe qui requiert la participation d'experts qui ont une compréhension claire du phénomène de l'étude (Okoli et al., 2004). La méthode Delphi est résolument et exclusivement qualitative puisqu'elle ne prétend pas à l'analyse statistique.

Ainsi, la méthode Delphi est relativement lourde et fastidieuse tant pour les analystes que pour les experts (quatre tours de questionnaire). Elle apparaît, à certains égards, davantage intuitive que rationnelle. Seuls les experts qui sortent de la norme sont amenés à justifier leur position. On peut aussi considérer que l'opinion des déviants est, en termes prospectifs, plus intéressante que celle de ceux qui rentrent dans le rang.

En prenant compte les avantages et les défis de la méthode Delphi, nous n'avons pas utilisé toute les caractéristiques du processus Delphi (en particulier le fait de trouver un consensus par itérations successives et surtout d'engager des discussions entre les experts qui peuvent réviser leur analyse dans le contexte des Delphis complets). Le focus de notre méthode Delphi était sur le choix et la qualité des experts, ainsi que leur contribution pour évaluer la pertinence des :

— termes extraits par un processus de fouille de textes pour caractériser des maladies modèles.
— pairs d'associations entre les termes décrivant des signes cliniques et les hôtes, obtenus avec un processus de fouille de textes, qui sont les mieux adaptées pour des recherches automatisées sur le Web.

Pour notre enquête Delphi, les experts ont été choisis d'abord et avant tout pour leur expertise pour chacune des maladies modèles. La sélection des experts a été réalisée sur la base d'une carte des acteurs établie pour la circonstance (Gustafson et al., 2013). Dans ce contexte, en déployant un questionnaire en ligne (Annexe F), nous avons sollicité 21 experts pour la PPA, 7 pour la FA, 7 pour la FCO et 5 pour le SBV.

Dans une étape préliminaire (Delphi 1 de l'Annexe F), les experts ont proposé des termes qui caractérisent les maladies modèles et qui peuvent être utilisés pour détecter leur émergence sur le Web (termes spécifiques et termes hautement spécifiques). La sous-section 3.2.1.4 présente les résultats de cette partie de l'enquête.

Dans la deuxième partie du questionnaire (Delphi 2 de l'Annexe F.2), les experts ont évalué une liste représentative de termes extraits avec un processus de fouille de textes (sous-section 3.2.1.1). Par exemple, pour la FA les experts ont évalué les termes : « livestock deaths », « general clinical signs », « onset of weakness », « fever outbreak », etc. Ces termes ont été évalués selon leur pertinence de caractériser les maladies modèles. Les résultats de cette partie de l'enquête Delphi sont présentés dans cette sous-section 3.2.1.4. Le consensus général entre les experts de Delphi 2, est évalué selon le coefficient de Kendall (concordance pour plusieurs variables qualitatives de plusieurs classes de réponses ; dans notre cas, non spécifique, faiblement spécifique, moyennement spécifique, spécifique et hautement spécifique) (Hallgren, 2012).

Dans la troisième partie du questionnaire (Delphi 3 de l'Annexe F.3), les experts ont évalué un nombre représentatif des paires d'associations entre les termes décrivant des hôtes et des signes cliniques extraits par fouille de textes (*e.g.*, pour la FA, l'association entre des signes cliniques cutanés-muqueux et des bovins). Cette contribution est détaillée dans la sous-section 3.2.2 qui suit. Le consensus général entre les experts pour le processus de

Delphi 3, est évalué selon la statistique Kappa de Fleiss (Viera et al., 2005) (concordance pour plusieurs variables qualitatives de deux classes de réponses ; dans notre cas, spécifiques et hautement spécifiques).

Les valeurs de Kappa et du coefficient de Kendall entre 0 et 0,2 illustrent un accord très faible entre les experts, des valeurs entre 0,21 et 0,4 montrent un accord faible, des valeurs entre 0,41 et 0,6 mettent en relief un accord modéré, des valeurs de 0,61 à 0,8 illustrent un accord fort et des valeurs de 0,81 à 1,0 montrent un accord presque parfait (Hausberg et al., 2012).

La méthode Delphi a été calibrée selon une première enquête menée pour la PPA. Pour cette maladie modèle, les processus Delphi 2 et 3 ont été fusionnés dans une seule partie (Arsevska et al., 2016d) ; pour la FA, FCO et SBV, les processus Delphi 2 et 3 ont été mis en œuvre de manière distincte (Arsevska et al., 2016c).

3.2.1.4 Résultats

3.2.1.4.1 Performance de la nouvelle mesure de classement de termes. La Figure 3.9 présente les résultats de l'aire sous la courbe (AUC) ROC pour les quatre maladies modèles et les 2400 termes reclassés selon la pondération spécifique obtenue avec la nouvelle mesure *w(t)*. Dans ce cadre, des valeurs α_1 et α_2 propres à chacune des maladies sont appliquées [1].

La Figure 3.10 présente les résultats de l'aire sous la courbe (AUC) ROC pour les quatre maladies modèles et les 2400 termes reclassés selon la pondération générique pour la nouvelle mesure *w(t)*. Pour ces expérimentations, des valeurs génériques α_1 et α_2 sont appliquées [2].

Notons que les résultats obtenus selon les deux types d'approches (spécifiques et génériques) sont du même ordre (Figures 3.9 et 3.10) même si des spécificités subsistent.

Globalement, pour la PPA, la mesure *w(t)* selon la pondération spécifique était faiblement précise pour distinguer les termes pertinents des non pertinents (AUC ROC moyenne de 0,57). Cependant, la mesure *w(t)* distinguait précisément les termes pertinents jusqu'à la position 310 (AUC ROC > 0,7). L'AUC ROC pour les termes jusqu'à la position 40 et entre 100 et 160 était > 0,80 (AUC ROC maximal = 0,90) (Figure 3.9). Des résultats similaires ont été obtenus avec la pondération générique de la mesure *w(t)* (AUC ROC moyenne de 0,57). Cette pondération a principalement stabilisé les fluctuations de la performance de la mesure *w(t)*, surtout pour des termes jusqu'à la position 150 (AUC ROC > 0,80) (Figure 3.10).

1. α_1 = 0,57, α_2= 0,43 pour la PPA ; α_1 = 0,83, α_2= 0,17 pour la FA ; α_1 = 0,76, α_2= 0,24 pour la FCO ; α_1 = 0,88, α_2= 0,12 pour SBV
2. α_1=0,76 et α_2=0,24

Pour la FA, la mesure *w(t)* selon la pondération spécifique était modérément précise pour distinguer les termes pertinents des non pertinents (AUC ROC moyenne de 0,60). Cependant, la mesure *w(t)* distinguait précisément les termes pertinents entre la position 50 et 210 (AUC ROC > 0,70) et entre la position 500 et 2100 (AUC ROC > 0,70) (Figure 3.9). Après l'application de la pondération générique, la tendance de la performance de la mesure *w(t)* est stabilisée. Nous avons observé une AUC ROC > 0,70 pour les termes entre la position 60 et 240 ; entre 420 et 700 ; et entre 800 et 2100 (Figure 3.10). La performance de la mesure *w(t)* pour la FA n'a pas dépassé une AUC ROC > 0.80 dans les deux cas de pondération.

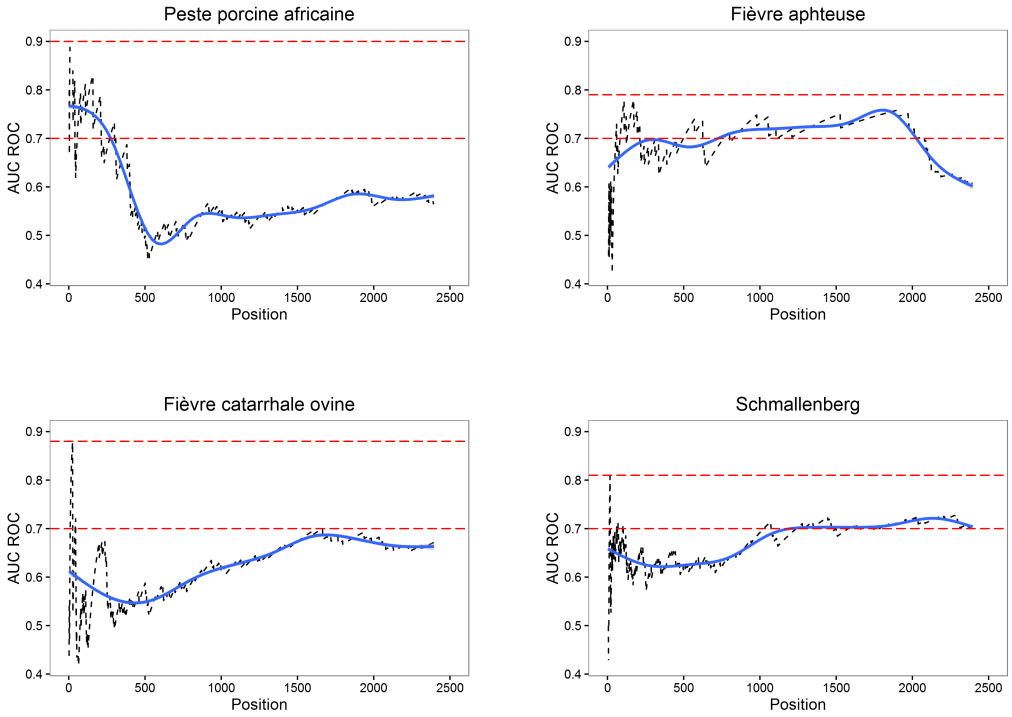


Figure 3.9 – L'aire sous la courbe (AUC ROC) selon les *n* premiers termes reclassés selon la pondération spécifique de la nouvelle mesure *w(t)*. La ligne en pointillé représente les différentes valeurs AUC, la ligne bleue présente la tendance selon la méthode de lœss. Les lignes horizontales en pointillé rouge représentent le classement de termes d'AUC > 0.7 à l'AUC maximale

Pour la FCO, la mesure *w(t)* selon la pondération globale et la pondération spécifique était modérément précise pour distinguer les termes pertinents des non pertinents (AUC ROC moyenne de 0,67). Pour les deux pondérations, la mesure *w(t)* distinguait précisément les termes pertinents entre la position 10 et 60 (AUC ROC de 0,71 au 0,84) et entre 1200 et 1700 (AUC ROC de 0,70 à 0,72) (Figures 3.9 et 3.10).

Pour le SBV, la mesure *w(t)* selon la pondération spécifique était modérément précise pour distinguer les termes pertinents des non pertinents (AUC ROC de 0,70). Selon la pondération générique cette mesure était moins précise pour distinguer les termes pertinents des non pertinents (AUC ROC moyenne de 0,67). Avec la pondération spécifique la

performance de la mesure *w(t)* n'a pas dépassé une AUC ROC > 0.81. Cette performance c'est améliorée avec la pondération générique (AUC ROC max de 0,88). La mesure *w(t)* distinguait précisément les termes pertinents des non pertinents pour les termes entre la position 13 et 20 (AUC ROC de 0,72 à 0,81) et entre 1000 et 2400 (AUC de 0,7 au 0,72) (Figure 3.9). Selon la pondération générique, la mesure *w(t)* distinguait précisément les termes pertinents entre la position 10 et 50 (AUC ROC de 0,70 au 0,88) et entre 1600 et 1700 (AUC ROC = 0,70) (Figure 3.10).

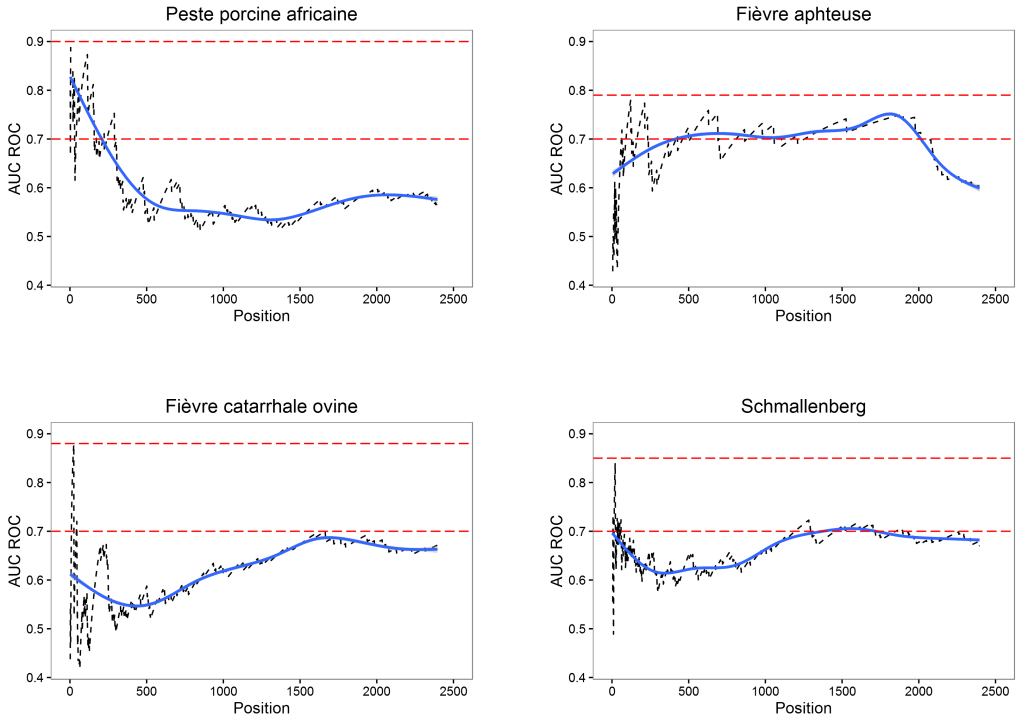


Figure 3.10 – L'aire sous la courbe (AUC ROC) selon les *n* premiers termes reclassés selon la pondération générique de la nouvelle mesure *w(t)*. La ligne en pointillé représente les différentes valeurs AUC, la ligne bleue présente la tendance selon la méthode de lœss. Les lignes horizontales en pointillé rouge représentent le classement de termes d'AUC > 0.7 à l'AUC maximale

Notons que notre fonction de rang *w(t)* a amélioré les classements de *BioTex* (cf. Tableau 3.2). Plus précisément, pour la PPA, l'AUC de base est de 0.45 alors que la valeur obtenue est 0.57 après l'introduction de la nouvelle mesure *w(t)*. Nous pouvons relever des améliorations encore plus significatives pour les trois autres maladies modèles (cf. Tableau 3.2).

Tableau 3.2 – Résultats pour l'aire sous la courbe (AUC) selon le rang de *BioTex* et les classements par la mesure $w(t)$ avec une pondération spécifique et générique

Maladie	Rang *BioTex*	Mesure $w(t)$ avec pondération spécifique	Mesure $w(t)$ avec pondération générique
Peste porcine africaine	0,45	0,57	0,57
Fièvre aphteuse	0,43	0,60	0,60
Fièvre catarrhale ovine	0,38	0,67	0,67
Schmallenberg	0,40	0,70	0,67

3.2.1.4.2 Termes proposés par des experts (Delphi 1). L'enquête Delphi nous a permis de bénéficier de la connaissance commune de nombreux experts. Dans la première partie du Delphi (Delphi 1 de l'Annexe F), les experts ont contribué à la proposition des termes spécifiques et hautement spécifiques qui caractérisent les maladies modèles. La Figure 3.11 présente les termes proposés par des experts, groupés par catégorie (signes cliniques, nom de la maladie, l'hôte, etc.). La liste détaillée de tous les termes proposés est présente en Annexe G de ce manuscrit.

Pour la PPA, les experts ont proposé 90 termes, dont 37 termes spécifiques (41%) et 53 termes très spécifiques (59%). Parmi ces termes, 30 termes (33%) sont propres à des signes cliniques liés à la mortalité des animaux, 18 termes décrivent des signes cliniques hémorragiques, la fièvre et la combinaison de ces signes cliniques chez le suidé (2%), ainsi que 8 synonymes du nom de la PPA (9%).

Pour la FA, les experts ont proposé 62 termes, dont 33 termes spécifiques et 29 termes très spécifiques. Parmi ces termes, 22 termes (35%) sont liés à des signes cliniques cutanés/muqueux et 10 termes (16%) des signes cliniques généraux chez les ruminants.

Pour la FCO, les experts ont proposé 64 termes, dont 44 spécifiques (69%) et 20 très spécifiques (31%). Parmi ces termes, 23 termes (34%) décrivent des signes cliniques cutanés/muqueux chez les ruminants et 23 termes (36%) synonymes du nom de la FCO.

Pour le SBV, les experts ont proposé 38 termes, dont 22 spécifiques (58%) et 16 hautement spécifiques (42%). Parmi ces termes, 29 termes sont propres à des malformations et déformations congénitales chez les ruminants nouveaux–nés (76%) et 5 termes décrivent des signes cliniques reproducteurs chez les ruminants (13 termes).

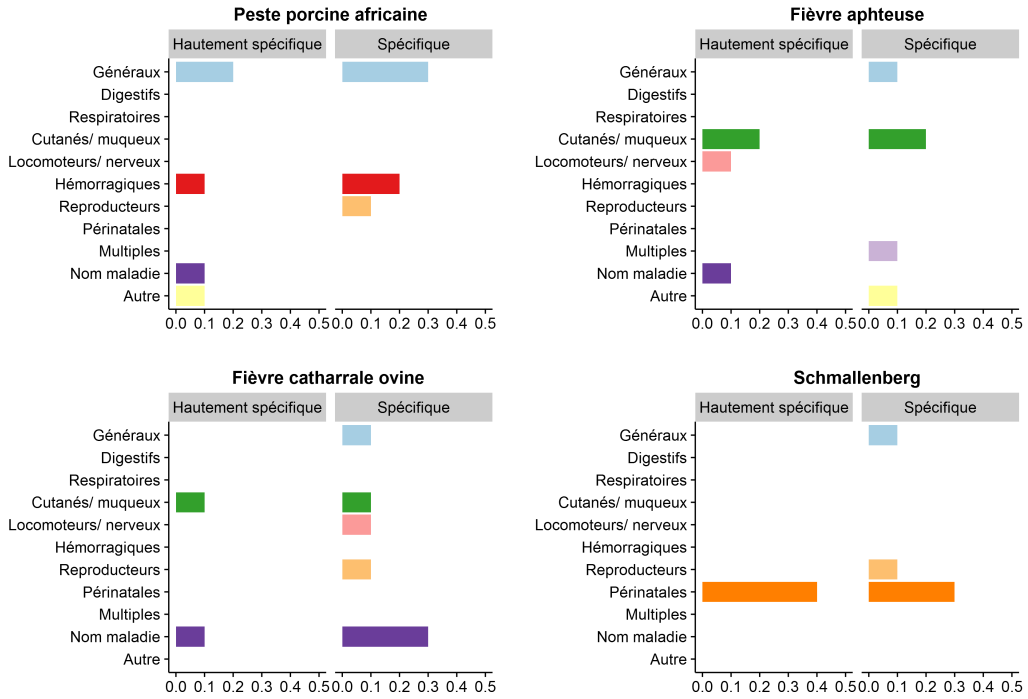


Figure 3.11 – Termes proposés par des experts pour caractériser l'émergence des maladies modèles. L'axe X présente le ratio de réponses par signe clinique ou l'hôte (l'axe Y)

3.2.1.4.3 Spécificité des termes extraits par un processus de fouille de textes (Delphi 2). Dans la deuxième partie du Delphi (Delphi 2 de l'Annexe F.2), les experts ont évalué les termes obtenus par le processus de fouille de textes représentatifs des maladies modèles. Les Figures 3.12, 3.13, 3.14 et 3.15 présentent les évaluations des experts pour chacune des maladies modèles.

Peste porcine africaine (PPA). Les experts ont évalué dix termes pour la PPA qui décrivent des signes cliniques généraux et hémorragiques, ainsi que des termes synonymes de la peste porcine (Figure 3.12). Le consensus entre les experts pour la spécificité des termes était modéré (coefficient de Kendall = 0,432, $p < 0{,}005$).

La majorité des experts ont relevé des termes qui décrivent la peste porcine (« swine fever kills », 20 sur 21 experts ; « extensive free range pig suspected swine fever », 16 sur 21 experts) comme étant moyennement à hautement spécifiques pour caractériser la PPA et identifier les premiers signes de son émergence sur le Web. La peste porcine est un terme qui englobe deux maladies porcines très similaires au niveau clinique : la peste porcine africaine et la peste porcine classique. Il n'est donc pas surprenant que la majorité des experts ait considéré que ces termes soient spécifiques pour détecter toute suspicion d'émergence de la peste porcine.

Termes décrivant des signes cliniques généraux. Le terme qui caractérise la mortalité, « high mortality » a été évalué par la majorité des experts comme étant hautement spécifique (12 sur 21 experts). Les autres termes ont été évalués comme étant moyennement à hautement spécifiques (« lethal pig disease », 19 sur 21 experts ; « dead wild boar », 18 sur 21

experts ; « fresh outbreak lethal pig disease », 17 sur 21 experts ; « wild pigs gross mortality », 17 sur 21 experts) pour détecter une émergence de la PPA sur le Web. Les termes décrivant la fièvre ont été évalués comme étant faiblement et moyennement spécifiques (« fever outbreak reported », 18 sur 21 experts ; « district pigs fever outbreak », 15 sur 21 experts), *i.e.,* ils ne sont pas suffisamment spécifiques pour détecter une émergence de la PPA sur le Web.

Termes décrivant des signes cliniques hémorragiques. Le terme qui caractérise les signes cliniques hémorragiques « devastating haemorrhagic fever », a été évalué comme étant moyennement à hautement spécifique (19 sur 21 experts). L'association entre signes cliniques hémorragiques et les porcins, « pig farms haemorrhagic fever », a été la plus discriminante et a été évaluée par la majorité des experts comme étant hautement spécifique pour caractériser l'émergence de PPA (15 sur 21 experts).

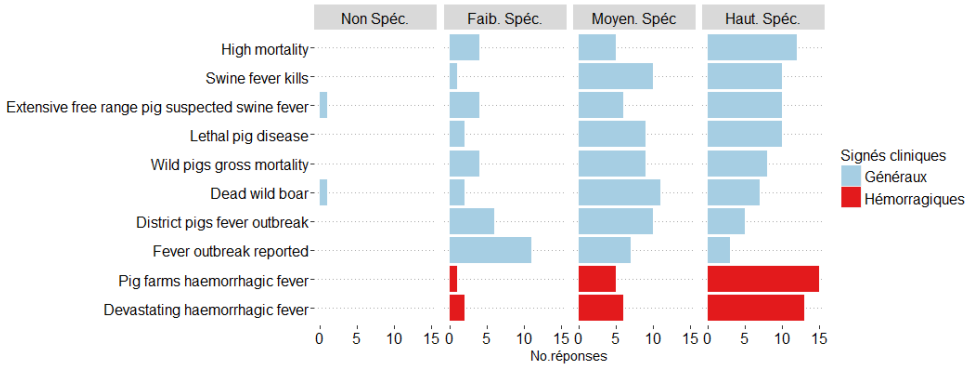


Figure 3.12 – Évaluation des termes pour la peste porcine africaine obtenus par le processus de fouille de textes

Fièvre aphteuse (FA). Les experts ont évalué sept termes pour la FA qui décrivent des signes cliniques cutanés/ muqueux et généraux (Figure 3.13). Le consensus entre les experts pour la spécificité des termes était faible (coefficient de Kendall = 0,304, $p < 0{,}005$).

Termes décrivant des signes cliniques cutanés et muqueux. La majorité des experts a noté la formation de vésicules comme étant hautement spécifique et très hautement spécifique pour caractériser l'émergence de FA (« veisicular stomatitis » 5 sur 7 experts, « vesicular disease » 5 sur 7 experts, « swine vesicular disease » 4 sur 7 experts) ; la maladie des muqueuses comme moyennement et hautement spécifique (« mucosal disease » 5 sur 7 experts) ; la formation de papules (« papular stomatitis ») comme faiblement et moyennement spécifique pour caractériser l'émergence de la FA (7 sur 7 experts).

Termes décrivant des signes cliniques généraux. Le terme lié à la perte de production (« production losses ») a été relevé par la majorité des experts (4 sur 7 experts) comme étant moyennement spécifique. Le terme lié à la mortalité (« low mortality ») a été relevé comme étant non spécifique et faiblement spécifique pour caractériser l'émergence de la FA (6 sur 7 experts).

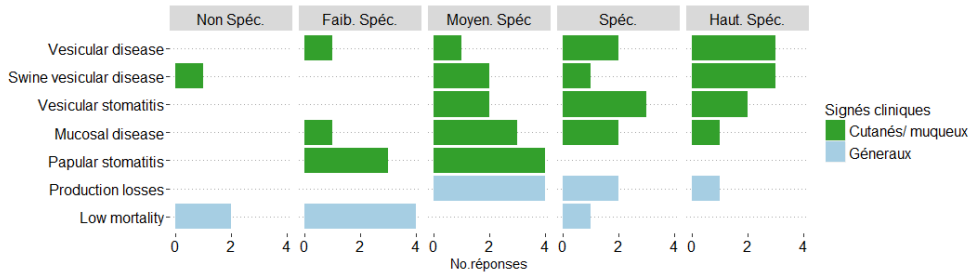


Figure 3.13 – Évaluation des termes pour la fièvre aphteuse obtenus par le processus de fouille de textes

Fièvre catarrhale ovine (FCO). Les experts ont évalué six termes pour la FCO qui décrivent des signes cliniques généraux et reproducteurs. Le consensus entre les experts pour la spécificité des termes était non significatif (coefficient de Kendall = 0,057, $p < 0,005$).

La majorité des experts a relevé les termes décrivant des signes cliniques généraux comme étant non spécifiques à faiblement spécifiques pour caractériser l'émergence de la FCO (« livestock deaths », 5 sur 7 experts ; « general clinical signs », 5 sur 7 experts ; « onset of weakness », 6 sur 7 experts ; « fever outbreak », 4 sur 7 experts).

La majorité des experts a identifié les termes décrivant des signes cliniques reproducteurs comme étant non spécifiques à faiblement spécifiques pour caractériser l'émergence de la FCO (« embryonic death », 6 sur 7 experts ; « occurrence of abortion », 5 sur 7 experts) (Figure 3.14).

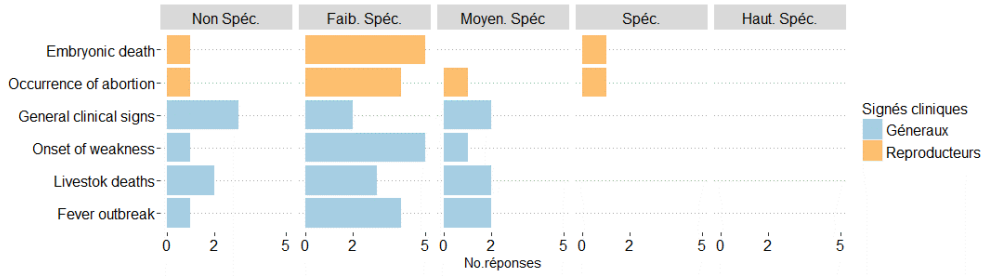


Figure 3.14 – Évaluation des termes pour la fièvre catarrhale ovine obtenus par le processus de fouille de textes

Schmallenberg (SBV). Les experts ont évalué 18 termes représentatifs de SBV qui décrivent des signes cliniques généraux, digestifs, respiratoires, reproducteurs et périnatals (Figure 16). Le consensus entre les experts pour la spécificité des termes était élevé (coefficient de Kendall = 0,752, $p < 0,005$).

La majorité des experts a évalué les termes caractérisant des malformations et déformations congénitales comme étant hautement et très hautement spécifiques (5 sur 5 experts), la mortalité postnatale et des avortements sont considérés comme modérément et hautement spécifiques (4 sur 5 experts).

La majorité des experts a évalué les termes décrivant des signes cliniques généraux et respiratoires comme étant non spécifiques et faiblement spécifiques pour caractériser l'émergence de SBV (« nonspecific febrile syndrome » ; 3 sur 5 experts, « mild transient disease », 3 sur 5 experts ; « acute bronchopneumonia », 5 sur 5 experts, respectivement). Les experts ont évalué les termes qui décrivent les signes cliniques digestifs (« watery diarrhoea ») comme étant non spécifiques à hautement spécifiques.

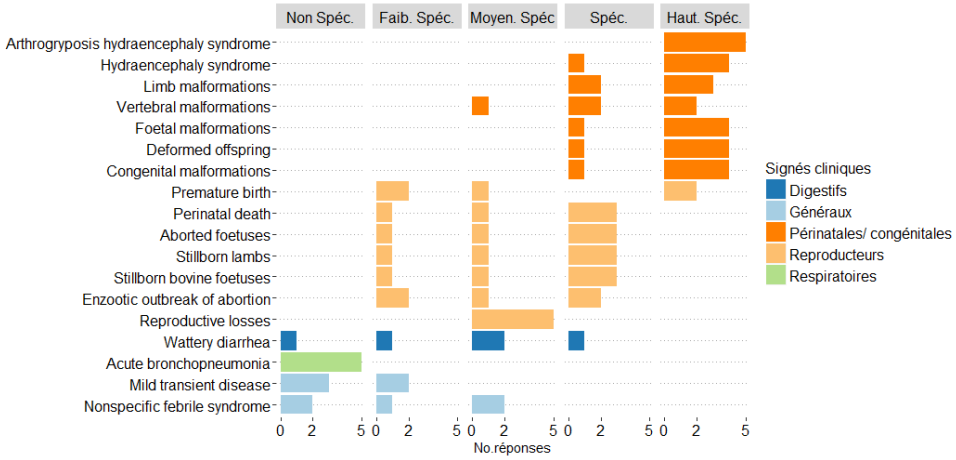


Figure 3.15 – Évaluation des termes pour le Schmallenberg obtenus par le processus de fouille de textes

3.2.1.5 Discussion

Dans cette sous-section de la thèse, nous avons proposé un processus de fouille de textes pour l'extraction automatique de la terminologie à partir d'un corpus de documents qui décrivent des foyers de maladies animales infectieuses. La terminologie obtenue pourra servir à construire un vocabulaire spécifique pour des recherches automatiques sur le Web. Le processus proposé permet d'extraire les termes mono et polylexicaux à partir d'un corpus spécifique du domaine. Le principal avantage de la méthode est son applicabilité à d'autres langues (français, espagnol) et à des données textuelles d'autres maladies.

Nous avons également proposé une nouvelle mesure *w(t)* de classement automatique des termes pour la veille. Nous avons pondéré cette mesure : *i)* d'une manière spécifique selon la maladie donnée ; et *ii)* d'une manière globale selon la qualité des sources. Les résultats montrent que l'AUC optimale pour l'ensemble des maladies se situe entre 0,70 et 0,90 ce qui met en valeur la qualité du classement automatique des termes avec la mesure *w(t)*. Nos résultats indiquent également que la pondération globale de la mesure *w(t)* n'a pas influencé sur sa performance. Ceci indique que pour les futures maladies à étudier, nous pouvons appliquer une pondération globale indépendante du type de la pathologie.

Dans certains cas, nous avons également constaté des variations de la performance de la mesure *w(t)*. Ces variations ont été probablement provoquées par le déséquilibre entre les termes considérés comme pertinents et ceux considérés comme non pertinents. Comme indiqué dans la littérature, le problème de classes déséquilibrées n'est pas rare dans les processus de fouille de textes. Il existe de nombreuses situations où l'une des deux catégories (en général, la plus intéressante pour l'analyse) est moins fréquente (He et al., 2009). Japkowicz et al., 2002 rapportent que ce déséquilibre de catégories compromet le processus de classification et l'analyse ROC, car le modèle a tendance à mettre l'accent sur la classe plus fréquente et à ignorer les événements rares. Non seulement l'estimation de la performance de classification est touchée par une distribution asymétrique des classes, mais également l'évaluation de leur exactitude est compromise, parce que le faible nombre de données aboutit à des estimations médiocres de l'exactitude du modèle. Les travaux récents proposent un cadre systématique et unifié pour traiter le problème des classes de données déséquilibrées, qui s'appuie sur la génération de nouveaux exemples artificiels à partir des classes, selon une approche bootstrap lissée (Lunardon et al., 2014 ; Menardi et al., 2014).

Les résultats d'extraction automatique de termes (EAT) dépendent fortement des sources utilisées (Laroche et al., 2011). En général, cette dépendance n'est pas liée à la taille du corpus mais à sa qualité (Zesch et al., 2010). En utilisant 13 versions successives de Wikipédia, les derniers auteurs ont mis en relief que la croissance du corpus n'a pas d'effets significatifs sur la performance de l'EAT et la similarité sémantique des termes extraits. Selon eux, des corpus plus petits (donc plus facile à obtenir) peuvent être utilisés sans nuire à la performance. Nous partageons le même constat.

Nos prochaines expérimentations porteront sur l'influence du contenu du corpus de référence sur la qualité de l'extraction. Pendant la période de référence de nos travaux, entre 2011 et 2014, le corpus de documents pertinents pour la PPA et le SBV décrivait en détail leur émergence. Pour la PPA, il s'agissait d'une nouvelle extension des foyers en Russie et dans les pays Baltes. Pour SBV, les documents pertinents décrivent en détail l'apparition d'un nouveau virus en Europe.

Peu parmi de sources pour la FCO et la FA détaillent des signes cliniques chez des animaux infectés. Les principales sources pour ces deux maladies décrivent des événements de FCO dans les Balkans ainsi que des foyers de FA en Afrique. Le sujet principal de ces sources était sur les conséquences de ces maladies sur la productivité animale et les pertes économiques chez les éleveurs. Cette différente disponibilité d'information médiatique en langue locale, explique probablement le faible nombre de termes spécifiques pour la FA et la FCO.

Un deuxième facteur influençant la performance de la mesure *w(t)* est le choix du « gold standard » (Zou et al., 2007). Par exemple, en épidémiologie clinique pour estimer la performance d'une mesure à l'aide d'analyse ROC, le statut de la maladie pour chaque patient est théoriquement estimée sans erreur. L'état réel de la maladie « gold standard »

peut être disponible à partir d'un suivi clinique et dans certains cas, s'il n'y a pas de possibilités de suivi, il est jugé par des experts. Dans ce dernier cas, une erreur de mesure peut influencer les résultats de l'analyse ROC et l'AUC. Plus précisément, une erreur de mesure peut arriver quand un gold standard est absent ou qu'un standard imparfait est utilisé pour l'évaluation. Par exemple, en santé publique il n'y a pas de gold standard pour le diagnostic de l'infection de la tuberculose latente et l'exactitude des tests de diagnostic disponibles demeurent donc incertaines (Sadatsafavi et al., 2010). Souvent, en santé animale pour déterminer le statut sanitaire du cheptel dans les pays avec peu de moyens disponibles, l'estimation de la prévalence des maladies se réalise selon la perception des éleveurs. Ces études participatives ont rarement pu valider la précision issue des rapports des éleveurs (Morgan et al., 2014). Avec nos analyses nous avons remarqué les mêmes contraintes que celles mentionnées ci-dessus.

En effet, dans nos expérimentations, nous avons considéré comme étant pertinents (« gold standard ») un ensemble de termes spécifiques pour la veille, tels que des noms de maladies modèles, ses signes cliniques et ses hôtes. Toutefois, en réalité ce gold standard a été choisi pour répondre à une question spécifique : l'identification de la terminologie qui caractérise des maladies modèles et qui peut être utilisée pour des recherches sur le Web. Ceci n'exclut pas la qualité des autres termes à d'autres fins épidémiologiques. Par exemple, dans la liste de termes extraits pour la PPA, des termes comme : « illegal importation », « spread through trade », « unprocessed meat products », « pig meat exports », etc. décrivent des facteurs de risque pour la diffusion de la PPA. D'autres termes, tels que « soft ticks », « soft tick ornithodoros », etc. caractérisent le vecteur du virus de la PPA. Plusieurs termes de cette liste donnent des zones de risque, tels que « swine fever in lusaka », « eastern africa », « zambian capital », « african swine fever in kenya », etc.

Finalement, la majorité des termes proposés par les experts ont confirmé l'efficacité de l'approche de fouille de textes mise en œuvre dans les travaux de thèse pour identifier une terminologie pour la veille. Par exemple, pour la PPA, parmi les termes proposés par les experts, nous pouvons noter une prédominance des signes cliniques hémorragiques (*e.g.*, « haemorrhagic fever in pigs », « haemorrhagic disease of pigs », « haemorrhagic fever in boars », « bloody diarrhea in pig breedings », etc.) et généraux chez les suidé (*e.g.*, « hyperthermia in pig breedings », « high fever in pigs », « high fever and mortality in pigs », « haemorrhagic syndrome and mortality of pigs », « dead wild boar », « raising mortality in pigs »), ainsi que les sigles pour le nom de la maladie (*e.g.*, « african swine fever », « ASFV », « DNA arbovirus in pigs ») (Annexe G.1). Ces termes correspondent aux termes obtenus par le processus de fouille de textes (*e.g.*, « haemorrhagic fever », « haemorrhagic disease », « dead pigs », « dead wild boar », « deadly pig disease », « fever outbreaks », « suspected swine fever », « african swine fever outbreak » « fever virus infection », etc.) (Annexe E.1).

De manière similaire, les experts ont proposé comme hautement spécifiques pour caractériser SBV, des termes décrivant des malformations et déformations congénitales chez

les ruminants nouveau nés (*e.g.,* « malformed foetuses », « arthrogryposis hydranencephaly syndrome », « congenital malformations», « anomaly in newborn lambs », etc.), ainsi que des signes cliniques reproductifs (*e.g.,* « stillbirth », « abortions », etc.), et des signes cliniques généraux (*e.g.,* « mild acute diarrhea », « drop in milk production », « transient drop in milk production », etc.) (Annexe G.4). Ces termes correspondent également aux termes obtenus par notre processus de fouille de textes (*e.g.,* « arthrogryposis hydranencephaly syndrome », « congenital malformations », « deformed lambs », « malformed offspring», « watery diarrhea », « mild transient disease », « transient drop in milk production », etc.) (Annexe E.4).

A contrario, en prenant en compte le nombre limité et peu spécifique des termes obtenus par fouille de textes pour la FA et la FCO (Annexes E.2 et E.3), les experts ont été une meilleure source de termes. Par exemple, pour la FA, les experts ont identifié l'hypersalivation, la mortalité chez les jeunes animaux, ainsi que des ulcères et vésicules des sabots, de la bouche, de la langue, et de la mamelle, principalement chez les bovins (*e.g.,* « hypersalivation and lameness in bovines », « blisters and lameness in bovine and porcine farms », « blisters on the groin in pigs », « lesions of the hoofs », « mortality of suckling piglets », « brutal decrease of production in dairy cows », etc.) (Annexe G.2). Ces termes spécifiques n'ont pas été obtenus avec la fouille de textes.

De manière similaire, pour la FCO, la majorité des experts a proposé des termes, tels que les lésions buccales et les ulcères, les œdèmes faciaux, la cyanose de la langue et d'hypersalivation, principalement chez les ovins et les bovins (*e.g.,* « ptyalisme in ruminants », « bucal lesions in sheep (but also bovines) », « abundant nasal discharge in ruminants », « facial oedemas », « lameness in bovines », « limited mouvement in sheep », « arthritis in ruminants », « cyanosys of the tongue in ruminants », « gengival ulcers and hypersalivation in ruminants », etc.). Ces termes n'ont pas été obtenus avec la fouille de textes (Annexe G.3). Cependant, les termes sigles du nom de la maladie proposés par des experts (*e.g.,* « bluetongue disease », « ovine catarrhal fever » , « BTV », « disease transmitted by Culicoides», etc.), ainsi que le vocabulaire décrivant la mortalité et la fièvre (*e.g.,* « hyperthermia between sheep », « fever ovine or bovine » , « mortality ovine or bovine », etc.), correspondent aux termes obtenus avec la fouille des textes *e.g.,* « bluetongue disease », « btv infection », « bt outbreaks » ou *e.g.,* « livestock deaths », « fever outbreak », etc.

L'extraction automatique de termes par fouille de textes peut être suffisante pour construire un lexique propre à différentes maladies. Cependant, ces termes ne sont pas suffisants pour la veille s'ils ne sont pas associés de manière pertinente. Par exemple, les termes « sanglier » et « mortalité » extraits d'un corpus sont pertinents pour caractériser la PPA. Pourtant, leur utilisation indépendante engendre une collecte *via* le Web de nombreuses pages non pertinentes pour la veille. Ainsi, une de nos contributions que nous décrivons dans la sous-section suivante est une méthode pour la construction automatique des associations entre les termes spécifiques. Par exemple, la combinaison entre les termes « sanglier » et « mortalité » permet de collecter des articles plus adaptés qui traitent de

l'apparition de la PPA et d'autres maladies potentiellement épizootiques. La manifestation d'un signe clinique chez un hôte est un facteur significatif permettant le diagnostic d'une pathologie significative à une nouvelle émergence de maladie. Finalement, cette méthode est complétée par l'évaluation de la qualité de notre approche à l'aide d'experts des domaines (Delphi 3).

3.2.2 Nouvelles mesures d'association entre les termes

Les nouvelles mesures statistiques que nous proposons dans cette sous-section s'appuient sur des techniques de fouille de textes et de fouille du Web (Article 2 de l'Annexe A.2). Ces mesures ont deux buts principaux. Le premier est d'aider les utilisateurs à réaliser une meilleure recherche d'information sur le Web. Le deuxième est l'amélioration de la qualité de bases de connaissances que l'on peut ainsi enrichir en nouvelles associations entre les termes.

Par exemple, pour la VSI, un utilisateur pourrait effectuer une requête originale sur le Web avec le terme « ASF ». Plusieurs définitions sont possibles pour ce terme, tels que « peste porcine africaine », « african swine fever », « swine fever », « warthog disease », etc. Un deuxième exemple concerne une requête originale sur le Web avec la combinaison des termes « pig mortality » pour détecter des signaux précoces d'émergence de la PPA. Plusieurs combinaisons entre les termes signifient le même signe clinique chez cet hôte, tels que « pig death », « high porcine mortality », « pig unknown death », etc.

En déterminant la définition adaptée, notre méthode permet d'améliorer significativement la recherche d'information sur le Web par l'expansion de la requête originale. Cette expansion pourrait par exemple, être une disjonction (opérateur « OR ») du terme et de sa définition afin de retourner un nombre de documents plus important (amélioration du rappel). La conjonction du terme et de la définition (opérateur « AND ») permettrait quant à elle d'obtenir des documents plus pertinents (amélioration de la précision). Dans notre cas, nous ne recherchons pas les définitions des termes dans les textes mais nous nous intéressons au classement des associations propres aux termes. Ainsi, comme nous allons le montrer dans cette section, nos travaux consistent à utiliser des approches de fouille du Web et de fouille de textes pour établir une fonction de rang pour les associations entre les termes (c'est-à-dire « couple de termes »). Ce rang permettra à l'utilisateur de sélectionner des associations entre les termes décrivant des signes cliniques et les hôtes qui sont les mieux adaptées pour des recherches sur le Web.

3.2.2.1 Approche de fouille du Web

Dans la littérature, de nombreuses mesures de qualité sont utilisées afin d'effectuer un classement par intérêt décroissant. Ces mesures sont issues de domaines variés : recherche de règles d'associations (Lallich et al., 2004 ; Azé, 2003), extraction de la terminologie (Roche, 2004), etc.

Notre approche consiste à sélectionner des associations entre les termes à partir d'une liste de termes obtenue par un processus de fouille de textes. Le but est donc d'effectuer un classement par pertinence en utilisant des mesures statistiques ; les paires d'associations les plus pertinentes (hautement spécifiques) devant être placées en début de liste.

Pointwise Mutual Information et Information Retrieval (PMI-IR) est un algorithme qui utilise le moteur de recherche AltaVista pour déterminer des synonymes appropriés pour une requête donnée sur le Web (Turney, 2001). Pour un mot donné, PMI-IR choisit un synonyme à partir d'une liste donnée. Les mots sélectionnés correspondent aux questions du TOEFL (« Test of English as a Foreign Language »). L'objectif est d'identifier les synonymes qui permettent d'obtenir un meilleur score. Afin de calculer le score, PMI-IR utilise plusieurs mesures fondées sur la proportion de documents où les termes sont présents ensembles (*e.g.*, les termes « haemorrhagic fever » et « domestic pigs »). La formule de Turney s'inspire de l'Information Mutuelle (Church et al., 1990). Notre travail applique un tel principe.

Tout d'abord nous proposons des mesures D_{Web} fondées sur le coefficient de $Dice$. D'autres mesures statistiques comme l'Information Mutuelle (« Mutual Information », MI_{Web}) (Church et al., 1990) et l'Information Mutuelle au Cube (« Cubic Mutual Information », CMI_{Web}) (Nazar et al., 2008 ; Vivaldi et al., 2001) peuvent être associées à des techniques de fouille du Web. Ces mesures du Web ont un bon comportement (Roche et al., 2010).

La mesure D_{Web} calcule la relation entre les termes qui décrivent les hôtes *(h)* et les signes cliniques *(cs)*. Dans ce contexte, nous mesurons le nombre de pages où les termes h et cs apparaissent tous les deux (*i.e., hit (h AND cs)*). Nous obtenons ce nombre de pages de résultats avec le moteur de recherche *Exalead*. Pour calculer cette forme de dépendance entre ces termes, nous prenons en compte le nombre de pages où chaque terme apparaît (*i.e., hit (h)* and *hit (cs)*). La formule suivante définit D_{Web} avec l'opérateur « AND » :

$$D_{Web}^{AND}(h, cs) = \frac{2 \times hit(h\ AND\ cs)}{hit(h) + hit(cs)} \tag{3.2}$$

Nous avons choisi le moteur de recherche Exalead parce qu'il applique les mêmes fonctions (en particulier l'opérateur « NEAR ») utilisés par Turney, 2001. NEAR est un opérateur qui retrouve les pages Web où les termes h et cs sont tous les deux présents dans le document dans une fenêtre de 16 mots. Nous pouvons adapter la formule précédente avec la nouvelle formule détaillée ci-dessous :

$$D_{Web}^{NEAR}(h, cs) = \frac{2 \times hit(h\ NEAR\ cs)}{hit(h) + hit(cs)} \tag{3.3}$$

De plus, nous pouvons utiliser MI et CMI pour établir la relation entre *h* et *cs* :

$$MI_{Web}^{AND}(h, cs) = \frac{hit(h\ AND\ cs)}{hit(h) \times hit(cs)} \tag{3.4}$$

$$CMI_{Web}^{AND}(h, cs) = \frac{hit^3(h\ AND\ cs)}{hit(h) + hit(cs)} \tag{3.5}$$

Nous pouvons adapter ces mesures avec l'opérateur « NEAR ». La mesure originale MI (Church et al., 1990) utilise un logarithme (*i.e.,* $\log_2 P(x, y)/(P(x) \times P(y))$. La fonction logarithmique est strictement croissante, ainsi elle ne modifie pas le classement des associations *h* - *cs*. Il est également important de noter que MI a tendance à extraire des dépendances rares et spécifiques (Vivaldi et al., 2001).

Dans notre approche, nous n'utilisons pas d'autres mesures de fouille du Web comme « Google Similarity Distance » (Cilibrasi et al., 2007) qui fonctionne avec des paramètres spécifiques et parfois difficiles à établir, par exemple le nombre total de pages Web indexées par les moteurs de recherche.

3.2.2.2 Approche de fouille de textes

Les mesures statistiques D, MI et CMI peuvent être utilisées par l'approche de fouille de textes. Dans ce contexte, la mesure $Dice(D)$ est définie par le nombre de fois (*i.e.,* fonction *nb*) où *h* et *cs* apparaissent dans le même contexte divisé par la somme du nombre total de fois où ils apparaissent dans le corpus pour chaque maladie modèle. Nous pouvons adapter cette définition avec d'autres mesures statistiques (MI et CMI). Dans ce cas, le choix du contexte est crucial. Dans notre cas, les deux termes *h* et *cs* sont très rares dans une phrase donnée alors nous utilisons un contexte plus large (*i.e.,* résumé ou article entier). Dans le contexte de ces définitions, les formules suivantes sont appliquées aux données textuelles :

$$D_{text}(h, cs) = \frac{2 \times nb(h\ AND\ cs)}{nb(h) + nb(cs)} \tag{3.6}$$

$$MI_{text}(h, cs) = \frac{nb(h\ AND\ cs)}{nb(h) \times nb(cs)} \tag{3.7}$$

$$CMI_{text}(h, cs) = \frac{nb^3(h\ AND\ cs)}{nb(h) \times nb(cs)} \tag{3.8}$$

3.2.2.3 Combinaison entre les approches de fouille du Web et fouille de textes

Les approches de fouille du Web (sous-section 3.2.2.1) et de fouille textes (sous-section 3.2.2.2) ont un comportement complémentaire. Nous proposons d'appliquer des méthodes de fouille de textes pour les associations plus courantes. En cas d'associations rares et un nombre de documents réduit, il est préférable d'utiliser une mesure statistique plus globale calculée sur le Web, *i.e.,* en considérant le Web comme un corpus de documents. Pour mettre en pratique ces principes, nous proposons une mesure globale, CMI_{global} (formule 3.9). La combinaisons CMI_{global} que nous proposons est fondée sur le critère CMI (Saneifar et al., 2015 ; Vivaldi et al., 2001) et l'utilisation de l'opérateur « AND ». Dans cette formule, la valeur 1 permet de préférer l'approche de fouille de textes pour la fonction de classement globale, lorsque la situation le permet (c-à-dire, que h et cs sont présents dans un même contexte dans nos corpus).

$$CMI_{global}(h, cs) = \begin{cases} 1 + CMI_{text}(h, cs) \text{ si } CMI_{text}(h, cs) \neq 0 \\ CMI_{Web}^{AND}(h, cs) \text{ sinon} \end{cases} \tag{3.9}$$

Le Tableau 3.3 présente des exemples des paires d'associations entre les termes décrivant l'hôte et les termes caractérisant des signes cliniques pour la FCO et le SBV. Notez que les termes sont issus de la liste de termes extraits avec *BioTex* et sélectionnés par l'utilisateur spécialiste comme pertinents pour caractériser ces maladies modèles.

Tableau 3.3 – Exemples de paires d'associations classées avec CMI_{global}

Position	**Fièvre catarrhale ovine (FCO)** Signes cliniques/ hôtes	**Schmallenberg (SBV)** Signes cliniques/ hôtes
1	general clinical signs/ pregnant ewes	stillborn bovine foetuses/ bison
2	livestock deaths/ sheep	aborted foetuses/ sheep
3	embryonic death/ cattle	deformed offspring/ sheep
4	general clinical signs/ sheep	stillborn bovine foetuses/ deer
5	livestock deaths/ cattle	aborted foetuses/ cattle
6	livestock deaths/ deer	deformed offspring/ cattle
7	fever outbreak/ sheep	stillborn bovine foetuses/ calves
8	embryonic death/ sheep	deformed offspring/ lambs
9	fever outbreak/ cattle	acute bronchopneumonia/ bison
10	embryonic death/ pregnant ewes	stillborn lambs/ goat

3.2.2.4 Évaluation et protocole expérimental

3.2.2.4.1 Approche collaborative pour évaluer la spécificité des paires d'associations. Afin d'évaluer la qualité de nos mesures statistiques, nous avons sollicité des experts du

domaine à l'aide de la méthode Delphi (Delphi 3 du questionnaire en Annexe F.3). La méthode Delphi est détaillée dans la sous-section 3.2.1.3.2.

Pour chacune des maladies modèles, les experts ont évalué la spécificité des paires d'associations entre termes. Les termes sont représentatifs des groupes de signes cliniques et des hôtes obtenus par le processus de fouille de textes. Rappelons que les paires d'associations spécifiques caractérisent très probablement la maladie donnée. Les paires d'associations hautement spécifiques caractérisent, dans presque tous les cas, la maladie donnée. Les paires d'associations entre les termes proposés aux experts, pour chacune des maladies modèles sont présentées en Annexe F.3 (partie Delphi 3).

Les experts ont évalué 14 paires d'associations pour la FA. Les paires sont représentatives deux groupes de signes cliniques (généraux et cutanées-muqueux) et sept hôtes (« cattle », « buffaloes », « small ruminants », « pigs », « wild boars », « camels », « deer »).

Pour la FCO les experts ont évalué 12 paires d'associations. Les paires caractérisent deux groupes de signes cliniques (généraux et reproductifs) et six hôtes (« cattle », « sheep », « goats », « pregnant ewes », « newborn calves », « deer »).

Les experts ont évalué 40 paires d'associations pour le SBV. Les paires représentent quatre groupes de signes cliniques (généraux, respiratoire, digestifs, reproductifs et congénitales) et huit hôtes (« cattle », « sheep », « goats », « calves », « lambs », « goat kids », « deer », « bison »).

Pour la PPA les experts ont évalué six paires d'associations (« dead wild boar », « district pigs fever outbreak », « extensive free range pig suspected swine fever », « lethal pig disease », « pig farms haemorrhagic disease »). Ces associations illustrent deux groupes de signes cliniques (généraux et hémorragiques) chez deux hôtes (les sangliers et les porcs domestiques). Notons que pour la PPA, les enquêtes Delphi 2 et 3 ont été menées ensemble. C'est la raison pour laquelle, les résultats pour la PPA sont présentés dans la sous-section 3.2.1.3.2.

3.2.2.4.2 Précision des associations obtenues avec les mesures statistiques. Les résultats des évaluations des experts (Delphi 3) ont servi pour évaluer la performance des mesures statistiques pour les vingt meilleures paires d'associations pour chacune des maladies modèles :

— La précision de classement correspond à la précision des mesures statistiques pour favoriser des associations hautement spécifiques (HT) sur les vingt associations qui donnent un meilleur classement *(mc)*.
— La précision d'acquisition des pages Web correspond au nombre de pages Web pertinentes (sous-section 3.1.2) lorsque les vingt associations avec un meilleur classement *(mc)* ont été utilisées comme requêtes de recherche (de manière rétrospective, sur Google, période de 2011 à 2014). Une page Web pertinente est une page

qui décrit des foyers de maladies modèles ainsi qu'une situation sanitaire préoccupante concernant des foyers de ces maladies.

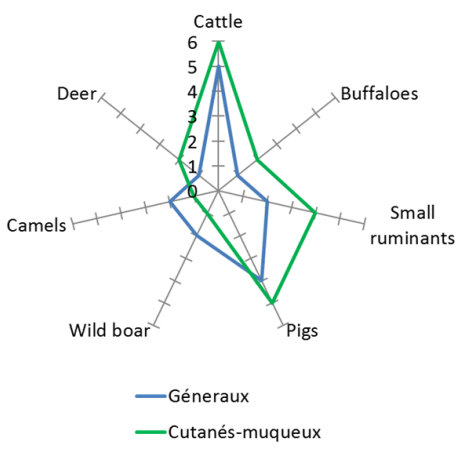
3.2.2.5 Résultats

3.2.2.5.1 Spécificité des paires d'associations décrivant des hôtes et des signes cliniques. Pendant une émergence de FA, la majorité des experts a évalué comme étant hautement spécifiques l'apparition des signes cliniques cutanés - muqueux chez les bovins (6 sur 7 experts), les porcins (5 sur 7 experts) et les petits ruminants (4 sur 7 experts) ; ainsi que l'apparition de signes cliniques généraux chez les bovins (5 sur 7 experts) et les porcins (4 sur 7 experts) (Figure 3.16a).

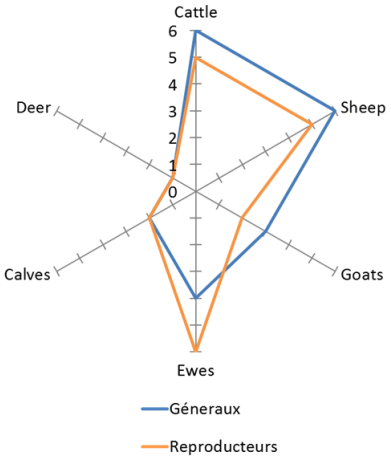
Pour la FCO, la majorité des experts a évalué comme étant hautement spécifiques l'apparition des signes cliniques généraux (6 sur 7 experts) et reproducteurs (5 sur 7 experts) chez les bovins et caprins (Figure 3.16b).

Pour le SBV, la majorité des experts a évalué comme étant hautement spécifiques l'apparition des signes cliniques reproducteurs chez les bovins, les ovins (4 sur 5 experts, respectivement) et les caprins (3 sur 5 experts) ; les malformations, déformations et mortalité postnatale chez les veaux (3 sur 5 experts), les agneaux et les chevreaux (4 sur 5 experts) ; et les signes cliniques généraux (3 sur 5 experts) et digestifs (4 sur 5 experts) chez les bovins (Figure 3.16c).

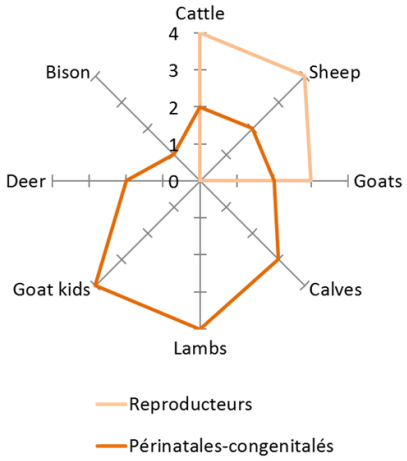
Le consensus entre les experts pour les associations entre les termes décrivant des signes cliniques et des hôtes pour la FA était non significatif (statistique Kappa = 0,103, $p < 0,005$), pour la FCO était faible (statistique Kappa = 0,229, $p < 0,005$), et pour le SBV était modéré (statistique Kappa = 0,525, $p < 0,005$).



(a) Fièvre aphteuse



(b) Fièvre catarrhale ovine



(c) Schmallenberg

Figure 3.16 – Pairs d'associations entre les termes caractérisant les hôtes et les signes cliniques, évaluées par des experts comme étant hautement spécifiques

3.2.2.5.2 Précision de mesures statistiques pour favoriser des paires d'associations hautement spécifiques. Au total, 54 paires d'associations pour la PPA, 31 paires pour la FA, 30 associations pour la FCO et 61 paires pour le SBV sont issues des meilleures paires d'associations obtenus avec les mesures statistiques.

Les meilleures mesures statistiques pour la PPA sont CMI^{Web}_{AND} et CMI_{global} avec une précision supérieure à 0,90 pour classer les paires d'associations hautement spécifiques parmi les 20 meilleures associations.

Pour la FA, la mesure CMI_{global} est la plus précise pour le classement des paires d'associations hautement spécifiques parmi les 20 meilleures associations (12 sur 20 associations), suivi par la mesure CMI^{Web}_{AND} (11 sur 20 associations).

Pour la FCO, la meilleure mesure statistique est $Dice^{Web}_{NEAR}$ qui a classé 8 associations hautement spécifiques parmi les 20 meilleures associations.

Pour le SBV, la meilleure mesure statistique est CMI^{Web}_{AND} (12 paires d'associations hautement spécifiques sur 20 associations).

Le Tableau 3.4 présente les résultats de la précision des mesures statistiques pour favoriser des associations hautement spécifiques pour les quatre maladies modèles. Même si le comportement différé selon les maladies, ces résultats montrent un bon comportement global du CMI^{Web}_{AND} et CMI_{global}.

Tableau 3.4 – Précision des mesures statistiques

Mesure statistique	Peste porcine africaine	Fièvre aphteuse	Fièvre catarrhale ovine	Schmallenberg
$Dice^{Web}_{AND}$	0,90	0,40	0,25	0,45
$Dice^{Web}_{NEAR}$	0,85	0,40	0,40	0,50
MI^{Web}_{AND}	0.70	0,45	0,35	0,50
MI^{Web}_{NEAR}	0,65	0,45	0,35	0,40
CMI^{Web}_{AND}	**1,00**	**0,55**	**0,25**	**0,60**
CMI^{Web}_{NEAR}	0,90	0,50	0,35	0,40
CMI_{global}	**0,80**	**0,60**	**0,30**	**0,50**

3.2.2.5.3 Précision des paires d'associations pour la tâche de collecte des pages Web pertinentes. Les requêtes de recherche fondées sur 54 paires d'associations pour la PPA, 31 paires pour la FA, 30 paires pour la FCO et 61 paires pour le SBV retournent 1539 pages Web pour la PPA, 587 pages pour la FA, 733 pages pour la FCO, et 564 pages pour le SBV. Pour la PPA, nous avons de plus évalué comme requêtes de recherche sur le Web toutes les paires d'associations (n=506) entre les différentes combinaisons de termes décrivant des signes cliniques et des hôtes. Ces résultats sont décrits dans les travaux d'Arsevska et al., 2016c.

Les résultats de la précision d'acquisition des pages Web pertinentes pour toutes les paires d'associations parmi les 20 meilleures obtenues avec les mesures statistiques sont détaillées en Annexe H.

Le Tableau 3.5 synthétise les résultats de la précision des paires d'associations hautement spécifiques (parmi les 20 meilleures associations) pour acquérir des pages Web pertinentes. Ces resultats sont discutés dans la section suivante.

Tableau 3.5 – Précision des mesures statistiques pour acquérir de pages Web pertinentes

Mesure statistique	Peste porcine africaine	Fièvre aphteuse	Fièvre catarrhale ovine	Schmallenberg
$Dice^{Web}_{AND}$	0,57	0,92	0,05	0,22
$Dice^{Web}_{NEAR}$	0,64	0,95	0,07	0,22
MI^{Web}_{AND}	0,07	0,92	0,08	0,14
MI^{Web}_{NEAR}	0,15	0,93	0,08	0,15
CMI^{Web}_{AND}	0,58	0,90	0,05	0,16
CMI^{Web}_{NEAR}	0,65	0,92	0,08	0,22
CMI_{global}	0,18	0,90	0,09	0,34

3.2.2.6 Discussion

Dans cette section du manuscrit, nous avons présenté sept mesures statistiques pour l'association automatique de termes composés de plusieurs mots, qui caractérisent des hôtes et des signes cliniques de nos quatre maladies modèles. Ce sont des termes issus du processus de fouille de textes détaillé en section 3.2.1 et appliqué à partir d'un corpus de documents qui caractérisent des foyers de nos maladies modèles. L'objectif principal de ces mesures est d'aider les utilisateurs à sélectionner les paires d'associations entre les termes qui décrivent le mieux des hôtes et des signes cliniques pour une maladie donnée. Ces mesures permettent de réduire l'effort humain nécessaire à la validation des termes candidats pour des recherches automatiques sur le Web.

Par la méthode Delphi, les experts ont contribué à l'évaluation des différents termes et des paires d'associations entre termes décrivant des hôtes et des signes cliniques pour des maladies modèles. Cependant, le consensus variable obtenu par les différents groupes d'experts, suggère plusieurs points à améliorer. Par exemple, les résultats de la statistique Kappa (Delphi 3) et le coefficient de Kendall (Delphi 2), montrent un consensus faible entre les experts pour la FCO et la FA. Notons que l'enquête en ligne n'était probablement pas la méthode la plus appropriée pour solliciter des experts. Une telle limite est aussi valable pour la formulation des questions et la sélection des termes représentatifs qui a nécessairement pris du temps. Il était également difficile d'estimer la bonne compréhension des objectifs du Delphi par les experts. Les résultats pour la FCO et la FA suggèrent également que les termes proposés aux experts sont parfois peu représentatifs ou peu spécifiques. Par conséquent, ceci a influencé une subjectivité des experts.

A contrario, la majorité des experts ont eu un consensus important pour des termes de la PPA et SBV. En effet, pour ces maladies modèles, nous avons obtenu un large corpus de termes par le processus de fouille de textes. Ceci a permis aux experts d'avoir un choix plus large de termes décrivant des hôtes et des signes cliniques pour ces maladies modèles. Le consensus des experts pour les termes décrivant la FA et la FCO est cependant faible. Ceci peut s'expliquer par le faible nombre et le manque de spécificité des termes obtenus par le processus de fouille de textes. Dans nos futurs travaux, nous envisageons de solliciter des experts par des entretiens directs pour confirmer et vérifier la sélection préliminaire des termes.

Nos résultats montrent également une performance variable des mesures statistiques pour favoriser les paires d'associations hautement spécifiques entre les signes cliniques et les hôtes. Notons que les évaluations ont été menées pour un nombre représentatif des paires d'associations (54 pour la PPA, 31 pour la FA, 30 pour la FCO et 62 pour le SBV). Ceci suggère que nos futurs travaux peuvent s'orienter vers l'évaluation d'un nombre plus grand des paires d'associations.

Nous considérerons que la performance des mesures statistiques était principalement influencée par les caractéristiques des termes inclus dans les paires d'associations. Ces caractéristiques dépendent principalement de leur spécificité pour caractériser la maladie

donnée. En effet, la spécificité de chaque association était évaluée en s'appuyant sur un consensus établi par des experts (questionnaires Delphi 2 et 3). Il semble évident que pour évaluer la spécificité d'une paire d'associations, le nombre de termes n'est pas un facteur déterminant pour les experts, mais le contexte épidémiologique des termes est crucial. Par exemple, les mesures statistiques pour la PPA ont donné une précision de classement d'associations hautement spécifiques supérieure à 0,65. Ainsi, les associations classées par les mesures statistiques comme pertinentes ont également été caractérisées par les experts comme hautement spécifiques. Notons que la précision de classement d'associations hautement spécifiques pour la FA et le SBV obtenue par des mesures statistiques était moyenne (précision de classement entre 0,40 et 0,60).

Les mesures statistiques pour la FCO ont donné quant à elles une précision de classement d'associations hautement spécifiques inférieure à 0,4 (sous-section 3.2.2.4.2). Ce résultat peut s'expliquer par la spécificité faible des paires d'associations pour caractériser l'émergence de cette maladie.

Pour la PPA, même si les termes liés aux fièvres hémorragiques (« pigs farms haemorrhagic fever », « devastating haemorrhagic fever », sous-section 3.2.1.3.2) ont été évalués par la majorité des experts comme étant hautement spécifiques. Ces associations n'ont pas permis de collecter sur le Web des articles sanitaires (pertinents) décrivant l'émergence de maladies exotiques (incluant la PPA). Ceci s'applique aux paires d'associations comme « devastating haemorrhagic fever AND domestic pig populations » ou « devastating haemorrhagic fever AND wild pigs », »pigs farms » AND « haemorrhagic fever », etc. (Annexe H.1). Ces résultats peuvent s'expliquer par le nombre de mots présents dans les associations (plus de 5 mots). Au contraire, les paires d'associations liées à la fièvre et la mortalité composés de quatre mots ont une haute précision pour collecter sur le Web des articles pertinents pour la PPA. Par exemple, une précision de 1.00 (c'est-à-dire toutes les pages retournées sont pertinentes)a été obtenue pour les paires d'associations « fever outbreaks » AND « wild boars », « fever outbreaks » AND « domestic pigs », « fever outbreaks » AND « pig farm ». Les paires d'associations, telles que « high mortality » AND « wild boars », « high mortality » AND « wild boars », « high mortality » AND pig population » obtiennent une précision de 0,50, 0,33 et 0,25, respectivement, pour la collecte sur le Web d'articles pertinents. Ceci suggère que nos futurs travaux peuvent s'orienter vers la définition d'un seuil de nombre de mots pour des recherches automatiques sur le Web (Arsevska et al., 2014b).

Pour le SBV, les paires d'associations entre les malformations congénitales et les ruminants, caractérisées par la majorité des experts comme hautement spécifiques (sous-section 3.2.2.4.1), sont adaptées pour constituer des requêtes sur le Web. Les associations comme « limb malformations » AND « goat kids », « deformed offspring » AND « lambs », « deformed offspring » AND « calves », obtiennent une précision élevée pour collecter des articles sanitaires pertinents (précision maximale de 1,00). Sont moins spécifiques, les paires d'associations entre les signes cliniques reproductifs et les ruminants,

par exemple « reporoductive losses » AND « cattle », « aborted foetuses » AND « sheep », « aborted foetuses » AND « cattle », etc.

Pour la FA, les associations liées aux signes cliniques cutanés et muqueux chez les bovins et les porcins, évaluées par la majorité des experts comme hautement spécifiques (sous-section 3.2.2.4.1), sont adaptées pour la collecte sur le Web des articles pertinents (précision supérieure de 0,75). Nos résultats ont également montré que certaines requêtes liées à la maladie vésiculeuse chez les ruminants sauvages (cerfs) sont également adaptées pour des recherches automatiques sur le Web. Par exemple, les requêtes « vesicular disease » AND « deer », « production losses » AND deer », ont eu une précision supérieure à 0,77 pour identifier des articles pertinents. Ces requêtes ont essentiellement identifié des articles non liés à la FA mais sur d'autres maladies exotiques telles que la stomatite vésiculeuse ou la découverte d'un nouveaux poxvirus aux Etats-Unis. Par exemple, la stomatite vésiculeuse est une virose qui affecte les chevaux, les ruminants comme les bovins, les ovins et les membres de la famille du cerf et du lama ainsi que les porcs. En plus de causer de l'inconfort aux animaux infectés et d'entraîner des pertes de production pour les animaux vivants, la maladie revêt une grande importance en raison de sa ressemblance avec la FA qui affecte les ruminants et les porcs et qui peut dévaster les élevages (Rodriguez et al., 2000).

Pour la FCO, les associations les plus adaptées pour constituer des requêtes sur le Web sont celles liées aux signes cliniques généraux et reproductifs chez les ruminants domestiques telles que « onset of weaknes » AND « cattle », « embryonic death » AND « sheep », « embryonic death » AND « cattle » (precision de 1,00) et moins les associations, telles que « embryonic death » « deer », « livestock deaths » AND « cattle », « livestock deaths » AND « deer », « livestock deaths » AND « goats », « fever outbreak » AND « cattle », « fever outbreak » AND « sheep » (précision de 0,13 à 0,33) (Annexe H.3). Les articles pertinents collectés par des termes décrivant des signes cliniques chez les cerfs n'ont pas seulement identifié des cas de FCO mais également des cas de la maladie hémorragique épizootique (EHD) du cerf aux États-Unis. Les signes cliniques de cette maladie sont semblables à ceux du virus de la FCO. La maladie est très répandue dans le Sud-Est et le Centre Est des États-Unis où il y a des épizooties chaque année (Stevens et al., 2015).

Finalement, remarquons que parmi les paires d'associations représentatives des maladies modèles ayant le meilleur classement automatique, seuls 20 paires d'associations pour la PPA (37%), 18 paires pour la FA (58%), 12 paires pour la FCO (40%) et 16 paires pour le SBV (26%) ont identifié des pages Web pertinentes.

Notons que l'évaluation de la précision d'acquisition de pages Web était fondée sur les dix premières pages Web collectées (sauf pour la PPA), cf. Arsevska et al., 2016b. Ce corpus restreint de pages Web n'est pas suffisamment exhaustif pour donner des conclusions définitives. Dans nos futurs travaux, nous avons l'intention d'évaluer un nombre plus large de paires d'associations de termes. Pour ces évaluations, nous envisagerons également d'inclure les termes proposés par des experts.

Un deuxième facteur qui a probablement influencé la performance des différentes requêtes réalisées était la tendance du moteur de recherche à ne pas proposer d'articles anciens ($>$ 1 an). Dans ce contexte, dans nos futurs travaux, il semble nécessaire d'évaluer des requêtes de recherche en temps réel et éventuellement avec des associations peu complexes et plus génériques.

Les termes et leurs associations représentent les principaux aspects pour identifier des articles pertinents sur le Web. Cependant, cette étape du processus de fouille de textes, ne permet pas de gérer la masse d'information apportée par les pages Web acquises. Généralement, si une classification de documents est effectuée manuellement, sa réalisation est donc coûteuse en terme de temps. En effet, chaque texte (ou une partie) doit être lu manuellement pour attribuer une catégorie adaptée (classe). C'est la raison pour laquelle le domaine de la classification automatique de documents est de plus en plus utilisé dans la gestion de cette importante masse d'information disponible.

Dans la section suivante de ce manuscrit, nous présentons la troisième contribution du processus de fouille de textes pour la VSI en utilisant des méthodes de classification automatique des documents collectés sur le Web. Dans cette étape du processus, nous nous intéressons à la catégorisation de documents selon différentes classes : *i)* pertinentes et *ii)* non pertinentes. Nous nous intéressons aux algorithmes d'apprentissage supervisé et plus particulièrement aux méthodes fondées sur la représentation vectorielle de documents. Nous utiliserons deux algorithmes de classification automatique de documents afin de comparer les principales approches du domaine.

3.3 Classification automatique des documents

La classification de documents a pour objectif de regrouper les textes similaires, c'est-à-dire thématiquement proches, au sein d'un même ensemble. L'intérêt d'une telle démarche est d'organiser les connaissances de façon à pouvoir effectuer, par la suite, des tâches de recherche d'information (RI) ou d'extraction d'information (EI) efficaces (Figure 3.17).

3.3.1 Approche mise en œuvre

Dans l'approche que nous présentons dans cette section, deux étapes sont clairement identifiées. L'étape de représentation des données et la classification de documents.

Représentation vectorielle des documents textuels. L'exploitation et le traitement automatique des documents, en particulier pour les tâches de classification, nécessitent une première étape consistant à les représenter. Pour cela, la méthode la plus courante consiste à projeter les données textuelles (par exemple, les mots) dans un espace vectoriel (Salton, 1983). De nombreux travaux utilisent une telle approche. Citons par exemple Vinot et al., 2003 qui représentent les données à l'aide d'un modèle vectoriel pour une

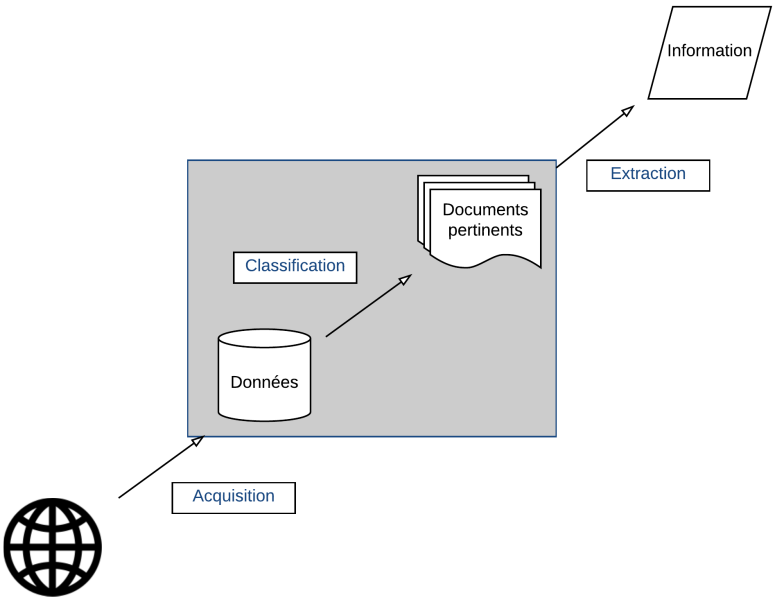


Figure 3.17 – Focus sur la deuxième étape du processus de fouille de textes pour la veille sanitaire sur le Web

tâche de classification de documents à contenus racistes. Delichère et al., 2002 et Bruno et al., 2002 utilisent quant à eux un modèle vectoriel pour calculer des scores de similarité entre différents documents. En épidémiologie, cette représentation est utilisée pour classifier des données textuelles non structurées tels que des notes médicales (MacRae et al., 2015 ; Martinez et al., 2015) ou des articles de médias électroniques (Lee et al., 2015 ; Zhang et al., 2009).

Le modèle vectoriel le plus couramment utilisé dans la littérature est le « Vector Space Model » de Salton, 1983, qui permet d'obtenir des performances proches d'une indexation manuelle. Le modèle de Salton consiste à représenter un corpus par une matrice telle que les lignes soient relatives aux descripteurs linguistiques (mots, termes, etc.) et les colonnes aux documents. Une cellule d'une telle matrice comptabilise la fréquence d'apparition d'un descripteur dans un document. Ainsi, la matrice formée peut être utilisée pour effectuer diverses tâches automatiques de fouille de textes (Munzert, 2015 ; Lombardo et al., 2006). En raison de sa nature simple et sa capacité à produire une représentation appropriée d'un corpus, nous nous appuyons sur la représentation vectorielle de documents de Salton, 1983 dans notre système de VSI.

Classification de documents. Le principe d'une classification automatique de textes est d'utiliser un modèle afin de classer un document dans une catégorie pertinente. Dans le domaine de la classification automatique, on distingue deux types d'approches : la classification supervisée et la classification non supervisée. Ces deux méthodes diffèrent sur la façon dont les classes sont générées. En effet dans le cas de la classification non supervisée, les groupes de documents (clusters) sont calculés automatiquement par la machine, tandis qu'ils sont, dans l'approche supervisée, définis en amont (étiquetés) par

un expert (Witten et al., 2005).

Dans le cadre de cette thèse, nous nous intéressons à la catégorisation de documents collectés du Web par apprentissage supervisé. Les données étiquetées sont les documents préalablement associés à une classe par un expert en épidémiologie, c'est-à-dire des documents « pertinents » qui décrivent des foyers de maladies ; et des documents « non pertinents » qui sont des bilans ou des descriptions générales de maladies.

Il existe de nombreuses méthodes de classification de données textuelles avec apprentissage supervisé. Dans cette thèse, nous nous appuyons sur les algorithmes (classificateurs) : Naïve Bayes (NB), et Machines à Vecteurs de Support (SVM). Ces deux algorithmes de classification sont utilisés avec succès par d'autres systèmes de biosurveillance, tels que les projets HealthMap (Freifeld et al., 2008), BioCaster (Conway et al., 2009), Argus (Torii et al., 2011) et le projet FMD-lab (Zhang et al., 2009), ainsi que la classification des messages des réseaux sociaux (Cui et al., 2015), des tweets (Adebayo, 2013) ou des rapports médicaux (Ye et al., 2014). Dans cette étape de notre processus, nous avons utilisé NB, « Discriminative Multinomial Naïve Bayes » (Su et al., 2008) et une version de SVM, « Sequential Minimal Optimization » en utilisant le kernel polynomial (Schölkopf et al., 1999).

Une évaluation rigoureuse des différentes méthodes de classification avec apprentissage supervisé s'appuie généralement sur l'utilisation du processus de « validation croisée ». Elle consiste à segmenter le corpus initial en n parties de même taille. En général, le nombre n de parties est fixé à dix : neuf parties pour l'apprentissage et une de test. Ainsi, les différents documents constituant le corpus deviennent alternativement corpus de test et d'apprentissage. Une telle méthode permet d'avoir une robustesse dans le choix de l'algorithme et des paramètres (Witten et al., 2005).

3.3.2 Évaluation et protocole expérimentale

Pour apprendre les modèles de classification, nous avons utilisé un corpus d'apprentissage préalablement établi à partir d'articles de Google (cf. sous section 3.1.2). Ces documents ont été manuellement étiquetés en catégories : « nouveaux cas », « bilan » et « général ». Dans le cas d'un corpus réduit (545 articles pour la PPA, 275 pour la FA, 220 pour la FCO et 352 pour le SBV), nous avons appliqué une méthode de validation croisée en dix plis.

Afin d'évaluer la performance des classificateurs, nous avons mesuré la capacité de NB et SVM à correctement catégoriser les articles selon le contenu dans les différentes classes (« nouveaux cas », « bilan » et « général »). Dans ce contexte, nous avons calculé la précision, le rappel, et la F-mesure. La précision (formule 3.10) indique la proportion des documents bien classés. Le rappel (formule 3.11) indique l'exhaustivité de la classification. Une précision de 100% signifie donc que tous les documents trouvés sont pertinents, un

rappel de 100% que tous les documents pertinents ont été trouvés. La F-mesure (formule 3.12) est la moyenne harmonique entre la précision et le rappel.

$$Précision_i = \frac{\textit{nombre des documents correctement attribués à la classe i}}{\textit{nombre total des documents attribués à la classe i}} \tag{3.10}$$

$$Rappel_i = \frac{\textit{nombre des documents correctement attribués à la classe i}}{\textit{nombre total des documents appartenant à la classe i}} \tag{3.11}$$

$$F - mesure = 2 \times \frac{\textit{précision} \times \textit{rappel}}{\textit{précision} + \textit{rappel}} \tag{3.12}$$

3.3.3 Résultats et Discussion

En comparant les performances de NB et SVM, nous relevons que les deux classificateurs obtiennent globalement de bonnes performances de prédiction, NB étant légèrement plus performant que SVM. De manière générale, les classificateurs ont mieux prédit les articles de catégories « nouveaux cas » (F-mesure de 0,64 à 0,842) et « général » (F-mesure de 0,724 à 0,859) en comparaison avec les articles de la catégorie « bilan » (F-mesure inférieure de 0,584) (Tableau 3.6).

Les meilleurs résultats pour la catégorie « nouveaux cas » ont été obtenus pour la FCO avec NB (F-mesure de 0,842) et SVM (F-mesure de 0,83) ; pour la catégorie « non pertinentes » pour la FA avec SVM (F-mesure de 0,859) et la PPA avec NB (F-mesure de 0,831) et SVM (F-mesure de 0,810) ; et pour la catégorie « bilan » pour la PPA avec SVM (F-mesure de 0,584) et NB (F-mesure de 0,503) (Tableau 3.6).

Dans des expérimentations similaires, pour le système de biosurveillance Argus, Torii et al., 2011 ont montré une performance supérieure du classificateur NB, pour les textes composés de 900 à 2700 mots (AUC de 0,841). Les textes de plus de 2700 mots dégradent la performance du classificateur NB. Notons que lorsque le texte contient plus de 2700 mots, la performance du classificateur SVM est supérieure (AUC de 0,836). Ceci est conforme à la propriété connue de SVM qui peut exploiter un grand nombre de descripteurs (mots). Cependant, le classificateur NB constitué par des descripteurs « généraux » donne une performance stable (Torii et al., 2011). Ceci peut expliquer la bonne performance de ce classificateurs dans nos expérimentations.

Tableau 3.6 – Performance des classificateurs

Maladie	Algorithme	Catégorie	Rappel	Précision	F-mesure
Peste porcine africaine	NB	Nouveaux cas	0,724	0,766	0,744
		Bilan	0,487	0,530	0,503
		Général	**0,860**	**0,804**	**0,831**
	SVM	Nouveaux cas	0,657	0,680	0,669
		Bilan	0,489	0,726	0,584
		Général	**0,864**	**0,763**	**0,810**
Fièvre aphteuse	NB	Nouveaux cas	0,750	0,685	0,716
		Bilan	0,318	0,438	0,386
		Général	**0,871**	**0,848**	**0,859**
	SVM	Nouveaux cas	0,655	0,625	0,640
		Bilan	0,341	0,484	0,400
		Général	**0,823**	**0,776**	**0,799**
Fièvre catarrhale ovine	NB	Nouveaux cas	**0,837**	**0,846**	**0,842**
		Bilan	0,053	0,059	0,056
		Général	0,761	0,741	0,751
	SVM	Nouveaux cas	**0,793**	**0,869**	**0,830**
		Bilan	0,013	0,019	0,016
		Général	0,771	0,689	0,727
Schmal-lenberg	NB	Nouveaux cas	0,635	0,662	0,648
		Bilan	0,364	0,471	0,410
		Général	**0,746**	**0,703**	**0,724**
	SVM	Nouveaux cas	0,644	0,650	0,637
		Bilan	0,353	0,460	0,401
		Général	**0,735**	**0,692**	**0,713**

Concernant la performance faible des deux classificateurs pour les articles de la catégorie « bilan », une des explications pourrait être liée au contenu des articles qui contiennent des informations sanitaires relatives à plusieurs classes. Par exemple, certains articles décrivent des mesures de contrôle (catégorie « bilan ») et des foyers de maladies (catégorie « nouveaux cas »). D'autres articles décrivent à la fois différentes mesures de contrôle et d'éradication des foyers (catégorie « bilan ») ou l'influence sociologique ou économique des foyers pour le pays et les éleveurs (catégorie « bilan ») mais également une description générale de la maladie (catégorie « général »). Des défis similaires relatifs à la classification des classes multiples ont également été relevés par Zhang et al., 2009 pour la classification d'articles sur la fièvre aphteuse (FA) dans le cadre du projet FMD BioPortal.

Une méthode d'amélioration de la performance des classificateurs a été proposée par Zhang et al., 2009. Dans les expérimentations réalisées, les auteurs ont noté que la performance des classificateurs SVM et NB varie selon le type de représentation utilisée. Dans le cas de représentation vectorielle de documents, la classification avec SVM a légèrement surpassé celle de NB (F-mesure de 0,722 versus 0,688 ; respectivement). Une

amélioration de la performance de deux classificateurs est obtenue avec une combinaison de la représentation vectorielle, des syntagmes nominaux et des entités nommées (EN, « Named Entities »). Dans ce cas, NB obtient une meilleure performance (F-mesure de 0,757 et 0,744, pour NB et SVM, respectivement). Des travaux de Tolle et al., 2000 et de Wei et al., 2004 mettent en relief que l'utilisation des syntagmes nominaux (groupes de mots nominaux) comme descripteurs est souvent plus précise. Par exemple, le syntagme « mortalité porcine » est sémantiquement plus pertinent que la prise en compte des mots « mortalité » et « porcine » de manière indépendante. Ainsi, dans la suite de notre travail, nous envisagerons d'utiliser les termes obtenus par *BioTex* (sous-section 3.2.1.1) comme descripteurs pour la tâche de classification proposée dans cette section.

Une troisième technique, qui est un prolongement de l'association des syntagmes nominaux, s'appuie sur l'utilisation des EN. Ceci consiste à sélectionner les noms propres d'un article qui relèvent de catégories bien définies. Ce processus peut utiliser une hiérarchie lexicale sémantique ainsi qu'un processus d'annotation sémantique et syntaxique (Conway et al., 2009 ; Doan et al., 2009 ; Jimeno et al., 2008 ; Volkova et al., 2010 ; Zhang et al., 2007). En utilisant cette technique, la performance de la classification aurait tendance à s'améliorer (F-mesure de 0,824 pour NB et 0,856 pour SVM) ; en comparaison aux résultats de la classification avec la représentation vectorielle des documents (F-mesure de 0,789 pour NB et 0,801 pour SVM) (Doan et al., 2009). De la même manière, Zhang et al., 2007 ont constaté une amélioration de la performance de NB et SVM à partir d'un corpus de rapports de ProMED et de pages Web de *Google news* en utilisant des entités nommées avec une F-mesure plus élevée (0,76 v.s 0,70).

La représentation des EN, utilisée par le système de biosurveillance BioCaster, est fondée sur la combinaison de rôles (catégories sémantiques des maladies associées à des noms et des verbes). La combinaison des EN et des rôles améliore la performance de NB et SVM pour la classification (F-mesure de 0,838 pour NB et 0,857 pour SVM) en comparaison aux méthodes classiques. Par ailleurs, la combinaison des EN et des catégories sémantiques de noms et verbes contribuent à une amélioration plus importante de la classification (F-mesure de 0,853 pour NB et 0,912 pour SVM) en comparaison avec la méthode de combinaison des EN et des rôles (Doan et al., 2009).

Par ailleurs, une autre raison qui aurait pu influencer les résultats de classification est le déséquilibre des catégories d'intérêt au sein du corpus d'apprentissage (Amrine et al., 2014). Lors de nos expérimentations, les articles de la catégorie « bilan » ont été sous-représentés contrairement aux catégories « nouveaux cas » et « général ». L'utilisation des méthodes de ré-échantillonnage (Elrahman et al., 2013 ; Heredia-Langner et al., 2015), pourrait améliorer la performance de la classification. Par exemple, un tel principe est appliqué pour la classification de tweets, de messages des réseaux sociaux (Adebayo, 2013 ; Tuarob et al., 2014 ; Zuccon et al., 2015) ou des dossiers médicaux (Doan et al., 2012).

En termes de temps de traitement, nous estimons que la classification automatisée que

nous proposons a des avantages significatifs pour le Dispositif de la VSI en France. Nous souhaitons approfondir son évaluation et l'étendre pour l'intégrer dans le dispositif de VSI (le système PADI-Web) en temps-réel que nous avons réalisé (cf. section 3.5).

Parmi la classification des documents, les techniques de fouille de textes sont aujourd'hui largement utilisées pour découvrir des connaissances pertinentes à partir des données textuelles. Dans ce contexte, la tâche d'extraction d'information (« information extraction », EI) consiste à identifier les éléments pertinents dans des ressources textuelles peu ou non structurées (Piskorski et al., 2013). Pour la VSI en France, nous nous intéressons à la meilleure utilisation des sources d'informations textuelles non structurées pour aider l'expert en charge de la veille à connaître la situation sanitaire au niveau international. La masse d'informations aujourd'hui disponible, même sur un sujet bien défini (*e.g.*, la situation épizootiologique actuelle de la PPA dans les pays Baltes et l'Europe de l'Est), est très importante et généralement les experts n'ont pas les moyens d'effectuer une lecture exhaustive. Par ailleurs, ils ne peuvent pas omettre l'information importante. Pour aider ces experts, dans la section suivante, nous proposons une méthode d'extraction automatique des événements sanitaires à partir d'articles médiatiques publiés sur le Web.

3.4 Extraction automatique d'information

L'extraction d'information (EI) consiste à analyser un texte de manière automatique afin d'en extraire un ensemble d'informations jugées pertinentes (Poibeau, 2003).

L'événement sanitaire étant l'objet central de nos travaux, il est nécessaire de définir plus précisément ce concept. Afin de proposer une représentation plus formelle d'un événement, nous nous appuyons sur les quatre principaux indicateurs utilisés pour décrire une épizootie : la cause, l'espèce, le lieu et le temps (Collier et al., 2008). Nous définissons un événement comme la combinaison d'une propriété sémantique (maladie, hôte, signes cliniques), un intervalle temporel (date) et une entité spatiale (lieu) (Serrano et al., 2013). Dans notre cas, la propriété sémantique est définie par les hôtes et les signes cliniques suite à une émergence sanitaire, la composante temporelle constitue la date ou période d'occurrence d'un événement et l'entité spatiale correspond à son lieu d'occurrence.

En se basant sur ces postulats, dans cette section, nous proposons une approche d'extraction automatique d'information depuis des pages Web collectées (Figure 3.18), ce qui constitue notre quatrième contribution. Il s'agit d'une méthode d'EI à partir de documents sanitaires ayant un format non structuré.

Un des objectifs clé de cette étape est la standardisation de l'information extraite à partir de données non structurées disponibles sur le Web. Ceci permettra de comparer les informations extraites avec celles issues des données structurées proposés par les organismes de surveillance traditionnelles (OIE, FAO et ADNS). En résumé, ceci pourra au final produire un ensemble de données comparables et adaptées pour une analyse épidémiologique complète des événements sanitaires au niveau international.

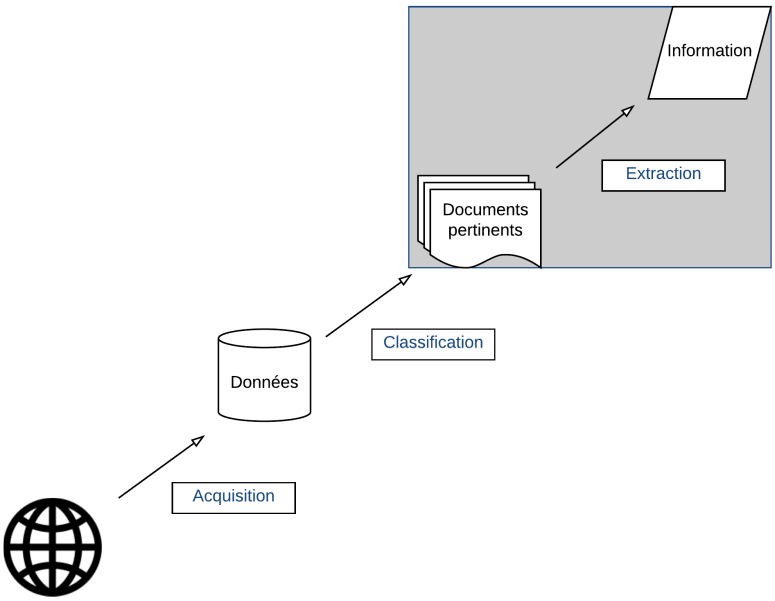


Figure 3.18 – Focus sur la troisième étape du processus de fouille de textes pour la veille sanitaire sur le Web

Deux approches pour l'EI émergent : l'extraction fondée sur les techniques linguistiques et les systèmes statistiques.

Les tâches les plus communes en EI sont l'extraction d'entités nommées et des relations entre entités et événements, NER (« named entity recognition ») (Song et al., 2015). Le NER traite le problème d'identification (détection) et de classification d'éléments prédéfinis, tels que des organisations (*e.g.,* « Organisation mondiale de la santé »), des lieux (*e.g.,*« Russie »), des expressions temporelles (*e.g.,* « 1er septembre 2011 »), des expressions numériques (« 20 cas de la peste porcine africaine »), etc.

Les méthodes linguistiques à base de dictionnaires et d'ontologies, bien que généralement très précises, ont pour principales faiblesses d'être spécifiques à un domaine donné et d'avoir un taux de rappel plutôt faible et un coût de développement manuel élevé (Chiang et al., 2005). Par exemple, « This » est à la fois le nom d'une ville en France et un pronom en anglais. L'extraction fondée sur des dictionnaires est une des techniques les plus utilisées en EI du domaine biomédical (Song et al., 2015). La première étape est la création d'un dictionnaire. Ce dernier est ensuite utilisé pour extraire l'information recherchée. Les dictionnaires peuvent varier entre ceux créés par les utilisateurs ou ceux déjà existants, par exemple la nomenclature SNOMED pour extraire les noms de pathologies humaines (Jindal et al., 2013). Le thésaurus AGROVOC peut être utile pour extraire des entités et leurs relations pour les pathologies en santé animale (Pazienza et al., 2012). GeoNames [3] est un gazetier multilingue de données géo-spatiales avec plus de

3. http://www.geonames.org/

dix millions localisations, en incluant leurs coordonnées spatiales, pays et régions administratives (Uzaman et al., 2010). HeidelTime est un outil d'annotation temporelle multilingue pour l'extraction et la normalisation des expressions temporelles (Strötgen et al., 2015). Au cours des dernières années, les dictionnaires sont souvent remplacés par des ontologies, adaptables aux différents domaines. Citons par exemple les ontologies pour les données de l'industrie alimentaire (Touhami et al., 2015 ; Dibie et al., 2016), ainsi que pour des maladies humaines, animales et leurs pathogènes, etc. (Collier, 2010, Volkova et al., 2010).

Du côté des approches statistiques, Ahn, 2006 propose de combiner plusieurs classificateurs pour l'extraction des événements. L'apprentissage statistique permet de prendre en compte de nombreux contextes d'apparition mais nécessite une grande quantité de données annotées pour être performant. En outre, les méthodes d'apprentissage de patrons ou les approches semi-supervisées semblent intéressantes comme par exemple le système développé par Serrano et al., 2013.

Comme toutes ces approches prises séparément restent imparfaites, nous proposons une approche hybride permettant d'exploiter les points forts des méthodes classiques. Notre approche est fondée sur les deux approches actuelles en EI : la première s'appuie sur des règles linguistiques construites manuellement et la deuxième se fonde sur un apprentissage automatique de patrons linguistiques. Dans un contexte épidémiologique, une règle simple pourrait être : « un nombre qui apparaît juste après un nom au pluriel représente le nombre de cas », ou encore, « le nom de la localisation quand elle est précédée des mots *infected in* représente le nom du lieu du foyer ». Afin de minimiser le travail des experts pour obtenir un nouvel ensemble de règles, notre approche combine également des méthodes d'apprentissage supervisé à partir d'un corpus de documents annotés par des experts (Tang et al., 2008).

3.4.1 Approche combinée d'extraction automatique d'événements

Comme mentionné précédemment, notre approche est fondée d'abord sur l'extraction automatique des règles en utilisant des techniques de fouille de données. Pour chaque information, nous nous appuyons d'abord sur des dictionnaires qui fournissent des candidats (termes à extraire) à notre méthode : le dictionnaire Geonames, utilisé pour identifier de noms de localisations (pays, régions, villes, villages, etc.) ; HeidelTime, utilisé pour marquer toutes les dates dans le texte ; des dictionnaires de noms de maladies, des signes cliniques et des hôtes obtenues avec le processus de fouille de textes (sous-chapitre 3.2.1.1). Les candidats pour un nombre d'espèces affectées (nombre de cas) sont listés avec une expression régulière qui reconnaît les chiffres (au format numérique ou sous forme de texte). Ensuite, les types de règles découvertes automatiquement à partir d'un corpus manuellement annoté sont utilisés comme traits (descripteurs) dans un modèle de classification, dans notre cas, l'algorithme SVM.

3.4.1.1 Extraction automatique de règles

La découverte automatique de règles est fondée sur la sélection de règles d'association. Cette technique de fouille de données permet de découvrir des corrélations entre différents éléments dans de grands volumes de données. Elle est utilisée, par exemple, pour l'identification des signaux d'émergence des maladies à partir de flux RSS (Collier et al., 2012, Keller et al., 2009a) ou à partir de tweets (Tuarob et al., 2014, Velardi et al., 2014, Santos et al., 2014).

Dans notre contexte, nous voulons identifier les règles qui indiquent si les entités candidates identifiées, entre autres, sur la base de dictionnaires, sont pertinents ou non pertinentes. Par exemple, si un foyer apparaît à une date donnée et que l'article mentionne que les autorités déclarent un événement sanitaire trois jours plus tard, alors la date correspondante sera étiquetée en tant que correcte. Un candidat non pertinent n'a pas de lien avec l'événement correspondant.

Dans ce cadre, chaque candidat est associé à un ensemble d'éléments de plusieurs types. Pour expliquer les différents types d'éléments à identifier, nous considérons la phrase suivante, où le mot ASF qui est candidat au processus d'extraction d'information a été reconnu comme nom de maladie (African swine fever) :

« 12 pigs have been infected by ASF in Poland »

Éléments liés au mot. Chaque mot se situant près du candidat est encodé comme un élément qui décrit à la fois le mot lui-même et sa position relative par rapport au candidat. Par exemple, l'élément (*infected, -2*) signifie que le mot « *infected* » se trouve deux mots (tockens) avant le nom candidat d'une maladie.

La position d'un élément peut également être exprimée selon la position relative de la phrase ou du paragraphe où il se trouve. Par exemple, (*infected, -1sent*) identifie que le mot « *infected* » se situe dans une phrase précédant la phrase candidate.

Nous pouvons également adapter la définition de la position d'un élément pour qu'elle devienne plus souple, par exemple *(infected, -1 to -3)* signifie que le mot « *infected* » se trouve d'un à trois mots avant un candidat.

Éléments liés à l'abstraction des mots. En plus du mot lui-même, chaque mot peut être abstrait grâce à plusieurs types d'informations.

Par exemple, un mot peut être associé à sa fonction grammaticale ou à son lemme (la forme canonique d'un mot ; *e.g.,« pig »* est le lemme « *pigs* »). Ils sont tous les deux obtenus grâce à l'étiquetage morphosyntaxique produit par TreeTagger (Schmid, 1994).

Par exemple, l'élément (*location, +2*) signifie que le nom de la localisation a été trouvé deux tockens après le candidat « *ASF* » et les éléments *(verb past participle, -2)* et (*lemma : infect, -2*) mentionnent, respectivement, qu'un participe passé s'est trouvé à deux tockens avant notre candidat et ce verbe est « *to infect* ».

Éléments liés à la position. Nous avons également notifié la position du candidat dans le texte entier. Ces positions peuvent également être exprimées en termes de position au sein des paragraphes. Par exemple, l'élément (*position, 0% - 10%*) signifie que le candidat « *ASF* » se trouve dans les dix premiers pour cents du document (10%), tandis que (*position, PAR1*) signifie qu'il est dans le premier paragraphe du document.

Ainsi, chaque candidat du corpus pourra être associé à un ensemble d'éléments décrivant les informations contextuelles. À partir de ces données, nous avons appliqué un algorithme de découverte de règles d'association capable de générer automatiquement des règles qui représentent des descripteurs pour chaque classe (correct ou non correct) (Soderland, 1999).

Par exemple, dans la classe qui correspond au « *nombre d'animaux infectés* », la règle suivante a été découverte automatiquement : *(Killed, -1 to -3), (position, PAR1)*, confiance de 83% et fréquence de 26%. Cette règle comprend trois informations principales :

1. « *(killed, -1 to -3), (position, PAR1)* » : Ces éléments décrivent le contenu de la règle. Ici, le mot « *killed* » se trouve à un mot sur 3 avant un nombre (*i.e.*, candidat pour un nombre de cas), dans le premier paragraphe du document.
2. « *Confiance de 83%* » représente la fiabilité de la règle, qui est directement liée à sa qualité. La confiance est la probabilité que le candidat qui couvre cette règle soit un candidat pertinent (correct). En d'autres termes, la règle peut être traduite de la manière suivante : *si le mot « killed »* apparaît de 1 à 3 mots avant un nombre candidat dans le premier paragraphe, alors ce candidat représente le nombre d'animaux infectés dans 83% des cas.
3. « *Fréquence de 26%* » indique le nombre correct de candidats respectant cette règle. Plus la règle est fréquente plus elle sera utile sachant qu'elle couvre un nombre plus important de cas.

Comme les résultats des algorithmes retournent un grand nombre de règles, nous avons effectué une sélection de ces dernières pour chaque classe de candidats. Nous avons alors conservé les règles qui ont la meilleure confiance et qui ne sont pas redondantes (deux règles sont considérées comme redondantes si elles couvrent les mêmes cas).

3.4.1.2 Classification de nouveaux candidats

Une fois les règles extraites pour chaque type (*i.e.*, localisation, date, nombre d'animaux infectés, etc.) et pour chaque classe (*i.e.*, correct, incorrect), une classification est effectuée pour classer les candidats (extraits de l'article).

Dans ce contexte, les règles d'association extraites automatiquement sont utilisées en tant que descripteurs dans l'approche d'apprentissage supervisé. L'algorithme de classification choisi est SVM. L'algorithme est ensuite évalué en suivant un principe de validation croisée. Les résultats du modèle SVM donnent une prédiction (correcte ou incorrecte) ainsi qu'une probabilité estimée. Par exemple, prédire la localisation d'un candidat comme correcte avec une probabilité estimée de 75% signifie que le modèle SVM estime que la probabilité que le candidat se réfère à la localisation d'un événement est de 75%.

Il est essentiel de remarquer que d'autres types de classificateurs que SVM peuvent être utilisés pour l'étape de classification. Certains d'entre eux ont été testés (Arbres de décision, Forêts d'arbres décisionnels, Naïve Bayes). SVM a été sélectionné au regard de son bon comportement dans la littérature (Conway et al., 2009 ; Doan et al., 2009 ; Torii et al., 2011 ; Zhang et al., 2009) et aux nos données.

3.4.1.3 Classification de noms de lieux

Contrairement à d'autres types d'informations, la classification de lieux requiert une étape supplémentaire. En effet, pour la localisation d'un candidat donné, il n'est pas rare que ce nom corresponde à plusieurs entités géographiques (ambiguïtés), localisées dans des lieux différents (parfois dans des continents différents). Par exemple, selon le gazetier GeoNames, le toponyme « Paris », se réfère à plus de soixante localisations différentes au niveau mondial. Plus de 46% de tous les toponymes de GeoNames ont plus d'une référence (Klopotek et al., 2013).

Dans ce but, pour prédire les meilleures localisations parmi les localisations potentielles, nous avons appris un autre modèle de classification SVM. Pour la description des entités géographiques, nous avons utilisé quatre autre descripteurs :

— la fréquence des entités géographiques du candidat dans le document parmi tous les candidats des localisations dans le document,
— la fréquence des entités géographiques du pays du candidat dans le document parmi tous les candidats de localisation dans le document,
— la position des entités géographiques du candidat dans le document parmi tous les autres candidats,
— le nombre total de candidats d'entités géographiques dans le document.

En suivant le même principe que la classification fondée sur les règles, chaque entité géographique ambiguë est alors prédite comme correcte ou non.

3.4.2 Évaluation et protocole expérimental

Pour mettre en œuvre notre méthode fondée sur la découverte automatique de règles, il était nécessaire de constituer un corpus de données étiquetées avec un nombre suffisant d'exemples (corrects et incorrects) pour chaque type d'entité.

Nous avons donc collecté 532 articles de média (articles) qui correspondent aux foyers déclarés à l'OIE en 2014 et 2015. Ce corpus a été utilisé à la fois pour l'apprentissage et l'évaluation. Pour chaque déclaration à l'OIE, une requête a été créé sur *Google news* pour acquérir des articles de média électronique qui ont été publiés entre la date d'observation d'un événement et sa date de rapport et dont le titre contient à la fois le nom de la maladie et le nom du pays.

Pour chaque recherche, les dix premiers articles ont été collectés (ou bien tous les articles lorsque la recherche a retourné moins de dix articles). Parmi les 532 articles, 352 articles ont été manuellement étiquetés en tant qu'articles pertinents (voir les critères d'évaluation d'un article pertinent dans la sous-section 3.1.2). Pour chaque article pertinent, les informations des candidats ont été identifiées automatiquement (en utilisant les dictionnaires définis pour chaque type d'information) et ont été manuellement annotées comme étant soit correctes ou incorrectes.

Pour évaluer la qualité de l'extraction, pour chaque classe, nous avons mesuré l'exactitude (« accuracy »), la précision, le rappel et la F-mesure (Piskorski et al., 2013). Les définitions de ces mesures sont présentées dans le Chapitre 3.3.2. Ci-dessous nous définissons l'exactitude de la manière suivante :

$$Exactitude = \frac{nombre\ de\ candidats\ correctement\ classés}{nombre\ total\ de\ candidats} \tag{3.13}$$

Pour l'évaluation détaillée dans la section suivante, une validation croisée a été réalisée avec dix plis.

3.4.3 Résultats et Discussion

Les premiers résultats sur le corpus annoté de 352 articles montrent que les scores d'exactitude pour l'extraction d'informations spatiales dans les dépêches est d'environ 0,8 alors que l'exactitude propre à l'extraction des autres entités est comprise entre 0,85 et 0,96 (Tableau 3.7).

Tableau 3.7 – Performance de l'approche de fouille de textes pour l'extraction d'entités sanitaires

Entités	Exactitude méthode de base	Exactitude méthode combinée	Précision méthode combinée	Rappel méthode combinée	F-mesure méthode combinée
Lieux	0,60	0,80	0,81	0,80	0,80
Date	0,86	0,88	0,82	0,88	0,83
Maladie	0,97	0,96	0,95	0,96	0,95
Nombre de cas	0,45	0,85	0,86	0,85	0.85
Espèce	0,97	0,96	0,94	0,96	0,95

Nous pouvons constater que notre approche combinée (*i.e.*, règles apprises par fouille de données) obtient une très bonne précision globale et que comme attendu, le rappel

est meilleur en comparaison avec la méthode de base (*i.e.,* règles fondées sur des patrons simples et des dictionnaires). Ce qu'il faut retenir de ces expérimentations est que la méthode combinée obtient une exactitude nettement supérieure (près de 30 points par rapport à la méthode de base), ce qui dénote une amélioration globale de la qualité d'extraction pour tout type d'entités.

Les résultats montrent que notre approche était la plus performante pour l'extraction des noms de maladies et des espèces. Ceci peut s'expliquer par le vocabulaire relativement stable utilisé pour décrire une maladie ou des hôtes (Collier, 2012). Cependant, même si le vocabulaire pour les maladies connues est stable, il existe une émergence non négligeable de nouvelles maladies chez de nouveaux hôtes (*e.g.,* la découverte du virus de Schmallenberg) (Santamaria et al., 2012). Nos dictionnaires de termes créés par le processus de fouille de textes et complétés par les propositions des experts nous permet d'avoir une liste de termes plus exhaustive. Par exemple, comme Schmallenberg est une maladie animale infectieuse récente, pour elle, il n'y a aucune définition dans MeSH ou AGROVOC (Ferreira et al., 2012). Ainsi, notre approche nous permet d'avoir des termes non présents dans les ressources usuelles.

Nos résultats pour l'extraction des entités spatiales sont également meilleurs que ceux obtenus par Volkova et al., 2010 et Keller et al., 2009a. Volkova et al., 2010 appliquent une méthode d'extraction des entités nommées (espèces, noms de maladies et agents pathogènes) en utilisant des ressources existantes (OIE, Wikipédia) et une approche d'apprentissage sémantique des relations pour l'extraction des événements (maladie dans un lieu donné). En utilisant ces techniques, une précision de 0,76 et un rappel de 0,56 pour l'extraction de noms de maladies et une exactitude de 0,65 pour l'identification des événements ont été obtenus. Keller et al., 2009a ont utilisé une méthode d'extraction d'événements en appliquant un « réseau neuronal artificiel » appris à partir d'un corpus d'expressions d'entités spatiales. Les résultats obtenus (c'est-à-dire la précision de 0,61, rappel de 0,68 et F-mesure de 0,64) sont largement inférieurs aux nôtres.

Cependant, nos résultats sont très proches de ceux obtenus par Chanlekha et al., 2010 et Collier, 2012 qui ont utilisé trois algorithmes d'apprentissage automatique (SVM, CRF et Arbres de décision) pour extraire des événements, avec une meilleure performance d'algorithme obtenue avec les CRF[4] (précision de 0,86, rappel de 0,58 et F-mesure de 0,86). Ces auteurs suggèrent que le véritable lieu d'un événement est souvent mentionné autour d'un candidat (comme nous le proposons également). Selon Collier, 2012, les résultats variables de la performance de l'extraction d'entités spatiales peuvent s'expliquer par la variété des contextes dans lesquels les expressions géographiques pour des foyers apparaissent. L'information contextuelle pour décider si l'un des nombreux lieux mentionnés dans un article est le lieu correct d'un événement demande une interprétation contextuelle du document dans son ensemble (sans se restreindre à la seule phrase courante).

4. Conditional random fields

Concernant l'extraction des entités temporelles, nous avons choisi le système HeidelTime (fondé sur des règles). Les évaluations d'HeidelTime par Strötgen et al., 2010 montrent une exactitude de 0,96 pour la reconnaissance correcte des dates et 0.86 pour l'extraction correcte de valeurs appropriées aux dates. D'autres systèmes qui utilisent les règles d'extraction, comme SUTime (Chang et al., 2013) ont obtenu des résultats du même ordre, tels qu'une F-mesure de 0,90, une précision de 0,89 et un rappel de 0,91. Le système TimeText (Jindal et al., 2013) retourne une exactitude de 0,84 pour l'extraction des entités temporelles à partir des dossiers médicaux. Les principales erreurs d'extraction des entités temporelles rapportées par ces auteurs sont liées à la désambiguisation des expressions temporelles qui n'ont pas une valeur associée comme : « this time, that time, a couple of weeks, the ensuing days », etc. ou des expressions comme : « digital age, each season », etc. Cependant, Strötgen et al., 2010 relèvent une amélioration de la performance de la version étendue d'HeidelTime pour plusieurs langues et notamment pour le français et l'espagnol. Ceci nous garantira une bonne évolution de notre système (détaillé en section 3.5) qui prend en compte ces langues.

Dans cette section du manuscrit, nous avons proposé une méthode combinant deux approches pour l'extraction automatique d'événements. Nos résultats montrent que la méthode combinée améliore significativement la qualité des événements extraits. Les résultats obtenus sont encourageants et nous invitent à explorer de nouveaux modes d'hybridation afin de mieux extraire les lieux, les dates et les nombres de cas (le but étant d'améliorer le taux de rappel sans trop perdre en précision). Finalement, l'approche que nous avons présentée dans cette section est implémentée dans la plateforme de fouille de textes sur le Web, PADI-web.

3.5 Plateforme pour l'extraction automatique d'information sanitaire sur le Web

Cette section décrit l'implémentation du processus de fouille de textes pour la VSI en santé animale dans la plateforme PADI-web [5], conçue dans le cadre de cette thèse. Cette partie du manuscrit constitue donc la cinquième contribution de ce travail, la détection en temps réel des émergences des maladies animales infectieuses au niveau international.

La plateforme PADI-web est développée de façon générique. Elle peut être considérée comme un système de biosurveillance (informel) qui permet : *i)* de collecter automatiquement et en temps réel des documents non structurés (articles de média) obtenus par des requêtes sur le Web, *ii)* de nettoyer le contenu non désiré et *iii)* d'extraire et *iv)* de visualiser l'information sanitaire.

L'objectif principal du système est la détection précoce des dangers sanitaires au niveau international. À notre connaissance, PADI-web est le premier systèmes de biosurveillance en santé animale destiné aux autorités vétérinaires d'un pays. Il n'a pas pour objectif de

5. Platform for automated extraction of disease information from the Web

remplacer l'information sanitaire en provenance de systèmes de surveillance traditionnelles mais de les compléter.

Actuellement, le système est appliqué aux quatre maladies animales infectieuses : la PPA, la FCO, la FA, le SBV et en complément l'influenza aviaire (IA). Le système a été initialement développé pour traiter les documents en anglais ; depuis mars 2016, le système PADI-web a été adapté à deux autres langues : le français et l'espagnol.

3.5.1 Acquisition automatique de données

Pour l'acquisition automatique des articles du Web, le système PADI-web utilise les données issues de flux RSS (« Really Simple Syndication ») dont le contenu est produit automatiquement, en fonction des ajouts et mises à jour provenant d'un site Web. Par exemple, si un site Web dispose d'un flux RSS et publie un nouvel article, il apparaîtra sur le flux avec son titre, la date, l'heure, une brève description et l'URL menant à la page Web de l'article. Plusieurs systèmes de veille comme HealthMap (Freifeld et al., 2008) et BioCaster (Collier et al., 2008) utilisent des flux RSS pour obtenir des données.

Par rapport aux autres systèmes de veille qui explorent une liste des pages Web spécifiques, par exemple, les médias spécialisés ou généraux, sites ministériels, etc. (Collier et al., 2008 ; Rortais et al., 2010 ; Uzaman et al., 2010), PADI-web est plus générique et se base sur des mots-clés qui génèrent des flux RSS. Ces derniers représentent des alertes liées à l'occurrence des maladies et des associations entre termes correspondant à des syndromes observés sur des hôtes spécifiés (Annexe I). Actuellement, le système PADI-web contient plus de soixante flux RSS créés pour *Google news* agrégateur d'articles en anglais.

3.5.2 Pré-traitement et pré-filtrage des données

Quand un nouvel article est détecté par PADI-web, avant d'être stocké dans une base de données, il est pré-traité et normalisé (suppression des balises HTML et Java Script, reconnaissance de la langue, etc.). Une étape de pré-filtrage permet de sélectionner les articles potentiellement pertinents. Pour ce faire, le système s'appuie sur un dictionnaire de mots-clés (Annexe I.1) extraits par le processus de fouille de textes à partir de documents pertinents sur des foyers de maladies ainsi que des termes qui décrivent les maladies, les hôtes et les signes cliniques (Annexe I.2, I.3 et I.4). Ces dictionnaires sont également utilisés pour la tâche d'extraction d'information. Pour catégoriser les articles, le système s'appuie sur une base de données relationnelle (MySQL) qui va associer l'article avec les différents dictionnaires de maladies, signes cliniques et hôtes (cf. Figure 3.19).

Le dictionnaire de termes des maladies (*e.g.*, warthog disease) est associé à une catégorie générique (*e.g.*, african swine fever). Actuellement, ce dictionnaire contient 19 termes pour les cinq maladies (Annexe I.2).

Le dictionnaire de termes propres aux hôtes (*e.g.*, sow, piglet, pig, boar, etc.) est associé à une catégorie d'espèce (*e.g.*, porcine). Cinq catégories d'espèces sont définies : *i)* aviaire, *ii)* bovine, *iii)* ovine, *iv)* caprine, *v)* porcine et *vi)* autres. Actuellement, ce dictionnaire contient 65 termes (Annexe I.3).

Le dictionnaire de termes de signes cliniques (*e.g.*, mortality, letality, death, etc.) est lié à une catégorie de syndrome (mortality). Huit catégories de syndromes sont ainsi définies : *i)* généraux (fièvre, mortalité ou autre), *ii)* respiratoire, *iii)* digestifs, *iv)* locomoteurs/ nerveux, *v)* cutanés/ muqueux, *vi)* hémorragiques, *vii)* reproducteurs et *viii)* périnatales/ congénitales. Actuellement, ce dictionnaire contient 140 termes (Annexe I.4).

L'interface actuelle du PADI-web permet à l'analyste de faire des recherches avancées en combinant différents critères (maladies, hôtes, symptômes, sources et dates de publication des articles) (cf. Figure 3.20).

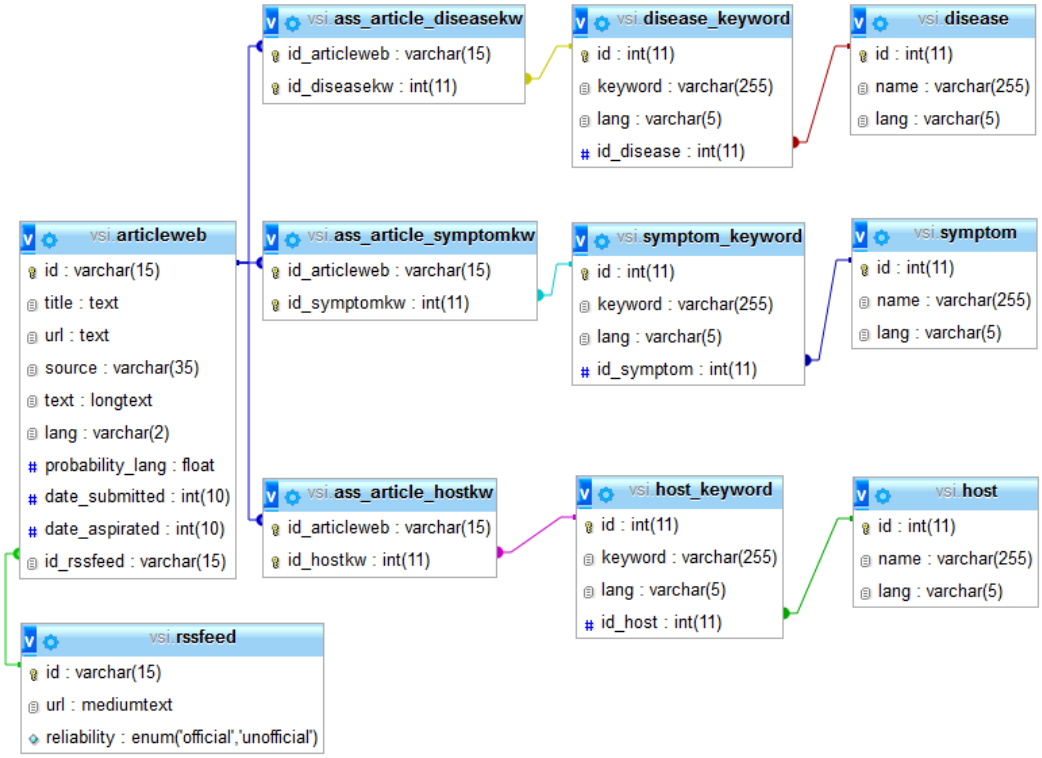


Figure 3.19 – Modèle relationnel d'association d'un article avec les différents dictionnaires pour le système PADI-web

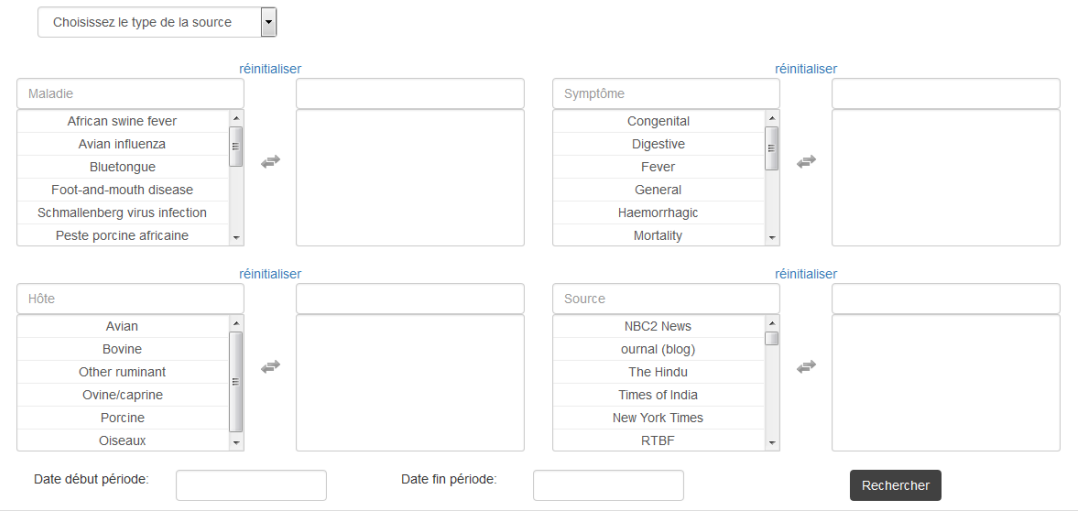


Figure 3.20 – Critères pour des recherches avancés sur l'interface du système PADI-web

3.5.3 Extraction automatique d'information

L'approche de fouille de textes permet l'extraction d'information sanitaire, telle que la maladie, les hôtes, les signes cliniques, le lieu d'événement, les dates, le nombre de cas, etc. (cf. section 3.4).

Un événement sanitaire doit obligatoirement être associé à un lieu (représenté par des coordonnées géographiques, une zone administrative, un pays), une date, ainsi qu'une espèce touchée, avec soit une maladie confirmée, soit des signes cliniques. Un article peut contenir plusieurs événements sanitaires. Le degré de confiance attribué à chaque lieu est déterminé automatiquement en utilisant l'algorithme SVM (section 3.4) et varie entre 0 et 1.

L'information extraite est organisée sous format tabulaire et structurée en prenant en compte les mêmes variables que pour les données structurées provenant de sources officielles (*e.g.*, ADNS, OIE). Cette organisation de l'information facilite l'analyse des données par des épidémiologistes. L'Annexe I.5 présente un exemple d'information extraite pour la PPA pour la période du 1er janvier au 28 juin 2016.

3.5.4 Visualisation de l'information sanitaire

Actuellement le système PADI-web comprend une interface Web qui permet à l'épidémiologiste une analyse descriptive des événements sanitaires des dernières 24 heures, ainsi que de tous les autres événements sanitaires générés depuis la mise en place du système (archives). L'analyse descriptive consiste en une représentation géographique des derniers événements sanitaires pour un article donné (Figure 3.21) ainsi qu'un histogramme avec la fréquence d'articles publiés par mois pour une maladie donnée (Figure 3.22).

Finalement, l'analyste est un contributeur direct au processus de fouille de textes pour PADI-web et peut influencer et améliorer sa performance en : *i)* paramétrant les flux RSS et les dictionnaires pour une meilleure acquisition d'articles pertinents, *ii)* en évaluant le contenu d'articles recueillis ainsi que les annotations automatiques, pour une meilleure extraction de l'information.

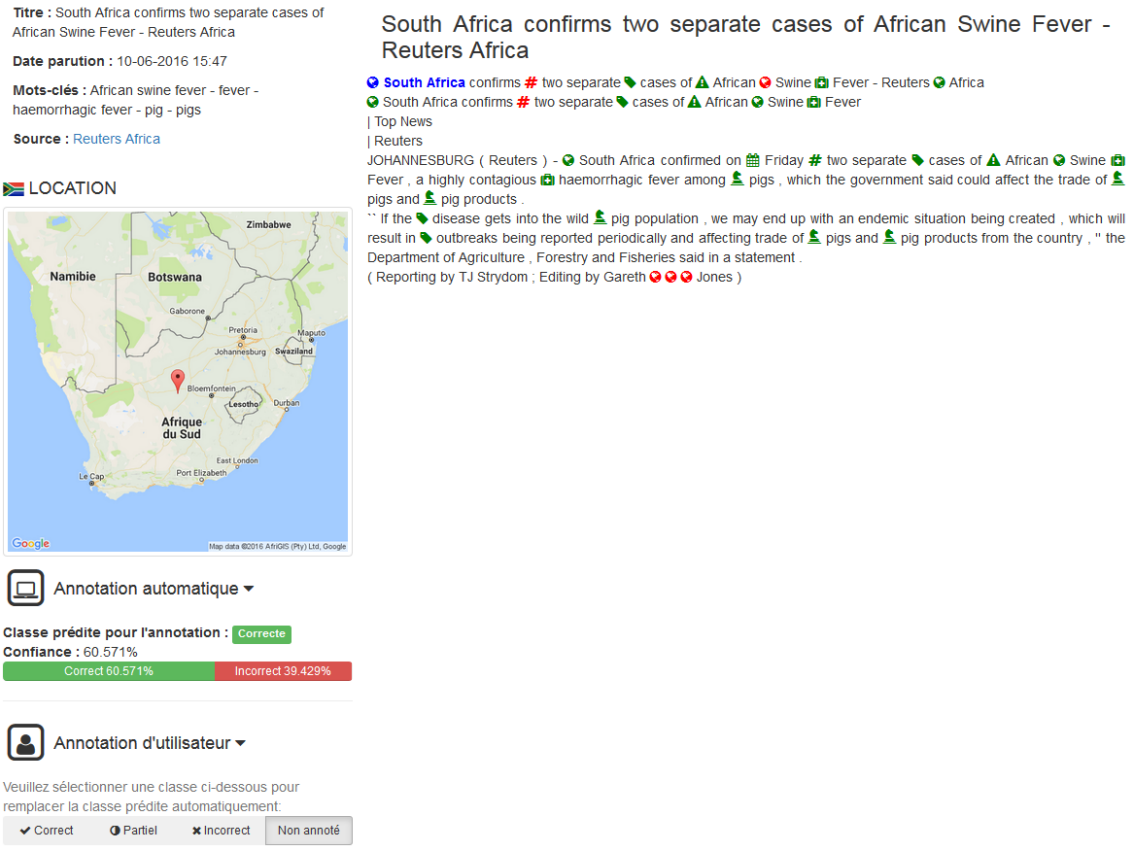


Figure 3.21 – Exemple d'une annotation et visualisation d'un article pertinent par le système PADI-web

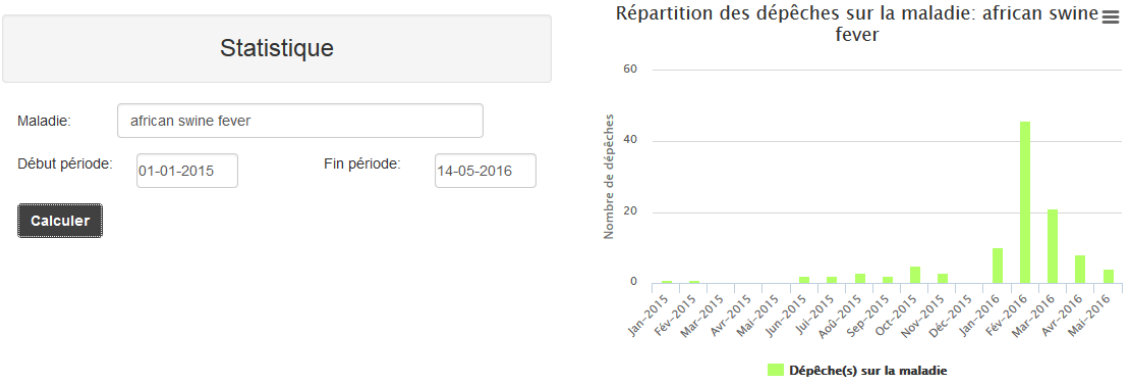


Figure 3.22 – Exemple d'un série chronologique obtenue avec le système PADI-web

3.5.5 Évaluation du système PADI-web

Le système PADI-web que nous avons élaboré est fonctionnel depuis janvier 2016 avec cependant des modifications qui ont été réalisées jusqu'en août 2016. Comme mentionné précédemment, notre système est entièrement automatisé et ne nécessite aucune intervention humaine. Cependant, pour les évaluations préliminaires et pour améliorer l'efficacité du système, il est nécessaire de quantifier sa performance à détecter des signaux de maladies émergentes (Barboza et al., 2014).

Nous avons donc évalué le système PADI-web sur quatre maladies modèles pour la période du 1er janvier au 28 juin 2016. En complément, nous avons évalué la performance du système PADI-web pour l'IA (Influenza aviaire), maladie pour laquelle des foyers sont déclarés en continu à l'échelle mondiale. Nous avons ainsi composé des flux RSS qui contiennent le nom de la maladie et les signes cliniques chez différentes espèces d'oiseaux. Cette liste de termes a été obtenue depuis le thésaurus AGROVOC et a été complétée par des termes extraits par fouille de textes à partir des alertes pour l'IA issues d'HealthMap (de manière similaire avec les autres maladies modèles présentées en section 3.2.1.1).

Pour les évaluations, nous avons retenu comme critère de sélection qu'un événement important (signal) issu du système PADI-web doit obligatoirement contenir une information sur la maladie et sur une espèce animale (Por = porcine, Bov = bovine, Ovi = ovine, Cap = caprine et Avi = avian). Pour cette période, notre système a détecté au total 9582 événements au niveau international (excluant la France) à partir de 1559 articles différents (Figure 3.23).

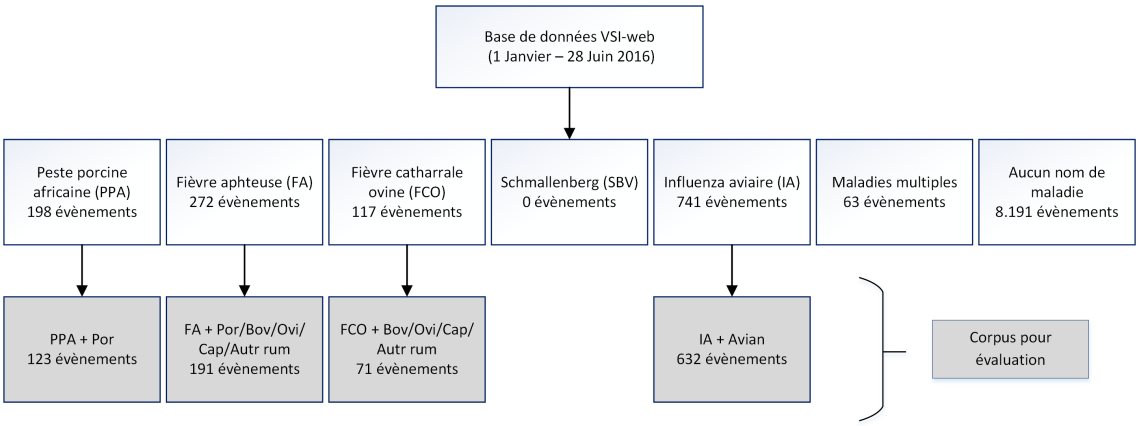


Figure 3.23 – Critères pour la sélection des événements du système PADI-web

Les signaux de système PADI-web ont été comparés avec trois autres sources, une source formelle et deux sources informelles :

— déclarations de l'OIE (source formelle, déclarations par des autorités compétentes d'un pays et vérification de l'information par des analystes) ;

— alertes sanitaires ProMED (source informelle, information soumise par des abonnés et vérification de l'information par des analystes) ;
— alertes sanitaires HealthMap (source informelle, collecte automatique des données sur le Web et vérification des signaux par des analystes).

Une déclaration immédiate a été considérée comme une apparition ou réapparition d'une maladie à déclaration obligatoire, à un changement dans la distribution ou sa présence chez une espèce hôte inhabituelle ou encore la détection d'une nouvelle maladie, selon le règlement international défini par l'OIE.

Un rapport de suivi concerne les foyers secondaires au premier foyer (foyer - index) d'une déclaration immédiate. Les déclarations immédiates et les rapports de suivi proviennent de la base de données FAO Empres-i.

Les données de HealthMap et ProMED ont été obtenues pour la même période à partir de la base de données de HealthMap. Un signal (événement) correspond à une détection d'une maladie donnée pour un lieu donné. Les signaux ont été divisés en deux catégories : les signaux du système ProMED (HealthMap | ProMED) et ceux provenant d'autres sources non officielles comme des agrégateurs de nouvelles, tels que *Google news* et Baidu (Healthmap | Autre) (Figure 3.24). Les analystes d'HealthMap vérifient chaque événement détecté par le système et lui attribuent des scores de 1 à 5. Les scores moins importants signifient des émergences peu probables avec un moindre risque de survenue et les scores plus élevés signifient des émergences probables avec des conséquences plus importantes pour la santé ou l'économie ou encore des agents pathogènes très contagieux (Bahk et al., 2015). Une description détaillée de ces deux systèmes est donnée dans le chapitre 2 et dans les travaux de Brownstein et al., 2008 ; Madoff, 2004.

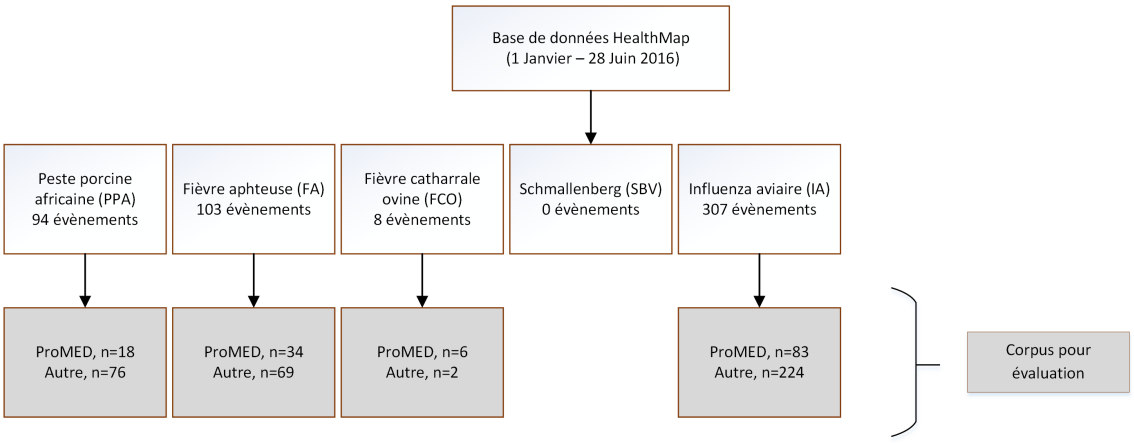


Figure 3.24 – Critères pour la sélection des événements des autres systèmes informels

A. Précision des systèmes informels pour alerter sur des événements sanitaires d'importance internationale pour des maladies modèles.

Pour évaluer un système informel (PADI-web, HealthMap et ProMED) dans sa globalité, nous avons calculé sa valeur prédictive positive (VPP) qui est la probabilité de détecter un vrai positif (VP) pour un événement d'importance internationale (déclaration immédiate à l'OIE, rapport de suivi, information publiée dans la littérature scientifique avec comité de lecture, ou un avis d'experts) parmi tous les événements détectés :

$$VPP = \frac{VP}{VP + FP} \tag{3.14}$$

Les faux positifs (FP) représentent le nombre d'événements signalés par les sources informelles qui ne correspondent à aucune déclaration immédiate ou rapport de suivi de l'OIE, ou à aucune donnée publiée dans la littérature scientifique, ou à aucune confirmation par des avis d'experts.

B. Sensibilité et réactive du système informel pour détecter des émergences des maladies modèles.

La capacité du système informel à alerter pour des événements sanitaires d'importance internationale a été évaluée au regard des déclarations immédiates à l'OIE (vrais positifs, VP). Les faux négatifs (FN) correspondent au nombre de déclarations immédiates de l'OIE qui n'ont pas été détectées par les sources informelles. Les vrais négatifs (VN) n'ont pas pu être déterminés.

La sensibilité (Se) du système a été définie comme sa capacité à détecter un événement sanitaire notifié dans une déclaration immédiate à l'OIE :

$$Se = \frac{VP}{VP + FN} \tag{3.15}$$

La réactivité du système a été évaluée par la différence entre les dates de publication d'un même événement par les voies officielles (formelles) et non officielles (informelles). Un délai positif signifie qu'un événement a été signalé par la source informelle après la déclaration officielle à l'OIE.

3.5.6 Résultats

Du 1er Janvier au 28 Juin 2016, 1392 foyers ont été extraits de la base de données Empres-i, soit :

— 794 foyers pour la PPA,
— 53 foyers pour la FA,

— 14 foyers pour la FCO,
— 531 foyers pour l'IA (excluant les cas humains).

La base de données d'Empres-i contient également des données en provenance d'autorités nationales, des experts de la FAO sur le terrain et des laboratoires de référence (Tableau 3.8). La maladie de Schmallenberg (SBV) n'étant pas à déclaration obligatoire à l'OIE, aucune information officielle n'était disponible. De plus, aucune alerte n'a été signalée pour l'émergence du SBV par PADI-web, HealthMap et ProMED. En conséquence, cette maladie n'a pas été prise en compte dans les évaluations.

Tableau 3.8 – Type et répartition géographique des foyers de la base de données Empres-i

Maladie	Type	Afrique	Amérique	Asie	Europe	Total
Peste porcine africaine	OIE DI [a]	6	-	-	5	11
	OIE suivi	-	-	-	783	783
Fièvre aphteuse	OIE DI [a]	1	-	14	-	15
	OIE suivi	15	-	22	-	37
	LR [b]	-	-	1	-	1
Fièvre catarrhale ovine	OIE DI [a]	1	4	-	3	8
	OIE suivi	-	-	-	6	6
Influenza aviaire	OIE DI	4	2	15	5	26
	OIE suivi	244	38	63	1	346
	FAO agent	12	-	3	-	15
	Auth. nat. [c]	121	-	22	-	143
	LR [b]	-	-	1	-	1

a. Déclaration immédiate
b. Laboratoire de référence
c. Autorité nationale

A. Événements signalés et précision des systèmes informels.

Les Figures 3.25, 3.26, 3.27 et 3.28 représentent les événements signalés par des systèmes informels pour la PPA, la FA, la FCO et l'IA. En moyenne, le système PADI-web a eu une précision peu élevée pour détecter des événements d'importance internationale pour les maladies modèles. La VPP était moindre pour la FA (27%) et plus élevée pour l'IA (54%). Les événements détectés par HealthMap ont eu une précision supérieure à 88% pour la PPA, la FA et l'IA grâce aux rapports du système ProMED (VPP > 96%) (Tableau 3.9).

Cependant, contrairement à HealthMap, l'information produite par le système PADI-web est complètement automatisée. En effet, avant d'être mises en ligne, les informations ProMED et HealthMap sont validées par des experts.

Tableau 3.9 – Précision des systèmes informels pour alerter sur des événements internationaux majeurs pour des maladies modèles

Maladies	Performance	PADI-web	HealthMap | ProMED [a]	Healthmap | Autre [a]
Peste porcine africaine	No.événements	123	18	76
	VP	37	18	67
	FP	86	0	9
	VPP	30%	100%	88%
Fièvre aphteuse	No.événements	191	34 [b]	69 [b]
	VP	51	33	62
	FP	118	1	7
	VPP	27%	97%	90%
Fièvre catarrhale ovine	No.événements	71	6	2
	VP	32	6	0
	FP	39	0	2
	VPP	45%	100%	0%
Influenza aviaire	No.événements	632	83	224
	VP	342	82	209
	FP	290	1	15
	VPP	54%	99%	93%

a. Les faux positifs (FP) d'HealthMap sont des événements annotés comme étant faiblement probables par les analystes (note 1)
b. Période de 28 février à 28 juin 2016

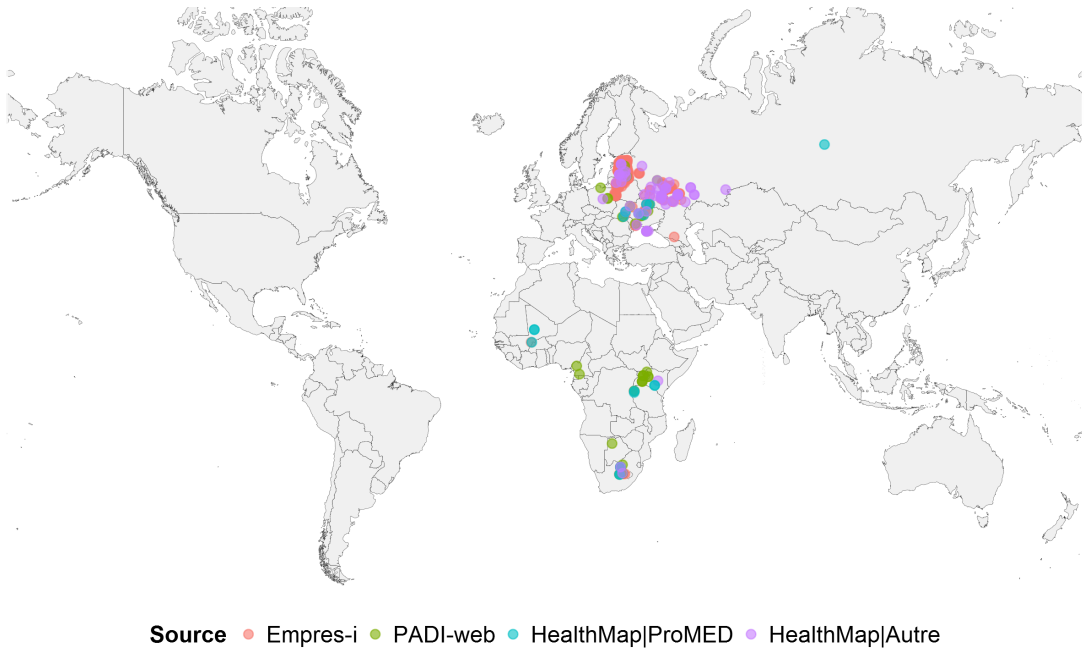


Figure 3.25 – Événements pour la peste porcine africaine au niveau international (de janvier à juin 2016) détectés par les systèmes informels

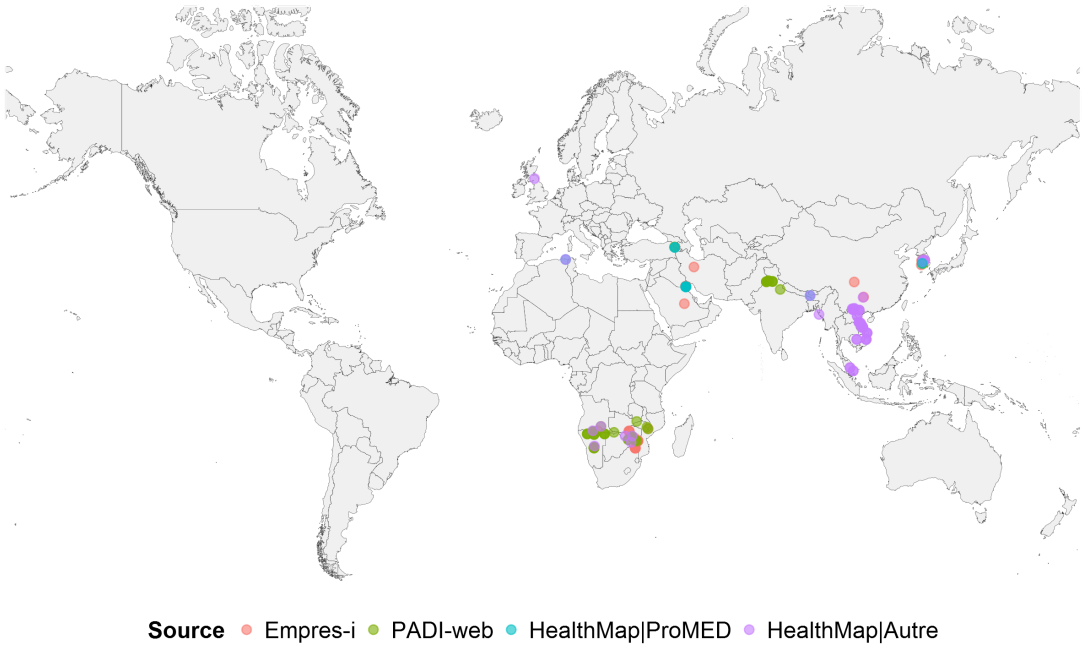


Figure 3.26 – Événements pour la fièvre aphteuse au niveau international (de janvier à juin 2016) détectés par les systèmes informels

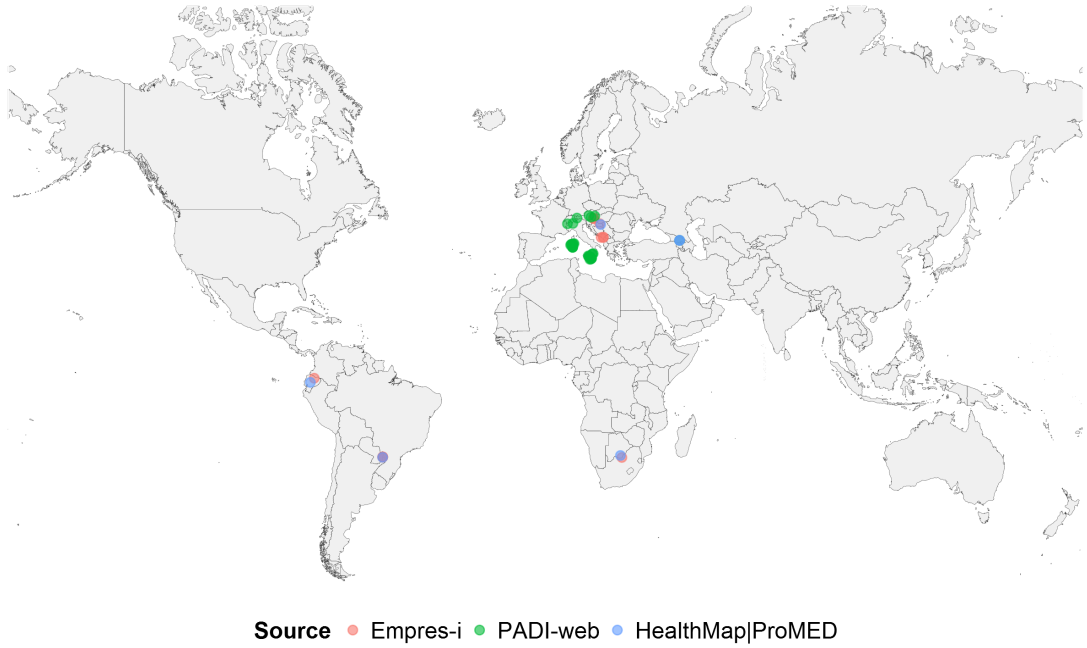


Figure 3.27 – Événements pour la fièvre catarrhale ovine au niveau international (de janvier à juin 2016) détectés par les systèmes informels

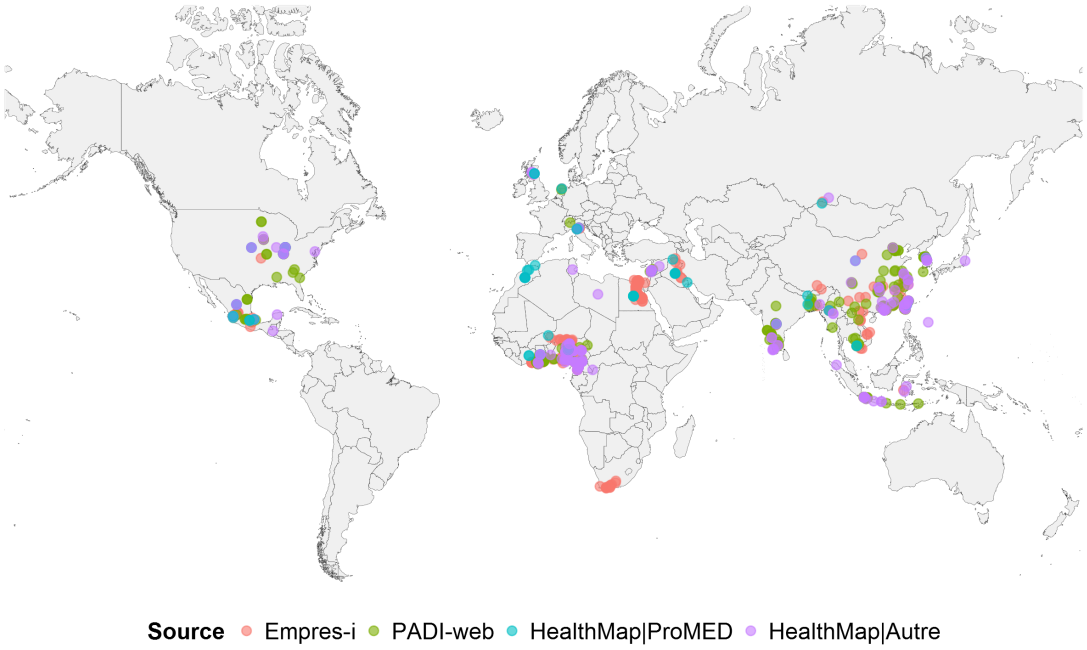


Figure 3.28 – Événements pour l'influenza aviaire au niveau international (de janvier à juin 2016) détectés par les systèmes informels

B. Sensibilité et délais des systèmes informels pour détecter des émergences.

Peste porcine africaine. Au niveau international, du 1er janvier au 28 juin 2016, 11 foyers primaires de PPA ont été notifiés à l'OIE, tous dans des pays ou régions déjà infectés : Europe de l'Est et pays Baltes (cinq foyers) et Afrique subsaharienne (six foyers). Le délai entre la date de survenue d'un événement jusqu'à sa déclaration immédiate a varié de 1 à 53 jours (j) (Figure 3.29).

Huit événements (Se=73%) signalés par PADI-web correspondaient aux 11 foyers primaires avec un décalage d'alerte de zéro à cinq j après les déclarations immédiates. Tous les événements (Se=100%) signalés par HealthMap | ProMED correspondaient aux 11 foyers primaires avec un décalage de zéro à quatre j après les déclarations immédiates. Quatre événements (Se=36%) signalés par le HealthMap | Autre correspondaient aux 11 foyers primaires avec un décalage de -1 à 21 j par rapport aux déclarations immédiates à l'OIE (Tableau 3.10).

Pour les déclarations immédiates de PPA en Europe, le seul pays ayant déclaré des foyers primaires a été l'Ukraine (cinq foyers primaires). Le système PADI-web a détecté quatre de ces cinq foyers primaires (Se=80%) avec un décalage de zéro à deux j après les déclarations immédiates. Healthmap | ProMED a détecté les cinq foyers primaires (Se=100%) avec un décalage de zéro à deux j après les déclarations immédiates. HealthMap | Autre a signalé trois des cinq foyers (Se=60%) avec un décalage de -1 à 21 j avant et après les déclarations immédiates.

Des suspicions de PPA en Crimée ont été signalées dans les médias ukrainiens et détectées par PADI-web (trois événements) et HealthMap (huit événements). Les analystes d'HealthMap ont attribué à ces événements un score faible (faible probabilité de survenue). HealthMap a également détecté cinq événements en Russie, dans les régions de Tatarstan, Kirov, Penza et Pskov et un événement en Moldavie, pays n'ayant jamais déclaré de cas de PPA. Ces événements ont été évalués par les analystes comme ayant une probabilité faible de survenue.

Pour l'Afrique, deux foyers primaires ont été déclarés en Afrique du Sud, deux au Burundi, un au Kenya et un au Mali, soit un total de six foyers primaires. PADI-web a détecté les deux foyers en Afrique du Sud avec un décalage de zéro à trois j, et les deux événements au Burundi avec un décalage de cinq j (Se=67%) après la déclaration immédiate à l'OIE. HealthMap | ProMED a détecté les six foyers correspondant aux déclarations immédiates à l'OIE (Se=100%) avec un décalage d'un à cinq j. HealthMap | Autre a signalé trois des six déclarations (Se=50%) avec un décalage de zéro à sept j.

Notons que PADI-web a détecté sept foyers de PPA en février et mars 2016 en Ouganda, pour lesquels il n'y a pas eu de déclaration immédiate. En effet, la PPA est enzootique dans ce pays, avec une vague épizootique signalée pour la dernière fois en septembre 2011.

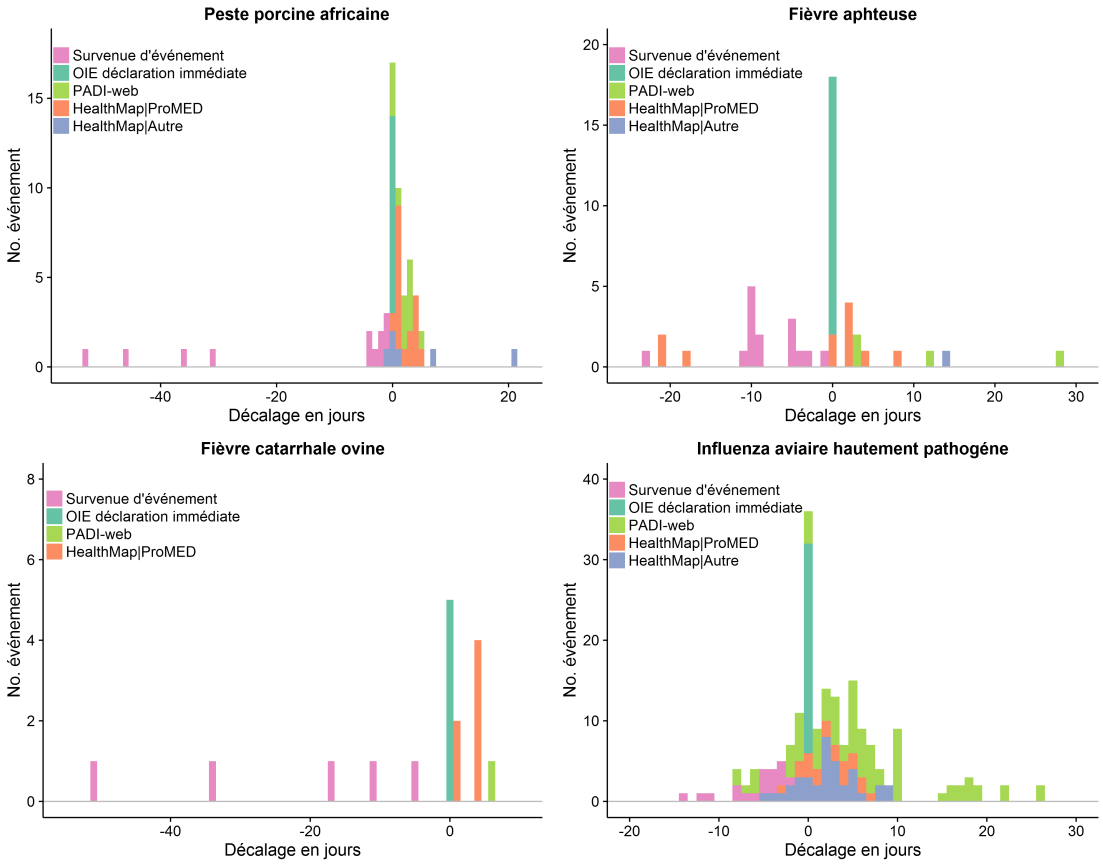


Figure 3.29 – Délais entre la première observation des foyers de maladies modèles, leur déclaration immédiate et leur rapport par des sources informelles de janvier à juin 2016

Tableau 3.10 – Performance des systèmes informels pour détecter des signaux d'émergence pour les maladies modèles

Maladies	Performance	PADI-web	HealthMap | ProMED	HealthMap | Autre [a]
Peste porcine africaine	VP	8	11	4
	FN	3	0	7
	Sensibilité	73%	100%	36%
	Délais (m, CI) [a]	+2j (0 à +5j)	+2j (0 à +5j)	+4j (-1 à +21j)
Fièvre aphteuse	VP	3	4	2
	FN	1	0	NA
	Sensibilité	75%	100%	NA
	Délais (m, CI) [a]	+12j (+3 à +28j)	-4j (-21 à +8j)	+ 22,5j (+14 à 35j)
Fièvre catarrhale ovine	VP	1	7	0
	FN	6	0	7
	Sensibilité	14%	100%	0
	Délais (m, CI) [a]	+6j	+3j (+1 à +4j)	0
Influenza aviaire hautement pathogène	VP	13	14	10
	FN	4	3	9
	Sensibilité	76%	82%	53%
	Délais (m, CI) [a]	+ 8j (-2 à +42j)	-9j (-82 à +7j)	+2j (-5 à 9j)

a. m=moyenne, CI=Intervalle de confiance

Fièvre aphteuse. Du 1er janvier et 28 juin 2016, 15 foyers primaires ont été notifiés à l'OIE en Arménie, Angola, Koweït et République du Corée. Le délai entre la date de survenue d'un foyer primaire - ou sa confirmation par un laboratoire et sa déclaration a été d'un à 23 j (1 j pour la République de Corée, 23 j pour l'Arménie) (Figure 3.29).

Le système PADI-web a détecté trois des quatre événements (Se=75%), les foyers au Koweït n'ayant pas été trouvés ; le délai de détection a été de trois à 28 j après la déclaration à l'OIE (Tableau 3.10, Figure 3.29).

HealthMap | ProMED a rapporté les émergences pour toutes les déclarations immédiates avec un décalage important de -21 j pour les foyers au Koweït, c'est-à-dire avant leur déclaration à l'OIE. Pour tous les autres événements, ProMED a partagé l'information avec un délai de zéro à huit j après la déclaration (Tableau 3.10, Figure 3.29).

Entre le 28 février et le 28 juin 2016, les autres sources d'information pour HealthMap n'ont signalé que deux événements, l'un pour le foyer survenu en Angola (délai de 14 j) et l'autre pour les foyers d'Arménie (délai de 35 j). Dans les deux cas, il s'agissait d'articles présentant le bilan de la situation. Cela n'exclut pas la possibilité que d'autres articles aient été publiés auparavant, dans la langue locale par exemple.

Fièvre catarrhale ovine. Du 1er janvier et 28 Juin 2016, 14 foyers de FCO ont été notifiés à l'OIE, parmi lesquels sept sous forme de déclaration immédiate : quatre pour des foyers-index (Croatie, Géorgie, Botswana et Équateur) et trois pour des clôtures d'événements (Brésil et Géorgie). Le délai entre la date de survenue d'un foyer-index et sa déclaration

a été de cinq à 54 j après l'observation des premiers signes cliniques ou la confirmation par un laboratoire (cinq j pour Botswana, 11 j pour la Géorgie, 34 j pour la Croatie et 54 j pour l'Équateur) (Figure 3.29).

Le système PADI-web n'a détecté que le foyer en Géorgie, six j après la déclaration à l'OIE. L'information détectée était un article sur les conséquences économiques de la FCO dans ce pays (Se=14%) (Tableau 3.10, Figure 3.29).

HealthMap | ProMED a détecté tous les foyers des déclarations à l'OIE avec un décalage de un à quatre j (Se=100%). Pour la déclaration de clôture des événements du Brésil, le réseau ProMED a rapporté cet événement avec un délai de quatre j. Cet événement est en fait survenu 310 j avant sa déclaration officielle (Tableau 3.10, Figure 3.29).

HealthMap | Autre a détecté trois événements liés à la FCO au Royaume-Uni. De la même manière, PADI-web a détecté cinq événements au Royaume-Uni et un autre en Bulgarie. Toutes ces détections correspondaient en fait à la préparation de vaccination préventive contre la FCO. De plus, le système PADI-web a détecté les mesures de vaccination prises en Sicile et Sardaigne contre les foyers de la FCO.

Influenza aviaire hautement pathogène. Du 1er janvier et 28 juin 2016, l'OIE a reçu des déclarations pour 26 foyers primaires (19 déclarations immédiates) : Russie, Royaume-Uni, Pays-Bas, Italie, Cameroun, Niger, Irak, Hong Kong, Liban, Cambodge, Inde, Myanmar, Bangladesh, République de Corée et États-Unis. Le délai entre la date de première observation de ces événements et leur déclaration officielle a varié entre deux et 131 j (moyenne de 33 j). La Figure 3.29 ne présente pas les délais de -30 j pour des déclarations immédiates (six foyers en Irak et un au Niger).

Le système PADI-web a détecté des événements correspondant à 13 déclarations immédiates (Se=76%) avec un décalage de -8 j pour des foyers à Hong Kong et -2 j pour les foyers au Liban, jusqu'à > 30 j pour des foyers aux États-Unis (max 42 j). De nombreux articles ont signalé l'information pour les foyers d'IA aux États-Unis, avec les premiers signes détectés par le système PADI-web le même jour pour des foyers survenus en Indiana et deux j après pour des foyers s'étant produits au Missouri (Tableau 3.10).

Le système HealthMap | ProMED a détecté des événements correspondant à 14 déclarations immédiates (Se=82%) avec un décalage de 89 j avant, et jusqu'à sept j après la déclaration immédiate des foyers primaires. L'information des foyers en Irak a été partagée avec la communauté de ProMED et HealthMap 80 et 89 avant la déclaration officielle, trois j avant pour l'Inde et un j avant pour les foyers primaires au Liban et au Royaume-Uni (Tableau 3.10).

Les autres sources d'information utilisées par HealthMap ont détecté des événements correspondant à dix déclarations immédiates (Se=53%). Ces sources ont précocement signalé des foyers en Inde (-4 j), au Cameroun (-3 et -5 j), en République de Corée et au Liban (-1 j) et au Royaume-Uni (-1 et -2 j) (Tableau 3.10).

3.5.7 Discussion

Ce travail nous a permis d'évaluer la performance du système PADI-web en comparant avec la performance des autres systèmes de veille sanitaire internationale : ProMED (fondé sur une expertise humaine) et HealthMap (semi-automatique).

Les premiers résultats montrent que par rapport à un « gold standard » (déclarations à l'OIE), la précision du système PADI-web pour signaler des événements sanitaires d'importance internationale est inférieure à 50% pour les maladies modèles étudiées. A titre de comparaison, les systèmes HealthMap et ProMED ont eu pour ces mêmes maladies, une précision de plus de 88%. Précisions que notre système est complètement automatisé, contrairement aux autres approches.

La précision du système PADI-web a probablement été influencée par l'étape de catégorisation d'articles dans le processus général de fouille de textes. En effet, pour catégoriser les articles, le système PADI-web utilise actuellement des filtres fondés sur des mots-clés [6]. À cette étape du processus, aucun algorithme de classification automatique n'est appliqué. Cela influence l'algorithme qui considère comme pertinents les articles contenant un nom de maladie, d'une espèce animale et des mots-clés spécifiques tels que « disease », « outbreaks », « case », etc. En conséquence, la majorité des articles retenus par le système PADI-web sont associés à la catégorie « bilans » d'une situation sanitaire, des mesures de prévention ou de contrôle, et pas nécessairement liés à une information précise sur des foyers. Par exemple, pour l'IA, 267 événements (41%) détectés étaient des bilans. Pour les articles sur la FA, cette catégorie a constitué 27% de tous les événements (43/161), 68% (48/70) pour le BTV et 70% pour la PPA (86/123).

Les exemples suivants présentent la mise en œuvre des mesures de prévention contre l'introduction et la diffusion de la FA et d'IAHP :

« *DES MOINES, Iowa — The pork industry is working to make sure an ample vaccine supply is available in the event of a foot and mouth disease (FMD) outbreak…* »

« *Mammals capable of transmitting avian influenza to poultry … Researchers have known for some time that a wide range of wild mammals, including raccoons and foxes - common visitors to farms and waterways across Delmarva - can be carriers of avian influenza…* »

Un deuxième groupe d'articles ayant influencé la précision du système de PADI-web était ceux rapportant des cas humains. Par exemple le système PADI-web a retenu comme pertinents les articles suivants pour l'IAHP :

« *Human infection with avian influenza A(H7N9) virus – China …. The majority (25 cases, 89%) reported exposure to live poultry or live poultry markets ; the exposure history of three cases is unknown or no clear exposure to poultry…* »

6. Le processus de classification décrit en section 3.3 n'est pas encore intégré à PADI-web

« Egypt reports 4th human H5N1 avian influenza case, 2nd from Giza .... Almost all cases of H5N1 infection in people have been associated with close contact with infected live or dead birds, or H5N1-contaminated environments... »

Pour la FA, le système PADI-web a retenu comme pertinents des articles traitant de cas humains de « Hand, foot-and-mouth disease (HFMD) » une maladie qui n'a pas de lien avec la FA (en anglais, la FA est nommée foot-and-mouth disease) :

« Lane County : Beware of hand, foot and mouth disease .... FLORENCE, Ore. - More than two dozen students in the Siuslaw School District have shown symptoms of hand, foot and mouth disease in recent weeks, Lane County Public Health said Friday. The symptoms include a fever, which can be accompanied by a sore throat, poor appetite and a vague feeling of being unwell, the county said. This disease should note be confused with foot-and-mouth disease in cattle... »

Cependant, le système PADI-web s'est montré plus performant (précision de 27% pour la FA, 30% pour la PPA, 45% pour la FCO et 52% pour la IA HP) que d'autres systèmes informels dans des circonstances similaires. Ainsi, Barboza et al., 2013 ont montré une faible précision de ces autres systèmes pour détecter des foyers animaux pour l'IAHP et le virus H5N1, à la fois pour des systèmes automatiques tels qu'Argus (précision de 6%), MedISys (3%), PULS (24%) et BioCaster (11%) ou semi-automatique tels que HealthMap (11%). Le plus performant de ces systèmes est ProMED avec une précision de 38%. Selon les auteurs, la faible performance des systèmes informels était liée au faible nombre de rapports officiels - qui ne représentent pas la situation épizootiologique réelle (Barboza et al., 2013 ; Tsai et al., 2013). Pour nos évaluations, nous avons pris comme gold standard les données des organismes officiels, des autorités nationales, des laboratoires de référence, des rapports et des évaluations de la part d'experts du terrain et des analystes. De manière similaire, Zeldenrust et al., 2008 proposent un gold standard issu des rapports du système ProMED ; pour le système BioCaster un gold standard est automatiquement créé à partir de rapports évalués par des analystes (Conway et al., 2010) ou par des annotations automatiques sur des bases de données de pathogènes (Buza et al., 2015 ; Schriml et al., 2009).

Nos travaux ont permis d'établir que les performances de détection précoce étaient influencées par le type de maladie, la langue et le lieu de survenue des foyers, mais également qu'elles variaient pour différents types de déclaration d'une même maladie (par exemple, la déclaration immédiate pour de nouvelles émergences n'a pas le même impact que la déclaration immédiate pour une clôture d'évènement sanitaire). Rappelons que la précocité de détection des événements sanitaires d'importance et la sensibilité sont des attributs cruciaux d'un système informel. La sensibilité, définie comme la proportion de déclarations immédiates détectées par un système informel, est élevée pour le système PADI-web, avec plus de 73% de détection de déclarations immédiates, sauf pour la FCO où la sensibilité est de 14%. Cette faible sensibilité pour la FCO peut s'expliquer avec le peu d'événements épidémiologiques exceptionnels durant les six premières mois de

2016. Les sources d'HealthMap qui dépendent d'agrégateurs de nouvelles n'ont pas non plus identifié d'événements d'importance pour la FCO.

Considérant que notre système est automatique et dépend actuellement d'un seul agrégateur d'articles (*Google news*), ces résultats sont tout à fait encourageants. De plus, malgré des alertes survenues dans la plupart des cas après la déclaration officielle des foyers, le système PADI-web reste très réactif, notamment pour l'IA.

En outre, contrairement aux autres systèmes, PADI-web a permis la détection d'événements avec une étendue thématique (par exemple, des informations liées à des mesures de vaccination) et géographique (par exemple, la PPA en Ouganda) plus vaste.

Nous avons mis en évidence une réactivité et une sensibilité importantes pour le système ProMED. Cela illustre l'importance des réseaux d'experts locaux pour la VSI. Les rapports de ProMED pour l'IA en santé animale ont été publiés jusqu'à 82 j avant la déclaration officielle auprès de l'OIE. Ces résultats contrastent avec ceux présentés par Barboza et al., 2013, avec une sensibilité de 33% de ProMED à détecter les foyers d'IAHP, et cela toujours après la déclaration officielle à l'OIE. Par contre Tsai et al., 2013 ont rapporté un décalage de 4 j en faveur de ProMED par rapport aux déclarations d'IA à l'OMS. De plus, Cowen et al., 2006 ont montré que le système ProMED a détecté 47% des événements de 2 à 14 j avant la déclaration auprès de l'OIE. Pour les autres systèmes, Bahk et al., 2015 montent une sensibilité du système HealthMap de 65% avec un délai d'alertes de 1 à 3 j après la déclaration immédiate des foyers. Cependant, dans la majorité des cas, les rapports étaient publiés de 0 à 10 j après la déclaration immédiate, avec une sensibilité de 20% à 40% pour HealthMap, PULS, MedISys et Argus selon Barboza et al., 2013.

Notre travail a montré que la prise en compte des sources officielles est cruciale pour la détection des foyers, principalement pour les délais de déclaration. Cependant, les sources informelles sont plus réactives dans le partage d'information sanitaire pour des émergences d'importance internationale. Par ailleurs, ces sources informelles fournissent des informations plus vastes thématiquement et géographiquement. L'application de ces résultats dans le contexte de la veille en santé animale internationale montre que les systèmes formels et informels peuvent être complémentaires pour mieux détecter les émergences de maladies exotiques, plus rapidement et plus précisément. Ces résultats ouvrent de nouveaux champs d'investigations dont les applications pourraient être d'importance pour les utilisateurs. En effet, le fait que notre système soit plus performant pour la détection de certaines maladies ou pour certaines régions pourrait permettre d'optimiser l'utilisation de PADI-web.

Notons que le nombre d'évènements inclus dans l'estimation de la sensibilité de notre système était malheureusement insuffisant pour permettre des analyses stratifiées notamment en raison de l'existence de potentiels facteurs de confusion. Rappelons qu'en raison de contraintes, telles que la disponibilité des ressources humaines et l'opérabilité du système PADI-web, la sensibilité du système n'a pu être évaluée que sur la période du janvier à juin 2016.

Aussi, la langue source et la zone géographique pourraient constituer un éventuel facteur déterminant de la performance de notre système. Par exemple, on peut imaginer que le français soit plus susceptible d'être la langue source d'évènements déclarés en Afrique francophone, l'arabe au Moyen-Orient, l'espagnol en Amérique du Sud, le russe en Europe de l'Est, le Chinois en large partie d'Asie, etc. Le système PADI-web est actuellement fondé sur des recherches en anglais. Au contraire, ProMED et HealthMap ont un accès vers des articles de média publiés en plus de cinq langues, BioCaster plus de 13 et MedISys plus de 40 langues. Par exemple, Barboza, 2014 ont montré une meilleure performance de BioCaster pour la détection des foyers de l'IA HP en chinois que dans d'autres langues. Selon les auteurs, le processus de fouille de textes du système BioCaster est adapté à la veille des médias en Asie de Sud Est avec la prise en compte des langues locales. Les spécificités liées à la maladie, la liberté de la presse, l'accès à internet, pourraient également être considérés comme des éléments déterminants. Nous pouvons également imaginer, par exemple, des différences de qualité de détection d'une même maladie selon le lieu de sa survenue, c'est-à-dire dans une zone enzootique ou indemne. Ainsi, à l'heure actuelle, il est difficile d'évaluer l'influence de ces différents facteurs dans un système de veille automatisé.

Dans le cadre d'une utilisation d'information en provenance de différents systèmes EBS et IBS, les implications potentielles seraient plus importantes. En particulier, l'intégration de l'information sanitaire extraite à partir de PADI-web, ProMED, HealthMap, OIE, Empres-i, ADNS, dans un outil unique pourrait bénéficier de leur complémentarité et ainsi améliorer la sensibilité tout en augmentant la spécificité de détection de signaux de maladies émergentes.

Chapitre 4

Conclusion et Perspectives

4.1 Synthèse des principales contributions

Ce travail de thèse apporte plusieurs contributions méthodologiques dans un contexte pluridisciplinaire (informatique et épidémiologique). Tout d'abord, nous avons proposé une démarche générale pour définir une méthode automatique de veille sur le Web pour l'étude des maladies animales infectieuses exotiques et émergentes. Devant la croissance de ces événements (Jones et al., 2008), cette fonction devient de plus en plus indispensable pour les autorités compétentes nationales en termes de protection de la santé publique et de la santé animale. Par rapport aux autres systèmes de biosurveillance fondés sur la veille de nombreuses maladies infectieuses, notre approche se concentre sur des maladies spécifiques, établie par des experts en santé animale, sur une base épidémiologique. Avec ce type d'approche, nous avons pu acquérir automatiquement les données sur le Web liées aux maladies animales infectieuses ciblées, puis les classer en plusieurs catégories. Nous avons pu également extraire l'information épidémiologique importante issue de ces données afin de fournir aux utilisateurs une information pertinente en temps réel sur les principales émergences sanitaires au niveau international. Cependant, d'autres études complémentaires à celles présentées dans ce manuscrit devront être menées pour approfondir les différentes facettes de ce domaine complexe.

L'extraction automatique de la terminologie biomédicale (par la fouille de textes) à partir des articles pertinents (*e.g.*, « nouveaux cas ») permet de construire un vocabulaire utilisé ensuite pour l'acquisition, le filtrage et l'identification automatique de l'information sanitaire importante pour le veille. Ce vocabulaire est constitué de noms de maladies, d'hôtes et de signes cliniques, ainsi que des mots-clés décrivant des événements sanitaires. La terminologie extraite varie selon l'origine, comme par exemple les résumés d'articles scientifiques, qui ont été une source de termes utilisés par des spécialistes du domaine et les médias électroniques, qui ont été une source de termes utilisés au quotidien (chapitre 3.2.1 et l'Annexe E). La terminologie extraite a varié selon la maladie, la langue et la zone géographique couverte par les événements décrits dans les documents sanitaires. Par exemple, pendant la période étudiée, de nombreux articles ont été publiés sur l'émergence de la PPA dans les pays Baltes. De même, de nombreux articles décrivent la découverte d'un nouveau virus, le SBV, en 2011 en Allemagne et sa diffusion rapide

pendant les saisons vectorielles suivantes en Europe de l'Ouest. Au contraire, pendant la période étudiée une épizootie de BTV a touché les pays des Balkans et une épizootie de FA a impacté les pays du Maghreb. Cependant, le nombre de termes extraits se révèle limité probablement à cause de l'information publiée dans les langues locales. Dans nos futurs travaux, il serait intéressant de rechercher des facteurs de variation de la sensibilité du système et de la précocité des alertes parmi des indicateurs socio-économiques et politiques. En effet, de tels indicateurs peuvent être liés aux processus épidémiologiques eux-mêmes (Godfrey et al., 2011). De plus, il est possible qu'ils aient une influence sur les déclarations officielles et informelles (rumeurs), selon la liberté de la presse et le contrôle plus ou moins étroit exercé par les autorités sur le Web et les réseaux sociaux.

L'évaluation de la nouvelle mesure de classement des termes que nous avons introduite a montré que des termes utiles pour construire un vocabulaire épidémiologique pour la veille sont pertinents selon le type de sources ; c'est la raison pour laquelle nous avons proposé dans ces travaux de thèse, une nouvelle fonction de rang qui prend en compte ce type d'information.

L'évaluation des mesures statistiques d'association entre les termes décrivant des signes cliniques et les hôtes ont révélé une précision variable *vis-à-vis* des associations considérées comme spécifiques par les experts en santé animale. Cette précision dépendait de la maladie (de 65% à 100% pour la PPA, de 40% à 60% pour la FA et le SBV et de 25% à 40% pour la FCO) mais également de la mesure statistique employée.

En complément, la méthode Delphi nous a permis de bénéficier des connaissances de nombreux experts. Ils ont contribué à l'évaluation de termes extraits par la fouille de textes, ainsi qu'à la sélection des associations spécifiques entre les hôtes et les signes cliniques obtenues par des mesures statistiques combinant fouille du Web et fouille de textes. Les experts ont aussi proposé des termes caractérisant les maladies modèles et qui n'avaient pas été extraits de nos corpus (chapitre 3.2.1.4). Ces résultats montrent que le processus de veille doit prendre en compte la connaissance des experts. Cependant, l'avis des experts était parfois discordant pour évaluer certains termes. Cela peut s'expliquer par l'orientation de leur expertise professionnelle (terrain, laboratoire de diagnostic, recherche, enseignement, etc.) mais également par une compréhension différente des objectifs de l'étude. Nous envisageons à l'avenir de travailler avec moins d'experts et avec des entretiens directs.

La contribution des experts nous a montré qu'ils favorisent les signaux de l'émergence des maladies modèles chez les animaux domestiques (sauf pour les sangliers et la PPA). Ils ont confirmé et complété les associations suivantes pour caractériser les maladies modèles et détecter des signaux d'émergence sur le Web :

— pour la PPA, des termes et associations qui décrivent l'apparition de la mortalité, la fièvre et les signes cliniques hémorragiques chez les porcs domestiques, ainsi que la mortalité et la maladie hémorragique chez des sangliers ;

— pour la FA, les pertes de production, ainsi que des signes cutanés et muqueux chez les bovins et les porcs ;
— pour la FCO, des mortalités et de la fièvre chez les bovins et les ovins, ainsi que des signes cutanés et muqueux chez les ovins ;
— pour le SBV, des malformations et déformations congénitales des chevreaux, agneaux et veaux, y compris les avortements chez les bovins, ovins et caprins.

Nous considérons que la combinaison d'approches de fouille de textes et fouille du Web avec les connaissances des experts permet d'identifier des termes et des associations de termes décrivant les signes cliniques et les hôtes. Ceci facilite la détection des signes de maladies infectieuses connues ou nouvelles. Notre méthode est générique et peut avoir des applications en santé animale et en santé publique.

Nos résultats ont également permis d'étudier la précision des associations spécifiques comme les requêtes de recherche sur le Web. Les associations évaluées comme étant hautement spécifiques par des experts ont eu la meilleure précision pour la FA (supérieur à 90%) et une précision moindre pour les associations hautement spécifiques des autres maladies modèles (de 7% à 65% pour la PPA, de 14% à 34% pour la SBV et de 5% à 9% pour la FCO).

Ces résultats ont également permis d'analyser les limites des études rétrospectives sur le Web. Les moteurs de recherche ont en effet des capacités de stockage d'archives d'articles limitées. Plus les recherches sont profondes, plus la probabilité de trouver des signaux diminue. Dans ce contexte, nous avons évalué des associations hautement spécifiques comme des requêtes de recherche dans les conditions réelles d'utilisation avec le système PADI-web. Ces résultats ont montré une précision de notre système pour alerter des événements sanitaires d'importance internationale de 27% pour la FA, 30% pour la PPA, 45% pour la FCO et jusqu'au 54% pour l'IA.

Les résultats de la classification automatique de documents nous ont montré que l'utilisation de la représentation vectorielle de documents et ensuite l'apprentissage supervisé en utilisant les classificateurs NB et SVM est simple et efficace. La meilleure performance des algorithmes était pour des catégories « nouveaux cas » (F-mesure de 64% à 84%) et « généraux » (F-mesure de 72% à 86%) et moindre pour la catégorie « bilans » (F-mesure inférieure à 40%). Cela suggère que les travaux ultérieurs doivent évaluer l'influence de la fusion d'articles de la catégorie « bilans » vers les deux autres catégories. Une étude plus approfondie des descripteurs linguistiques qui influencent la classification devra également être conduite.

La méthode combinée d'extraction d'information nous a permis de profiter des points forts des deux types d'approches d'extraction d'information utilisées (découverte automatique de règles et apprentissage supervisé). En utilisant des dictionnaires de termes obtenus par fouille de textes et proposés par des experts, la précision de cette méthode pour extraire correctement des noms de maladies, des hôtes et des signes cliniques était

supérieure à 94%. L'extraction de lieux, dates et nombres de cas s'appuyant sur des gazetiers externes a eu une précision de 80% à 85%.

Au-delà de la démarche scientifique, nos résultats ont des implications pratiques dans le cadre du dispositif de la VSI en France. Le système PADI-web montre la faisabilité du développement d'une plateforme Web unique qui permettrait d'optimiser la détection de dangers sanitaires au niveau international à partir des données non structurées et permettra de remplacer le travail manuel et fastidieux des personnes en charge de la veille du Web.

Le choix d'utiliser des flux RSS pour collecter des articles sur le Web s'est montré très efficace car il nous a permis de collecter des données sous un format semi-structuré (format XML) qui permet un traitement plus facile du contenu. De plus, le choix de flux RSS fondés sur des combinaisons de mots-clés nous a permis d'avoir un vaste nombre d'articles et sources d'information qui sont mises à jour régulièrement.

Les évaluations du système PADI-web actuel illustrent que notre outil détecte efficacement les signaux de nouvelles émergences à partir des médias du Web (73% pour la PPA à 75% et 76% pour la FA et IAHP, respectivement). Ces résultats sont très encourageants en considérant que l'outil actuel s'appuie sur l'acquisition et la catégorisation de documents fondés sur des mots-clés dans une langue (l'anglais) et à partir d'un seul agrégateur d'articles (*Google news*).

Cependant, nos évaluations ont également permis de mettre en relief les limites de cet outil et en particulier la faible précision des signaux bruts détectés (27% pour la FA à 45% pour la FCO) et la faible réactivité pour la détection précoce de nouvelles émergences (-2 à +45 jours). La diversité des langues traitées mais également la nature de leur traitement est certainement en mesure d'influencer la performance d'un système de biosurveillance. Le système PADI-web prend en compte les articles publiés en anglais, même si très récemment nous avons intégré le français et l'espagnol. La démarche complète de sélection de mots-clés multilingue pour des recherches sur le Web, la sélection d'agrégateurs de nouvelles ciblés, ainsi que l'extraction de l'information sanitaire multilingue restent à mener dans le futurs travaux.

Finalement, l'expertise humaine est une étape clé dans le processus. Le système PADI-web permet aux utilisateurs d'évaluer des articles et des événements extraits ainsi que de modifier des requêtes de recherche selon leurs besoins. L'évaluation humaine est souvent nécessaire pour comprendre ce qui est inhabituel, découvrir des événements rares que le système peut avoir manqué, prendre la décision finale sur le contenu des rapports collectés et donner un sens épidémiologique à une situation sanitaire. Enfin, l'expertise pourra contribuer à l'identification des limites du système une fois qu'il sera entré dans sa phase opérationnelle à grande échelle.

Ces résultats illustrent l'importance que peut jouer l'intelligence épidémiologique (IE) dans le renforcement des capacités de la veille sanitaire et la prise de mesures pour prévenir l'introduction des pathogènes exotiques et nouveaux en France. Les travaux futurs

peuvent s'orienter vers la modélisation de la relation entre la détection anticipée et la mise en œuvre précoce de mesures de contrôle et l'impact que cette réactivité peut avoir sur la santé animale.

4.2 Perspectives

4.2.1 Acquisition de données

L'agrégateur d'articles *Google news* est librement accessible au public et il a accès à un très grand nombre de sources d'information (Collier, 2012). Certains systèmes d'IE tels que GPHIN ont des contrats avec des entreprises privées proposant des agrégateurs de nouvelles tels que *Factiva* et *LexisNexis* (Mykhalovskiy et al., 2006). Ces entreprises offrent une vaste gamme de sources d'information (plus de 9000) avec une grande variété de langues et un nettoyage du contenu des articles. Cependant, de plus en plus de systèmes utilisent des sources librement accessibles sur le Web, tels que le système BioCaster reposant sur plus de 1700 flux RSS à partir de différentes pages Web (Collier et al., 2008), ou MedISys qui utilise plus de 3600 pages Web manuellement sélectionnées par des experts (utilisateurs) ainsi que plus de 350 pages Web émanant des autorités compétentes en santé humaine (Rortais et al., 2010). De même, HealthMap utilise les données provenant de 200 000 sources à partir des agrégateurs d'articles tels que *Google news* par région géographique et *Baidu* et *Soso*, pour la Chine. Les projets « Outbreaks near me » et « Flu near you » aux États-Unis collectent et visualisent des alertes soumises en ligne par des utilisateurs concernant des symptômes inexpliqués ou de la grippe (Bahk et al., 2015). L'application mobile d'Empres-i collecte des alertes de maladies infectieuses auprès des éleveurs et des vétérinaires. Des sources supplémentaires d'information pourraient être constituées des médias sociaux tels que Twitter, Facebook et les blogs (« FluTrackers », « Flu Near You ») (Hartley et al., 2010).

En prenant en compte que la version actuelle de PADI-web ne collecte que les données en anglais pour des requêtes à partir de 60 flux RSS et 930 sources d'information (données de janvier à juin 2016), nos travaux futurs pourraient s'orienter vers la mise en place de flux RSS multilingues. Nous pourrions ajouter d'autres sources informelles tels que les agrégateurs *Baidu, SoSo, Yahoo news, Yandex*, etc., soit de contenu mondial, soit ciblées sur certaines régions. Le système PADI-web a également la capacité de collecter l'information sanitaire à partir de flux RSS officiels comme les notifications à l'OIE ou à l'Empres-i. Les travaux à venir pourraient s'orienter vers l'extraction d'informations à partir de ces données structurées. Nous pouvons ainsi imaginer une plateforme servant à collecter et à visualiser des rapports sur des événements sanitaires provenant de différentes sources et dans plusieurs langues.

4.2.2 Classification de données

Avec le système PADI-web, la catégorisation des articles s'appuie sur des listes de dictionnaires créés par des experts (utilisateurs) et des thésaurus et gazetiers (par exemple, GeoNames). Cette étape clé peut être améliorée en intégrant la classification automatique dans le processus. Ainsi, aétant données les resultats encourageants obtenus avec SVM et NB pour la classification des documents, ces deux algorithmes pourraient être intégrés et évalués avec des données en temps réel. Certains auteurs proposent également l'application de méthodes d'ensembles s'appuyant sur plusieurs algorithmes pour obtenir une meilleure classification (Doan et al., 2012 ; Heredia-Langner et al., 2015 ; Torii et al., 2011 ; Tuarob et al., 2014). Cependant, ces algorithmes demandent un travail supplémentaire d'annotation de nouveaux articles et leur mise à jour régulière dans un corpus actualisé. Une alternative pourrait être l'utilisation d'une méthode d'apprentissage actif (Novak et al., 2006) permettant de prendre en compte les articles collectés au cours du processus.

Un autre sujet de recherche est l'étude de l'impact de la taille du corpus sur les méthodes d'apprentissage En effet, les performances d'un algorithme de classification peuvent dépendre de la taille du corpus d'apprentissage et surtout du nombre d'articles annotés comme pertinents (Noto et al., 2008). Les performances des classificateurs NB et SVM et plus récemment le « deep learning » (LeCun et al., 2015) peuvent évoluer au cours d'un processus en temps réel, lorsque des articles pertinents sont ajoutés au corpus.

Enfin, le sujet le plus difficile est la détection des événements rares (signaux faibles). En effet, pour obtenir de bonnes performances globales, un classificateur doit se concentrer sur la détection à la fois d'événements fréquents et sans négliger les articles décrivant des événements rares. Ces événements sont susceptibles d'avoir de lourdes conséquences épidémiologiques, s'il s'agit d'une maladie nouvelle par exemple. Nous prévoyons d'approfondir ces travaux de recherche, en particulier sur une période temporelle plus étendue.

4.2.3 Analyse de l'information

4.2.3.1 Visualisation des événements sanitaires

Le système PADI-web contient une interface simple pour visualiser les articles collectés et les événements extraits. Les travaux futurs peuvent s'orienter vers l'amélioration de la représentation visuelle des événements extraits. Les interfaces actuellement déployées par la plupart des systèmes de biosurveillance utilisent des services de cartographie Web, tels que Google Maps (implémenté par BioCaster, IBIS, PULS, Empres-i et HealthMap). Des services de cartographies interactives permettent de visualiser des événements dans le temps et l'espace. Les utilisateurs peuvent ensuite étudier les relations spécifiques du domaine, agréger ou suivre leur évolution temporelle par pays, syndrome, maladie, etc.

L'interface doit également permettre de visualiser les événements détectés dans les autres langues avec une traduction en français ou en anglais.

L'analyse statistique des événements sanitaires est un défi pour tous les systèmes de veille actuels. Les limites sont que l'analyse statistique des événements se fait pour des besoins spécifiques mais pas de manière systématique. Les travaux à venir pourraient s'orienter dans trois directions principales : *i)* la détection des anomalies spatiales et temporelles ; *ii)* l'estimation de l'évolution des événements (vitesse de diffusion) ; et *iii)* la prédiction des zones à risque.

4.2.3.2 Détection des aberrations spatiales et temporelles

La détection des aberrations à partir des données épidémiologiques s'appuie sur l'identification des métriques qui sont liées à un objectif donné. Ainsi, cet objectif peut avoir un focus régional ou global, concerner la fréquence des alertes (journalière, hebdomadaire) et le taux de fausse alerte attendu. Ensuite, il faut définir des indicateurs pour une maladie ou un syndrome, tels que la fréquence des événements pour un période donnée et l'établissement d'un niveau de base au-dessus duquel l'alerte sera donnée.

Concernant la détection des anomalies à partir des événements rapportés dans les médias, un problème majeur est la duplication de l'information. Par exemple, la distribution des fréquences des rapports (signaux) détectés peut être fortement accrue suite à l'attention particulière des médias, ou des conséquences sanitaires et socio-économiques particulièrement graves.

Le deuxième facteur important pour la modélisation des séries temporelles est la période historique prise comme référence. Des algorithmes de détection sur des séries temporelles doivent préalablement être calibrés sur cette période de référence. Le système PADI-web possède actuellement des données historiques de moins d'un an, ce qui influence le choix d'un algorithme pour identifier des éventuelles aberrations.

En l'absence de contraintes, le choix de l'étendue de la période de référence dépendra de la volonté de l'utilisateur de modéliser de façon rigide l'effet du temps, ou au contraire de ne prendre en compte que les tendances récentes. La période de référence peut être fixe ou glissante. Dans ce dernier cas, elle est complétée à chaque fois qu'une nouvelle valeur est enregistrée. Les nouvelles valeurs sont intégrées en l'état ou suite à un traitement si elles ont été identifiées comme étant des valeurs aberrantes (Perrin, 2013). Dans le cas d'un historique mis à jour à chaque nouvelle valeur, un intervalle avant la valeur à tester (quardband) peut être gardé de manière à exclure le démarrage éventuel d'un événement de jeu de données de référence utilisée pour calculer la prédiction. Deux jours de quardband sont par exemple prévus dans les méthodes C2 et C3 (Hutwagner et al., 2003) ou dans la méthode fondée sur une carte de contrôle EWMA (Jackson et al., 2007). À l'heure actuelle, le seul système de biosurveillance qui a mis en œuvre et évalué des méthodes de détection des anomalies pour des articles issus du Web est le système BioCaster. Les

auteurs ont implémenté des méthodes C2, C3, (Hutwagner et al., 2003) la variation W2 (Copeland et al., 2007) qui prend un compte les tendances pour les jours de la semaine et pour des jours du weekend, la F-statistique et la méthode EWMA (Jackson et al., 2007 ; Hutwagner et al., 2005).

Les données obtenues avec le système PADI-web permettent également d'analyser des regroupements géographiques (agrégats spatiaux, temporels ou spatiaux-temporels). La principale approche utilisée en surveillance épidémiologique pour détecter un agrégat spatial ou spatio-temporel est la statistique de scan spatial et son adaptation la statistique de scan spatio-temporel, développés par Kulldorff et al., 1995 ; Kulldorff, 2001 ; Kulldorff et al., 2005 et adaptés par Assunção et al., 2009. La statistique de scan peut être prospective et employée avec des données qui évoluent avec l'apparition des nouveaux événements. Cette dernière permet la détection d'agrégats émergents, qui se finissent au temps présent. Elle donne également un ajustement de la statistique prospective pour tenir compte des analyses répétées sur différentes périodes temporelles (journalière ou hebdomadaire). A l'heure actuelle aucune étude n'a évalué des données spatiales issues des systèmes de biosurveillance. Dans ce contexte, les futurs travaux du système PADI-web peuvent s'orienter vers ce type d'analyses.

4.2.3.3 Vitesse de diffusion des pathogènes

La compréhension de la dynamique spatiale d'une maladie infectieuse est critique pour prédire où et à quelle vitesse elle se propage. Certaines méthodes en épidémiologie portent sur l'estimation de la vitesse de propagation générale, c'est-à-dire le taux moyen de propagation d'un pathogène ou d'une espèce invasive pendant toute sa durée et sa portée (Gilbert et al., 2010). D'autres méthodes portent plutôt sur la vélocité de la propagation des pathogènes dans une zone géographique prédéfinie, pour des émergences avec un caractère enzootique (Brunton et al., 2015) ou épizootique (Pioz et al., 2011 ; Tisseuil et al., 2016). Brunton et al., 2015 ont estimé la propagation de l'enzootie de la tuberculose bovine en créant des contours autour des foyers agrégés par hexagones au cours des périodes de deux ans et en calculant la différence moyenne entre chaque hexagone. L'estimation de la vitesse de diffusion décrite par Tisseuil et al., 2016 est fondée sur un modèle de régression linéaire de la semaine de première occurrence d'événement sur une fonction spline des coordonnées géographiques (Wood, 2003). Ce modèle est ensuite utilisé pour interpoler les données sur l'étendue géographique considérée. Les vitesses de diffusion sont déduites des données interpolées par l'intermédiaire du calcul d'un indicateur de friction de l'environnement (résistance à la diffusion de l'infection). Nous avons appliqué cette méthode pour l'étude de l'épizootie de dermatose nodulaire contagieuse dans les Balkans en 2016, dans la cadre du dispositif de VSI (Arsevska et al., 2016a).

4.2.3.4 Prédiction des zones à risque

Finalement, les données acquises avec le système PADI-web peuvent être exploitées pour générer des cartes des zones spatiales à risque prédites par des modèles, qui peuvent être mises à jour régulièrement. Plusieurs approches combinent des données spatio-temporelles sur les foyers avec des données environnementales (Anyamba et al., 2009), la densité animale (Glanville et al., 2014), la présence des vecteurs (Arsevska et al., 2015), la mobilité animale (Dawson et al., 2015) et les échanges commerciaux (Semenza et al., 2014). Récemment, Hay et al., 2013a ont proposé une méthode de prédiction des zones à risque pour des données mises à jour en temps réel. La méthode qui applique des approches classiques de combinaison des données spatiales et covariables environnementales a été intégrée dans l'interface du système HealthMap. Cette méthode n'a actuellement été appliquée que pour la dengue, en prenant en compte les données sur des foyers obtenues par des sources informelles et officielles (Hay et al., 2013b). Cette approche illustre la valeur des données multi-sources pour mieux comprendre la diversité des maladies et leur émergence.

4.2.4 Dissémination de l'information sanitaire

À l'heure actuelle les alertes aux utilisateurs ne sont pas incorporées dans le système PADI-web, même si ces notifications sont la dernière étape clé du processus de fouille de textes pour la veille épidémiologique. Bien que les normes telles que le protocole d'alerte commun existent (« Common Alerting Protocol »), à notre connaissance, aucune norme interopérable pour la structure des messages d'alertes (sémantique ou de vocabulaire) ne semble avoir été acceptée parmi les systèmes de veille internationaux. Les alertes emails sont le format le plus répandu (IBIS, ProMED), ainsi que les flux RSS (Empres-i, OIE) qui contiennent le type d'alerte, le nom de la maladie, le pays et la date de l'envoi du message. Un autre format plus adapté pourrait être le GeoRSS, un format XML léger qui encode l'information géographique (Collier et al., 2012). Un minimum d'éléments nécessaires pourrait inclure, par exemple, la date du message, la date de l'événement, le lieu de l'événement, le nom de la maladie, l'espèce touchée, une description de la source, le degré de certitude pour l'événement ou le niveau de confidentialité, le type de message (par exemple, nouvelle émergence ou ré-émergence ou suivi) et d'un identificateur unique. Une telle implémentation est réalisé dans le projet « RTBP Real-time biosurveillance program » (Waidyanatha et al., 2011). Nos méthodes de fouille de textes, permettent d'alimenter automatiquement de tels flux GeoRSS.

4.2.5 Contribution d'experts

Finalement, le nombre de facteurs nécessaires pour intégrer des propositions telles que celles décrites précédemment ne devrait pas être sous-estimé. Cela inclut des tâches

longues et parfois coûteuses comme la collaboration avec des statisticiens et des informaticiens pour travailler sur les composantes de la collecte des données, le filtrage des données et leur traitement, la personnalisation des résultats, ainsi que la visualisation et la dissémination de l'information sanitaire aux épidémiologistes.

Être capable de détecter des événements sanitaires n'est pas suffisant pour avoir un système de veille utile. Afin d'avoir une valeur, le système de veille doit être en mesure de différencier entre des rapports inhabituels (signaux faibles) et des rapports non pertinents (bruit) et fournir précocement cette information aux utilisateurs. Le système PADI-web doit être flexible pour s'adapter à l'évolution des maladies sans aucun biais pour un pays ou une langue. Cependant, dans la pratique, le jugement des experts sera presque toujours nécessaire pour analyser et interpréter les signaux.

4.3 Conclusion

En conclusion, au-delà du développement d'un outil pour la veille du Web, c'est bien le concept même de l'intelligence épidémiologique qui a pu être proposé dans le cadre de ces travaux de thèse. En anticipant la détection des événements sanitaires, l'intelligence épidémiologique complète efficacement la surveillance traditionnelle et contribue à la consolidation d'un dispositif de veille internationale en santé animale. L'approche de fouille de textes et la contribution d'experts du domaine, sont des moyens nécessaires pour couvrir l'ensemble des menaces sanitaires potentielles. La mise en œuvre des approches automatisées devrait être considérée comme une priorité afin de bénéficier au mieux de ces synergies tout en assurant une meilleure utilisation des ressources disponibles. Nos résultats suggèrent des applications concrètes. Les maladies animales émergentes et exotiques continueront de poser des menaces potentielles pour la santé animale nationale et mondiale. L'intelligence épidémiologique est désormais une composante essentielle des systèmes d'alerte précoce. Pour répondre à ces enjeux, la poursuite de la recherche dans ce domaine est nécessaire.

Annexe A

Articles

A.1 Identification of terms for detecting early signals of emerging infectious disease outbreaks on the Web

A.2 Identification of associations between clinical signs and hosts to monitor the Web for detection of animal disease outbreaks

Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web

Elena Arsevska[1,2]*, Mathieu Roche[3,4], Pascal Hendrikx[5], David Chavernac[1,2], Sylvain Falala[1,2], Renaud Lancelot[1,2], Barbara Dufour[6]

[1] French Agricultural Research and International Cooperation Organization (CIRAD), Unit for control of exotic and emerging diseases in animals (UMR CMAEE), Campus international de Baillarguet, 34398 Montpellier, France

[2] French National Institute for Agricultural Research (INRA), Unit for control of exotic and emerging diseases in animals (UMR CMAEE), Campus international de Baillarguet, 34398 Montpellier, France
{elena.arsevska, renaud.lancelot, david.chavernac, sylvain.falala}@cirad.fr

[3] French Agricultural Research and International Cooperation Organization (CIRAD), Unit for land, environment, remote sensing and spatial information (UMR TETIS), 500 rue Jean-François Breton, 34093 Montpellier, France

[4] Laboratory of Informatics, Robotics and Microelectronics (LIRMM), UMR 5506, French National Centre for Scientific Research (CNRS), Montpellier University, 34000 Montpellier, France
mathieu.roche@cirad.fr

[5] French Agency for Food, Environmental and Occupational Safety (ANSES), Unit for coordination and support to surveillance (UCAS), 14 rue Pierre et Marie Curie, 94706 Maisons-Alfort, France
Pascal.HENDRIKX@anses.fr

[6] Alfort Veterinary School (ENVA), 7 avenue du Général de Gaulle, 94704 Maisons-Alfort, France
bdufour@vet-alfort.fr

*Corresponding author: elena.arsevska@cirad.fr

Abstract

Timeliness and precision for detection of infectious animal disease outbreaks from the information published on the web is crucial for prevention against their spread. The work in this paper is part of the methodology for monitoring the web that we currently develop for the French epidemic intelligence team in animal health. We focus on the new and exotic infectious animal diseases that occur worldwide and that are of potential threat to the animal health in France.

In order to detect relevant information on the web, we present an innovative approach that retrieves documents using queries based on terms automatically extracted from a corpus of relevant documents and validated with a consensus of domain experts (Delphi method). As a decision support tool to domain experts we introduce a new measure for ranking of extracted terms in order to highlight the more relevant terms. To categorize documents retrieved from the web we use Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers.

We evaluated our approach on documents on African swine fever (ASF) outbreaks for the period from 2011 to 2014, retrieved from the Google search engine and the PubMed database. From 2,400 terms extracted from two corpora of relevant ASF documents, 135 terms were relevant to characterize ASF emergence. The domain experts identified as highly specific to characterize ASF emergence the terms which describe mortality, fever and haemorrhagic clinical signs in *Suidae*.

The new ranking measure correctly ranked the ASF relevant terms until position 161 and fairly until position 227, with areas under ROC curves (AUCs) of 0.802 and 0.709 respectively.

Both classifiers were accurate to classify a set of 545 ASF documents (NB of 0.747 and SVM of 0.725) into appropriate categories of relevant (disease outbreak) and irrelevant (economic and general) documents.

Our results show that relevant documents can serve as a source of terms to detect infectious animal disease emergence on the web.

Our method is generic and can be used both in animal and public health domain.

Keywords: web, disease outbreak, text mining, term extraction, query, Delphi method

1. Introduction

Textual information sources on the web, such as publically available news articles, official disease reports and newsletters, have been found informative for early detection of emerging infectious disease outbreaks. Over the years, several web focused, event-based biosurveillance systems (further in the text web monitoring systems) have been created in order to detect infectious disease outbreak information from articles published on the web (Collier et al., 2008; Freifeld et al., 2008; Mykhalovskiy and Weir, 2006; Steinberger et al., 2008).

Despite the great potential in detection of early signals of infectious disease emergence from diverse web sources, the timeliness in detection of relevant articles is challenging due to the vast amount of ever growing publications on the web. Barboza et al., (2013) have shown that due to the access to diverse information, the web monitoring systems can detect avian influenza epizootics 12.7 days before the official notification to the World Organisation for Animal Health (OIE). In January 2014, an online post on the ProMED-mail system, referred to a local news media in Lithuania which reported complaints by hunters on increased mortality in wild boars at the border line with Belorussia (ProMED-mail, 2014). These reports are probably among the first signals of the spread of African swine fever (ASF) to a new territory well before official government reports were issued (OIE, 2014).

Therefore, automated identification of relevant articles on the web is the first step toward an effective event-based biosurveillance. In order to increase specificity in detection of relevant articles on the web, the current web monitoring systems widely use disease related search terms and Boolean queries (using the operators AND, OR, AND NOT, e.g., “african swine fever” OR “swine fever” AND NOT “classical swine fever”) and proposed by domain experts (Mantero et al., 2011), trained analysts (Mykhalovskiy and Weir, 2006) or based on a medical ontology (Collier et al., 2010). However, up to this point, no detailed work exists on how the current web monitoring systems identify the terms to detect signals of infectious disease emergence, especially in animal health. Moreover, identification of disease related vocabulary in animal health, faces additional challenges, such as multiple clinical signs and multiple hosts (Santamaria and Zimmerman, 2011; Smith-Akin et al., 2007).

Limited number of studies exploited the text mining approaches in order to construct terminology for infectious animal diseases (Anholt et al., 2014; Arsevska et al., 2014; Furrer et al., 2015) and therefore we propose an innovative methodology for identification of terms to build queries for monitoring the web for new and exotic infectious animal disease outbreaks. The method is based on automatic extraction of terms from relevant corpora of disease outbreak articles and identification of relevant terms using domain expert knowledge. The method is based among other on machine learning techniques and on a new function for ranking of the automatically extracted terms.

The methodology that we propose is generic and can be applied both to animal and public health domain.

For our experiments we use data on ASF. We choose this disease, because it is highly contagious and mortal in porcine animals; it has neither vaccine nor treatment and due to trade barriers the affected countries suffer great economic losses. This disease, endemic in sub-Saharan Africa and Sardinia (island in Italy) is an emerging threat to the European countries after its introduction for the first time in 2007 in the Caucasus region of Europe (Sánchez-Vizcaíno et al., 2013).

This rest of the work is organised as follows: Section 2 presents the related work, Section 3 presents our methodology, Section 4 presents our experiments and the results, Section 5 discusses the results, and Section 6 concludes the paper.

2. Related work

The earliest automatic web monitoring system, the Global Public Health Intelligence Network (GPHIN) founded by the Public Health Agency of Canada in 1997, in order to detect disease outbreak articles of potential relevance uses mainly two news aggregator feeds, Al Bawaba which covers information from the Middle East and North Africa and Factiva which covers information from more than 32,000 web sources worldwide. Once detected, articles are selected using a scanning tool based on a custom-built taxonomy of search terms and Boolean queries, updated regularly by GPHIN's human analysts. The search terms have a relevance score automatically attributed to the retrieved article; and an article is rejected if it has a relevance score below an established threshold (Keller et al., 2009; Mykhalovskiy and Weir, 2006).

The Medical Information System (MedISys), founded by the Joint Research Centre (JRC) of the European Commission (EC) in 2004, retrieves articles from diverse web sources, such as more than 150 medical web sites and 1,000 news portals worldwide in 40 languages. The retrieval is based on a list of predefined multilingual terms for each disease included in the system (alert definitions) and a combination of search terms, as proposed by domain experts. The MedISys covers more than 200 alert definitions for different public health-related subjects. A dedicated algorithm developed by the JRC team scans in real time the incoming articles for the alert definitions and keeps the articles that satisfy those criteria. An article is kept by the system and displayed in the disease category for dengue if the term "dengue" appears at least three times in the text of the article. However, if the text of the article includes an irrelevant term (proposed by experts), such as the term "concert", the article will not be selected. An article is also selected if two combinations of relevant search terms appear in the text of the article, such as "dengue" and "outbreak" and regardless of how many times they appear in the news article (Mantero et al., 2011; Steinberger et al., 2008).

The Argus system, hosted at the Georgetown University Medical Centre, United States, since 2004 detects events on the web that might threaten the human, plant, and animal health globally, except United States. It collects, in an automated process, media news pages, including blogs and official sources, such as World Health Organisation (WHO) and OIE, and interprets their relevance according to a specific set of concepts, search terms and Boolean queries relevant to each infectious disease covered by the system. Argus does not use scientific journals as a primary source to identify emerging events. Regional experts, collectively fluent in more than 40 languages, review manually the acquired articles before they are posted in the system (Nelson et al., 2010).

HealthMap, founded by the Boston Children's Hospital in 2006, draws terms from a continually expanding dictionary of pathogens (human, plant, and animal diseases) and geographic names (country, province, state, and city). Using a Bayesian classifier, articles are categorized by disease and location, automatically tagged according to their relevance and then overlaid on an interactive geographic map (Freifeld et al., 2008). The web search criteria used by HealthMap include disease names (scientific and common), symptoms, keywords, and phrases in seven languages (Brownstein et al., 2008). The system integrates outbreak data from multiple electronic sources, including online news aggregators (e.g., Google News), Really Simple Syndication (RSS) feeds, expert curated accounts (e.g., ProMED-mail), multinational surveillance reports (e.g., Eurosurveillance), and validated official alerts (e.g., from WHO and OIE) (Keller et al., 2009).

The conceptual framework for the BioCaster project founded in 2006 is a multilingual ontology - a structured public health vocabulary and terms, such as names of diseases, agents, clinical signs, syndromes and hosts, as well their relations, such as clinical signs or pathogen agents which affect a particular host. Terms are identified for eight Asia-Pacific languages by domain experts in biology, epidemiology, genetics and computational linguistic and linked with sources such as ICD10, MeSH, SNOMED CT and Wikipedia (Collier et al., 2007). Via a purpose built news aggregator BioCaster analyses documents from 1,700 different RSS feeds which are automatically classified based on the content using a Naïve Bayes classifier. For the relevant documents a named entity recognition is performed for 18 term types based on the BioCaster ontology (Collier et al., 2010).

The International Biosurveillance System (IBIS) founded by the University of Melbourne in 2013 uses as a main source of data the Goole search engine (general search, news, blogs). In the focus of IBIS are plant, aquatic and terrestrial animal infectious diseases. The articles of potential relevance are retrieved using a combination of search terms in English language, first determined by domain experts and later updated by registered users of the system (Lyon et al., 2013a).

A comparison of the main web monitoring systems is shown in Table 1.

3. Method

The work presented in this paper is part of the global methodology for monitoring the web for infectious disease emergence that we currently develop for the French epidemic intelligence team in animal health. Our focus is the new and exotic animal infectious diseases that occur worldwide and of potential threat to the animal health in France. The objective is to acquire relevant documents with information about disease outbreaks from diverse web sources (step 1). To retrieve relevant documents from the web, we use queries based on terms extracted from a corpus of relevant documents using a text mining approach. Given the amount of documents available on the web, we use machine-learning techniques for classification of new retrieved documents. According to the content two categories of documents are of our interest: relevant ("disease") documents and irrelevant documents ("economy" and "general") (step 2). Further, using text mining approaches from the relevant documents we extract relevant information, i.e. name of the disease – if mentioned, date and location of the outbreak, affected hosts, clinical signs etc. One part of this step focuses on extraction of terms to build queries (step 3). Finally, domain experts' asses the overall process and verify the extracted information (step 4). Figure 1 shows the workflow of our proposed methodology.

In this paper we focus on the document classification and the extraction of terms in order to build queries for improved information retrieval. Further, we introduce a new measure for ranking of extracted terms that favours the terms from relevant web sources and the terms that are highly frequent in the corpora of relevant documents. We also use the domain expert knowledge to evaluate these steps. We give some definitions and we describe our approach in the following subsections.

Definitions

In this paper we use the following definitions:

- **Query:** a phrase that consists of terms to be typed in quotes into a web source to search for disease outbreak information (e.g., "african swine fever outbreak", "high mortality" AND "wild boar" etc.). Quotes indicate a search for the entire phrase;
- **Corpus of documents:** result web pages of a query, transformed in a text format and stored in a database, before being categorized as relevant or irrelevant, according to their content;
- **Relevant documents:** text documents that consist of valuable disease outbreak information (e.g., articles about an occurrence of unexplained clinical signs or unknown disease, suspicion or confirmation of a known pathogen etc.);
- **Relevant terms:** names of diseases (including their synonyms and acronyms), names of diseases in differential diagnosis, clinical signs and hosts, extracted from a corpus of relevant documents or

proposed by domain experts that can be used independently or in combinations to build queries to detect signals of an infectious disease emergence on the web.

3.1. Data acquisition (step 1)

Using queries related to a disease of interest (e.g., “african swine fever”), web pages are retrieved from diverse web sources, such as general search engines, aggregator news sites and specialised health-related pages. For example, as web sources for our experiments in Section 4, we used the Google search engine and the PubMed database. The Google search engine is exploited by several web monitoring systems (Brownstein et al., 2008; Hartley et al., 2010; Lyon et al., 2013a). The PubMed database is the largest open access electronic database for biomedical literature which is updated regularly and has over 24 million articles (Falagas et al., 2007). For example, in our experiments, we retrieved web pages related to the ASF the emergence, i.e. news articles and scientific abstracts.

To clean the web pages, we developed a php script based on regular expressions. This script removes all HTML tags from the collected web pages (including formatting). The web pages are cleaned and then stored as documents in a text format. Then, the documents can easily be used by other applications and for other purposes (text mining for example). This treatment follows four distinct steps: i) a veterinary epidemiologist - user specialist identifies the web sources of interest, ii) next, the sources are stored in a MySQL database including the URL of the source, iii) the php script reads the URL source from the database, and iv) for each of the sources, the php script creates a text document which will be saved in a specific directory on the server. Finally, the text documents contain the main information of the document (such as the title and the body of the text).

To prepare the data for statistical analysis, the text is transformed into a term-document matrix, where the rows represent the individual terms and the columns contain the texts of each retrieved document (Munzert et al., 2015). The cells are filled with counts of how often a particular term appears in a given text (Bag of words method). The numbers, stop words, extra white space, and punctuation characters are removed from the texts. The terms that appear infrequently (sparse terms) in the term-document matrix, are also removed (e.g., terms that appear in 20% or less documents).

3.2. Document classification (step 2)

As we are interested to automatically classify the retrieved documents from the web, according to the content into relevant, “disease” and irrelevant, “economy” and “general” documents, we use a supervised approach in machine learning. For such purposes one veterinary epidemiologist – a user specialist, labels a training corpus of documents that have observations for which the category is known *a priori*, (e.g., “disease”, “economy” and “general”) while the labelling of the remaining

documents is performed by a classification algorithm automatically (Ceri et al., 2013). The learnt models are then used to classify the future streams of newly retrieved documents (Liu, 2007).

We deploy two widely used machine learning algorithms (classifiers) for text classification: (i) an original Naïve Bayes (NB) approach - Discriminative Multinomial Naïve Bayes (DMNB) (Su et al., 2008), and a version of Support Vector Machine (SVM) - Sequential Minimal Optimisation (SMO), using a polynomial kernel (Schölkopf et al., 1999). These two classifiers have been deployed in several web monitoring systems (Freifeld et al., 2008; Torii et al., 2011; Zhang et al., 2009). The results of this classification are presented in Section 4.3.

3.3. Extraction of terms (step 3)

Once classified, the relevant documents, i.e. which consist of disease outbreak information, serve as a corpus for automatic extraction of terms. For this purpose we use *BioTex*[1], a tool that combines linguistic and statistic information adapted to biomedical area (Lossio-Ventura et al., 2016). To select appropriate terms *BioTex* uses two principles: i) implementation of a relevant combination of information retrieval techniques and statistical methods, e.g., term frequency - inverse document frequency (TF-IDF), OKAPI (a cross-platform and open-source set of components and applications that give localisation and translation documentation and software), and C-value measures; and ii) a use of a list of syntactic structures of the terms that have been learnt with relevant sources, e.g., medical subject headings (MeSH). The terms extracted with *BioTex* can be either simple - one term (e.g., "pig"), or composed, multi-term (e.g., "domestic pig") - as we did in our experiments (Section 4). Table 2 shows an example of extracted terms from the corpus of ASF relevant documents retrieved from the Google search engine and the PubMed database and ranked according to *Biotex*.

3.3.1. Identification of relevant terms

For each disease and each list of extracted terms, a veterinary epidemiologist – user specialist with a knowledge in information retrieval, identifies the relevant terms, i.e. terms which characterize the disease of interest, such as in our experiments (Section 4), the terms that characterise ASF: the disease name (e.g., "african swine fever episode", "asfv introduction"), the clinical signs (e.g., "group mortality", "sick pigs"), the hosts (e.g., "wild boars", "domestic pig populations") and the diseases in differential diagnosis (e.g., "swine fever kills", "swine fever reported"). Table 3 shows the terms identified as relevant to characterize ASF.

3.3.2. Ranking of terms

As mentioned before, one of the core steps of our proposed methodology is a new measure for ranking the terms extracted with *Biotex* from the corpora of relevant documents from diverse web sources. The

[1] http://tubo.lirmm.fr/biotex/

purpose of the new ranking measure is to help the user specialist in the selection of relevant terms. More precisely, the new measure, *w(t)* (formula (1)), takes into account the *BioTex* ranking of the terms, and the quality of the web source (i.e. sources that give more relevant terms, as identified by a veterinary epidemiologist; Section 3.3.1). With this original measure, the higher the rank of a term extracted with *BioTex* and the higher the quality of a web source (i.e. the weight attributed to each web source by a veterinary epidemiologist), the higher is the *w(t)* rank calculated for that term.

In the formula (1), *t* represents the term, S_i is the web source, $rank_{Si}(t)$ is the *BioTex* rank of the term *t* from a web source S_i, and α_i is the weight attributed to a source S_i.

$$w(t) = \sum \alpha_i \times \frac{1}{rank_{Si}(t)} \quad \text{with } \alpha_i \in [0,1] \text{ et } \sum \alpha_i = 1 \tag{1}$$

The terms that are common in all lists of extracted terms, keep the automatic ranking. If a term is not present in the extracted list from a source S_i we affect the following "artificial" ranking: *max(rankSi(t))+1*. The aim is to consider a very low ranking weight for terms present in only one *BioTex* list.

For example, in the experiments for ASF (Section 4), we used two web sources: the Google search engine (S_1) and the PubMed database (S_2). A relevant corpus of documents was constituted from both sources S_i and consequently two lists of 1,200 terms were extracted from each source S_i. 77 terms were identified as relevant from the source S1 (i.e. PubMed database) and 58 terms were identified as relevant from the source S2 (i.e. Google search engine). In order to confirm this proportion, we affect the following weights: α1=0.57 and α2=0.43.

For a term present in both *BioTex* lists, such as in our experiments, the term "wild boar" ranked automatically on the 1st position in the list of extracted terms from the corpus of relevant documents from the PubMed database, and on the 3rd position in the list of extracted terms from the corpus of relevant documents from the Google search engine; this term *t* had a *w(t)* of 0.713, calculated as: ((1/1) × 0.57) + ((1/3) × 0.43).

For a term present in one *BioTex* list, such as in our experiments, the term "asfv introduction", extracted only from the corpus of relevant documents from the PubMed database and ranked on the 4th position in the list of extracted terms; this term received a constant value, max($rank_{Si}(t)$)+1, meaning it was re-ranked to the 1201th position. This term *t*, had a *w(t)* of 0.143, calculated as: ((1/4) × 0.57 + (1/1201) × 0.43).

Table 2 shows an example of extracted terms from the corpus of ASF relevant documents retrieved from the Google search engine and the PubMed database and ranked according to the new *w(t)* measure.

3.4. Evaluation (step 4)

Finally, in order to evaluate the key steps of our approach, we evaluate the performance of the document classification, the new ranking of the terms based on the *w(t)* measure and the relevance of terms to characterise the disease of interest and to detect signals of its emergence from the information published on the web.

3.4.1. Document classification

To evaluate the performance of the document classification we measure the ability of the NB and SVM classifiers to correctly classify documents with different contents into the categories "disease", "economy" and "general". Additionally, the results of the document classification allow us to compare the performance of each of the classifiers.

For this purposes we calculate accuracy, precision, recall, and F-score (Liu, 2007).

Accuracy is the number of all correctly classified documents divided by the total number of documents.

Precision is the number of correctly classified documents for a category *(i)* divided by the total number of documents classified as a category *(i)* and represents the correctness of the classification.

Recall is the number of correctly classified documents for a category *(i)* divided by the total number of documents in category *(i)* and represents the completeness of the classification.

F-score is two times precision and recall for a category *(i)* divided by the sum of precision and recall for a category *(i)* and represents their harmonic mean.

In cases of a small data set we use a method of cross-validation (in our experiments we had a total of 545 ASF documents). More precisely, to learn a classifier the available data are partitioned into *n* equal-size disjoint subsets and then each subset is used as a test set and the remaining *n*-1 subsets are combined as a training set. This procedure is then run *n* times, which gives *n* accuracies. The final estimated accuracy of learning from this data set is the average of the *n* accuracies (Liu, 2007). In this manner, in our experiments we compared the performance of the NB and SVM classifiers through a 10-fold cross-validation.

3.4.2. Ranking of the terms

To evaluate the performance of the new *w(t)* measure to privilege the terms ranked high with *BioTex* and the terms from relevant web sources (Section 3.3.2)., we calculate the area under the receiver-operating characteristic curve (AUC). The AUC of the *w(t)* measure is equivalent to the probability

that the *w(t)* measure will rank a randomly chosen positive instance (relevant term) higher than a randomly chosen negative instance (irrelevant term) (Fawcett, 2006). The distribution of all values of AUC graphically allows us to select an optimal threshold for the *w(t)* measure.

3.4.3. Relevance of the terms

Finally, using a Delphi method expert elicitation, a panel of domain experts – specialists for the disease of interest evaluate a representative number of relevant terms, extracted with text mining from a corpus of relevant documents (identified by a veterinary epidemiologist; Section 3.3.1) according to their specificity to characterize the disease of interest and to identify early signals of its emergence on the web. For example, in our Experiments we proposed the domain experts terms that represent the ASF host, its clinical signs, the disease in differential diagnosis and the associations thereof (Section 4.6). The specificity of the terms can vary from non-specific to highly specific terms. Independently, the domain experts, propose other terms that characterize the disease of interest and can be used as queries to detect its emergence on the web. The main objective of the Delphi method is to reach a consensus through group responses. The major steps of the Delphi method are: i) identification of participants, ii) questioning the participants, iii) analysis of responses and identification of responses that are in disagreement with the group consensus, and iv) share of the group responses anonymously and allowing participants to revise answers (Vangay et al., 2014).

Delphi method expert elicitation has been successfully used to answer public and animal health issues when other knowledge was not sufficient or not available (Economopoulou et al., 2014; Debin et al., 2013; Gustafson et al., 2013).

To quantify the degree of agreement between the panel of domain experts (inter-rater reliability), we calculate the Krippendorff alpha coefficient (α) (Hallgren, 2012; Krippendorff, 2011). Values of α between 0.21 and 0.40 are “fair”, those between 0.41 and 0.60 are “moderate”, those between 0.61 and 0.80 are “substantial”, and those between 0.81 and 1.00 are “almost perfect” (Hausberg et al., 2012).

The following Section presents the experiments and the results of our work.

4. Experiments and Results

4.1. Data

Between June and September 2014, we retrieved documents related to African swine fever (ASF) outbreaks for the period between 2011 and 2014 from two web sources: the Google search engine and the PubMed database. The documents were published in English language.

The Google corpus of documents (news articles) was manually collected using the query: “african swine fever outbreak”, which resulted with 545 ASF news articles. As relevant were considered the news articles with a principal information of suspicion or confirmation of ASF, unknown disease or

unexplained clinical signs in animals of the pig species, with a description of the event, such as place, time, number and species affected, clinical signs etc. (category "disease", n=181). Irrelevant were: i) the news articles with a principal information about a socio-economic impact of an ASF outbreak to a country or a region, and a secondary information about the event (category "economy", n=92) and ii) the news articles with general information about ASF (category "general", n=272).

The PubMed corpus of documents (abstracts) was manually collected using the query: "(african swine fever [Title] AND has abstract [text] AND ("2011/01/01"[PDat]: "3000/12/31"[PDat])) AND has abstract [text] AND ("2011/01/01"[PDat]: "3000/12/31"[PDat])", which resulted in 118 ASF abstracts. As relevant were considered the abstracts indexed with the term "epidemiology"[2], and when not indexed, consisted of epidemiologic information about ASF (n=45).

4.2. Statistical analysis

The data were analysed with the statistical program R (packages tm, AUC, irr) (R Development Core Team, 2009) and the text mining software Weka (Witten and Frank, 2005).

4.3. Performance of the document classification

We evaluated the performance of the document classification on a predictive model from a test set of 545 ASF Google news articles (Section 4.1). As described in Section 3.2, we used NB and SVM classifiers and compared their predictive performance using a 10-fold cross-validation. We measured accuracy, precision, recall and F-score.

Both classifiers had good predictive abilities; NB was slightly better than SVM, i.e. NB had accuracy of 0.747 (correctly classified 409 instances from 545 instances) and SVM had accuracy of 0.725 (correctly classified 399 instances from 545 instances). Highest performance results were obtained for the news articles from the category "general", with F-score of 0.831 for NB and F-score of 0.81 for SVM. The NB over performed the SVM for the category "disease", with F-score of 0.744 for NB, and F-score of 0.669. Lowest performance results were obtained for the category "economy", with F-score of 0.503 for NB and F-score 0.584 for SVM.

Table 4 shows the performance results of the two classifiers.

4.4. Terms that characterise African swine fever

From 2,400 terms extracted with the text mining, 135 terms characterised ASF (Table 3). The corpus of relevant documents from the Google search engine was a source of 58 relevant terms:

- Name of the disease (ASF) (19 terms),
- Disease in differential diagnosis (swine fever) (14 terms),

[2] MeSH (Medical Subject Headings)

- Mortality, fever and haemorrhagic clinical signs (including haemorrhagic pathologic signs) (15 terms),
- ASF host (10 terms).

The corpus of relevant documents from the PubMed database was a source of 77 relevant terms:

- Name of the disease (ASF) (39 terms),
- Disease in differential diagnosis (swine fever) (4 terms),
- Mortality, fever and haemorrhagic clinical signs (including unspecified clinical signs) (16 terms), and
- ASF host (18 terms).

The terms hosts represented mainly two populations of *Suidae* affected by ASF, domestic *Suidae* - represented by the terms synonyms: pigs and swine (26 terms); and wild *Suidae* – represented by the terms synonyms: boars, warthogs, tampans, suids, swine and pigs (15 terms). The terms clinical signs represented mainly three groups of clinical signs: fever (13 terms), mortality (12 terms) and haemorrhagic (4 terms).

4.5. Quality of the ranking function

As described in Section 3.3.2., to evaluate the quality of the new ranking function, we measured the ability of the proposed *w(t)* measure to discriminate relevant from irrelevant terms in the list of 2,400 newly ranked terms. We also determined the terms with highest twenty and highest 200 AUC values (thresholds).

Figure 2 shows the values of the AUC for the list of 2,400 ranked terms according to the *w(t)* measure. The analysis of the AUC values associated to the new *w(t)* measure showed that the ranking was very relevant until position 161 (AUC of 0.802), with a highest value of AUC of 0.902 at position 29; and it was fairly relevant until position 227 (AUC of 0.709). Sixteen relevant terms were among the highest 227 ranked terms with the *w(t)* measure, i.e. seven terms described the name of the disease: “asfv introduction”, “asf outbreaks”, “asf infection”, “asf incursion”, “african swine fever episode”, “african swine fever outbreak”, “prevalence of asf”; one term described the haemorrhagic clinical signs: “devastating haemorrhagic fever” and eight terms described the ASF hosts: “wild boar(s)”, “pig farms”, “wild pig”, “slaughter pigs”, “warthog burrows”, “district pigs” and “domestic pig”.

4.6. Relevance of the terms

As described in Section 3.4.3, using a Delphi method and an online questionnaire (available from the first author), we elicited 21 domain experts - specialists for ASF. The experts were virologists, entomologists, epidemiologists, officials from governmental, international and farmer’s organisations

from Europe, Americas and Africa. The questionnaire was written in French and English language, proof read by the authors of this work.

The domain experts evaluated sixteen terms. Three terms represented the ASF clinical signs: fever - "fever outbreak reported", mortality - "high mortality" and haemorrhagic - "devastating haemorrhagic fever"; one term "wild boar" represented the ASF host; six terms were associations between hosts and clinical signs for ASF: fever - "district pigs fever outbreak", "extensive free range pig suspected swine fever", mortality - "fresh outbreak lethal pig disease" , "dead wild boar", "wild pigs gross mortality", and haemorrhagic - "pig farms haemorrhagic fever"; one term represented the disease in differential diagnosis: "swine fever kills". The domain experts also evaluated four terms irrelevant to ASF: "arthrogryposis hydranencephaly syndrome", "malformed offspring", "aborted foetuses" and "enzootic outbreak of abortions", which were extracted from a corpus of relevant documents not related to ASF (i.e. Schmallenberg disease).

For each term, the domain experts evaluated four levels of specificity: i) "non-specific" term did not characterise ASF and it was neither sensitive nor specific to detect ASF emergence on the web; ii) term with "low specificity" characterised ASF but also other human and animal diseases and it was sensitive but not sufficiently specific to detect ASF emergence on the web; iii) term with "medium specificity", characterised ASF and other porcine diseases and it was sensitive but with lower specificity to detect ASF emergence on the web; and iv) "highly specific" term characterised ASF and it was sufficiently sensitive and sufficiently specific to detect ASF emergence on the web.

We conducted two rounds of Delphi. Twenty one domain experts participated in the first round of Delphi and evaluated all previously described sixteen terms. Seven experts participated in the second round of Delphi and evaluated four terms for which a consensus was not reached in the first round of Delphi.

Figures 3A and 3B show the domain experts evaluations in the first and the second round of Delphi.

The inter-rater reliability in the first round of Delphi was $\alpha = 0.411$, meaning the overall consensus between the 21 domain experts about the specificity of all 16 proposed terms was moderate. The inter-rater reliability in the second round of Delphi was $\alpha = 0.276$, meaning the overall consensus between the seven domain experts about the specificity of all four proposed terms was low.

Terms "associations" and terms "differential diagnosis"

The term association which described haemorrhagic clinical signs, "pig farms haemorrhagic fever", was the most discriminant term, evaluated by the majority of the domain experts (15/21; 71%) as highly specific to characterize ASF and to identify early signals of its emergence on the web.

The terms associations which described mortality were evaluated by the majority of the domain experts as medium to highly specific, “lethal pig disease” (19/21; 90%), “dead wild boar” (18/21; 86%), “fresh outbreak lethal pig disease” (17/21; 81%), “wild pigs gross mortality” (17/21; 81%). These terms were considered as providing a good sensitivity, but a lower specificity for identification of early signals of ASF emergence on the web.

The term association which described fever, “district pigs fever outbreak”, was evaluated as low to medium specific (15/21; 71%), i.e., not specific enough to identify early signals of ASF on the web.

The majority of the domain experts evaluated the term which described swine fevers, “swine fever kills”, as medium to highly specific (20/21; 95%) to characterize ASF and to identify early signals of its emergence on the web. Similarly, as medium to highly specific was evaluated the term: “extensive free range pig suspected swine fever” (16/21; 78%). Swine fever is a term that encompasses two emerging porcine diseases: African swine fever and Classical swine fever. These two diseases show identical clinical signs and the only way to differentiate them is with laboratory diagnostics. Therefore it is of no surprise to us that the majority of domain experts considered that the terms related to swine fevers are relevant to detect any suspicion of emergence of swine fevers (low specificity), in order to reduce the risk of missed information (high sensitivity).

Terms “clinical signs” and terms “hosts”

In the first round of Delphi, the domain experts did not reach a consensus about the specificity of the terms which described the host and the clinical sign of ASF, i.e. “fever outbreak reported”, “high mortality” “devastating haemorrhagic fever” and “wild boar”. For these terms we conducted a second round of Delphi.

The term “high mortality” was evaluated as a highly specific term (12/21; 57%) in the first round of Delphi, while it was seen as non-specific (5/7; 71%) in the second round of Delphi. Similarly, the term “wild boar” in the first round of Delphi was evaluated as a term with low to medium specificity (15/21; 71%); it was evaluated as a term with low specificity in the second round of Delphi (4/7; 57%), i.e. not specific to identify early signals of ASF on the web.

The term “devastating haemorrhagic fever” was evaluated as medium to highly specific in the first round of Delphi (19/21; 90%) and low to medium specific in the second round (5/7; 71%). Similarly, the term “fever outbreak reported” was evaluated as a term with low to medium specificity in the first (18/21; 86%) and the same results were obtained in the second round of Delphi (6/7; 86%), meaning not specific enough to identify early signals of ASF on the web.

“Irrelevant” terms

The domain experts reached a consensus about the specificity of the irrelevant terms; they were all evaluated by the majority of the experts as non-specific to low specific to characterize ASF and detect its emergence on the web.

The term “arthrogryposis hydranencephaly syndrome” was evaluated as a non-specific term (14/21; 67%).

The term “enzootic outbreak of abortions” was evaluated as a term with low specificity (12/21; 57%).

The terms “malformed offspring” and “aborted foetuses” were evaluated as non-specific to low specific terms (19/21, 90% and 13/21, 62% respectively).

4.7. Expert proposals of relevant terms

In the first round of Delphi, the panel of domain experts were also invited to propose terms that are highly specific to characterize ASF and can be used to detect its emergence on the web. The main 53 proposals were:

- Name of the disease (ASF) (8 proposals),
- Disease in differential diagnosis (Classical swine fever) (1 proposal),
- Mortality in pigs and wild boars (20 proposals),
- Haemorrhagic disease including haemorrhagic fever in pigs and wild boars (7 proposals),
- Non-specified disease or clinical signs in pigs and wild boar (highly contagious disease, serious disease, sick animals, emerging disease) (5 proposals),
- Pathological signs in pigs (haemorrhagic and hypertrophic spleen, haemorrhagic lymph nodes, septicaemia) (5 proposals),
- Combined clinical signs in pigs (mortality, fever and haemorrhagic lesions, lack of appetite and prostration) (4 proposals),
- Digestive clinical signs (haemorrhagic diarrhea) in pigs (1 proposal),
- Reproductive clinical signs in sows (abortions) (1 proposal),
- Socio-economic changes (panic among pig farmers) (1 proposal).

Table 5 shows the full list of proposed terms by the domain experts.

5. Discussion

In this paper, we presented part of the methodology for monitoring the web for infectious disease emergence that we are developing for the French epidemic intelligence team in animal health. We evaluated the performance of two classifiers (NB and SVM) that we deploy in our methodology and we evaluated the relevance of the terms extracted with text mining from a corpus of relevant documents using a Delphi method expert elicitation. We finally evaluated the performance of a new

ranking measure, *w(t)*, which favours the terms extracted from all corpora of relevant documents and from relevant web sources.

The two classifiers that we deployed in our methodology showed good performance to correctly classify relevant ASF Google news articles (category "disease"), with NB slightly better than the SVM (F-score of 0.744 in NB compared to 0.669 in SVM). Studies conducted on data sets from disease related news articles from Google news, ProMed-mail disease reports and animal health web pages showed similar results (Freifeld et al., 2008; Torii et al., 2011; Zhang et al., 2009). As described by Torii et al. (2011) the stable performance of the NB is due to the presence of terminology strongly implicative to be positive in the articles (i.e. keywords that describe a disease event). Similarly, Freifeld et al., (2008) investigated the accuracy of a Bayesian classifier used in the HealthMap project in order to detect disease and location on a test corpus of ProMed-mail reports and Google news articles. They also suggested that the regular structure and the presence of data curated specifically for disease outbreak reporting in ProMed-mail reports influenced the better performance of the classifier with accuracy of 0.91 on ProMed-mail reports, and accuracy of 0.81 on Google news articles (Freifeld et al., 2008). Similarly, in the work conducted by Zhang et al. (2009), NB slightly outperformed SVM in the classification of news articles for foot-and-mouth disease acquired from animal health related web pages (accuracy of 0.72 for NB compared to 0.70 for SVM).

Lower performance results our classifiers showed for the ASF Google news articles from the category "economy" (F-score of 0.503 for NB compared to 0.584 for SVM). Zhang et al., (2009) reported that errors in classification can be caused by news that contain information related to more than one category. In our experiments each news article was manually categorized based on its content. Some news articles described a socio-economic event due to an ASF outbreak and they were categorized as "economy" news articles. Some other news articles first described different control or eradication measures due to an ASF outbreak, and when they missed precise description of the disease event, they were categorized as "economy" news articles. Another reason that might have influenced the precision results of the classifiers was the imbalance of the categories of interest within the test set (Amrine et al., 2014). In our experiments, the news articles from the category "economy" was under-represented compared to the categories "disease" and "general". In future, we intend to evaluate the methods of ensemble of classifiers (Elrahman and Abraham, 2013; Heredia-Langner et al., 2015), which have shown to increase the performance of the classifiers when evaluated on data from social media (Adebayo, 2013; Tuarob et al., 2014; Zuccon et al., 2015) and medical records (Doan et al., 2012).

The text mining approach that we deployed in our experiments, gave us the possibility to choose the number of terms to be extracted from a corpus of relevant ASF documents. For this work, we

evaluated 2,400 extracted terms. For the purposes of monitoring the web and building queries we exploited only the terms which described the name of the disease, the disease in differential diagnosis, the clinical signs and the hosts. In this work, we did not take into consideration other terms extracted from the corpus of relevant documents, such as names of places, villages, regions and countries, e.g., "eastern Europe", "Krasnodar region", "Elembele district". However, in future we intend to exploit these terms in order to detect emerging regions and zones and monitor the spatio-temporal characteristics of infectious disease outbreaks. In the list of extracted terms, few, but present, were terms that indicated risk factors for transmission of the ASF virus, such as: "kitchen waste", "illegal import of meat" etc. These terms can be used in future to detect disease related events that can refer to a potential infectious disease emergence.

The new *w(t)* measure that we introduced in this work (Section 3.3.2) highly accurately ranked (AUC > 0.802) the relevant terms until position 161 and fairly ranked the relevant terms until position 227 (AUC > 0.709). These results suggest that the user specialist can use the *w(t)* measure as a decision support tool in order to prioritize terms ranked to a higher position rather than other terms ranked to a lower position.

Finally, in order to identify relevant terms to build queries and detect infectious disease emergence on the web, in our methodology we used the knowledge of two types of domain experts: a veterinary epidemiologist – user specialist with knowledge in information retrieval, who did a primary selection of relevant terms; and the common knowledge of a panel of domain experts (Delphi method) which validated the terms selected by the veterinary epidemiologist and proposed other relevant terms.

With a consensus, the domain experts identified as relevant the terms – associations which described haemorrhagic, fever and mortality clinical signs in *Suidae*. In future we intend to test these terms as queries. With a consensus, the domain experts also evaluated as irrelevant the terms that originated from a corpus of relevant documents from a disease not related to ASF – which confirmed their knowledge in the domain. However, in the experiments, we noted an overall variability of the answers given by the domain experts (moderate consensus in the first round of the Delphi and low consensus in the second round of the Delphi). The variability of the answers was influenced by the terms that were attributed to several categories at the same time (such as hosts and clinical signs). Cox et al. (2012) reported similar constraints while using an expert opinion to assess the risk of emergence and re-emergence of infectious diseases in Canada. Upon our exchanges with the domain experts, we noted that certain among the experts evaluated the terms from a domain knowledge point of view, which influenced the results of the overall agreement. This variability of the answers further on shows the challenges to construct a terminology for infectious animal diseases (multiple hosts and multiple

clinical signs) (Arsevska et al., 2014; Dórea et al., 2015; Furrer et al., 2015; Santamaria and Zimmerman, 2011; Smith-Akin et al., 2007). Some works suggest that the Delphi method has highest efficiency when personal interviews are conducted and with a smaller number of domain experts – what we intend to do in future (Gustafson et al., 2013).

Finally, the proposed terms by the domain experts, which corresponded with the terms extracted with *BioTex*, confirm the suitability of our text mining approach as a source of terms for improved monitoring of ASF emergence on the web.

6. Conclusion

We believe that our methodology provides a new insight into the monitoring of infectious disease emergence on the web. We bring an integrated approach of automatic extraction of terms using text mining and a domain expert knowledge to identify relevant terms for improved information retrieval. The approach is generic and can be used by animal and public health authorities.

Our work shows that corpora of relevant documents from diverse web sources can serve as sources of terms to detect infectious disease emergence on the web. In our experiments, however, the ASF abstracts from the PubMed database gave 20% more relevant terms compared to the ASF Google news articles. As source of terms to build queries, we intend to keep both sources and use the PubMed database as a source of scientific and technical terms and the Google search engine as a source of everyday, common terms. We also intend regular updates and evaluations of the terms as epidemiological patterns of infectious diseases can vary in time and place.

For monitoring the web for ASF emergence, we propose as relevant, the terms: i) name of the disease in order to detect disease outbreaks where ASF is confirmed; ii) associations between *Suidae* and haemorrhagic clinical signs, mortality and fever in order to detect early signals of potential ASF emergence.

Acknowledgements

We would like to thank all the experts that contributed to this work. This work was supported by a grant from the French Ministry of Agriculture, Food and Forestry (DGAL), the French Agricultural Research Centre for International Development (Cirad) and the SONGES Project (FEDER and Languedoc-Roussillon).

References

Adebayo, S., 2013. Evolving epidemic intelligence: Towards improved health events detection over social media streams (Master dissertation). University St Andrews.

Amrine, D.E., White, B.J., Larson, R.L., 2014. Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for bovine respiratory disease. Comput. Electron. Agric. 105, 9–19.

Anholt, R.M., Berezowski, J., Jamal, I., Ribble, C., Stephen, C., 2014. Mining free-text medical records for companion animal enteric syndrome surveillance. Prev. Vet. Med. 113, 417–422. doi:10.1016/j.prevetmed.2014.01.017

Arsevska, E., Roche, M., Lancelot, R., Hendrikx, P., Dufour, B., 2014. Exploiting Textual Source Information for Epidemiosurveillance, in: B. S. Clos et Al. (Ed.). MTSR 2014: 8th Metadata and Semantics Research Conference, Springer International Publishing Switzerland, pp. 359–361. doi:10.13140/2.1.4049.1522

Barboza, P., Vaillant, L., Mawudeku, A., Nelson, N.P., Hartley, D.M., Madoff, L.C., Linge, J.P., Collier, N., Brownstein, J.S., Yangarber, R., Astagneau, P., on behalf of the Early Alerting, Reporting Project of the Global Health Security Initiative, 2013. Evaluation of Epidemic Intelligence Systems Integrated in the Early Alerting and Reporting Project for the Detection of A/H5N1 Influenza Events. PLoS ONE 8, e57252. doi:10.1371/journal.pone.0057252

Brownstein, J.S., Freifeld, C.C., Reis, B.Y., Mandl, K.D., 2008. Surveillance Sans Frontiéres: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. PLoS Med. 5(7): e151. doi:10.1371/journal.pmed.0050151

Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P., Quarteroni, S., 2013. Web Information Retrieval. Springer Berlin Heidelberg, Berlin, Heidelberg.

Collier, N., Doan, S., Kawazoe, A., Goodwin, R.M., Conway, M., Tateno, Y., Ngo, Q.-H., Dien, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M., Taniguchi, K., 2008. BioCaster: detecting public health rumors with a Web-based text mining system. Bioinformatics 24, 2940–2941. doi:10.1093/bioinformatics/btn534

Collier, N., Goodwin, R.M., McCrae, J., Doan, S., Kawazoe, A., Conway, M., Kawtrakul, A., Takeuchi, K., Dien, D., 2010. An ontology-driven system for detecting global health events, in: Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, pp. 215–222.

Collier, N., Kawazoe, A., Jin, L., Shigematsu, M., Dien, D., Barrero, R.A., Takeuchi, K., Kawtrakul, A., 2007. A multilingual ontology for infectious disease surveillance: rationale, design and challenges. Lang. Resour. Eval. 40, 405–413. doi:10.1007/s10579-007-9019-7

Cox, R., Revie, C.W., Sanchez, J., 2012. The Use of Expert Opinion to Assess the Risk of Emergence or Re-Emergence of Infectious Diseases in Canada Associated with Climate Change. PLoS ONE 7, e41590. doi:10.1371/journal.pone.0041590

Debin, M., Souty, C., Turbelin, C., Blanchon, T., Boëlle, P.-Y., Hanslik, T., Hejblum, G., Le Strat, Y., Quintus, F., Falchi, A., 2013. Determination of French influenza outbreaks periods between 1985 and 2011 through a web-based Delphi method. BMC Med. Inform. Decis. Mak. 13, 138. doi:10.1186/1472-6947-13-138

Doan, S., Collier, N., Xu, H., Duy, P., Phuong, T., 2012. Recognition of medication information from discharge summaries using ensembles of classifiers. BMC Med. Inform. Decis. Mak. 12, 36. doi:10.1186/1472-6947-12-36

Dórea, F.C., Dupuy, C., Vial, F., Revie, C., Lindberg, A., 2015. Standardising Syndromic Classification in Animal Health Data. Online J. Public Health Inform. 7(1): e123. doi:10.5210/ojphi.v7i1.5789

Economopoulou, A., Kinross, P., Domanovic, D., Coulombier, D., 2014. Infectious diseases prioritisation for event-based surveillance at the European Union level for the 2012 Olympic and Paralympic Games. Euro Surveill. 19(15):pii=20770

Elrahman, S.M.A., Abraham, A., 2013. A Review of Class Imbalance Problem. J. Netw. Innov. Comput. 1, 332–340.

Falagas, M.E., Pitsouni, E.I., Malietzis, G.A., Pappas, G., 2007. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. FASEB J. 22, 338–342. doi:10.1096/fj.07-9492LSF

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27, 861–874. doi:10.1016/j.patrec.2005.10.010

Freifeld, C.C., Mandl, K.D., Reis, B.Y., Brownstein, J.S., 2008. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. J. Am. Med. Inform. Assoc. JAMIA 15, 150–157. doi:10.1197/jamia.M2544

Furrer, L., Küker, S., Berezowski, J., Posthaus, H., Vial, F., Rinaldi, F., 2015. Constructing a Syndromic Terminology Resource for Veterinary Text Mining. Proc. Conf. Terminol. Artif. Intell. 2015. 61–70.

Gustafson, L.L., Gustafson, D.H., Antognoli, M.C., Remmenga, M.D., 2013. Integrating expert judgment in veterinary epidemiology: Example guidance for disease freedom surveillance. Prev. Vet. Med. 109, 1–9. doi:10.1016/j.prevetmed.2012.11.019

Hallgren, K.A., 2012. Computing inter-rater reliability for observational data: An overview and tutorial. Tutor. Quant. Methods Psychol. 8, 23.

Hartley, D., Nelson, N., Walters, R., Arthur, R., Yangarber, R., Madoff, L., Linge, J., Mawudeku, A., Collier, N., Brownstein, J., Thinus, G., Lightfoot, N., 2010. Landscape of international event-based biosurveillance. Emerg. Health Threats J. 3: e3. doi:10.3134/ehtj.10.003

Hausberg, M.C., Hergert, A., Kröger, C., Bullinger, M., Rose, M., Andreas, S., 2012. Enhancing medical students' communication skills: development and evaluation of an undergraduate training program. BMC Med. Educ. 12, 16. doi: 10.1186/1472-6920-12-16

Heredia-Langner, A., Rodriguez, L.R., Lin, A., Webster, J.B., 2015. Selecting a Classification Ensemble and Detecting Process Drift in an Evolving Data Stream, in: Proceedings of the International Conference on Data Mining (DMIN). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), p. 31.

Keller, M., Blench, M., Tolentino, H., Freifeld, C.C., Mandl, K.D., Mawudeku, A., Eysenbach, G., Brownstein, J.S., 2009. Use of Unstructured Event-Based Reports for Global Infectious Disease Surveillance. Emerg. Infect. Dis. 15, 689–695. doi:10.3201/eid1505.081114

Krippendorff, K., 2011. Agreement and Information in the Reliability of Coding. Commun. Methods Meas. 5, 93–112. doi:10.1080/19312458.2011.568376

Linge, J., Steinberger, R., Weber, T., van der Goot, E., Khunhairy, D., Stilianakis, N., 2009. Internet surveillance systems for early alerting of health threats. Euro Surveill. 14(13):pii=19162.

Liu, B., 2007. Web data mining: exploring hyperlinks, contents, and usage data, Data-centric systems and applications. Springer, Berlin ; New York.

Lossio-Ventura, J.A., Jonquet, C., Roche, M., Teisseire, M., 2016. Biomedical term extraction: overview and a new methodology. Inf. Retr. J. 19, 59–99. doi:10.1007/s10791-015-9262-2

Lyon, A., Grossel, G., Burgman, M., Nunn, M., 2013a. Using internet intelligence to manage biosecurity risks: a case study for aquatic animal health. Divers. Distrib. 19, 640–650. doi:10.1111/ddi.12057

Lyon, A., Mooney, A., Grossel, G., 2013b. Using AquaticHealth.net to Detect Emerging Trends in Aquatic Animal Health. Agriculture 3, 299–309. doi:10.3390/agriculture3020299

Mantero, J., Belyaeva, J., Linge, J., 2011. How to maximise event-based surveillance web-systems the example of ECDC/JRC collaboration to improve the performance of MedISys., JRC Scientific and Technical Reports. Publications Office, Luxembourg.

Munzert, S., Rubba, C., Meißner, P., Nyhuis, D., 2015. Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining, Wiley. ed. United Kingdom.

Mykhalovskiy, E., Weir, L., 2006. The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health. Can. J. Public Health. 97, 42–44.

Nelson, N.P., Brownstein, J.S., Hartley, D.M., 2010. Event-based biosurveillance of respiratory disease in Mexico, 2007–2009: connection to the 2009 influenza A (H1N1) pandemic. Euro Surveill. 15(30). pii: 19626.

OIE, World Organisation for Animal Health 2014. African siwne fever, Lithuania. Immediate notification . http://www.oie.int/wahis_2/public/wahid.php/Reviewreport/Review/viewsummary?reportid=14690 (accessed 1.15.16).

ProMED-mail, 2014. Undiagnosed deaths, swine - Lithuania: wild boar, RFI. http://www.promedmail.org/post/2175896 (accessed 1.15.16).

R Development Core Team, 2009. R: A language and environment for statistical computing.

Sánchez-Vizcaíno, J.M., Mur, L., Martínez-López, B., 2013. African swine fever (ASF): five years around Europe. Vet. Microbiol. 165, 45–50. doi:10.1016/j.vetmic.2012.11.030

Santamaria, S.L., Zimmerman, K.L., 2011. Uses of Informatics to Solve Real World Problems in Veterinary Medicine. J. Vet. Med. Educ. 38, 103–109. doi:10.3138/jvme.38.2.103

Schölkopf, B., Burges, C.J.C., Smola, A.J., 1999. Advances in Kernel Methods: Support Vector Learning. MIT Press.

Smith-Akin, K.A., Bearden, C.F., Pittenger, S.T., Bernstam, E.V., 2007. Toward a veterinary informatics research agenda: an analysis of the PubMed-indexed literature. Int. J. Med. Inf. 76, 306–312. doi:10.1016/j.ijmedinf.2006.02.009

Steinberger, R., Fuart, F, Best, C, Von Etter, P, Yangarber, R, 2008. Text Mining from the Web for Medical Intelligence, in: Mining Massive Data Sets for Security, NATO Science for Peace and Security Series - D: Information and Communication Security. IOS Press, pp. 295–310.

Su, J., Zhang, H., Ling, C.X., Matwin, S., 2008. Discriminative parameter learning for Bayesian networks, in: Proceedings of the 25th International Conference on Machine Learning. ACM, pp. 1016–1023.

Torii, M., Yin, L., Nguyen, T., Mazumdar, C.T., Liu, H., Hartley, D.M., Nelson, N.P., 2011. An exploratory study of a text classification framework for Internet-based surveillance of emerging epidemics. Int. J. Med. Inf. 80, 56–66. doi:10.1016/j.ijmedinf.2010.10.015

Tuarob, S., Tucker, C.S., Salathe, M., Ram, N., 2014. An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. J. Biomed. Inform. 49, 255–268. doi:10.1016/j.jbi.2014.03.005

Vangay, P., Steingrimsson, J., Wiedmann, M., Stasiewicz, M.J., 2014. Classification of Listeria Monocytogenes Persistence in Retail Delicatessen Environments Using Expert Elicitation and Machine Learning. Risk Anal. 34(10):1830-45. doi: 10.1111/risa.12218

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2 edition. ed. Morgan Kaufmann, Amsterdam ; Boston, MA.

Zhang, Y., Dang, Y., Chen, H., Thurmond, M., Larson, C., 2009. Automatic online news monitoring and classification for syndromic surveillance. Decis. Support Syst. 47, 508–517. doi:10.1016/j.dss.2009.04.016

Zuccon, G., Khanna, S., Nguyen, A., Boyle, J., Hamlet, M., Cameron, M., 2015. Automatic detection of tweets reporting cases of influenza like illnesses in Australia. Health Inf. Sci. 3(Suppl 1): S4. doi: 10.1186/2047-2501-3-S1-S4

Tables

Table 1. Comparison of the web monitoring infectious disease surveillance systems, according to the sources, methods of information retrieval and classification algorithms

System name	GPHIN	MedISys	Argus	BioCaster	HealthMap	IBIS
Country	Canada	EU	USA	Japan	USA	Australia
Owner/ developer	Public Health Agency	Joint research centre	Georgetown University Medical Centre	University of Tokyo	Harvard University	University of Melbourne
Year launched	1997	2004	2004	2006	2006	2013
Access policy	Restricted	Open to public	Restricted	Open to public	Open to public	Open to public
Languages	9	40	40	8	7	1
Diseases covered	Infectious human and animal diseases	Infectious human and animal diseases, syndromes	Infectious human, animal and plant diseases	Infectious human and animal diseases	Infectious human and animal diseases, syndromes	Infectious animal diseases (terrestrial and aquatic)
Web sources	News aggregators (Al Bawaba, Factiva)	News web pages	News web pages, news aggregators (Google news), social media (blogs)	News aggregators	News aggregators (Google news), expert reports (Promed-mail), news media and institutional web pages (OIE, WHO)	General search engine (Google)
Information retrieval	Search terms, queries	Search terms, queries	Search terms, queries	BioCaster ontology	Search terms, queries	Search terms, queries
Identification of terms for queries	Analysts	Domain experts	Analysts	Domain experts	Analysts	Domain experts, Users
Document classifiers	Naïve Bayes, Support Vector Machine	Naïve Bayes	Naïve Bayes, Support Vector Machine	Naïve Bayes	Naïve Bayes	Not available
References	(Hartley et al., 2010; Keller et al., 2009; Mykhalovskiy and Weir, 2006)	(Hartley et al., 2010; Linge et al., 2009; Mantero et al., 2011)	(Hartley et al., 2010; Torii et al., 2011)	(Collier et al., 2007, 2010; Hartley et al., 2010)	(Brownstein et al., 2008; Freifeld et al., 2008; Keller et al., 2009)	(Lyon et al., 2013a, 2013b)

Table 2. Example of a list of extracted terms with text mining (columns: *Biotex* rank, term, term category and web source); and list of terms ranked based on the *w(t)* measure (columns: term, term category, web source, *w(t)* measure and new *w(t)* rank)

***Biotex* rank**	**Term**	**Term category**	**Web source**	***w(t)* measure**	**New *w(t)* rank**
1	wild boar	relevant - host	Google	0.713	1
3	wild boar	relevant - host	PubMed	0.713	2
1	wild boars	relevant - host	PubMed	0.620	3
3	wild boars	relevant - host	Google	0.620	4
2	european union	irrelevant	PubMed	0.285	5
2	mr speaker	irrelevant	Google	0.215	6
4	asfv introduction	relevant- disease	PubMed	0.143	7
5	asf outbreaks	relevant - disease	PubMed	0.114	8
4	lusaka province	irrelevant	Google	0.108	9

Table 3. List of terms extracted with text mining identified as relevant to characterise African swine fever

Term category	**Web source**	**Term**
Disease	Google search engine	african swine fever news, african swine fever outbreak affects, african swine fever outbreak reported, african swine fever outbreaks, african swine fever reported, african swine fever situation, african swine fever spreads, asf case in boar, asf outbreaks, asf-infected wild, asf-infected wild boar, boar detected with african swine fever, cases of asf, disease african swine fever, new outbreaks of african swine fever, outbreaks of asf, presence of the asf, scientific opinion on african swine fever, spread of asf
	PubMed database	acute asf, african swine fever episode, african swine fever outbreak, african swine fever outbreaks, african swine fever virus infection, asf diffusion, asf eradication, asf eradication programme, asf incursion, asf infection, asf outbreak, asf outbreak data, asf outbreaks, asf persistence, asf prevalence, asf situation, asf spread, asf virus antibody, asfv introduction, asfv prevalence, asfv spread, continued occurrence of asf, control of asf, control strategy for asf, eradication of asf, identification of asfv, ineffective management of asf, introduction of asf, occurrence of asf, potential risk of asf, prevalence of asf, regional asf outbreaks, risk for asfv, risk of asf, risk of asfv, sporadic outbreak of asf, spread of asfv, suitability for asf, temporal distribution of asf
Hosts	Google search engine	backyard pigs, boar population, district pigs, informal piggeries, pig farm, pig population, swine farms, wild boar population, wild pigs, wild boar, wild boars
	PubMed database	domestic pig, domestic pig populations, domestic pigs, extensive free range pig, pig farms, pig population, pig populations, slaughter pigs, warthog population, warthogs and tampans, wild boar populations, wild pig, wild pig population, wild pig populations, wild pig species, wild suids, wild boar, wild boars
Clinical signs	Google search engine	**Mortality**: dead pigs, dead wild boar, deadly pig disease, pigs dead **Fever**: fever case, fever case found, fever kills, fever outbreak affects, fever outbreak reported, fever outbreaks, fever reported, fever spreads, fever strikes

		Haemorrhagic: haemorrhagic fever **Pathologic signs**: muscle haemorrhages
	PubMed database	**Mortality**: gross mortality, group mortality, high lethality, high mortality, lethal pig disease, mortality loss, mortality losses **Fever**: fever episode, fever outbreak, fever outbreaks, fever virus infection **Haemorrhagic**: haemorrhagic disease, severe haemorrhagic disease, devastating haemorrhagic fever, haemorrhagic fever **Unspecified/ unknown disease**: sick pigs
Differential diagnosis	Google search engine	attack of swine fever, suspected swine fever, suspected swine fever outbreak, swine fever case, swine fever case found, swine fever kills, swine fever outbreak affects, swine fever outbreak reported, swine fever outbreaks, swine fever reported, swine fever scare, swine fever situation, swine fever spreads, swine fever strikes
	PubMed database	swine fever episode, swine fever outbreak, swine fever outbreaks, swine fever virus infection

Table 4. Performance of the Naïve Bayes and Support Vector Machine classifiers

Classification algorithm		**Naïve Bayes**			**Support Vector Machine**		
Performance		Recall	Precision	F-score	Recall	Precision	F-score
Category	"disease"	0.724	0.766	0.744	0.657	0.68	0.669
	"economy"	0.478	0.530	0.503	0.489	0.726	0.584
	"general"	0.860	0.804	0.831	0.864	0.763	0.810
Average		**0.750**	**0.745**	**0.747**	**0.732**	**0.729**	**0.725**

Table 5. List of terms proposed by domain experts as highly specific to characterize African swine fever

Expert ID	**Term**	**Term category**
53	high fever and mortality in pigs	**Combined clinical signs**
33	haemorrhagic syndrome and mortality of pigs	
56	Lack of appetite and prostration in pigs	
33	Hyperthermia and haematological disorders + swine	
55	Bloody diarrhea in pig breeding	**Digestive clinical signs**
28	haemorrhagic fever in pigs	**Haemorrhagic clinical signs**
57	haemorrhagic fever in pigs	
49	haemorrhagic disease of pigs	
32	pigs bleeding	
42	Haemorrhages in pigs	
57	haemorrhagic fever in boars	
33	skin redness + swine (pig / boar ...)	
32	pigs mortality	**Mortality**
33	grouped mortality of wild boars and pigs	
35	high mortalities of pigs	
38	wild boar mortality	
48	dead wild boar	
49	disease with high pork mortality	

56	High mortality of pigs	
60	sudden death of pigs	
35	high mortalities of wild boars	
38	dead wild boar	
42	High mortality in pigs	
53	increased observation of fallen animals in hunting grounds	
57	high pig mortality	
41	fatal disease of swine	
53	wild boar found dead	
55	raising mortality in pigs	
54	wild boar found dead	
60	boars found dead	
38	mass mortality in pigs	
28	mortality of wild boar	
60	abortions in sows	**Reproductive clinical signs**
42	Classical swine fever	**Differential diagnosis**
42	African swine fever	**Disease**
54	african swine fever	
38	African swine fever	
41	ASFV	
49	DNA arbovirus in pork	
41	ASFV vaccine	
55	ASF virus detection in wild boar	
48	ASF	
49	disease with big muddy pig spleen	**Pathologic signs**
57	haemorrhagic nodes in pigs	
57	Haemorrhagic nodes in boars	
42	septicaemia in pigs	
33	friable spleen, hypertrophy or splenomegaly + swine	
28	panic among pig farmers	**Socio-economic**
53	sick wild boar roaming at daylight	**Unspecified/**
30	highly contagious disease in domestic pigs and wild boar	**unknown disease**
49	highly contagious disease affecting the pork	
32	emerging disease in pigs	
49	serious illness without pork vaccine	

Figures

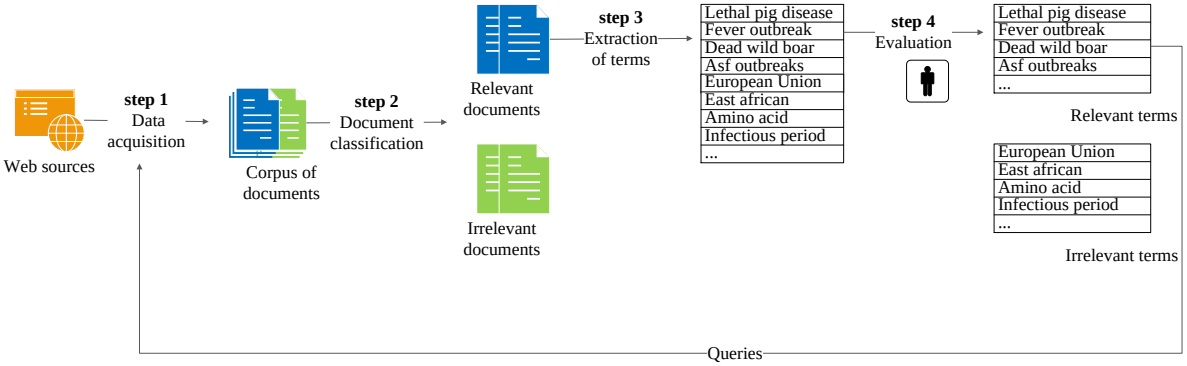


Fig 1. Workflow of the methodology for monitoring the web. This schema shows the example of identification of terms to detect signals of infectious disease emergence on the web.

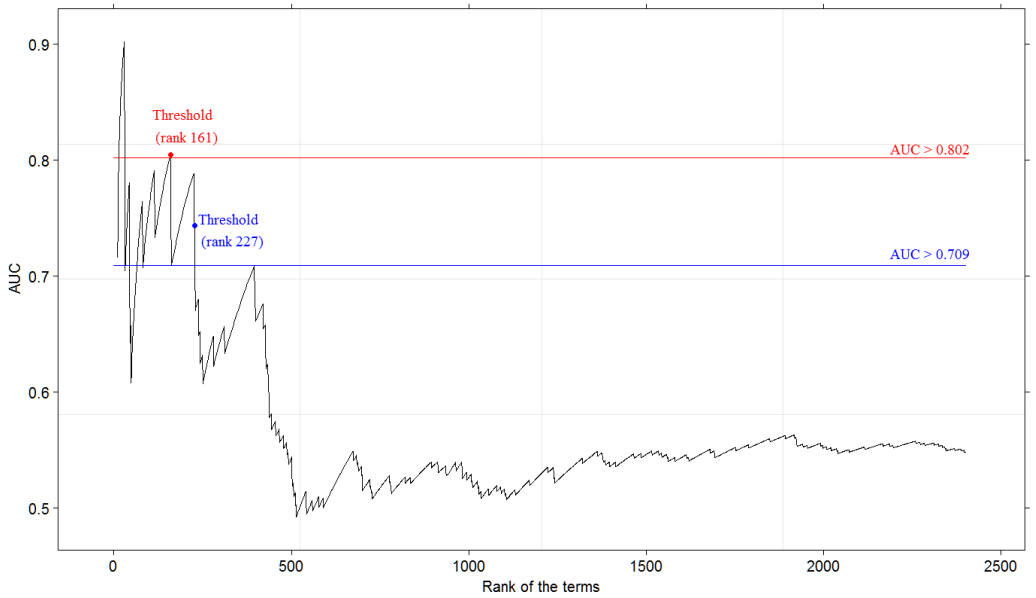


Fig 2. Values of the area under the receiver-operating characteristic curve (AUC) for the 2400 terms extracted from a corpus of ASF relevant documents and ranked according to the new w(t) measure. The term ranked on the position 227 is the last ranked term from the 200 terms with a highest AUC value (blue line). The term ranked on the position 161 is the last ranked term from the twenty terms with a highest AUC value (read line).

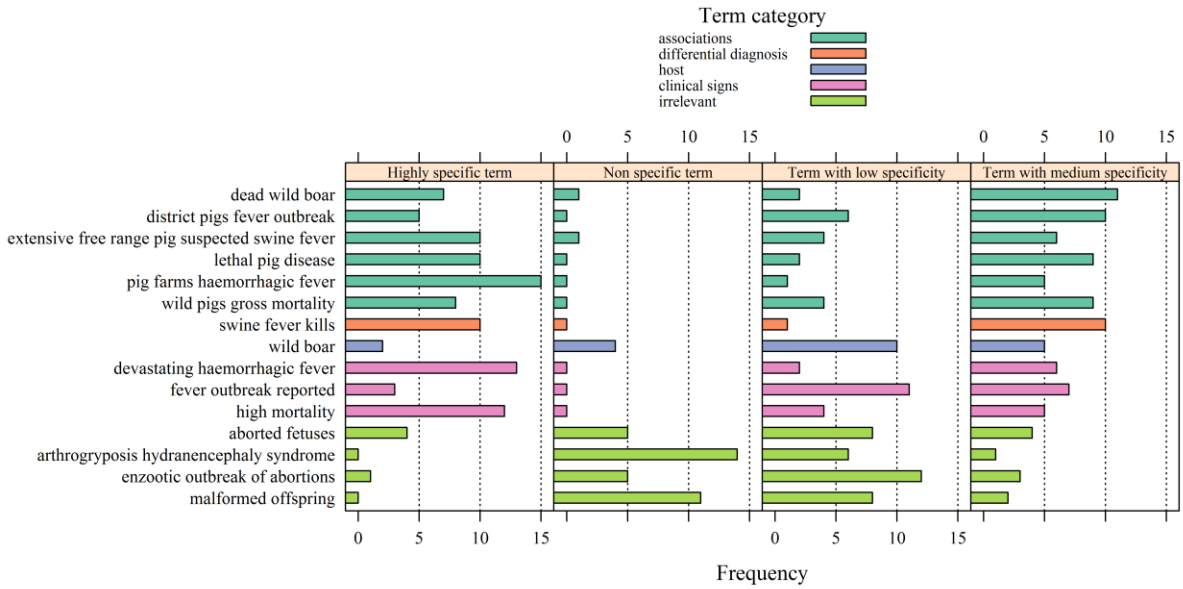


Fig 3A. Terms categorized according to the specificity to characterize African swine fever and to detect early signals of its emergence on the web, as evaluated by a panel of domain experts in the first round of Delphi.

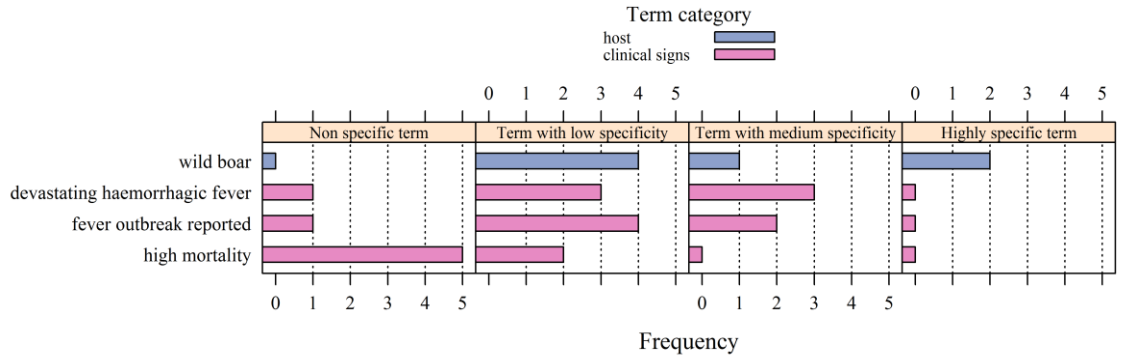


Fig 3B. Terms categorized according to the specificity to characterize African swine fever and to detect early signals of its emergence on the web, as evaluated by a panel of domain experts in the second round of Delphi.

Identification of associations between clinical signs and hosts to monitor the web for detection of animal disease outbreaks

Elena Arsevska[1, 2*], Mathieu Roche[3, 4], Pascal Hendrikx[5], David Chavernac[1, 2], Sylvain Falala[1, 2], Renaud Lancelot[1, 2], Barbara Dufour[6]

[1] French Agricultural Research and International Cooperation Organization (CIRAD), Unit for control of exotic and emerging diseases in animals (UMR CMAEE), Campus international de Baillarguet, 34398 Montpellier, France

[2] French National Institute for Agricultural Research (INRA), Unit for control of exotic and emerging diseases in animals (UMR CMAEE 1309), Campus international de Baillarguet, 34398 Montpellier, France

elena.arsevska@cirad.fr, renaud.lancelot@cirad.fr, david.chavernac@cirad.fr, sylvain.falala@cirad.fr

[3] French Agricultural Research and International Cooperation Organization (CIRAD), Unit for land, environment, remote sensing and spatial information (UMR TETIS), 500 rue Jean-François Breton, 34093 Montpellier, France

[4] Laboratory of Informatics, Robotics and Microelectronics (LIRMM), UMR 5506, French National Center for Scientific Research (CNRS), Montpellier University, 34000 Montpellier, France

mathieu.roche@cirad.fr

[5] French Agency for Food, Environmental and Occupational Safety (ANSES), Unit for coordination and support to surveillance (UCAS), 14 rue Pierre et Marie Curie, 94706 Maisons-Alfort, France

Pascal.HENDRIKX@anses.fr

[6] Alfort Veterinary School (ENVA), 7 avenue du Général de Gaulle, 94704 Maisons-Alfort, France

bdufour@vet-alfort.fr

*Corresponding author: elena.arsevska@cirad.fr

ABSTRACT

In a context of intensification of international trade and travels, the transboundary spread of emerging human or animal pathogens represents a growing concern. One of the missions of the national veterinary services is to implement international epidemiological intelligence for a timely and accurate detection of emerging animal infectious diseases (EAID) worldwide, and take early actions to prevent their introduction on the national territory. For this purpose, an efficient use of the information published on the web is essential. We present a comprehensive method for identification of relevant associations between terms describing clinical signs and hosts to build queries to monitor the web for early detection of EAID. Using text and web mining approaches, we present statistical measures for automatic selection of relevant associations between terms. In addition, expert elicitation is used to highlight the most relevant terms and associations among those automatically selected. We assessed the performance of the combination of the automatic approach and expert elicitation to monitor the web for a list of selected animal pathogens.

Keywords: information retrieval, disease emergence, text mining, web mining, ranking function, expert elicitation

INTRODUCTION

In recent years, the increased globalisation, movement of passengers and international trade has influenced the (re)emergence of new and exotic infectious diseases (Morens, Taubenberger, & Fauci, 2013). The traditional disease surveillance systems, organized via a multilevel health infrastructure, show delays in reporting disease outbreaks, starting from the first observation of clinical signs, laboratory confirmation until public communication (Chan et al., 2010). In consequence, the delays in reporting of disease outbreaks have themselves delayed the implementation of disease control measures, and thus influenced the spread of pathogens to uninfected territories (Khomenko et al., 2013).

As a complement to the traditional disease surveillance systems, several event-based surveillance systems (web monitoring systems) gather information about infectious disease outbreaks from automatically retrieved articles from the web (Collier et al., 2010; Freifeld, Mandl, Reis, & Brownstein, 2008; Steinberger, Fuart, Best, Von Etter & Yangarber, 2008). For this purpose, the current web monitoring systems use a specific vocabulary, such as names of diseases and clinical signs. However, it is not clear how these systems identify the vocabulary to mine the web, and especially for animal infectious diseases.

Innovative, data mining approaches have been successfully applied to clinical records in human medicine (Chapman, Dowling, & Wagner, 2004; Friedlin, Grannis, & Overhage, 2008) and to articles retrieved from the web, such as with the web monitoring systems HealthMap and BioCaster (Brownstein, Freifeld, Reis, & Mandl, 2008; Collier et al., 2008). However, the data mining approaches in animal health face challenges such as multiple vertebrate and possibly invertebrate (insects, ticks…) hosts and less formal vocabulary (Santamaria & Zimmerman, 2011; Smith-Akin, Bearden, Pittenger, & Bernstam, 2007). Indeed, a single pathogen agent can affect multiple animal hosts at the same time (cattle, sheep, goats, pigs) and can manifest with similar or different clinical signs. Furthermore, the clinical signs can vary from very typical, specific clinical signs (congenital malformations and deformations, blister-like sores on the skin and mucous membranes, haemorrhagic syndrome etc.) to less typical, non-specific clinical signs (fever, weakness, diarrhoea, etc.).

In this paper, we propose and evaluate a new method that combines text and web mining approaches to select relevant associations between terms describing hosts and clinical signs to build queries to mine the web and detect an emergence of an infectious animal disease outbreak. We focus on the new and exotic infectious animal diseases.

After the presentation of the related work in the next section, we present our method based on text and web mining approaches in section 3 and the results section 4. Finally, in the last section we discuss this paper and present our future work.

RELATED WORK

The Argus system, hosted at Georgetown University Medical Centre (USA), uses a simple method the to detect articles on the web, using search terms of multilingual disease names (Nelson et al., 2010; Nelson et al., 2012).

More complete, web search criteria are proposed by the HealthMap team, which include disease names (scientific and common), clinical signs, keywords, and phrases. The terms originate from a dictionary of pathogens (human, plant, and animal diseases) and geographic names (country, province, state, and city). HealthMap integrates outbreak data from multiple electronic sources,

including news feed aggregators (e.g., Google News), expert curated accounts (e.g., ProMED-mail), multinational surveillance reports (e.g., Eurosurveillance), and validated official alerts e.g., from WHO and OIE (Brownstein, Freifeld, Reis, & Mandl, 2008).

To detect articles on the web (e.g., Google News), the International Biosurveillance System (IBIS) uses the knowledge of registered users of the system that propose the search terms themselves. Registered users can edit the search terms and add or edit the tags in the articles that are relevant to the search terms (Lyon et al., 2013).

The GPHIN system developed by the Public Health Agency of Canada, retrieves automatically articles from news feed aggregators (e.g., Al Bawaba and Factiva), based on search terms and Boolean expressions, updated regularly by experts (Keller et al., 2009). The automated system filters out duplicate articles and establishes relevance scores for articles based on a value assigned to each search term. The automated tool discards articles with a score below an established threshold value to a search term. The scanned, filtered and categorized articles are evaluated by GPHIN's human analysts before being shared with the users of the system (Mykhalovskiy & Weir, 2006).

Similarly, MedISys, the web monitoring system created by the Joint Research Centre of the European Commission, retrieves articles from a list of web pages and news-feed aggregators. The retrieval is based on a predefined multilingual terms for each disease included in the system using a list of weighted terms; and/or combination of terms, as proposed by experts. For example, the system retrieves an article and displays it in the disease category for dengue if the term "dengue" appears at least three times in the text. However, if the text of the web article includes a term (based on a list of irrelevant terms), such as the term "concert", the article will not be retrieved. An article is also retrieved by the system if two combinations of search terms appear in the text, such as "dengue" and "outbreak" (Mantero, Belyaeva, & Linge, 2011).

The conceptual framework for the system BioCaster that has run from 2006 until 2012 was a multilingual ontology (Collier et al., 2007). The ontology was a structured public health vocabulary and relations of diseases, agents, clinical signs, syndromes and hosts. The BioCaster ontology adopted a thesaurus-like structure with synonym sets linking together terms across languages with similar meaning. Synonym sets used root terms. Root terms themselves were fully defined instances that provided bridges to external classification schemes and nomenclatures such as ICD10, MeSH, SNOMED CT and Wikipedia (Collier et al., 2010).

Zhang and Liu (2007) explored the detection of sentences containing disease outbreak information in ProMed-mail. They addressed the issue that the vocabulary used to report disease outbreaks overlaps with that used in other public health news, such as the treatment of diseases, and therefore search terms and text classifiers based on standard word features would not be effective in identifying disease outbreak articles. They used a dependency parser to identify sentence structures, and extracted verbs and adjectives directly modifying disease names as features. As additional features, they detected tense and negation status of verbs and extracted word n-grams (consecutive *n* words in text) and time terms (e.g., "today", "three months ago", etc.) from sentences (Torii et al., 2011).

Recently, Gesualdo et al., (2013) developed a minimally trained algorithm that exploits the abundance of health-related web pages to identify all jargon terms related to a specific technical term for avian influenza surveillance of Twitter. Then they translated an influenza case definition into a Boolean expression, with each clinical sign described by a technical term and all related

jargon terms, as identified by the algorithm. Subsequently, they monitored tweets that reported a combination of clinical signs meeting the case definition query and found a high correlation between the trend of their influenza-positive tweets and the trends identified by the US traditional avian influenza surveillance system.

Milinovich et al. (2014) found that the frequency of official notifications of exotic infectious diseases were correlated with a number of terms used as queries from the users on Google search engine. These terms varied from names of diseases or aetiological agent ("brucellosis, "Brucella"), colloquialisms ("flu", "help"), clinical signs ("cough", "cervical mucus") or medications or general health or treatment related queries ("clinical signs of dengue", "whooping cough treatment") to environmental ("flash flood", for leptospirosis) and behavioural vocabulary ("African tours", for malaria). Similarly, in our previous work we showed that queries consisting in the name of the animal disease such as "African swine fever outbreak" or "ASF outbreak" retrieve new articles about African swine fever (ASF) outbreaks, rather than general queries such as "fever outbreak" (Anonymous, 2014).

OUR PROPOSED METHOD

The work presented in this paper is a part of the methodology that we currently develop for the purposes of the French epidemic intelligence team in animal health (VSI). Since its creation in 2013, the VSI team focuses in detection of outbreaks of new and exotic animal infectious diseases of potential threat to France. Our contribution is in the methodology for monitoring the web, as one of the sources of information for the VSI team, and especially:

(i) automatic retrieval of documents based on search queries (queries);

(ii) automatic classification of documents based on their content (relevant – disease outbreak documents and irrelevant – any other document);

(iii) automatic extraction of information from relevant documents (name of a disease – if mentioned, date and location of an outbreak, affected hosts and clinical signs etc.); and

(iv) evaluation of the process, using domain expert knowledge.

In this paper, we present our work related to the first step of the methodology for monitoring the web. To retrieve documents from the web, we use queries based on terms, such as disease names, for a known disease, but also associations between terms describing clinical signs and hosts, for a new or unknown disease. We explore the use of web and text mining techniques and domain expert knowledge to identify relevant associations between terms describing clinical signs and hosts that can permit detection of new relevant documents about disease outbreaks. More precisely, our method consists of four steps (Figure 1):

1) extraction of terms from relevant documents, using a text mining approach;

2) identification of terms describing hosts and clinical signs, using expert knowledge;

3) association of terms describing hosts and clinical signs and calculation of statistical measures of association between terms, using text and web mining approaches; and

4) evaluation of the relevance of the terms and the associations thereof, using expert elicitation.

Figure 1. Workflow of the method for identification of relevant associations between terms describing hosts and clinical signs for monitoring disease emergence on the web

Extraction of Terms

A set of relevant (e.g., disease outbreak) documents for a certain disease serve as a basis for extraction of terms. Later, from each set of relevant documents, we automatically extract terms using *BioTex*[i], a tool that combines linguistic and statistic information adapted to biomedical area (Lossio Ventura et al., 2014). To select appropriate terms *BioTex* bases on two principles:

(i) implementation of a relevant combination of information retrieval techniques and statistical methods, e.g., term frequency - inverse document frequency (TF-IDF), OKAPI (a cross-platform and open-source set of components and applications that offer for localizing and translating documentation and software), or C-value measures;

(ii) use of a list of syntactic structures of the terms that have been learnt with relevant sources, e.g., medical subject headings (MeSH). The terms extracted with *BioTex* can be either simple, one term (e.g., "pig") or composed, multi-term (e.g., "domestic pig") expressions.

Identification and Association of Terms

In this second step, a domain expert identifies the terms describing hosts and clinical signs that characterize an emergence of a certain disease. For this purpose, the domain expert selects the terms from the list resulting from the former step.

Next, using text mining and web mining approaches, we propose statistical measures for automatic selection of relevant associations between terms describing clinical signs and hosts. We detail this contribution in the section below.

Evaluation

Economopoulou et al. (2014) elicited experts to prioritize the exotic infectious diseases of importance to the European Union. Mantero et al. (2011) used the same approach to identify keywords for the purposes of the MedISys public health web monitoring system of the European Commission. Expert elicitation was also adopted to evaluate the quality of different data mining approaches for identification of terminology for monitoring the web and for syndromic surveillance (Furrer et al., 2015; Steinberger et al., 2008). Therefore, we considered the expert elicitation as a necessary step for our method and we elicited a panel of domain experts to evaluate the level of relevance of the terms extracted with text mining as well as the associations thereof.

MEASURE OF THE ASSOCIATION BETWEEN TERMS

This section describes the data mining techniques that propose relevant associations between the terms describing hosts and clinical signs. This corresponds to the third step of our proposed method (see Figure 1).

Web Mining Approach

Pointwise Mutual Information and Information Retrieval (PMI-IR) is an algorithm using the AltaVista search engine to query the web to determine appropriate synonyms to a given query (Turney, 2001). For a given word, PMI-IR chooses a synonym among a given list. These selected terms correspond to the TOEFL questions. The aim is to identify the synonym that gives a better score. To obtain scores, PMI-IR uses several measures based on the proportion of documents where both terms are present on the web. Turney's formula is inspired from Mutual

Information (Church & Hanks, 1990). Our work applies this principle. First, we propose D_{web} measures based on the Dice's coefficient. Other statistical measures like Mutual Information (MI_{web}) (Church & Hanks, 1990) and Cubic Mutual Information (CMI_{web}) (Nazar, Vivaldi, & Cabré, 2008; Vivaldi, Màrquez, & Rodríguez, 2001) can be associated with web mining techniques.

The D_{web} measure computes the relationship between terms describing hosts (*h*) and clinical signs (*cs*). In this context, we measure the number of pages containing the terms *h* and *cs* together (i.e. *hit* (*h AND cs)*). We get this number of search pages with the search engine Exalead (http://www.exalead.fr). To calculate this association and the dependence of the terms, we also compute the number of pages where each term appears (i.e. *hit* (*h)* and *hit (cs)*). The following formula defines D_{web} with AND operator:

$$D_{Web}^{AND} = \frac{2 \times hit(h \text{ AND } cs)}{hit(h) + hit(cs)}$$

We chose Exalead because it offers the NEAR function like in Turney's approach (2001). NEAR is an operator that returns web pages where both terms *h* and *cs* are present in a 16-word window. We can adapt the previous formula with this new formula given below:

$$D_{Web}^{NEAR}(h, cs) = \frac{2 \times hit(h \text{ NEAR } cs)}{hit(h) + hit(cs)}$$

Moreover, we can use Mutual Information (MI) and Cubic Mutual Information (CMI) to estimate the dependency between *h* and *cs*:

$$MI_{Web}^{AND}(h, cs) = \frac{hit(h \text{ AND } cs)}{hit(h) \times hit(cs)}$$

$$CMI_{Web}^{AND}(h, cs) = \frac{hit^3(h \text{ AND } cs)}{hit(h) \times hit(cs)}$$

We can adapt these measures with the NEAR operator. The original MI measure (Church and Hanks, 1990) uses a logarithm (i.e. $\log_2 (P(x, y)/ (P(x) P(y)))$. The logarithm function is strictly increasing. So, the application or not of logarithm function does not change the ranking of the terms. Also, it is worth noting that *MI* tends to extract rare and specific dependencies (Vivaldi, Màrquez, & Rodríguez, 2001).

In our approach, we do not use other web mining measures such as "Google Similarity Distance" (Cilibrasi & Vitanyi, 2007) that needs specific parameters, e.g., the total number of web pages indexed by search engines.

Text Mining Approach

The *D*, *MI*, and *CMI* statistical measures can be used with the text mining approach. In this context, the Dice measure (*D*) is defined as the number of times (i.e. *nb* function) where *h* and *cs* appear in the same context over the sum of the total number of times each one appears in the corpus for each disease. We can adapt the text mining context with other statistical measures (*MI* and *CMI*). In this case, the choice of the context is crucial. In our data, both terms *h* and *cs* are

very rare in a given sentence so we use a larger context (i.e. an abstract or an article). To summarize, the following formulas give the text mining statistical measures:

$$D_{text}(h,cs) = \frac{2 \times nb(h \text{ AND } cs)}{nb(h) + nb(cs)}$$

$$MI_{text}(h,cs) = \frac{nb(h \text{ AND } cs)}{nb(h) \times nb(cs)}$$

$$CMI_{text}(h,cs) = \frac{nb^3(h \text{ AND } cs)}{nb(h) \times nb(cs)}$$

Combination of Text and Web Mining Approaches

Text mining and web mining approaches have a complementary behaviour. We apply text mining on specific texts in relation with each studied disease. Therefore, our global ranking function favours this type of approach. To overcome the case of rare associations, it is better to use a more global statistical measure calculated on the web, i.e. considering the web as a corpus. To take into account these principles, we propose a global measure called CMI_{global}. Based on our experiments (see following section) and the state-of-the-art review (Saneifar, Bonniol, Poncelet, & Roche, 2015; Vivaldi, Màrquez, & Rodríguez, 2001), the statistical measure CMI_{global} bases on the *CMI* criterion and the use of AND operator.

$$CMI_{global}(h,cs) = 1 + CMI_{text}(h,cs) \text{ if } CMI_{text}(h,cs) \neq 0 \text{ else } CMI_{Web}^{AND}(h,cs)$$

In this formula, the value 1 enables to favour the text mining value for the global ranking function (statistical measure) CMI_{global}. Table 1 shows examples of pairs of associations (hosts and clinical signs) for bluetongue and Schmallenberg virus infection.

Table 1: Examples of ranked associations with CMI_{global}

Rank	**Bluetongue**	**Schmallenberg virus infection**
	clinical signs / hosts	***clinical signs / hosts***
1	general clinical signs / pregnant ewes	stillborn bovine foetuses / bison
2	livestock deaths / sheep	aborted foetuses / sheep
3	embryonic death / cattle	deformed offspring / sheep
4	general clinical signs / sheep	stillborn bovine foetuses / deer
5	livestock deaths / cattle	aborted foetuses / cattle
6	livestock deaths / deer	deformed offspring / cattle
7	fever outbreak / sheep	stillborn bovine foetuses / calves
8	embryonic death / sheep	deformed offspring / lambs
9	fever outbreak / cattle	acute bronchopneumonia / bison
10	embryonic death / pregnant ewes	stillborn lambs / goat

EXPERIMENTS

Selection of Emerging Animal Infectious Diseases

For our experiments, we have chosen four animal infectious diseases of importance to animal health. These diseases have emerged in the last years in several countries in Europe and North Africa and they pose a risk of spread to other non-infected countries, including France.

African swine fever (ASF) is a highly contagious haemorrhagic viral disease of domestic and wild pigs. Typical clinical signs are high fever, loss of appetite, haemorrhages in the skin and internal organs, and death (Sánchez-Vizcaíno et al., 2013).

Foot-and-mouth disease (FMD) is a highly contagious viral disease of cattle, swine as well as sheep, goats, and other cloven-hoofed animals. Typical clinical signs are fever and blister-like sores on the tongue and lips, in the mouth, on the teats and between the hooves. The virus causes severe production losses and while the majority of affected animals recover, the virus often leaves them weakened and debilitated (Rodeia, 2008).

Bluetongue (BT) is a non-contagious, viral disease of domestic and wild ruminants transmitted by biting midges of the genus *Culicoides*. The clinical signs are most severe in sheep, resulting in weight loss, disruption in wool growth, and death (Wilson & Mellor, 2009).

In summer 2011, a new virus - Schmallenberg virus (SBV), appeared in The Netherlands and Germany. It quickly spread in many European countries. Affecting ruminants, it caused temporary, flu-like signs, with a short fever and drop in milk production, and sometimes abortion, stillbirth and severe foetal malformations (Wernike, Hoffmann, & Beer, 2013).

Textual Data

As sources of terms, we used a set of relevant documents (news articles and abstracts) for each of the studied diseases, which we manually retrieved in August and September 2014 from the Google search page and the PubMed database. We retrieved the news articles using the query in English-language “name of the disease” AND “outbreak” and we retrieved the abstracts using the query in English-language “name of the disease”, when present in the title or the body of the text. A relevant article described a disease outbreak event for each studied disease. A relevant abstract was indexed with the MeSH term “epidemiology”, and when not indexed, contained epidemiologic information for each studied disease. We used 181 news articles for ASF, 79 for BT, 84 for FMD, and 148 for SBV from Google and 45 abstracts for ASF, 116 for BT, 143 for FMD, and 53 for SBV from PubMed. The relevant documents principally covered topics about ASF outbreaks in Europe, FMD outbreaks in Northern Africa, BT outbreaks in the Balkans, and SBV outbreaks in Western Europe, for the period from 2011 to 2014.

Evaluation Protocol

Relevance of the Terms and the Associations Thereof

For each studied disease, we elicited a panel of domain experts (21 for ASF, 7 for FMD, 7 for BT, and 5 for SBV). Using an online questionnaire, the experts evaluated a representative number of terms extracted with text mining that described clinical signs and hosts for each studied disease. Tables 2, 3, and 4 show details of the terms describing clinical signs and hosts.

For the studied disease, the experts evaluated the relevance (specificity) of the terms describing the clinical signs. The relevance of the clinical signs was:

(i) very low, when their occurrence was very unlikely to characterise the disease emergence;

(ii) low, when they probably did not characterise the disease emergence;

(iii) medium, when they possibly characterised the disease emergence;

(iv) high, when they very probably characterised the disease emergence; and

(v) very high, when they characterised in almost all cases the disease emergence.

Next, the experts noted the groups of clinical signs that are most likely to appear in the hosts during an emergence of the disease.

Precision of the Associations obtained with the Statistical Measures

After consideration of the evaluation made by the experts, for each studied disease and for the terms evaluated by the experts, we noted the relevant associations. For these purposes, we analysed the relevance of 36 associations for FMD, 42 associations for BT and 144 associations for SBV. Further, we analysed the precision for the first 20 highest ranked associations obtained with the statistical measures D_{Web}^{AND}, D_{Web}^{NEAR}, MI_{Web}^{AND}, MI_{Web}^{NEAR}, CMI_{Web}^{AND}, CMI_{Web}^{NEAR}, and CMI_{global}. The precision of ranking was the number of relevant associations from the 20 highest ranked associations obtained with the statistical measures.

The precision of retrieval was the number of returned relevant pages (first 10 pages) for the 20 highest ranked associations obtained with the statistical measures and when tested as queries (in Google news, period from 2011 to 2014). A relevant page (RP) was a page that covered disease related information, including disease outbreak information. For these purposes, we analysed the relevance of 587 web pages for FMD, 733 web pages for BT and 564 web pages for SBV.

RESULTS

In the present section, we analyse the results for FMD, BT and SBV. We discuss ASF in another paper (Anonymous, 2014).

Relevance of the Terms and the Associations Thereof

Eighteen terms extracted with text mining described clinical signs and hosts that characterize a FMD emergence (Table 2). The experts evaluated 14 representative terms describing clinical signs and hosts. Two terms described general clinical signs (mortality, production losses) and five terms described mucous/ cutaneous clinical signs (vesicular and papular stomatitis, vesicular and mucosal disease, and swine vesicular disease). Seven terms described the hosts (cattle, small ruminants, buffaloes, pigs, wild boar, camels, and deer).

Table 2: List of terms extracted with text mining, which described clinical signs and hosts that characterize a foot-and-mouth disease emergence. In bold are the terms proposed to the experts for evaluation

Clinical signs	**Terms**
General	**production losses, low mortality**
Mucous/ cutaneous	**mucosal disease, papular stomatitis, swine vesicular disease, vesicular disease, vesicular stomatitis**
Hosts	**Terms**
	cattle herds, **wild boar**, cattle and **buffaloes**, Bactrian **camels**, dairy

	cattle, beef cattle, cattle and **pigs**, pig farms, **small ruminants**, cattle farms, **deer** farm

The majority of the experts (> 50%) evaluated the formation of vesicles as highly to very highly relevant; mucosal disease as medium to highly relevant, and production losses and formation of papules as medium relevant to characterise a FMD emergence (Figure 2A). During a FMD emergence, the majority of the experts noted as relevant the apparition of mucous/ cutaneous clinical signs in cattle, small ruminants and pigs and the apparition of general clinical signs in cattle and pigs (Figure 2B).

Figure 2A. Expert evaluation of the relevance of the terms describing clinical signs to characterize a foot-and-mouth disease emergence

Figure 2B. Expert evaluation of the relevance between the terms describing hosts and clinical signs to characterize a foot-and-mouth disease emergence

Twenty - two terms extracted with text mining described clinical signs and hosts that characterize a BT emergence (Table 3). The experts evaluated 13 representative terms describing clinical signs and hosts. Four terms described general clinical signs (deaths, general clinical signs, weakness and fever) and two terms described reproductive clinical signs (embryonic death and abortion). Seven terms described the hosts (cattle, sheep, goats, ewes, calves, red deer and roe deer).

Table 3: List of terms extracted with text mining, which described clinical signs and hosts that characterize a bluetongue emergence. In bold are the terms proposed to the experts for evaluation

Clinical signs	**Terms**
General	**livestock deaths, general clinical signs, onset of weakness, excess mortality, fever outbreak**
Reproductive	**embryonic death,** reproductive disorders, **occurrence of abortion**
Hosts	**Terms**
	red deer, adult **sheep**, **cattle** herds, **roe deer**, cattle population, **newborn calves**, newborn dairy calves, dairy calves, dairy ewes, **pregnant ewes**, cattle and **goats**, small ruminants, cattle production, goat farms

The majority of the experts (> 50%) evaluated the general clinical signs, including weakness, fever and mortality, the embryonic deaths and the abortions, as very low to low relevant to characterise a BT emergence (Figure 3A). During a BT emergence, the majority of the experts noted as relevant the apparition of general and reproductive clinical signs in cattle and sheep (Figure 3B).

Figure 3A. Expert evaluation of the relevance of the terms describing clinical signs to characterize a bluetongue emergence

Figure 3B. Expert evaluation of the relevance between the terms describing hosts and clinical signs to characterize a bluetongue emergence

Sixty - six terms extracted with text mining described clinical signs and hosts that characterize a SBV emergence (Table 4). The experts evaluated 29 representative terms describing clinical signs and hosts. Seven terms described congenital malformations or deformations (arthrogryposis hydranencephaly syndrome, foetal malformations, vertebral malformations, severe congenital malformations, hydranencephaly syndrome, limb malformations, and deformed offspring). One term described digestive clinical signs (watery diarrhoea). Two terms described general clinical signs (nonspecific febrile syndrome, mild transient disease). Seven terms described reproductive, including postnatal clinical signs in new-born (perinatal death, premature birth, aborted foetuses, reproductive losses, enzootic outbreak of abortion, stillborn bovine foetuses, stillborn lambs) and one term described respiratory clinical signs (acute bronchopneumonia). Nine terms described the SBV hosts (cows, goats, sheep, calves, and goat kids, lambs, bison, red deer and fallow deer).

Table 4: List of terms extracted with text mining, which described clinical signs and hosts that characterize a Schmallenberg virus emergence. In bold are the terms proposed to the experts for evaluation

Clinical signs	**Terms**
Congenital malformations, deformations	congenital malformations, deformed lambs, malformed offspring, ovine congenital malformations, **arthrogryposis hydranencephaly syndrome**, severe foetal malformations, **foetal malformations**, **vertebral malformations**, severe congenital malformation, **hydranencephaly syndrome**, **limb malformations**, congenital malformation, malformed calves, **severe congenital malformations**, malformed lambs, malformed progeny, **deformed offspring**
Digestive	**watery diarrhoea**
General	**nonspecific febrile syndrome**, **mild transient disease**, febrile syndrome
Reproductive	**perinatal death**, **premature birth**, lamb losses, **aborted foetuses**, **reproductive losses**, **enzootic outbreak of abortion**, outbreak of abortion, **stillborn bovine foetuses**, substantial reproductive losses, **stillborn lambs**
Respiratory	**acute bronchopneumonia**
Hosts	**Terms**
	newborn calves, sheep holdings, cow herds, **red deer**, goat holdings, lambs and **calves**, kids and calves, adult dairy cows, cows and ewes, **lambs** and **goats**, adult cows, cattle herds, dairy cows, bovine foetuses, goat population, dairy **cattle**, sheep herds, small ruminants, bison population, European **bison**, **fallow deer**, bovine foetus, **goat kids**, goat farms, goat foetus, deer populations, **sheep** farm, cows and calves, sheep herd, sheep population, cattle farms, newborn lambs, small ruminant, cattle population

The majority of the experts (> 50%) evaluated the congenital malformations and deformations as very highly relevant; the postnatal mortality and stillbirth as highly relevant; and the reproductive losses as medium relevant to characterize a SBV emergence (Figure 4A). During an

SBV emergence, the majority of the experts noted as relevant the apparition of congenital malformations and deformations and postnatal mortality and stillbirth in calves, kids and lambs, reproductive clinical signs in cattle, sheep and goats and digestive and general clinical signs in cattle (Figure 4B).

Figure 4A. Expert evaluation of the relevance of the terms describing clinical signs to characterize a Schmallenberg virus emergence

Figure 4B. Expert evaluation of the relevance of the terms describing hosts and clinical signs to characterize a Schmallenberg virus emergence

Precision of the Associations obtained with the Statistical Measures

Fourteen from 36 associations for FMD were relevant. The statistical measure CMI_{global} had the highest precision in ranking the relevant associations (12/ 20 associations), followed by CMI_{Web}^{AND} (11/ 20 associations) (Table 5). The precision of the relevant associations for FMD to retrieve relevant pages from the web for all statistical measures was > 0.92 (Table 6). From the relevant associations, the highest precision results to retrieve relevant pages from the web had the associations: vesicular stomatitis / cattle (10/ 10 pages), vesicular disease / cattle (9/ 10 pages), production losses / cattle (10/ 10 pages), and the associations: production losses / pigs (8/ 10 pages), vesicular disease / pigs (7/ 8 pages), and vesicular stomatitis / pigs (3/ 4 pages). From the associations evaluated as less relevant, the highest precision had the associations: vesicular stomatitis / deer (10/ 10 pages), and vesicular disease / deer (4/ 4 pages).

Eleven from 42 associations for BT were relevant. The statistical measures had a precision for ranking the relevant associations < 0.4 with the highest precision for D_{Web}^{NEAR} (8/ 20 associations) (Table 5). The precision of the relevant associations to retrieve relevant pages from the web for all statistical measures was ≤ 0.1 (Table 6). From the relevant associations, the highest precision to retrieve relevant pages from the web had the associations: livestock deaths / cattle (2/ 10 pages). From the associations evaluated as less relevant, the highest precision to retrieve relevant pages from the web had the associations: embryonic death / cattle (10/ 10 pages), embryonic death / sheep (4/ 4 pages), fever outbreak / cattle (3/ 10 pages), and fever outbreak / sheep (2/ 10 pages).

Forty-five from 144 associations for SBV were relevant. The statistical measure CMI_{Web}^{AND} had the highest precision for ranking the relevant associations (12/ 20 associations) (Table 5). The precision of the relevant associations for SBV to retrieve relevant pages from the web was the highest for the associations obtained with the statistical measure CMI_{global} (precision of 0.4) (Table 6). From the relevant associations, the highest precision to retrieve relevant pages from the web had the associations: deformed offspring / lambs (6/ 6 pages), deformed offspring / calves (7/ 8 pages), and limb malformations / goat kids (3/ 3 pages). The associations: watery diarrhoea / cattle, and reproductive losses / cattle retrieved three relevant pages from seven pages. The associations: aborted foetuses / cattle, aborted foetuses / sheep, aborted foetuses / calves retrieved relevant pages with precision < 0.25. From the associations evaluated as less relevant, the highest precision results to retrieve relevant pages from the web had the associations: deformed offspring / sheep (8/ 10 pages) and deformed offspring / cattle (5/ 6 pages).

Table 5: Precision of ranking of the associations obtained with the statistical measures

Statistical measure	FMD	BTV	SBV
D_{Web}^{AND}	0.40	0.25	0.45
D_{Web}^{NEAR}	0.40	0.40	0.50
MI_{Web}^{AND}	0.45	0.35	0.50
MI_{Web}^{NEAR}	0.45	0.35	0.40
CMI_{Web}^{AND}	0.55	0.25	0.60
CMI_{Web}^{NEAR}	0.50	0.35	0.40
CMI_{global}	0.60	0.30	0.50

Table 6: Precision of document retrieval based on relevant associations obtained with the statistical measures

Statistical measure	FMD	BTV	SBV
D_{Web}^{AND}	1.00	0.05	0.25
D_{Web}^{NEAR}	1.00	0.07	0.25
MI_{Web}^{AND}	1.00	0.10	0.21
MI_{Web}^{NEAR}	0.98	0.10	0.21
CMI_{Web}^{AND}	0.92	0.05	0.19
CMI_{Web}^{NEAR}	0.92	0.10	0.22
CMI_{global}	0.93	0.10	0.40

DISCUSSION AND FUTURE WORK

The work presented in this paper is part of the global methodology that we develop for the French epidemic intelligence team in animal health where we focus on monitoring the web to detect disease emergence (Arsevska et al., 2014). We currently have a list of more than twenty exotic diseases (including ASF, FMD, BT and SBV) and ten groups of clinical signs in five animal species.

In one of our previous works (Arsevska et al., 2014), we did experiments with data for African swine fever (ASF). We compared the predictive performance of (i) a Discriminative Multinomial Naïve Bayes (DMNB) algorithm and (ii) a Sequential Minimal Optimization (SMO) algorithm, using a 10-fold cross-validation method. Both classification algorithms reported good predictive performance for text documents about ASF, with precision, recall and F-score of 0.75 for

DMNB, and precision, recall and F-score of 0.73 for SMO (Table 7). Similar results were obtained elsewhere (Conway et al., 2009; Freifeld et al., 2008; Zhang & Liu, 2007). In the future, we intend to use machine-learning techniques to extract other specific information from non-structured texts retrieved from the web, such as outbreak location, date, hosts, their numbers and clinical signs.

Table 7: Classification results for African swine fever documents

Classification algorithm		**Naïve Bayes**			**Support Vector Machine**		
Performance		Recall	Precision	F-score	Recall	Precision	F-score
Class	disease	0.72	0.77	0.74	0.66	0.68	0.67
	economy	0.48	0.53	0.50	0.49	0.73	0.58
	general	0.86	0.87	0.83	0.86	0.76	0.81
Average		0.75	0.75	0.75	0.73	0.73	0.73

In this paper, we explored the use of the data mining approaches to automatically extract relevant terms (clinical signs and hosts) from a corpus of relevant documents and to automatically select relevant associations (between clinical signs and hosts) for monitoring the web for disease emergence. We were mostly interested in the terms describing clinical signs and hosts for known exotic animal infectious diseases. Considering their specificity, we got a highest number of specific terms (clinical signs and hosts) for SBV. Indeed, the relevant documents used for extraction of SBV terms described in detail the emergence of the newly discovered pathogen. In contrast, we got a limited number of specific terms (clinical signs) for BT because the relevant documents described the consequences of this disease for animal productivity, reproduction and lifespan rather than the specific clinical signs.

The statistical measures varied in precision of the ranking of the relevant associations, as well as in precision of retrieval of relevant documents from the web. The highest precision of the ranking was obtained with the statistical measures for CMI_{Web}^{AND} (SBV: 0.6) and for CMI_{global} (FMD: 0.6). The highest precision results for the retrieval of relevant pages from the web were obtained from the relevant associations for CMI_{global} (FMD: 0.93), and the lowest for CMI_{global} (SBV: 0.4; BT: 0.1).

The precision was highly influenced by the criteria selected to note the relevance of each association. Indeed, each association was evaluated based on the experts' answers (taking into consideration the highest notes) but also the semantic context of the terms in the association. For example, for SBV, most experts evaluated the congenital malformations and deformations in offspring (lambs, calves, goat kids) as relevant to characterize a SBV emergence. Accordingly, the associations such as deformed offspring / lambs and deformed offspring / calves – noted as relevant, retrieved a high number of relevant pages from the web. However, the associations such as deformed offspring / cattle and deformed offspring / sheep – noted as less relevant, also retrieved a high number of relevant pages. As these associations were highly ranked with the statistical measures, we do not exclude their value to build queries. These observations provide evidence that the statistical measures used in this work are a valuable decision-supporting tool to facilitate the evaluation of the associations.

The two evaluations we conducted for the ranking and the retrieval of relevant web documents for BT did not exceed a precision higher than 0.4 and 0.1, respectively. One reason for these results was the evaluation by the experts of the terms describing clinical signs – as low specific to characterize a BT emergence (these evaluations correspond to what is known in the literature). This assertion was confirmed by the low performance of the relevant associations as queries. However, the associations that had the highest precision to retrieve relevant documents from the web, such as: fever outbreak / sheep and fever / outbreak / cattle, livestock deaths / cattle, embryonic death / cattle, embryonic death / sheep, corresponded to what was described in the relevant documents – source of terms. For example, the majority of the articles discussed the mortality and reproductive disorders in sheep, goats and cattle for the 2014 BT epizootics in the Balkans, and the abstracts, discussed the BT clinical signs such as fever and weakness, reproductive disorders and abortions in ruminants.

The precision of our evaluations was also influenced by the number and type of associations used in the experiments. Indeed, in this first work, we only evaluated the representative associations proposed to the experts. In the future, we intend to evaluate all associations from all combinations of extracted terms (describing clinical signs and hosts). For example, we already evaluated 506 associations obtained with text mining for ASF, where the precision of ranking was 0.5 for CMI_{Web}^{AND} and 0.65 for CMI_{global}, and the average precision of retrieval of relevant documents from the web was > 0.83.

The expert elicitation method allowed us to benefit from the common knowledge of many experts. Most experts agreed on the specificity of the terms for each studied disease, and these evaluations corresponded to what is found in the literature. The experts also proposed new terms important for our future work. For example, for BT they proposed associations such as buccal lesions and ulcers, facial oedemas, cyanosis of the tongue and hypersalivation, principally in sheep and cattle. For FMD they proposed associations, such as lameness and hypersalivation, mortality in young animals, along with vesicle and ulcer lesions of the hoofs, mouth, tongue, and udder, principally in cattle. The proposals by the experts for SBV and ASF corresponded to the extracted terms with the text mining. In future, we intend to evaluate the expert proposals as queries.

Finally, we intend to build queries using the following associations - obtained with text mining:

(i) for FMD, production losses, vesicular stomatitis and vesicular disease in cattle and pigs;
(ii) for BT, deaths and fever in cattle and sheep; and
(iii) for SBV, malformations and deformations in goat kids, lambs and calves, including deformed offspring in cattle, sheep and goats, as well as reproductive losses and abortions in cattle, sheep and goats.

We consider that the combination of text and web mining approaches and expert knowledge enable to identify relevant associations between terms describing clinical signs and hosts that can improve the detection of infectious disease emergence on the web. Our method is generic and can have applications to both animal and public health.

ACKNOWLEDGEMENTS

Authors would like to thank the experts that participated in the elicitation. This work was funded by the French Ministry of Agriculture, Food and Forestry (MAAF), the French Agricultural

Research Centre for International Development (CIRAD), and the SONGES Project (FEDER and Languedoc-Roussillon)[ii].

REFERENCES

Arsevska, E., Roche, M., Lancelot, R., Hendrikx, P., & Dufour, B. (2014). Exploiting Textual Source Information for Epidemiosurveillance. In B. S. Clos et al. (Ed.), Eight Metadata and Semantics Research Conference (pp. 359–361). Springer International Publishing Switzerland. http://doi.org/10.13140/2.1.4049.1522

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., & Mandl, K. D. (2008). Surveillance Sans Frontiéres: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. *PLoS Medicine*, *5*(7). http://doi.org/10.1371/journal.pmed.0050151

Chan, E., Brewer, T., Madoff, L., Pollack, M., Sonricker, A., Keller, M., Freifeld, C., Blench, M., Mawude, A., & Brownstein, J. (2010). Global capacity for emerging infectious disease detection. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(50), 21701–21706. http://doi.org/10.1073/pnas.1006219107

Chapman, W. W., Dowling, J. N., & Wagner, M. M. (2004). Fever detection from free-text clinical records for biosurveillance. *Journal of Biomedical Informatics*, *37*(2), 120–127. http://doi.org/10.1016/j.jbi.2004.03.002

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Cilibrasi, R. L., & Vitanyi, P. (2007). The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, *19*(3), 370–383.

Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., Ngo, Q-H., Dien, D., Kawtrakul, A., Takeuchi, K., Takeuchi, K., Shigem, M., & Taniguchi, K. (2008). BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics*, *24*(24), 2940–2941. http://doi.org/10.1093/bioinformatics/btn534

Collier, N., Goodwin, R. M., McCrae, J., Doan, S., Kawazoe, A., Conway, M., Kawtrakul, A., Takeuchi, K., & Dien, D. (2010). An ontology-driven system for detecting global health events. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 215–222). Association for Computational Linguistics.

Collier, N., Kawazoe, A., Jin, L., Shigematsu, M., Dien, D., Barrero, R. Takeuchi, K., & Kawtrakul, A. (2007). A multilingual ontology for infectious disease surveillance: rationale, design and challenges. *Language Resources and Evaluation*, *40*(3-4), 405–413. http://doi.org/10.1007/s10579-007-9019-7

Conway, M., Doan, S., Kawazoe, A., & Collier, N. (2009). Classifying disease outbreak reports using n-grams and semantic features. *International Journal of Medical Informatics*, *78*(12), e47–e58. http://doi.org/10.1016/j.ijmedinf.2009.03.010

Economopoulou, A., Kinross, P., Domanovic, D., & Coulombier, D. (2014). Infectious diseases prioritisation for event-based surveillance at the European Union level for the 2012 Olympic and Paralympic Games. *European Communicable Disease Bulletin*, *19*(15).

Freifeld, C. C., Mandl, K. D., Reis, B. Y., & Brownstein, J. S. (2008). HealthMap: global infectious disease monitoring through automated classification and visualization of

Internet media reports. *Journal of the American Medical Informatics Association: JAMIA*, *15*(2), 150–157. http://doi.org/10.1197/jamia.M2544

Friedlin, J., Grannis, S., & Overhage, J. M. (2008). Using Natural Language Processing to Improve Accuracy of Automated Notifiable Disease Reporting. *AMIA Annual Symposium Proceedings*, *2008*, 207–211.

Furrer, L., Küker, S., Berezowski, J., Posthaus, H., Vial, F., & Rinaldi, F. (2015). Constructing a Syndromic Terminology Resource for Veterinary Text Mining. *Proceedings of the Conference Terminology and Artificial Intelligence 2015*, 61–70.

Gesualdo, F., Stilo, G., Agricola, E., Gonfiantini, M. V., Pandolfi, E., Velardi, P., & Tozzi, A. E. (2013). Influenza-Like Illness Surveillance on Twitter through Automated Learning of Naïve Language. *PLoS ONE*, *8*(12), e82489. http://doi.org/10.1371/journal.pone.0082489

Keller, M., Blench, M., Tolentino, H., Freifeld, C., Mandl, K., Mawudeku, A., Eysenba, G., & Brownstein, J. (2009). Use of Unstructured Event-Based Reports for Global Infectious Disease Surveillance. *Emerging Infectious Diseases*, *15*(5), 689–695. http://doi.org/10.3201/eid1505.081114

Khomenko, S., Beltran-Alcrudo, D., Rozstalnyy, A., Gogin, A., Kolbasov, D., Pinto, J., Lubroth, J., & Martin, V. (2013). African swine fever in the Russian Federation: risk factors for Europe and beyond. *Empres Watch, 28*. Retrieved June 12, 2014 from http://www.fao.org/docrep/018/aq240e/aq240e.pdf

Lossio Ventura, J. A., Jonquet, C., Roche, M., & Teisseire, M. (2014). Towards a Mixed Approach to Extract Biomedical Terms from Text Corpus. *International Journal of Knowledge Discovery and Bioinformatics*, *4*(1), 1–15. http://doi.org/10.4018/ijkdb.2014010101

Lyon, A., Mooney, A., & Grossel, G. (2013). Using AquaticHealth.net to Detect Emerging Trends in Aquatic Animal Health. *Agriculture*, *3*(2), 299–309. http://doi.org/10.3390/agriculture3020299

Mantero, J., Belyaeva, J., & Linge, J. (2011). *How to maximise event-based surveillance web-systems the example of ECDC/JRC collaboration to improve the performance of MedISys.* JRC Scientific and Technical Reports. Luxembourg: Publications Office.

Milinovich, G. J., Avril, S. M. R., Clements, A. C. A., Brownstein, J. S., Tong, S., & Hu, W. (2014). Using internet search queries for infectious disease surveillance: screening diseases for suitability. *BMC Infectious Diseases*, *14*(1). http://doi.org/10.1186/s12879-014-0690-1

Morens, D. M., Taubenberger, J. K., & Fauci, A. S. (2013). H7N9 Avian Influenza A Virus and the Perpetual Challenge of Potential Human Pandemicity. *mBio*, *4*(4), e00445–13. http://doi.org/10.1128/mBio.00445-13

Mykhalovskiy, E., & Weir, L. (2006). The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health. *Canadian Journal of Public Health*, *97*(1), 42–44.

Nazar, R., Vivaldi, J., & Cabré, M. T. (2008). A Suite to Compile and Analyze an LSP Corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Nelson, N. P., Brownstein, J. S., & Hartley, D. M. (2010). Event-based biosurveillance of respiratory disease in Mexico, 2007–2009: connection to the 2009 influenza A (H1N1) pandemic. *Euro Surveill*, *15*(30), 19626.

Nelson, N. P., Yang, L., Reilly, A. R., Hardin, J. E., & Hartley, D. M. (2012). Event-based internet biosurveillance: relation to epidemiological observation. *Emerging Themes in Epidemiology*, *9*(1), 4. http://doi.org/10.1186/1742-7622-9-4

Rodeia, S. P. (2008). Assessment of the Risk of Introducing Foot and Mouth Disease into the EU and the Reduction of this Risk through Interventions in Infected Countries: a review and follow-up. *Transboundary and Emerging Diseases*, *55*(1), 3–4. http://doi.org/10.1111/j.1865-1682.2007.01018.x

Sánchez-Vizcaíno, J. M., Mur, L., & Martínez-López, B. (2013). African swine fever (ASF): five years around Europe. *Veterinary Microbiology*, *165*(1-2), 45–50. http://doi.org/10.1016/j.vetmic.2012.11.030

Saneifar, H., Bonniol, S., Poncelet, P., & Roche, M. (2015). From Terminology Extraction to Terminology Validation: An Approach Adapted to Log Files. *Journal of Universal Computer Science*, *21*(4), 604–635.

Santamaria, S. L., & Zimmerman, K. L. (2011). Uses of Informatics to Solve Real World Problems in Veterinary Medicine. *Journal of Veterinary Medical Education*, *38*(2), 103–109. http://doi.org/10.3138/jvme.38.2.103

Smith-Akin, K. A., Bearden, C. F., Pittenger, S. T., & Bernstam, E. V. (2007). Toward a veterinary informatics research agenda: an analysis of the PubMed-indexed literature. *International Journal of Medical Informatics*, *76*(4), 306–312. http://doi.org/10.1016/j.ijmedinf.2006.02.009

Steinberger, R., Fuart, F, Best, C, Von Etter, P, & Yangarber, R. (2008). Text Mining from the Web for Medical Intelligence. In F. Fogelman-Soulié et al. (Eds.), *Mining Massive Data Sets for Security* (pp. 295–310). IOS press.

Torii, M., Yin, L., Nguyen, T., Mazumdar, C. T., Liu, H., Hartley, D. M., & Nelson, N. P. (2011). An exploratory study of a text classification framework for Internet-based surveillance of emerging epidemics. *International Journal of Medical Informatics*, *80*(1), 56–66. http://doi.org/10.1016/j.ijmedinf.2010.10.015

Turney, P. D. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In L. Raedt & P. Flach (Eds.), *Machine Learning: ECML 2001: 12th European Conference on Machine Learning Freiburg, Germany, September 5–7, 2001 Proceedings* (pp. 491–502). Berlin, Heidelberg: Springer Berlin Heidelberg.

Vivaldi, J., Màrquez, L., & Rodríguez, H. (2001). Improving Term Extraction by System Combination Using Boosting. In L. De Raedt & P. Flach (Eds.), *Machine Learning: ECML 2001* (Vol. 2167, pp. 515–526). Berlin, Heidelberg: Springer Berlin Heidelberg.

Wernike, K., Hoffmann, B., & Beer, M. (2013). Schmallenberg virus. *Developments in Biologicals*, *135*, 175–182. http://doi.org/10.1159/000312546

Wilson, A., & Mellor, P. (2009). Bluetongue in Europe: past, present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, (364), 2669–2681. http://doi.org/10.1098/rstb.2009.0091

Zhang, Y., & Liu, B. (2007). Semantic Text Classification of Emergent Disease Reports. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases* (pp. 629–637). Berlin, Heidelberg: Springer-Verlag. http://doi.org/10.1007/978-3-540-74976-9_67

[i] http://tubo.lirmm.fr/biotex/

[ii] http://textmining.biz/Projects/Songes

Annexes

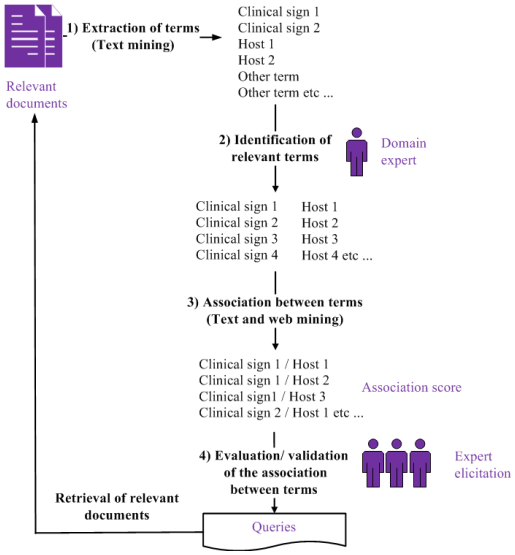


Figure 1. Workflow of the method for identification of relevant associations between terms describing hosts and clinical signs for monitoring disease emergence on the web

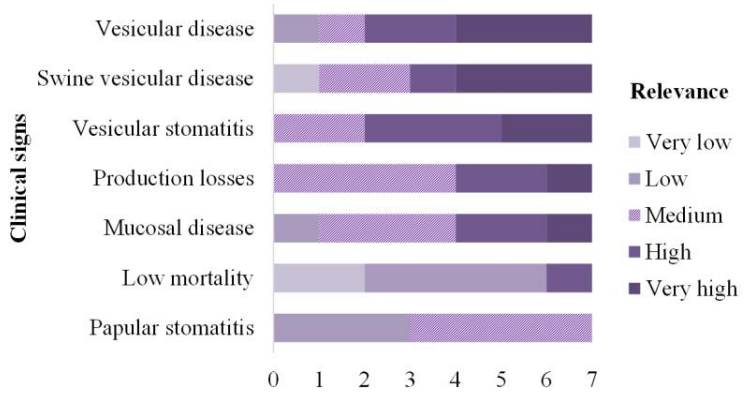


Figure 2A. Expert evaluation of the relevance of the terms describing clinical signs to characterize a foot-and-mouth disease emergence

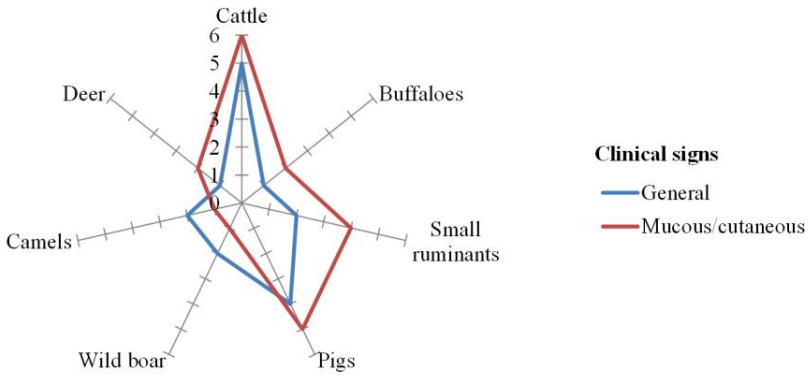


Figure 2B. Expert evaluation of the relevance between the terms describing hosts and clinical signs to characterize a foot-and-mouth disease emergence

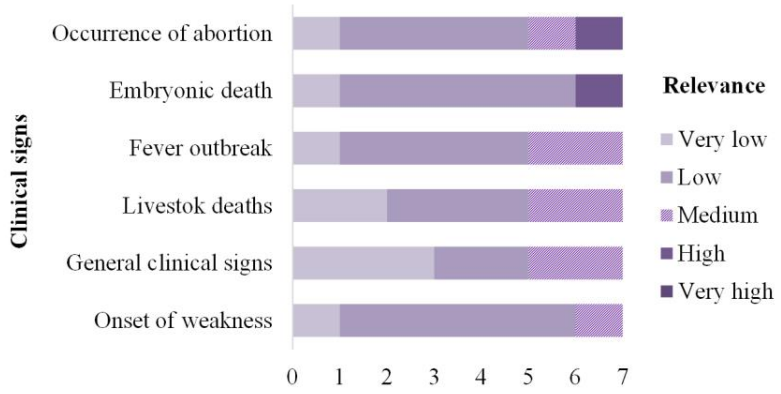


Figure 3A. Expert evaluation of the relevance of the terms describing clinical signs to characterize a bluetongue emergence

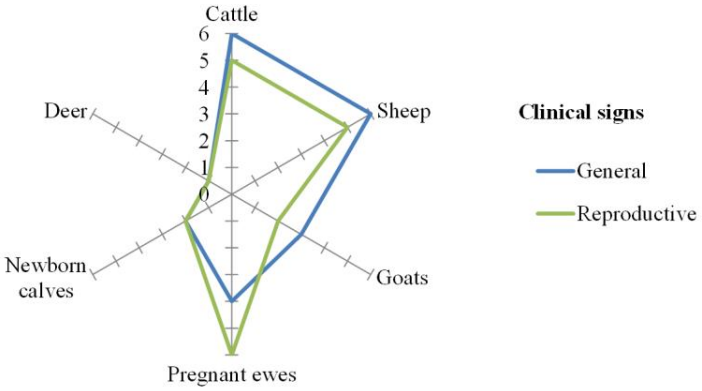


Figure 3B. Expert evaluation of the relevance between the terms describing hosts and clinical signs to characterize a bluetongue emergence

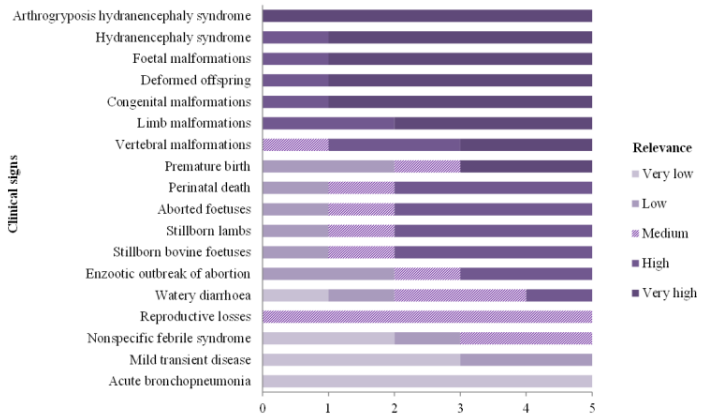


Figure 4A. Expert evaluation of the relevance of the terms describing clinical signs to characterize a Schmallenberg virus emergence

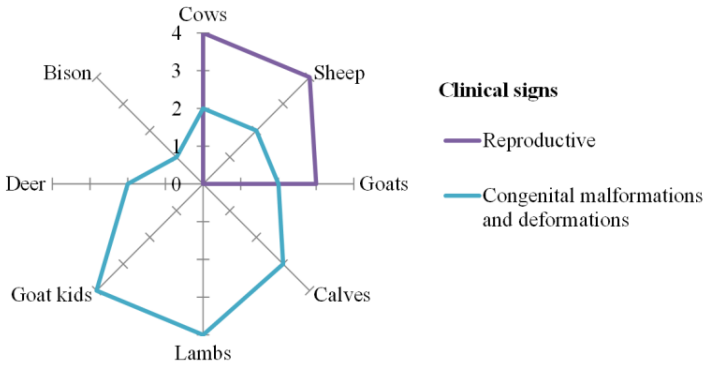


Figure 4B. Expert evaluation of the relevance of the terms describing hosts and clinical signs to characterize a Schmallenberg virus emergence

Annexe B

Caractéristiques des systèmes de surveillance

Pour assurer l'efficacité du recueil des données et de leur analyse, les systèmes de surveillance doivent pouvoir se baser sur un processus formalisé d'intelligence épidémiologique, fondé sur la surveillance des indicateurs et la surveillance des évènements. Le Tableau B.1 présente les principales caractéristiques de ces deux composantes de la surveillance.

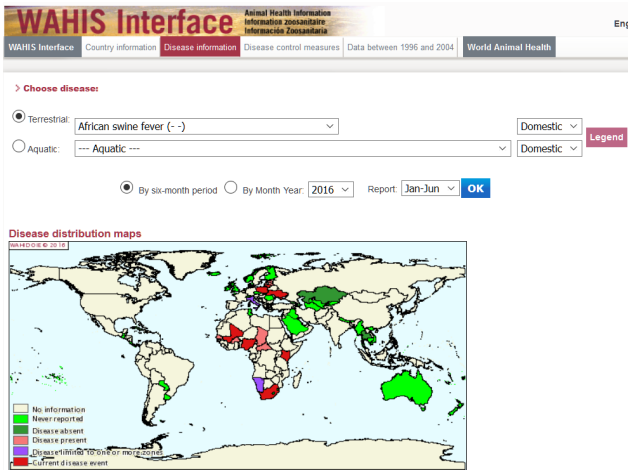
Tableau B.1 – Principales caractéristiques des la surveillance fondée sur des indicateurs et sur les événements

Caractéristiques	**Indicateurs (SBI)**	**Événements (SBE)**
Chronologie de la saisie de données	- L'information est saisie dès qu'elle est disponible. - La disponibilité est immédiate, hebdomadaire ou mensuelle. - Saisie humaine (modération humaine).	- L'information est saisie dès qu'elle survient. - La disponibilité varie selon le moment où les données sont fournies par ceux qui ont l'information. - Saisie humaine (modération humaine) ou automatique.
Structure des rapports	- Formulaires de rapports. - Analyse des données à intervalles réguliers.	- Formulaires de rapports flexibles pour des données qualitatives et quantitatives. - Analyse des données à tout moment.
Délais de détection	- Dépend de l'intervalle de temps entre le moment où l'événement se produit jusqu'à ce qu'un diagnostic soit porté. - Dépend du délai de rapport de l'événement selon l'organisation hiérarchique de la surveillance. - Retard possible entre l'observation d'un événement, la confirmation et la déclaration officielle.	- Dépend du moment où se produit l'événement jusqu'à ce que les premières rumeurs apparaissent, ce qui se produit avant qu'une déclaration officielle par les autorités sanitaires ne soit disponible. - Dépend de la capacité du système de surveillance à interpréter correctement le signal au moment où il est identifié. - Retard possible entre l'identification d'un événement et le rapport non-officiel.
Seuil pour la génération de signaux	- Génération de signaux différente : ex., déclaration des foyers par les autorités sanitaires, calcul de l'incidence/ prévalence des cas, identification d'agrégats dans le temps ou l'espace ou alertes automatiques pour de dépassements des seuils en temps ou espace.	- Génération de signaux différente : ex., indexation humaine de foyers de maladies ou des méthodes automatisées d'identification d'agrégats dans l'espace.
Déclenchement des mesures des suivi	- Un foyer confirmé, donne lieu à une déclaration officielle et à une analyse.	- Un foyer confirmé, ou des rumeurs d'un foyer, donne lieu à une collecte d'information supplémentaire et à une vérification.

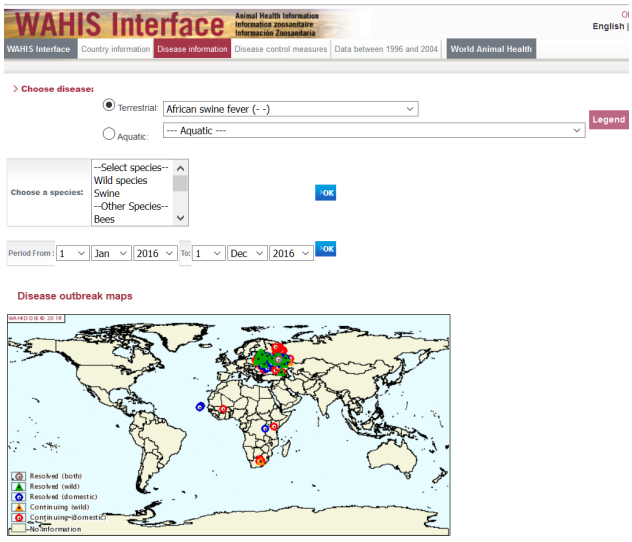
Annexe C

Exemples d'interfaces pour les systèmes de surveillance

C.1 Systèmes de surveillance fondés sur des indicateurs



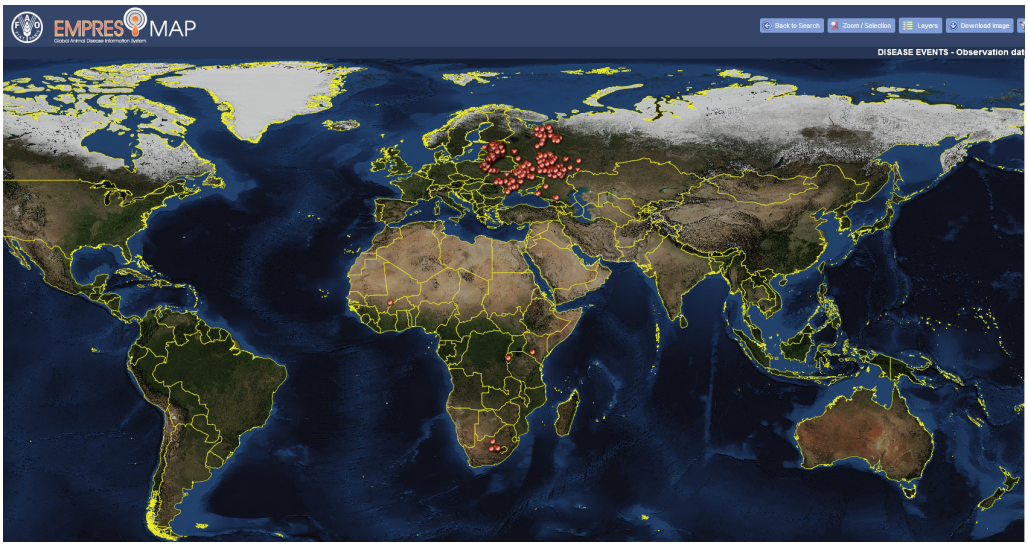
(a) Distribution mondiale de la peste porcine africaine de janvier à juin 2016



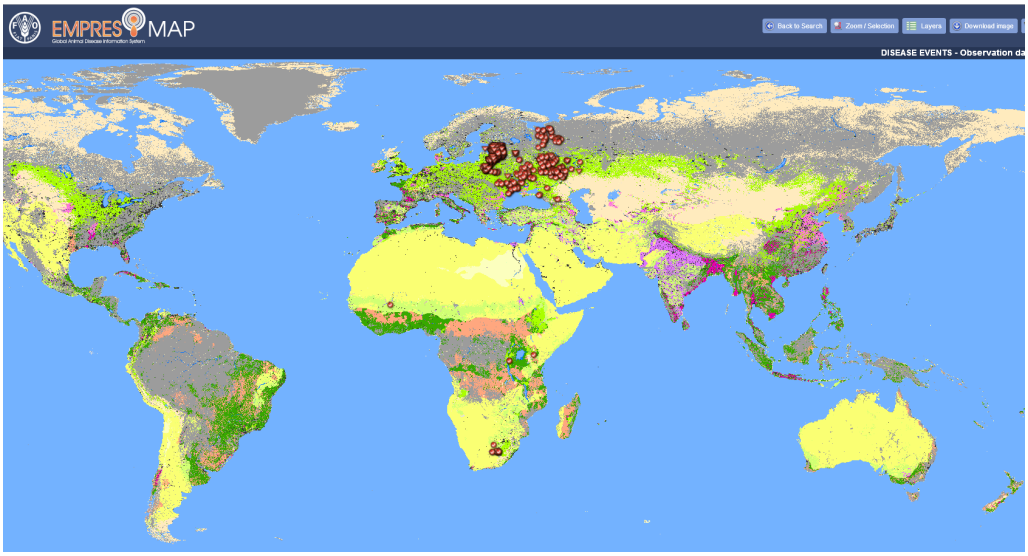
(b) Distribution mondiale des foyers de la peste porcine africaine de janvier à juin 2016

L'interface WAHIS de l'OIE (www.oie.int/wahid) permet la visualisation des foyers déclarés par les pays membres, laquelle comprend, entre autres, des informations sur la maladie et les espèces atteintes, la zone géographique atteinte, les mesures de lutte appliquées et les examens de laboratoire réalisés ou en cours. Les utilisateurs peuvent consulter aussi des rapports semestriels ou annuels concernant la présence et l'absence des maladies de la liste de l'OIE, ainsi que les mesures de prévention et de lutte appliquées, la population animale, les laboratoires de référence, etc.

Figure C.1 – Système WAHIS



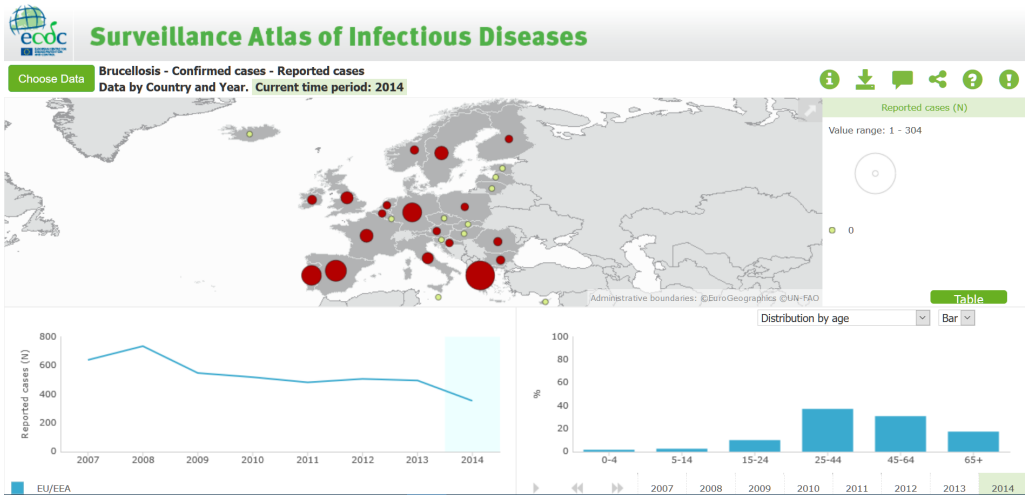
(a) Distribution mondiale des foyers de la peste porcine africaine au niveau international de janvier à juin 2016



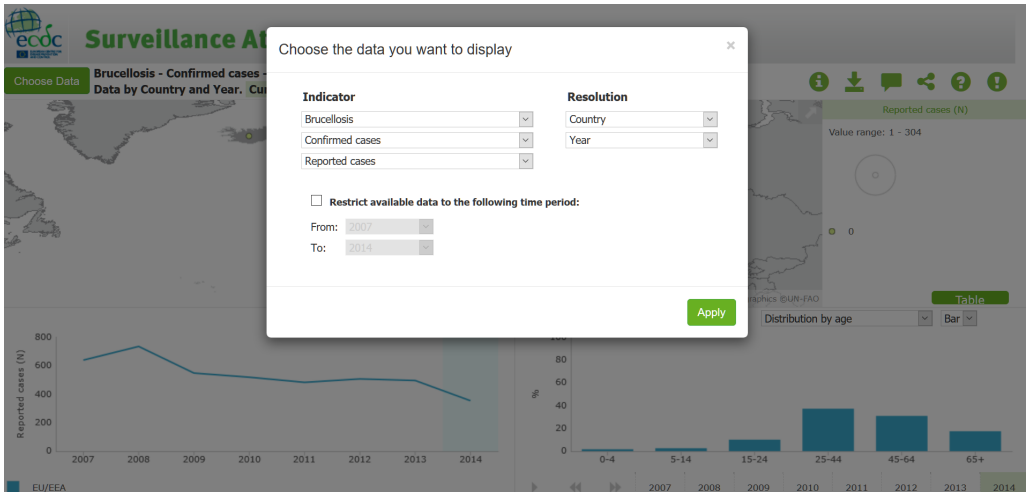
(b) Distribution mondiale des foyers de la peste porcine africaine et la densité porcine de janvier à juin 2016

Le système Empres-i (http://empres-i.fao.org/eipws3g/) de la FAO permet aux utilisateurs de visualiser et télécharger facilement des informations sur les foyers. Les utilisateurs peuvent sélectionner l'information d'intérêt selon des critères de recherche comme la maladie, la date, l'espèce, etc. Les données peuvent ensuite être facilement exportées en format structuré. Empres-i produit également des cartes et des graphes chronologiques des foyers. Les cartes peuvent être personnalisées par l'ajout de couches, comme les couches de densité animale, les zones d'élevage, le commerce, etc.

Figure C.2 – Système Empres-i



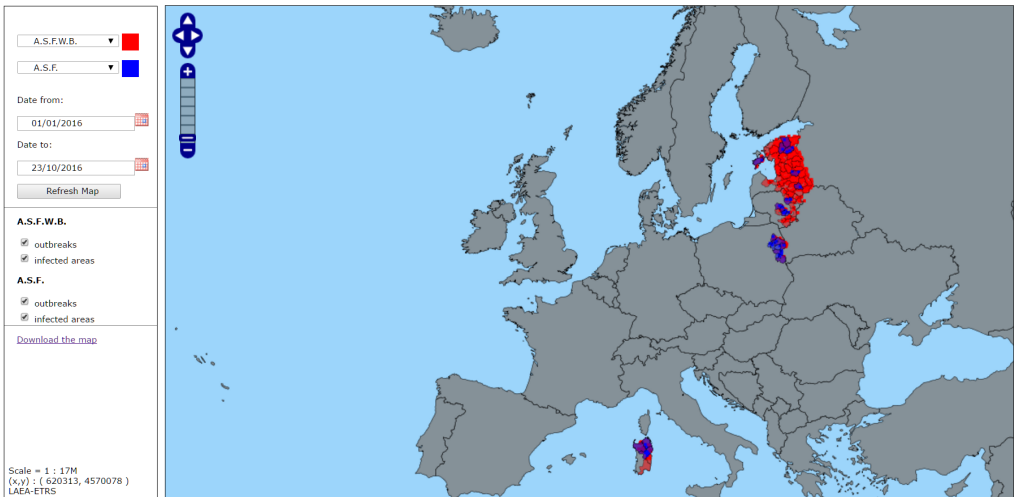
(a) Nombre de cas de la brucellose dans l'Union Européenne en 2014



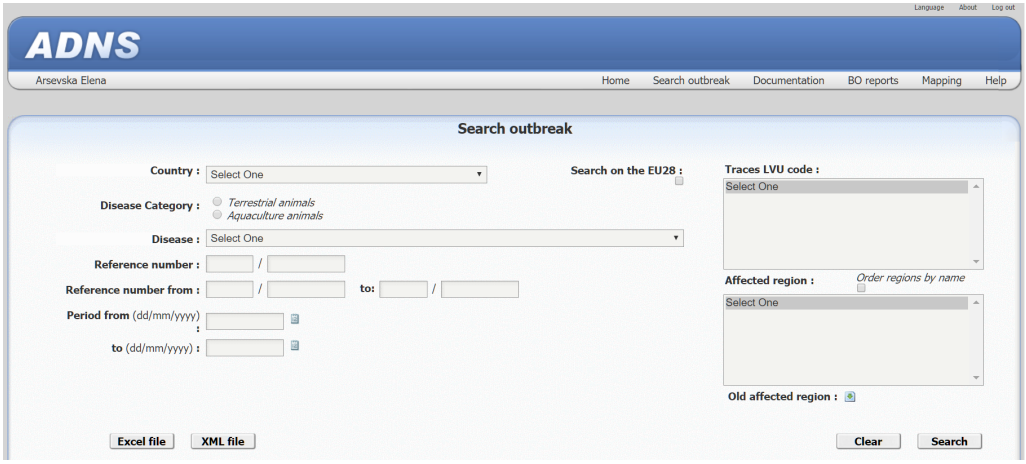
(b) Critères de recherche de l'information sanitaire

L'Atlas de surveillance des maladies infectieuses (http://ecdc.europa.eu/en/data-tools/atlas/Pages/atlas.aspx) du système TESSY vise à améliorer l'accessibilité aux données officielles pour des maladies infectieuses humaines dans l'UE. L'interface fournit des cartes de distribution et des graphes chronologiques de cas confirmés, ainsi que des tableaux exportables en format structuré. La recherche de l'information s'effectue selon la maladie, la région ou la période. Les données sont exportables en format structuré.

Figure C.3 – Atlas de surveillance des maladies infectieuses



(a) Distribution des foyers de la peste porcine africaine dans l'Union Européenne de janvier à octobre 2016

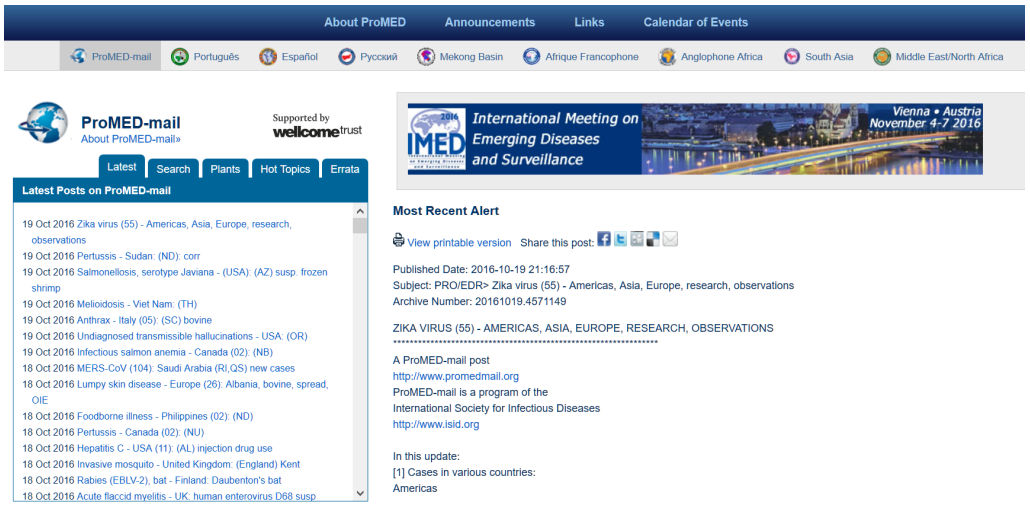


(b) Critères de recherche de l'information sanitaire

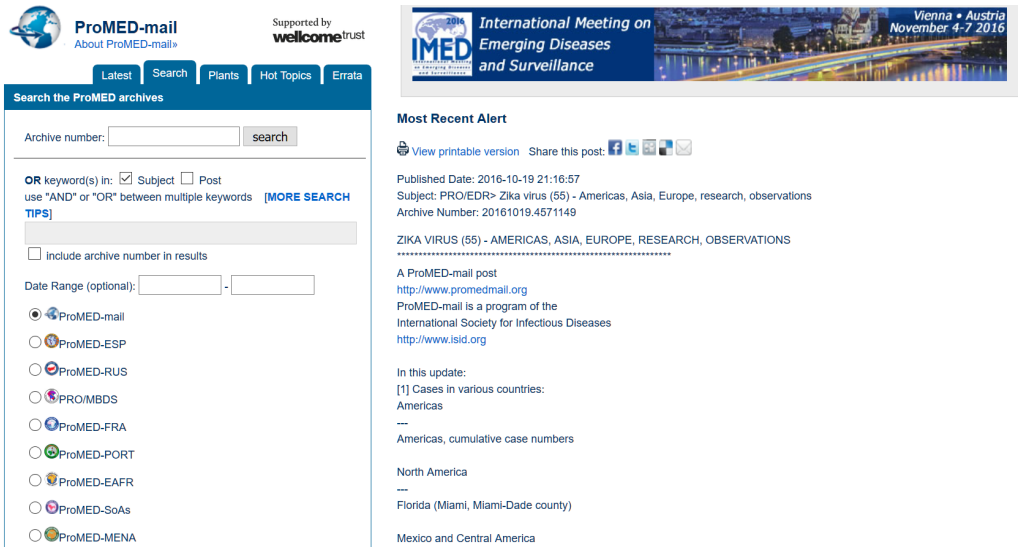
Le système ADNS (https://webgate.ec.europa.eu/ADNS) de la Commission Européenne permet à un groupe restreint d'utilisateurs de visualiser des foyers et des zones affectés par les maladies animales exotiques dans l'UE. Le système permet une recherche d'information sanitaire par pays, région, maladie, date ou espèce. Les données peuvent ensuite être facilement exportées en format structuré.

Figure C.4 – Système ADNS

C.2 Systèmes de surveillance fondés sur des événements



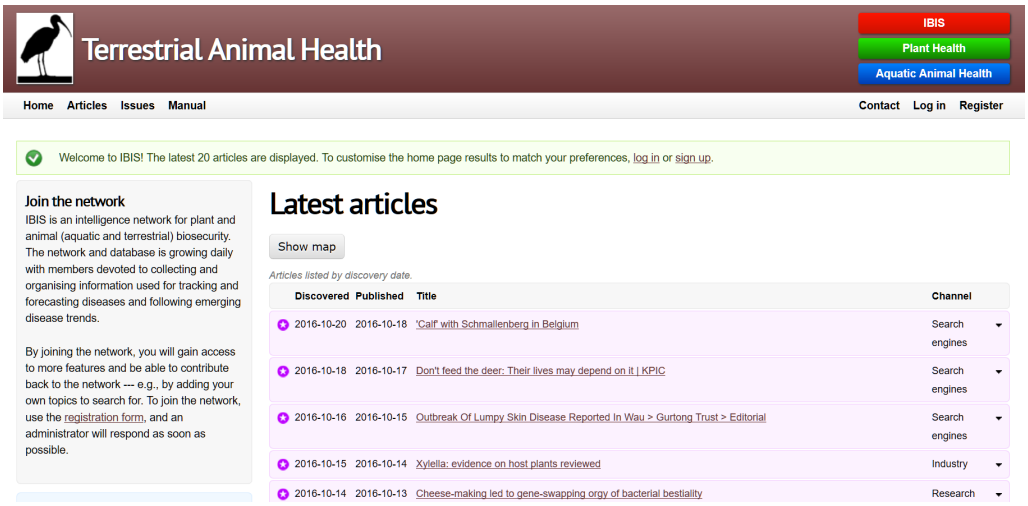
(a) Liste des dernières informations sanitaires publiées dans le système ProMED



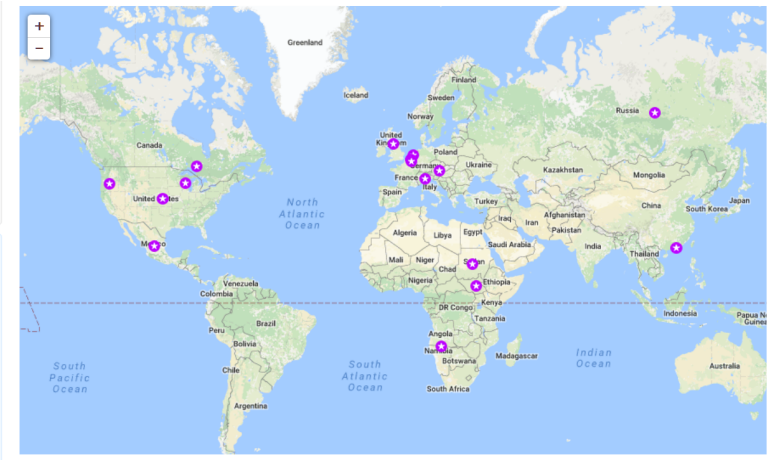
(b) Critères de recherche de l'information sanitaire

L'objectif du système ProMED (`http://www.promedmail.org/`) consiste à promouvoir la communication au sein de la communauté scientifique internationale. Cette communauté comprend les médecins, épidémiologistes, vétérinaires, et autres professionnels de santé. Les modérateurs du système exploitent également différentes sources d'information pour produire des rapports. Les rapports sont classés par paire maladie-lieu. Depuis peu, les rapports de ProMED peuvent être visualisés *via* l'interface du système HealthMap.

Figure C.5 – Système ProMED



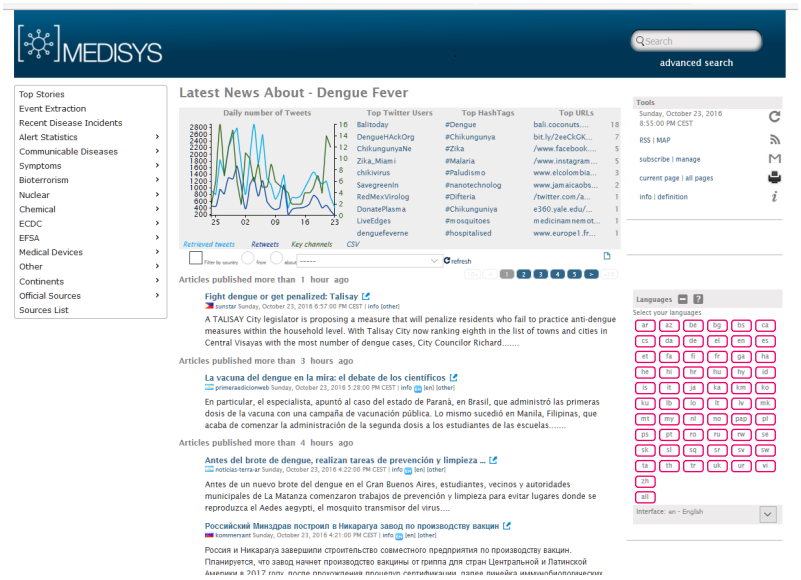
(a) Liste des derniers articles identifiés par le système IBIS



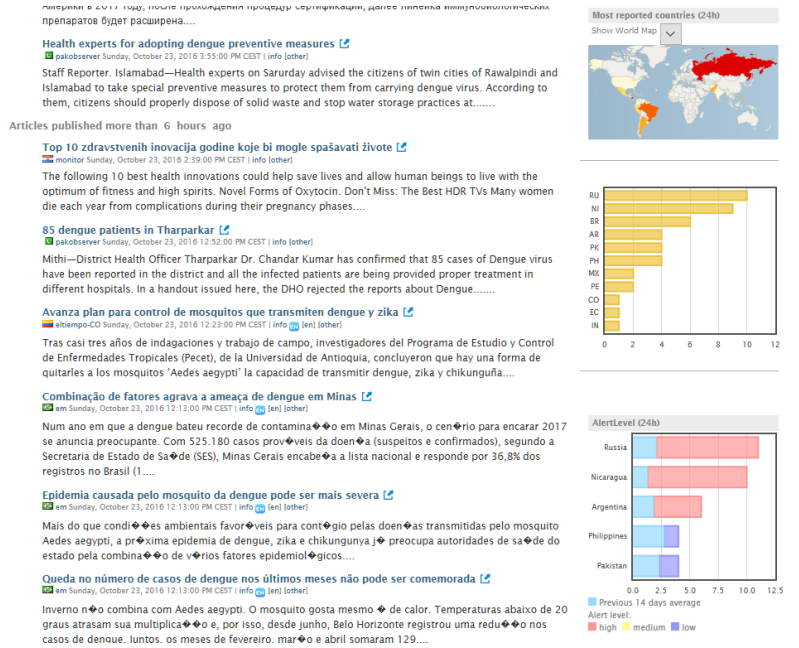
(b) Distribution mondiale des événements identifiés par le système IBIS

Le système IBIS (crée en 2013), est dédié à la veille électronique (sur le Web) des maladies infectieuses pour des animaux aquatiques et terrestres, ainsi que des plantes (`http://biointel.org/`). Tous les utilisateurs peuvent parcourir le contenu et générer des rapports et des cartes selon une maladie, termes de recherche, période, pays, etc. Les utilisateurs authentifiés peuvent annoter, modifier et catégoriser les rapports.

Figure C.6 – Système IBIS



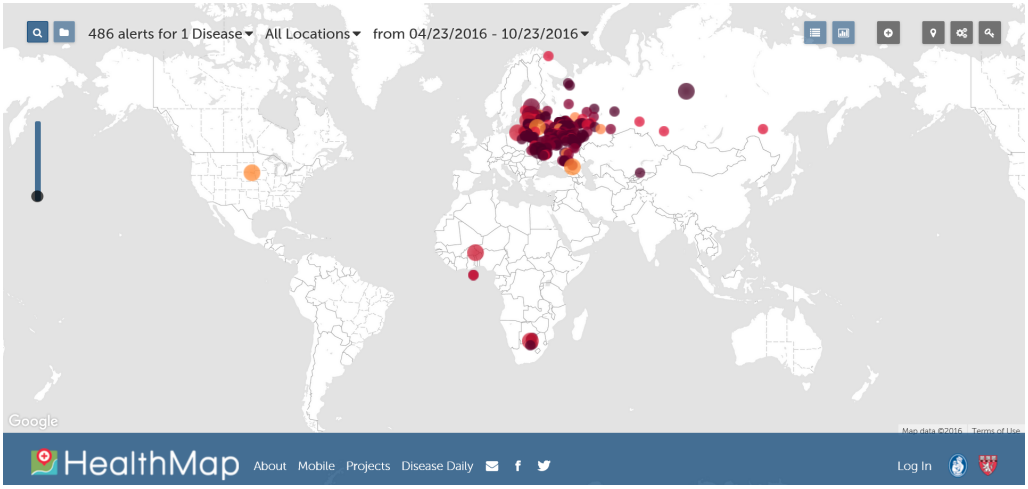
(a) Liste derniers articles identifiés par le système MedISys



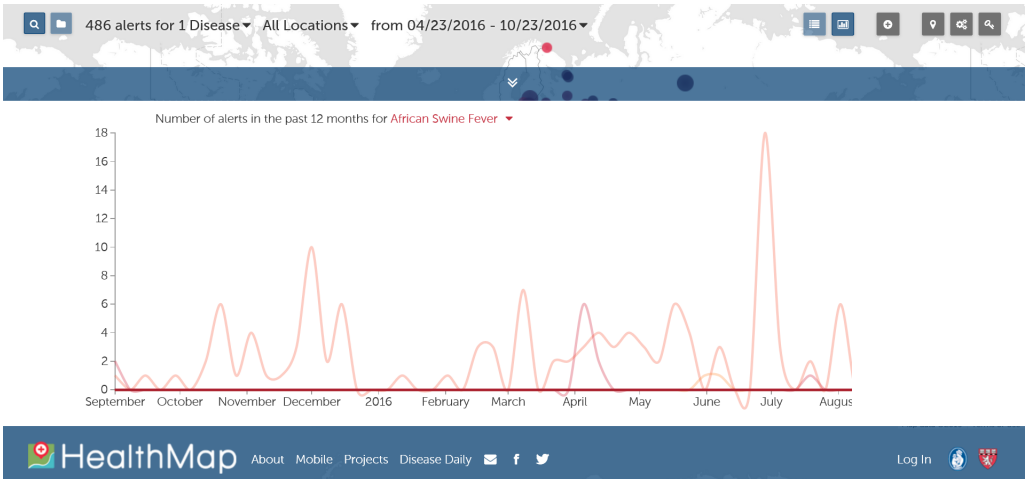
(b) Statistique pour les derniers articles par pays détectés par le système MedISys

Le système MedISys (http://medisys.newsbrief.eu) affiche les articles recueillis sur le Web dans plus de 200 catégories. Les articles sont classés en fonction du risque et/ou du pays. Les articles peuvent encore être filtrés par la langue, la source et le pays. En outre, le système re-groupe les articles sur le même sujet, identifie les doublons et extrait des événements *via* l'outil « Pattern-based Understanding and Learning System » (PULS) développé par l'Université d'Helsinki. L'interface permet une visualisation géographique des dangers sanitaires selon un niveau d'alerte, par pays et par période (dernier mois).

Figure C.7 – Système MedISys



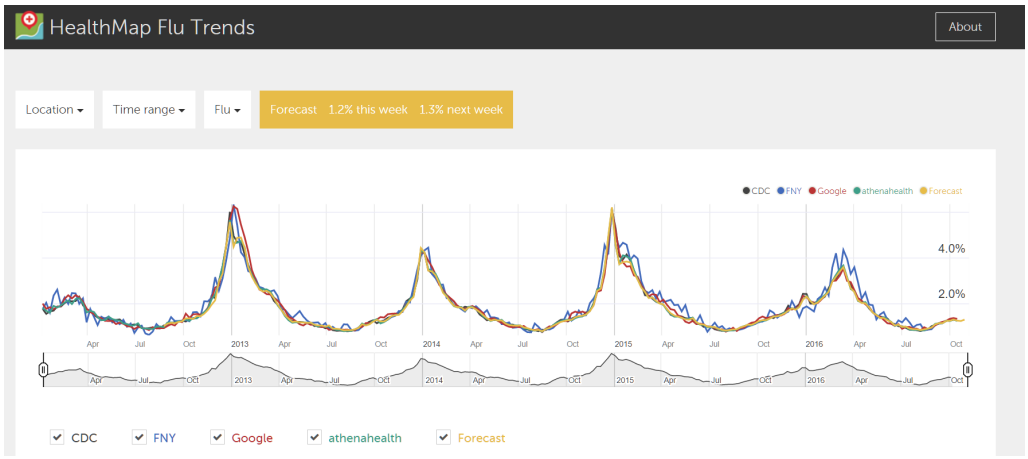
(a) Distribution mondiale des événements de la peste porcine africaine détectés par le système HealthMap d'avril à octobre 2016



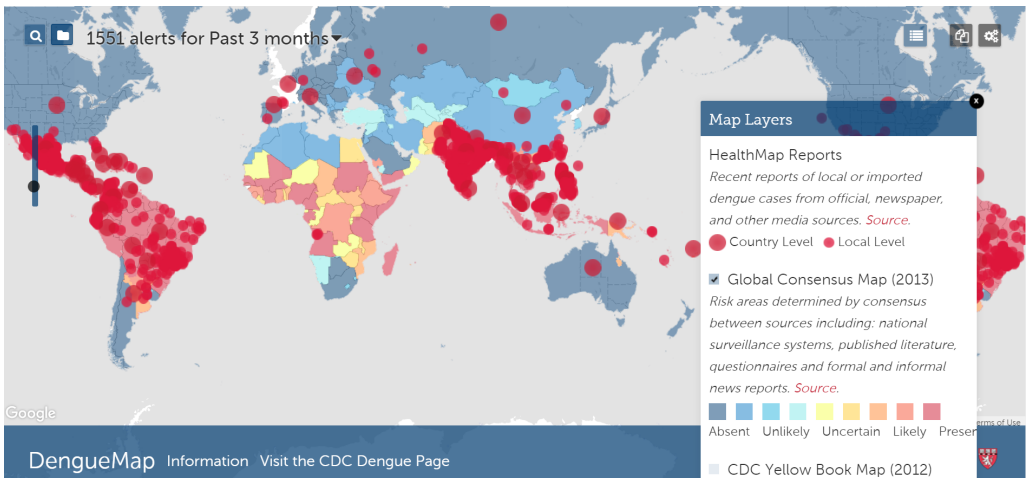
(b) Graphe chronologique des événements identifiés par le système HealthMap d'avril à octobre 2016

HealthMap (www.healthmap.org) dispose de cartes et de graphes chronologiques des événements détectés par le système. HealthMap permet aussi aux utilisateurs d'ajouter des informations et de les classer par importance ou qualité de l'information. Le système permet également la recherche par pays, maladie, période et source. Il permet également l'extraction de l'information structuré pour les derniers six mois.

Figure C.8 – Système HealthMap



(a) Prédiction des nouveaux cas de grippe aux États-Unis par le système HealthMap



(b) Cartographie des zones à risque de dengue du système HealthMap

Depuis peu, HealthMap en collaboration avec le Centre de Contrôle de Maladies (CDC) d'États-Unis, travaille sur la prédiction de la grippe aux États - Unis (http://compepi.org/project/healthmap-flu-trends/). HealthMap et CDC également travaillent sur la prédiction des zones à risque pour la dengue, au niveau mondial (http://www.healthmap.org/dengue/en/).

Figure C.9 – Autre projets

Annexe D

Caractéristiques des maladies modèles

Le Tableau D.1 de cet annexe présente les principales caractéristiques des maladies retenues comme modèle. Des détails supplémentaires sur la répartition géographique de ces maladies et leur importance au niveau international sont décrits dans le chapitre 3.1.1.

Tableau D.1 – Principales caractéristiques des maladies modèles

Caractéristiques		**Peste porcine africaine (PPA)**	**Fièvre aphteuse (FA)**	**Fièvre catarrhale ovine (FCO)**	**Schmallenberg (SBV)**
Déclaration obligatoire		Oui	Oui	Oui	Non
Transmission		Direct et Indirect	Direct et Indirect	Indirect	Indirect
Espèces sensibles [a]		POR	BOV, OVI, CAP, POR	BOV, OVI, CAP	BOV, OVI, CAP
Signes cliniques					
Généraux	Mortalité	Oui	Oui	Oui	Non
	Asthénie, prostration, anorexie	Oui	Oui	Oui	Oui
Respiratoires	Superficiels	Non	Non	Oui	Non
	Profonds	Oui	Oui	Oui	Non
Digestifs	Diarrhée	Oui	Non	Non	Oui
	Ptyalisme, sialorrhée	Non	Oui	Oui	Non
Locomoteurs/nerveux	Boitiers	Non	Oui	Oui	Non
	Parésie, paralysie	Oui	Non	Non	Non
	Modifications du comportement	Oui	Non	Non	Non
	Convulsions	Oui	Non	Non	Non
Cutanés/ muqueux	Vésicules	Non	Oui	Non	Non
	Ulcères	Non	Oui	Oui	Non
	Œdèmes	Non	Non	Oui	Non
	Purpura, hémorragies	Oui	Non	Non	Non
Reproducteurs	Avortements, mortinatalité	Oui	Oui	Oui	Oui
Périnatales/ congénitales	Malformations, déformations	Non	Non	Non	Oui
Vaccine disponible		Non	Oui (sérotype spécifique)	Oui (sérotype spécifique)	Non
Références		(ANSES, 2012)	(ANSES, 2012)	(ANSES, 2012)	(Wernike et al., 2013)

a. BOV=bovine, OVI=ovine, CAP=caprine, POR=porcine

Annexe E

Termes extraits avec *BioTex*

Cet annexe présente les termes sélectionnés par un expert du domaine, qui sont des termes de référence pour caractériser les quatre maladies modèles de ce travail. Les termes sont obtenus avec un processus de fouille de textes avec l'aide du logiciel *BioTex*. Les détails de l'extraction des termes sont présentés dans le chapitre 3.2.1.1.

E.1 Peste porcine africaine

Tableau E.1 – Termes extraits avec *BioTex* et sélectionnés comme pertinents pour caractériser la peste porcine africaine

Catégorie	Source	Terme
Maladie	***Google***	african swine fever news, african swine fever outbreak affects, african swine fever outbreak reported, african swine fever outbreaks, african swine fever reported, african swine fever situation, african swine fever spreads, asf case in boar, asf outbreaks, asf-infected wild, asf-infected wild boar, boar detected with african swine fever, cases of asf, disease african swine fever, new outbreaks of african swine fever, outbreaks of asf, presence of the asf, scientific opinion on african swine fever, spread of asf
	PubMed	acute asf, african swine fever episode, african swine fever outbreak, african swine fever outbreaks, african swine fever virus infection, asf diffusion, asf eradication, asf eradication programme, asf incursion, asf infection, asf outbreak, asf outbreak data, asf outbreaks, asf persistence, asf prevalence, asf situation, asf spread, asf virus antibody, asfv introduction, asfv prevalence, asfv spread, continued occurrence of asf, control of asf, control strategy for asf, eradication of asf, identification of asfv, ineffective management of asf, introduction of asf, occurrence of asf, potential risk of asf, prevalence of asf, regional asf outbreaks, risk for asfv, risk of asf, risk of asfv, sporadic outbreak of asf, spread of asfv, suitability for asf, temporal distribution of asf
Hôte	***Google***	backyard pigs, boar population, district pigs, informal piggeries, pig farm, pig population, swine farms, wild boar population, wild pigs, wild boar, wild boars
	PubMed	domestic pig, domestic pig populations, domestic pigs, extensive free range pig, pig farms, pig population, pig populations, slaughter pigs, warthog population, warthogs and tampans, wild boar populations, wild pig, wild pig population, wild pig populations, wild pig species, wild suids, wild boar, wild boars
Signe clinique	***Google***	*Mortalité* : dead pigs, dead wild boar, deadly pig disease, pigs dead *Fièvre* : fever case, fever case found, fever kills, fever outbreak affects, fever outbreak reported, fever outbreaks, fever reported, fever spreads, fever strikes *Hémorragique* : haemorrhagic fever *Pathologique* : muscle haemorrhages
	PubMed	*Mortalité* : gross mortality, group mortality, high lethality, high mortality, lethal pig disease, mortality loss, mortality losses *Fièvre* : fever episode, fever outbreak, fever outbreaks, fever virus infection *Hémorragique* : haemorrhagic disease, severe haemorrhagic disease, devastating haemorrhagic fever, haemorrhagic fever *Non-specifiquue* : sick pigs
Diagnose différentielle	***Google***	attack of swine fever, suspected swine fever, suspected swine fever outbreak, swine fever case, swine fever case found, swine fever kills, swine fever outbreak affects, swine fever outbreak reported, swine fever outbreaks, swine fever reported, swine fever scare, swine fever situation, swine fever spreads, swine fever strikes
	PubMed	swine fever episode, swine fever outbreak, swine fever outbreaks, swine fever virus infection

E.2 Fièvre aphteuse

Tableau E.2 – Termes extraits avec *BioTex* et sélectionnés comme pertinents pour caractériser la fièvre aphteuse

Catégorie Source	Terme	
Maladie	***Google***	fmd-control area, possible foot-and-mouth outbreak, foot-and-mouth outbreak at pig farm, foot-and-mouth case, foot-and-mouth outbreak at hog farm, first outbreak of foot-and-mouth disease, foot-and-mouth case at hog farm, recent outbreak of foot-and-mouth disease
	PubMed	fmd transmission, fmd outbreak, risk of fmd, control of fmd, introduction of fmd, fmdv infection, fmd control, fmd in sheep, prevalence of fmd, transmission of fmd, impact of fmd, foot-and-mouth disease seropositivity, outbreaks of fmd, first characterisation of fmdv, outbreak of fmd, fmd infection, endemic fmd, fmd lesions, situation of fmd, antibodies against fmd, fmd outbreak reports, fmd virus dispersal, potential introduction of fmd, fmd control strategies, fmdv introductions, favorable environment for fmdv, fmd epidemics, fmdv-infected livestock, foot-and-mouth outbreak, serious outbreak of fmd, foot-and-mouth-disease in ruminants
Hôte	***Google***	NA
	PubMed	cattle herds, wild boar, cattle and buffaloes, bactrian camels, dairy cattle, beef cattle, cattle and pigs, pig farms, small ruminants, large ruminants, cattle farms, deer farm
Signe clinique	***Google***	*Généraux* : production losses, low mortality
	PubMed	*Cutanés/ muqueux* : swine vesicular disease, mucosal disease, papular stomatitis, vesicular disease, vesicular stomatitis

E.3 Fièvre catarrhale ovine

Tableau E.3 – Termes extraits avec *BioTex* et sélectionnés comme pertinents pour caractériser la fièvre catarrhale ovine

Catégorie	Source	Terme
Maladie	***Google***	bluetongue reported, new strain of bluetongue virus, bluetongue vaccination, first outbreak of bluetongue, bluetongue antibodies, bluetongue cases, outbreak of bluetongue, first occurrence of bluetongue, occurrence of bluetongue, bluetongue outbreak confirmed, new, bluetongue outbreaks, outbreaks of bluetongue, first outbreak of btv, outbreak of bluetongue virus, outbreak of btv, bluetongue outbreak spreads, buetongue antibodies detected, voluntary bluetongue vaccination, fresh outbreak of bluetongue disease, bluetongue outbreak, first bluetongue case
	PubMed	bluetongue disease, prevalence of btv, bt spread, btv infection, btv epidemics, bluetongue virus antibody, emergence of bt, bluetongue vaccination, btv antibodies, bt vaccination, presence of btv rna, transmission of btv, outbreaks of bt, outbreaks of btv, spread of btv, recent incursion of btv, seasonal transmission of btv, outbreak of bluetongue, bluetongue incidence, distribution of btv, bluetongue disease results, control of btv, detection of bluetongue, dissemination of btv, fulminant btv infection, multiple novel btv, novel btv serotypes, btv outbreaks, antibodies against btv, incursion of btv, recent emergence of bluetongue virus, bt virus, recent emergence of bluetongue, emergence of bluetongue, btv incursions, clinical bluetongue disease, bluetongue virus infection, bt outbreaks, btv vaccination campaigns, presence of bluetongue, btv vaccination, incidence of btv, introduction of bluetongue, bluetongue infection, btv seroprevalence, btv antibody, bt epidemics, btv situation, bluetongue epidemic, bluetongue epizooty, bluetongue large epizooty
Hôte	***Google***	*Autre ruminant* : white-tailed deer
	PubMed	*Ovine/ caprine* : adult sheep, dairy ewes, pregnant ewes, cattle and goats, small ruminants *Bovine* : cattle herds, cattle population, newborn calves, newborn dairy calves, dairy calves *Autre ruminant* : red deer, roe deer, wild ruminant populations, cloven-hoofed ungulates, american camelids, ruminant populations, domestic ruminant species, domestic ruminant
Signe clinique	***Google***	Mortalité : livestock deaths *Hémorragique* : hemorrhagic disease, hemorrhagic disease widespread *Pathologique* : internal hemorrhaging in wild game, fatal internal hemorrhaging, internal hemorrhaging
	PubMed	*Mortalité* : excess mortality *Fièvre* : fever outbreak *Reproducteurs* : embryonic death, reproductive disorders, occurrence of abortion *Non-specifiquue* : sudden onset of weakness, general clinical signs, onset of weakness *Pathologique* : acute superficial myocardial hemorrhage, neutrophilic leukocytosis with monocytosis, marked pulmonary edema, normocytic normochromic anemia, superficial myocardial hemorrhage, hemorrhagic lesions, myocardial hemorrhage, neutrophilic leukocytosis, normochromic anemia, pulmonary congestion, pulmonary edema

E.4 Schmallenberg

Tableau E.4 – Termes extraits avec *BioTex* et sélectionnés comme pertinents pour caractériser le Schmallenberg

Catégorie	Source	Terme
Maladie	***Google***	schmallenberg disease, new schmallenberg virus outbreak, cases of schmallenberg, schmallenberg virus outbreak, sbv infection of wild deer, schmallenberg virus confirmed
	PubMed	antibodies against sbv, sbv infection, sbv infections, congenital sbv, sbv spread, current outbreak of Schmallenberg,, first confirmation of Schmallenberg, rapid spread of sbv, spread of Schmallenberg, antibodies against Schmallenberg, confirmation of schmallenberg, introduction of Schmallenberg, outbreak of Schmallenberg, presence of sbv, infections with sbv, sbv-associated abortion, schmallenberg virus infection, spread of sbv, sbv circulation, infections with schmallenberg, sbv-infected dairy herds, sbv-infected sheep herds, discovery of sbv, malformed sbv-positive offspring, schmallenberg virus epidemic, sbv-positive offspring, active sbv circulation, detection of schmallenberg, schmallenberg virus infections, introduction of sbv, first detection of sbv, first report of schmallenberg, rapid spread of schmallenberg, report of schmallenberg, risk for sbv, sbv circulation area, sbv detection area, sbv local circulation, schmallenberg virus outbreaks, subsequent sbv detection, suspicions of sbv, sbv detections, sbv outbreak, sbv outbreaks, sbv-affected farms, sbv-affected holdings, sbv-infected farms, cases of sbv, suspicion of sbv, sbv case, first case of Schmallenberg, case of Schmallenberg, first report of sbv, first suspicion of sbv
Hôte	***Google***	*Ovine/ caprine* : sheep farm, sheep holdings, goat holdings *Bovine* : bovine foetus *Autre ruminant* : deer populations
	PubMed	*Ovine/ caprine* : sheep holdings, red deer, lambs and calves, lambs and goats, goat population, sheep herds, small ruminants, goat kids, goat farms, goat foetus, sheep herd, sheep population, newborn lambs, small ruminant *Bovine* : kids and calves, adult dairy cows, cows and ewes, cow herds, newborn calves, adult cows, cattle herds, dairy cows, bovine foetuses, dairy cattle, cows and calves, cattle farms, cattle population *Autre ruminant* : bison population, european bison, fallow deer, deer populations
Signe clinique	***Google***	*Périnatales/ congénitales* : deformed offspring, lamb losses, stillborn lambs
	PubMed	*Fièvre* : nonspecific febrile syndrome, febrile syndrome *Neurologiques / locomoteurs* : neurologic signs *Digestifs* : watery diarrhea, watery diarrhoea *Reproducteurs* : substantial reproductive losses, aborted fetuses, reproductive losses, enzootic outbreak of abortion, outbreak of abortion *Périnatales/ congénitales* : congenital malformations, deformed lambs, malformed offspring, ovine congenital malformations, arthrogryposis hydranencephaly syndrome, severe foetal malformations, foetal malformations, vertebral malformations, severe congenital malformation, hydranencephaly syndrome, limb malformations, congenital malformation, malformed calves, severe congenital malformations, malformed lambs, malformed progeny, perinatal death, premature birth, lamb mortality, stillborn bovine foetuses *Non-spécifique* : mild transient disease *Respiratoire* : acute bronchopneumonia

Annexe F

Questionnaires Delphi

F.1 Delphi 1

Identification des signaux précoces d'émergence de la MALADIE* sur l'Internet

Merci de choisir « Répondre au questionnaire ».

Objectif : définir les expressions qui pourraient être utilisées pour des recherches automatiques sur l'Internet dans le but d'identifier des signaux précoces d'une émergence de MALADIE.

Ce questionnaire a deux parties :

1. Nous vous demandons de proposer des expressions qui d'après vous caractérisent la MALADIE et pourraient être utilisées pour identifier des signaux précoces de son émergence sur l'Internet.

2. Nous vous demandons d'évaluer une liste des expressions qui sont extraites automatiquement depuis des articles concernant les foyers de la maladie, selon leur spécificité de caractériser la MALADIE et en outre d'identifier des signaux précoces de son émergence sur l'Internet.

Les réponses à ce questionnaire sont confidentielles. La réponse aux questions prend 10 minutes.

Delphi partie 1
(Pour la peste porcine africaine, la fièvre aphteuse, la fièvre catharrale ovine, le Schmallenberg)

Listez des expressions qui d'après vous caractérisent la MALADIE* et pourraient être utilisées pour identifier sur l'Internet des signaux précoces de son émergence.
Si vous proposez des symptômes ou signaux faibles, merci d'indiquer l'hôte (espèce, catégorie d'âge etc.).
Parmi ces expressions, en ajoutant un « x » dans la 2eme colonne, identifiez les expressions les plus spécifiques pour identifier sur l'Internet des signaux précoces d'une émergence de la MALADIE.

	Expression	**Expressions les plus spécifiques (entrez « x »)**
1		
2		
3		
4		
5		etc…..

* MALADIE signifié soit la peste porcine africaine, soit la fièvre aphteuse, soit la fièvre catharrale ovine, soit le Schmallenberg

F.2 Delphi 2

Delphi partie 2 pour la peste porcine africaine

Evaluation des expressions extraites automatiquement
Évaluez chacune de ces expressions suivantes et leur spécificité pour caractériser la MALADIE et à identifier de signaux précoces de son émergence.

Qualificatifs de spécificité :
Très faible : une expression non spécifique, très improbable pour caractériser la PPA,
Faible : une expression faiblement spécifique, ne caractérise probablement pas PPA,
Moyenne : une expression moyennement spécifique, caractérise possiblement la PPA,
Très élevée : une expression très spécifique, caractérise dans presque tous les cas la PPA

	Très faible	Faible	Moyenne	Très élevée
aborted fetuses	○	○	○	○
arthrogryposis hydranencephaly syndrome	○	○	○	○
dead wild boar	○	○	○	○
district pigs fever outbreak	○	○	○	○
enzootic outbreak of abortion	○	○	○	○
extensive free range pig suspected swine fever	○	○	○	○
fever outbreak reported	○	○	○	○
fresh outbreak lethal pig disease	○	○	○	○
high mortality	○	○	○	○
lethal pig disease	○	○	○	○
malformed offspring	○	○	○	○
pig farms haemorrhagic fever	○	○	○	○
swine fever kills	○	○	○	○
wild boar	○	○	○	○
wild pigs gross mortality	○	○	○	○
devastating haemorrhagic fever	○	○	○	○

Commentaires. Veuillez écrire votre réponse ici :

Merci de vos réponses. Vous serez prochainement informé(e)s des résultats de cette étude.

Delphi partie 2 pour la fièvre aphteuse

Évaluation des expressions extraites automatiquement
Évaluez chacune de ces expressions suivantes et leur spécificité pour caractériser la MALADIE et à identifier de signaux précoces de son émergence.

Qualificatifs de spécificité :
Très faible : une expression non spécifique, très improbable pour caractériser la MALADIE (spécificité quantitative 0-0,2),
Faible : une expression faiblement spécifique, ne caractérise probablement pas la MALADIE (spécificité quantitative 0,21-0,4),
Moyenne : une expression moyennement spécifique, caractérise possiblement la MALADIE (spécificité quantitative 0,41-0,6),
Élevée : une expression spécifique, caractérise très probablement la MALADIE (spécificité quantitative 0,61-0,8),
Très élevée : une expression très spécifique, caractérise dans presque tous les cas la MALADIE (spécificité quantitative >0,81).

EXPRESSIONS INDICATIVES AUX SYMPTÔMES GÉNÉRAUX
Choisissez la réponse appropriée pour chaque élément (entrez « x » dans la colonne qui correspond à votre réponse):

	Très faible	**Faible**	**Moyenne**	**Élevée**	**Très élevée**
Mortalité faible	○	○	○	○	○
Pertes de production	○	○	○	○	○

EXPRESSIONS INDICATIVES AUX SYMPTÔMES CUTANÉO-MUQUEUX
Choisissez la réponse appropriée pour chaque élément (entrez « x » dans la colonne qui correspond à votre réponse):

	Très faible	**Faible**	**Moyenne**	**Élevée**	**Très élevée**
Stomatite vésiculeuse	○	○	○	○	○
Stomatite papuleuse	○	○	○	○	○
Maladie des muqueuses	○	○	○	○	○
Maladie vésiculeuse	○	○	○	○	○
Maladie vésiculeuse du porc	○	○	○	○	○

Delphi partie 2 pour la fièvre catharrale ovine

Évaluation des expressions extraites automatiquement
Évaluez chacune de ces expressions suivantes et leur spécificité pour caractériser la MALADIE et à identifier de signaux précoces de son émergence.

Qualificatifs de spécificité :
Très faible : une expression non spécifique, très improbable pour caractériser la MALADIE (spécificité quantitative 0-0,2),
Faible : une expression faiblement spécifique, ne caractérise probablement pas la MALADIE (spécificité quantitative 0,21-0,4),
Moyenne : une expression moyennement spécifique, caractérise possiblement la MALADIE (spécificité quantitative 0,41-0,6),
Élevée : une expression spécifique, caractérise très probablement la MALADIE (spécificité quantitative 0,61-0,8),
Très élevée : une expression très spécifique, caractérise dans presque tous les cas la MALADIE (spécificité quantitative >0,81).

EXPRESSIONS INDICATIVES AUX SYMPTÔMES GÉNÉRAUX
Choisissez la réponse appropriée pour chaque élément :

	Très faible	Faible	Moyenne	Élevée	Très élevée
Fièvre sur de nombreux animaux	○	○	○	○	○
Mortalité du bétail	○	○	○	○	○
Baisse de l'état général	○	○	○	○	○
Signes cliniques généraux	○	○	○	○	○

EXPRESSIONS INDICATIVES À LA PATHOLOGIE DE LA REPRODUCTION
Choisissez la réponse appropriée pour chaque élément :

	Très faible	Faible	Moyenne	Élevée	Très élevée
Apparition d'avortements	○	○	○	○	○
Décès embryonnaires	○	○	○	○	○

Delphi partie 2 pour le Schmallenberg - 1/2

Évaluation des expressions extraites automatiquement
Évaluez chacune de ces expressions suivantes et leur spécificité pour caractériser la MALADIE et à identifier de signaux précoces de son émergence.

Qualificatifs de spécificité :
Très faible : une expression non spécifique, très improbable pour caractériser la MALADIE (spécificité quantitative 0-0,2),
Faible : une expression faiblement spécifique, ne caractérise probablement pas la MALADIE (spécificité quantitative 0,21-0,4),
Moyenne : une expression moyennement spécifique, caractérise possiblement la MALADIE (spécificité quantitative 0,41-0,6),
Élevée : une expression spécifique, caractérise très probablement la MALADIE (spécificité quantitative 0,61-0,8),
Très élevée : une expression très spécifique, caractérise dans presque tous les cas la MALADIE (spécificité quantitative >0,81).

EXPRESSIONS INDICATIVES AUX SYMPTÔMES GÉNÉRAUX
Choisissez la réponse appropriée pour chaque élément :

	Très faible	Faible	Moyenne	Élevée	Très élevée
Syndrome fébrile non-spécifique	○	○	○	○	○
Maladie bénigne passagère	○	○	○	○	○

EXPRESSIONS INDICATIVES AUX SYMPTÔMES RESPIRATOIRES
Choisissez la réponse appropriée pour chaque élément :

	Très faible	Faible	Moyenne	Élevée	Très élevée
Bronchopneumonie aiguë	○	○	○	○	○

EXPRESSIONS INDICATIVES AUX SYMPTÔMES DIGESTIFS
Choisissez la réponse appropriée pour chaque élément :

	Très faible	Faible	Moyenne	Élevée	Très élevée
Diarrhée liquide	○	○	○	○	○

Delphi partie 2 pour le Schmallenberg – suite 2/2

EXPRESSIONS INDICATIVES À LA PATHOLOGIE DE LA REPRODUCTION
Choisissez la réponse appropriée pour chaque élément :

	Très faible	Faible	Moyenne	Élevée	Très élevée
Troubles de la reproduction	○	○	○	○	○
Avortements enzootiques	○	○	○	○	○
Veaux mort-nés	○	○	○	○	○
Agneaux mort-nés	○	○	○	○	○
Avorton	○	○	○	○	○
Mortalité périnatale	○	○	○	○	○
Naissance prématurée	○	○	○	○	○

EXPRESSIONS INDICATIVES AUX MALFORMATIONS ET DÉFORMATIONS CONGÉNITALES
Choisissez la réponse appropriée pour chaque élément :

	Très faible	Faible	Moyenne	Élevée	Très élevée
Malformations congénitales graves	○	○	○	○	○
Produits malformés	○	○	○	○	○
Malformations fœtales	○	○	○	○	○
Malformations des membres	○	○	○	○	○
Malformations vertébrales	○	○	○	○	○
Syndrome hydro-encéphalite	○	○	○	○	○
Syndrome arthrogryposehydro-encéphalie	○	○	○	○	○

F.3 Delphi 3

Delphi partie 3 pour la fièvre aphteuse

Quel(s) animal (animaux) sensible(s) – associerez-vous préférentiellement aux symptômes suivants ? (entrez « x » dans la colonne qui correspond à votre réponse)

	Symptômes généraux	**Symptômes cutanéo-muqueux**
Bovins		
Buffles		
Petits ruminants		
Porcins		
Sangliers		
Chameau bactrien		
Cerfs		

Commentaires. Veuillez écrire votre réponse ici :

Merci de vos réponses. Vous serez prochainement informé(e)s des résultats de cette étude.

Delphi partie 3 pour la fièvre catharrale ovine

Quel(s) animal (animaux) sensible(s) – associerez-vous préférentiellement aux symptômes suivants ? (entrez « x » dans la colonne qui correspond à votre réponse)

	Symptômes généraux	**Pathologie de la reproduction**
Bovins		
Ovins		
Caprins		
Brebis gestantes		
Veaux nouveaux - nés		
Cerf élaphe		
Chevreuil		

Commentaires. Veuillez écrire votre réponse ici :

Merci de vos réponses. Vous serez prochainement informé(e)s des résultats de cette étude.

Delphi partie 3 pour le Schmallenberg

Quel(s) animal (animaux) sensible(s) vous associerez plus avec vos réponses pour des symptômes préalablement évalués de votre part ? (entrez « x » dans la colonne qui corresponde à votre réponse)

	Symptômes généraux	**Symptômes respiratoires**	**Symptômes digestifs**	**Pathologie de la reproduction**	**Malformations et déformations congénitales**
Bovins					
Ovins					
Caprins					
Veaux					
Veaux nouveau-nés					
Foetus bovin					
Agneaux					
Agneaux nouveau-nés					
Chevreaux					
Foetus caprin					
Daim					
Cerf élaphe					
Bison					

Commentaires. Veuillez écrire votre réponse ici :

Merci de vos réponses. Vous serez prochainement informé(e)s des résultats de cette étude.

Annexe G

Termes proposés par les experts

G.1 Peste porcine africaine

Tableau G.1 – Termes proposés par un panel d'experts pour caractériser l'émergence de la peste porcine africaine

Catégorie	Pertinence	Terme
Généraux	***Hautement spécifique***	pigs mortality, grouped mortality of wild boars and pigs, high mortalities of pigs, wild boar mortality, dead wild boar, disease with high pork mortality, high mortality of pigs, sudden death of pigs, high mortalities of wild boars, dead wild boar, high mortality in pigs, increased observation of fallen animals in hunting grounds, high pig mortality, fatal disease of swine, wild boar found dead, raising mortality in pigs, wild boar found dead, boars found dead, mass mortality in pigs, mortality of wild boar
	Spécifique	strong mortality of boar, events of mortality in boars, dead wild boar, boars found dead, regular pig mortality, mortality of wild boar, pigs die, strong mortality in boar, dead wild boar at water spots (rivers, lakes, ponds), epizootic disease with high mortality in pigs, hyperthermia in pig breeding, strong hyperthermia in pigs, high fever in pigs
Digestifs	***Hautement spécifique***	bloody diarrhea in pig breeding
	Spécifique	raising vomiting in pigs
Nerveux/ locomoteurs	***Spécifique***	hunters observing unusual behaviour of wild boar towards humans and dogs, swine nervous disorders
Hémorragiques	***Hautement spécifique***	bloody diarrhea in pig haemorrhagic fever in pigs, haemorrhagic fever in pigs, haemorrhagic disease of pigs, pigs bleeding, Haemorrhages in pigs, haemorrhagic fever in boars, skin redness + swine (pig / boar ...), disease with big muddy pig spleen haemorrhagic nodes in pigs Haemorrhagic nodes in boars septicaemia in pigs friable spleen, haemorrhagic hypertrophy or splenomegaly + swine
	Spécifique	subcutaneous haemorrhage in pig breeding, haemorrhagic lesions in pigs, bleeding and pigs, internal haemorrhages in pigs
Reproducteurs	***Hautement spécifique***	abortions in sows
	Spécifique	abortion of pregnant sows, abortion may occur in pregnant sow, reproductive disorders suides
Multiples	***Hautement spécifique***	high fever and mortality in pigs, haemorrhagic syndrome and mortality of pigs, hyperthermia and haematological disorders + swine
CMaladie	***Hautement spécifique***	african swine fever, african swine fever, african swine fever, ASFV, DNA arbovirus in pork, ASFV vaccine, ASF virus detection in wild boar, ASF, classical swine fever
	Spécifique	swine fever
Autre	***Hautement spécifique***	sick wild boar roaming at daylight, highly contagious disease in domestic pigs and wild boar, highly contagious disease affecting the pork, emerging disease in pigs, serious illness without pork vaccine, lack of appetite and prostration in pigs,sharp drop in pork prices panic among pig farmers
	Spécifique	sick wild boar, contagious viral disease in pigs, disease in boars, pig-borne disease transmitted by pork products, health crisis in pigs, disease of high contagiousness in pigs or wild boars, serious epidemic in pigs, new disease in pigs, disease in pigs not responding to antibiotic treatment, pig disease requiring stamping mainly as a control measure

G.2 Fièvre aphteuse

Tableau G.2 – Termes proposés par un panel d'experts pour caractériser l'émergence de la fièvre aphteuse

Catégorie	Pertinence	Terme
Généraux	***Hautement spécifique***	brutal decrease of production in dairy cows, mortality of suckling piglets, mortality of sulking lambs and kids
	Spécifique	loss of appetite, mortality in piglets, high mortality in suckling animals, lamb mortality, contagious fever between herds moderate fever occurrence of high fever between cows
Cutanés/ muqueux	***Hautement spécifique***	bucal blisters, ulcers in the mouth in bovines, lesions of the hoofs, ulcers in the interdigital space in pigs, vesicle and ulcer in the mouth of bovines, contagious vesicles, blisters or ulcers in bovines and pigs, salivation, blisters on the mouth, blisters on the groin in pigs, blisters on the mouth and the lips in ruminants, vesicles in the udder of ruminants, hypersalivation in cows
	Spécifique	hypersalivation in bovines, highly contagious vesicular disease, blisters in animals, bucal lesions, ulcers in the interdigital space in bovines, udder blisters in cows, spread of blisters between bovines and pigs, excessive salivation, vesicle on the nose, ulcers in the udder in ruminants
Nerveux/ locomoteurs	***Hautement spécifique***	lameness in ruminants, lameness in pigs, lameness, lameness in cows
	Spécifique	lameness in animals, lameness, increased lameness in sheep
Multiples	***Hautement spécifique***	lameness and ptyalisme in bovines, hypersalivation and lameness in cows, blisters and lameness in bovine and porcine farms
	Spécifique	hypersalivation and lameness in bovines, fever and hypersalivation in bovines, outbreak of lameness and blisters in bovines or pigs, occurrence of high fever and hypersalivation between cows
Maladie	***Hautement spécifique***	foot and mouse disease, foot and mouth disease outbreak suspicion, foot and mouse disease, suspicion of case of foot and mouth disease, foot and mouth disease
	Spécifique	FMD vaccination, foot and mouth disease vaccination
Autre	***Hautement spécifique***	disease in ruminants and Suidae, high morbidity and contagiosity
	Spécifique	highly contagious animal disease, disease in domestic and wild ruminants and pigs, global distribution, high number of affected animals, considerable economic losses, illegal importation of animals, closed livestock markets

G.3 Fièvre catarrhale ovine

Tableau G.3 – Termes proposés par un panel d'experts pour caractériser l'émergence de la fièvre catarrhale ovine

Catégorie	Pertinence	Terme
Généraux	***Hautement spécifique***	Catarrhal fever, Hyperthermia between sheep,
	Spécifique	fever ovine or bovine, mortality ovine or bovine, morbidity and mortality in a sheep farm, mortality in sheep, decrease in ovine production, weakness in ruminants
Respiratoires	***Hautement spécifique***	dyspnea in ruminants
Cutanés/ muqueux	***Hautement spécifique***	ptyalisme in ruminants hypersalivation bucal lesions in sheep (but also bovines), nasal discharge in sheep, abundant nasal discharge in ruminants, facial oedemas, lacrimation/tearing in ruminants, oedematous/ swolen eyes in ruminants, oedemas in sheep
	Spécifique	ptyalisme, bucal ulcer ovine or bovine, nasal discharge ovine or bovine, oedema ovine or bovin, cyanosys of the tongue in ruminants, gengival ulcers and hypersalivation in ruminants, swelling of the neck in ruminants, teat leasons in bovines, hair loss in bovines
Nerveux/ locomoteurs	***Spécifique***	limited mouvement in sheep, arthritis in ruminants, lameness in ruminants, lameness in bovines
Reproducteurs	***Hautement spécifique***	abortions in ruminants
	Spécifique	abortion ovine or bovine, abortions in sheep, fertilty problems in sheep, infertility in ruminants, mortinatality/Still birth in ruminants, abortions
Nom de la maladie	***Hautement spécifique***	ovine catarrhal fever, bluetongue, BTV, ovine catarrhal fever, sheep catarrhal fever, blue tongue disease, BTV, bluetongue
	Spécifique	blue tongue, bluetongue, BTV, sheep catarrhal fever, blue tongue disease, midge disease ovine or bovine, disease transmitted by Culicoides, disease transmitted by midges, vector-borne disease in ovines, disease transmitted by Culicoides, bluetongue, bluetongue clinical signs, bluetongue virus antibody detection, bluetongue virus genome detection, clinical sign suspitious to Bluetongue, clinical sign suspitious to BT, ovine catarrhal fever
Autre	***Spécifique***	Culicoid, obligatory vaccination of ruminants

G.4 Schmallenberg

Tableau G.4 – Termes proposés par un panel d'experts pour caractériser l'émergence de Schmallenberg

Catégorie	Pertinence	Terme
Généraux	***Spécifique***	drop in milk production, transient drop in milk production
Digestifs	***Spécifique***	mild acute diarrhea
Reproducteurs	***Spécifique***	abortions, series of abortions, abortions, stillbirth, abortion
Périnatales/ congénitales	***Hautement spécifique***	congenital malformations, arthrogryposis hydranencephaly syndrome, malformed foetuses, scoliosis, arthrogryposis hydranencephaly syndrome, torticollis, arthrogryposis, arthrogryposis in calves, brachygnatia, brachygnatia, arthrogryposis in lambs, scoliosis, scoliosis in newborn calves, torticollis, scoliosis in newborn lambs, hydranencephaly
	Spécifique	malformed foetuses, arthrogryposis, anomaly in newborn calves, malformed new-borns, facial deformations, arthrogryposis, anomaly in newborn lambs, deformed calves, hydranencephaly, malformed newborn calves, deformed lambs, foetal teratology, malformed newborn lambs
Maladie	***Spécifique***	bunyavirus

Annexe H

Précision des paires d'associations obtenues avec des mesures statistiques

Les Figures H.1, H.2, H.3 et H.3 présentent les résultats de la précision (sous-chapitre 3.2.2.4.2) des 20 meilleures paires d'associations proposées par les sept mesures statistiques du chapitre 3.2.2. Notons que chaque paire d'association est constituée d'un hôte et d'un signe clinique pour les maladies modèles.

La précision est le nombre des pages Web pertinentes parmi toutes les pages Web collectés pour chaque paire d'association.

H.1 Peste porcine africaine

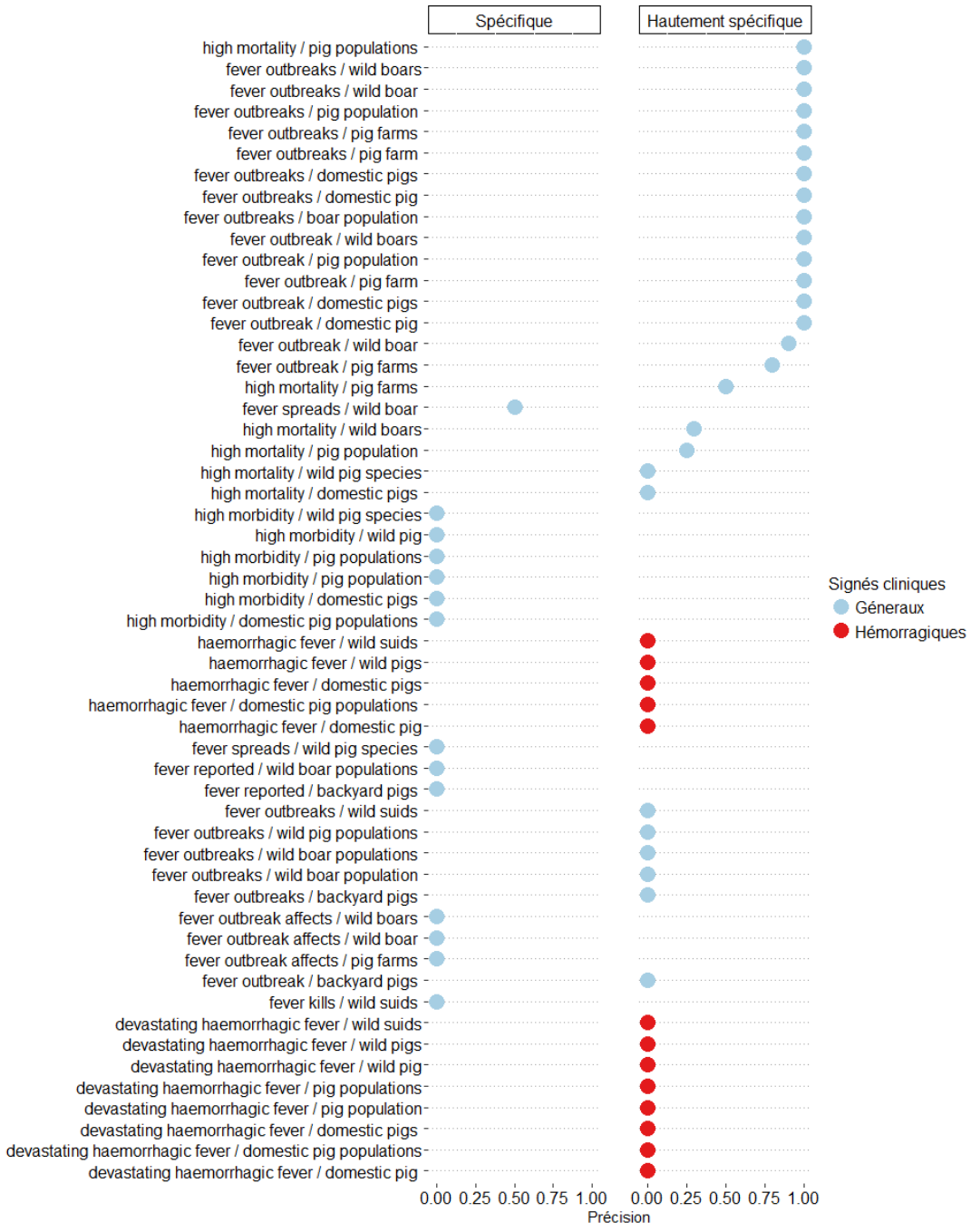


Figure H.1 – Précision des paires d'associations pour la peste porcine africaine de collecter des pages Web pertinentes sur le Web

H.2 Fièvre aphteuse

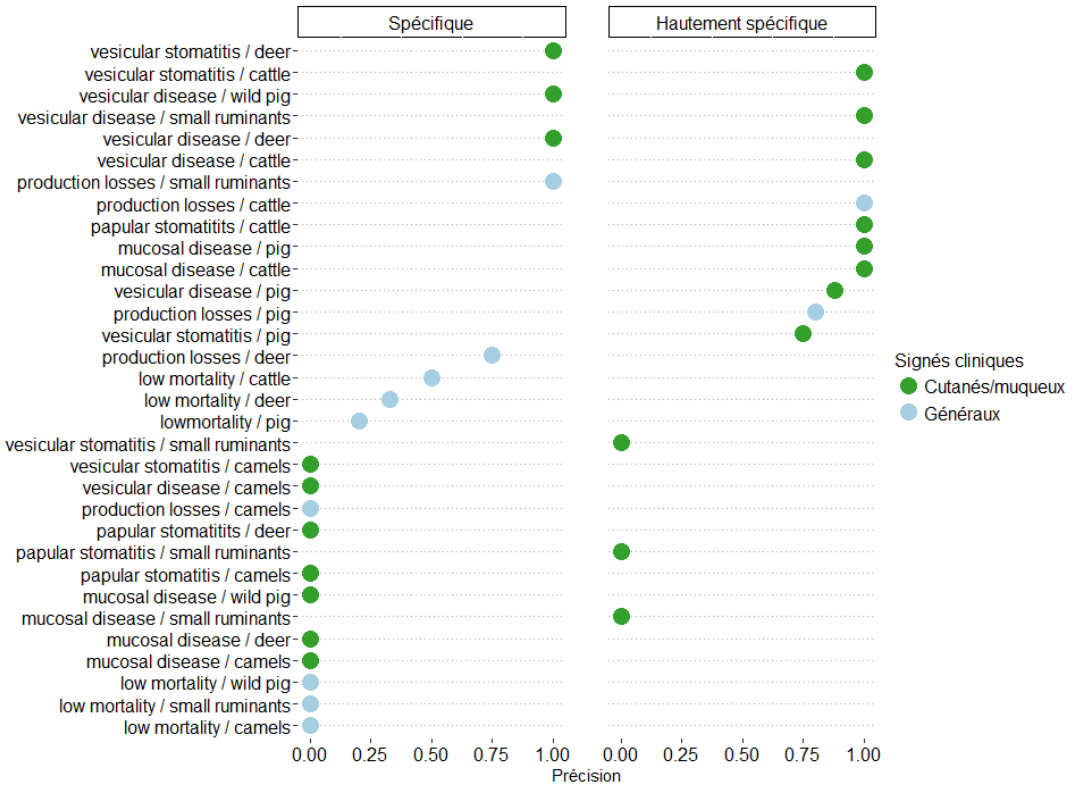


Figure H.2 – Précision des paires d'associations pour la fièvre aphteuse de collecter des pages Web pertinentes sur le Web

H.3 Fièvre catarrhale ovine

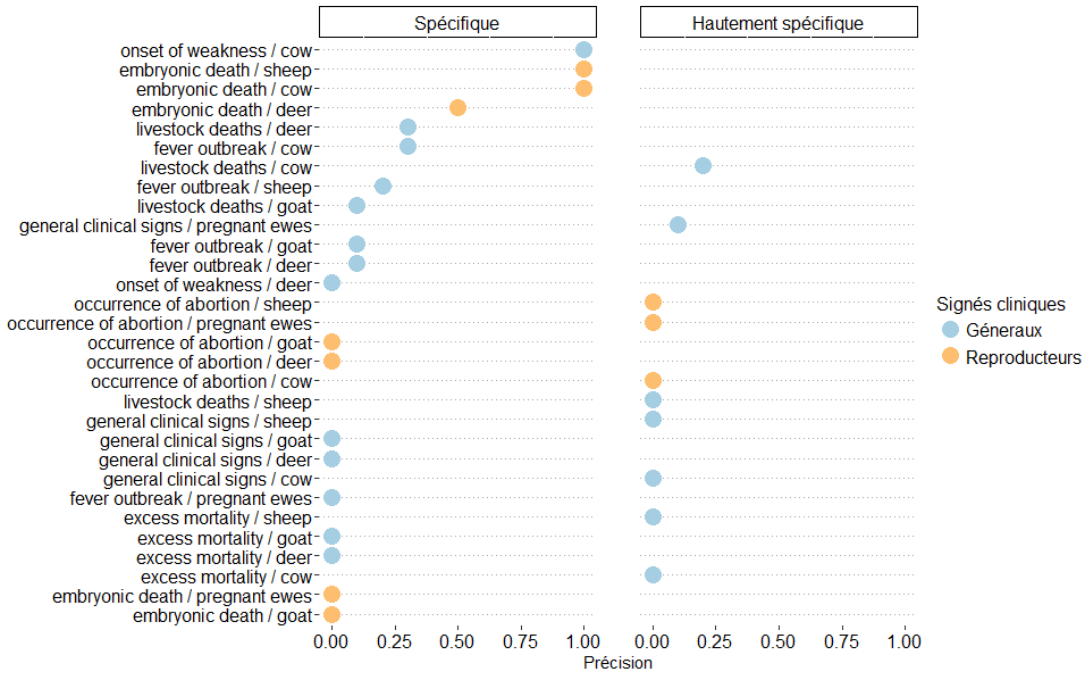


Figure H.3 – Précision des paires d'associations pour la fièvre catarrhale ovine de collecter des pages Web pertinentes sur le Web

H.4 Schmallenberg

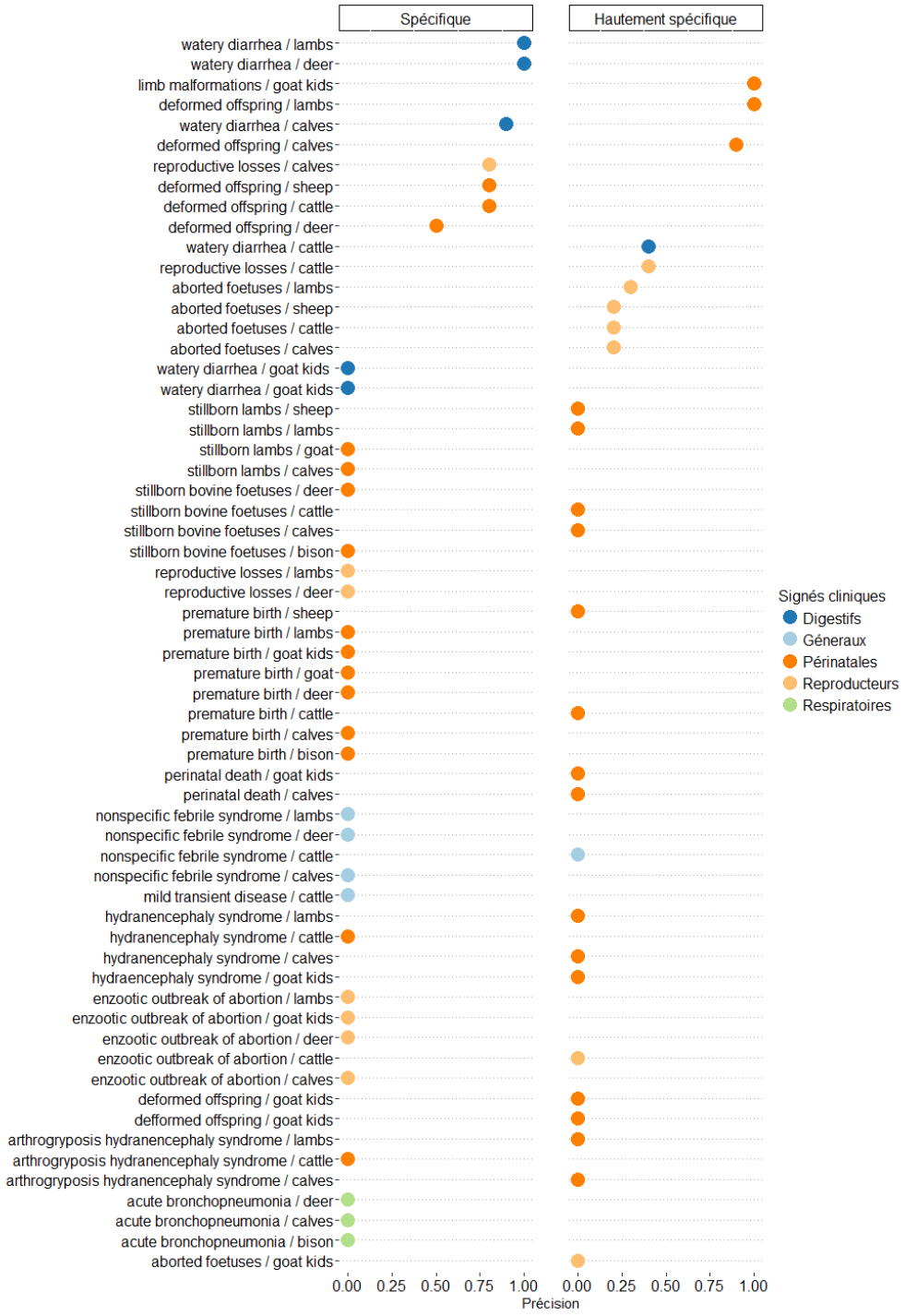


Figure H.4 – Précision des paires d'associations pour le Schmallenberg de collecter des pages Web pertinentes sur le Web

Annexe I

Dictionnaires pour le système PADI-web

Lorsqu'un article trouvé sur le Web est considéré comme pertinent, il doit obligatoirement contenir dans le titre ou dans le corps du texte, les mots-clés généreux ou leurs lemmes du Tableau I.1. Cette règle s'applique également pour des hôtes (Tableu I.3) et des signes cliniques (Tableau I.4). Le nom de la maladie (Tableau I.2) n'est pas obligatoire.

Actuellement le système PADI-web a des dictionnaires en trois langues (français, anglais, et espagnol), avec la langue anglaise opérationnelle. Notons que ces dictionnaires sont évolutifs, avec des ajouts possibles par l'utilisateurs et des liaisons possibles avec d'autres dictionnaires ou ontologies.

I.1 Dictionnaire des mot – clés généraux

Tableau I.1 – Dictionnaire des mots-clés généraux pour le système PADI-web

Anglais	**Français**	**Espagnol**
outbreak	foyer	foco
case	cas	caso
contamination	contamination	contaminación
loss	perte	pérdida
disease	maladie	enfermedad
infection	infection	infección
infestation	infestation	infestación
pathogen	pathogène	patógeno
virus	virus	virus
bacteria	bactérie	bacteria
illness	maladie	enfermedad
syndrome	syndrome	síndrome
introduction	introduction	introducción
incursion	incursion	incursión
identification	identification	identificación
spread	propagation	propagación
diffusion	diffusion	difusión
emergence	émergence	emergencia
discovery	découverte	descubrimiento
detection	dépistage	detección
restriction	restriction	restricción
alert	alerte	alerte
contamination	contamination	contaminación
attack	atteinte	ataque
diagnose	diagnose	diagnóstico
parasite	parasite	parásito
vector	vecteur	vector
arbovirus	arbovirus	arbovirus
arbovirosis	arbovirose	arbovirosis
dissemination	dissémination	diseminación

I.2 Dictionnaire des maladies

Tableau I.2 – Dictionnaire des maladies pour le système PADI-web

Maladie (en anglais)	**Termes** (en anglais)
Avian influenza	avian influenza, avian flu, fowl plague, bird flu
African swine fever	african swine fever, warthog disease, wart-hog disease, wart hog disease
Foot-and-mouth disease	foot-and-mouth disease, foot and mouth disease
Bluetongue	bluetongue, blue tongue, ovine catarrhal fever, sheep catarrhal fever
Schmallenberg	schmallenberg virus, schmallenberg infection, schmallenberg disease

I.3 Dictionnaire des hôtes

Tableau I.3 – Dictionnaire des hôtes pour le système PADI-web

Hôte (en anglais)	**Termes** (en anglais)
Avian	poultry, waterfowl, bird, turkey, duck, chicken
Bovine	cattle, cow, bull, heifer, calf, buffallo
Ovine	ovine, sheep, ram, ewe, lamb
Caprine	caprine, goat, buck, doe, goat kid, kid
Porcine	pig, boar, sow, swine, warthog, tampan, suidae, piglet, piggery, barrow, hog
Autre	deer, red deer, roe deer, white-tail deer, camel, small ruminant, bison

I.4 Dictionnaire des signes cliniques

Tableau I.4 – Dictionnaire des signes cliniques pour le système PADI-web

Signes cliniques (en anglais)	**Termes** (en anglais)
General-fever	fever, hyperthermia, pyrexia, febrile syndrome
General-mortality	dead, deadly, death, lethal, lethality, mortal, mortality, deaths
General-other	anorexia, depression, drop in milk production, fall of performance, fall of production, general clinical signs, inapetence, loss of appetite, nonspecific clinical signs, production disorders, production losses, prostration, weakness
Respiratory	dyspnoea, hard breathing
Digestive	diarrhea, vomiting
Skin/ mucous	aphthae, blisters, blue tongue, bluetongue, cyanosis, excessive salivation, hypersalivation, lacrimation, skin lesions, mucous lesions, mucosal disease, nasal discharge, oedemas, ptyalisme, papular stomatitis, stomatitis, vesicular stomatitis, swelling, tearing, ulcers, vesicles, vesicular disease
Haemorrhagic	bleeding, haemorrhages, haemorrhagic disease, haemorrhagic fever, haemorrhagic syndrome, petechiae
Nervous/ locomotive	arthritis, uncoordination, lameness, locomotion disorders, movement disorders, neurologic disorders, neurologic signs, paralysis, scoliosis
Reproductive	reproductive losses, fall of reproduction, fecundation disorders, fecundation problems, fertility disorders, fertility problems, reproductive disorders, reproductive problems, embryonic death, abortion(s), stillbirth, premature birth, stillborn
Perinatal/ congenital	AHS syndrome, arthrogryposis hydrancephaly syndrome, brachygnatia, perinatal death, hydranencephaly, congenital malformations, deformed lambs, malformed offspring, fœtal malformations, vertebral malformations, limb malformations, malformed calves, malformed lambs, deformed offspring, malformed progeny

I.5 Indicateurs épidémiologiques extraits du Web

Le Tableau I.5 d'information structurée contient l'information épidémiologique extraite à partir des données du Web. Les données sont acquises par le système PADI-web et catégorisées comme pertinentes.

Les indicateurs sanitaires sont conçus avec l'avis des experts du domaine et une revue bibliographique pour les systèmes de surveillance.

L'information extraite permettra une analyse spatiale et temporelle de événements sanitaires. Cette information permettra également le suivi de la performance du système PADI-web et du processus de fouille de texte en temps-réel.

Tableau I.5 – Exemple d'une information structurée extraite à partir des textes non-structurés du Web

article id	**publication date**	**location**	**latitude**	**longitude**	**admin**	**country**	**continent**	**event confidence**
118e93ee4e	13/06/2016	Mamusa	-27.22308	25.27706	North-West	South Africa	Africa	0.72
118e93ee4e	13/06/2016	Free State	-29	26	Orange Free State	South Africa	Africa	0.6
118e93ee4e	13/06/2016	Letsemeng	-29.35811	25.014931	Orange Free State	South Africa	Africa	0.58
148cc73f57	18/03/2016	Kirovohrad	48.5132	32.259701	Kirovohrad	Ukraine	Europe	0.875
3034773164	10/06/2016	North West	-26.5	26	North-West	South Africa	Africa	0.645083
3034773164	10/06/2016	Free State	-29	26	Orange Free State	South Africa	Africa	0.582167
30eace930b	08/03/2016	Ngozi Province	-2.875	29.924999	Ngozi	Burundi	Africa	0.585217
4198a3e4d0	26/05/2016	Sumy	50.9216	34.800289	Sumy	Ukraine	Europe	0.99
4198a3e4d0	26/05/2016	Sumskaja Oblast	51	34	Sumy	Ukraine	Europe	0.98
4316d582ea	09/06/2016	Chernivtsi	48.29149	25.94034	Chernivtsi	Ukraine	Europe	0.84
6cb2a7dec8	13/06/2016	Free State	-29	26	Orange Free State	South Africa	Africa	0.620091

suite...

dates	disease	species	clinical signs	keywords	source
2016-06-01,2016-06-06	African swine fever	Porcine	Fever	outbreak, cases, outbreaks	ThePigSite
2016-06-01,2016-06-06	African swine fever	Porcine	Fever	outbreak, cases, outbreaks	ThePigSite
2016-06-01,2016-06-06	African swine fever	Porcine	Fever	outbreak, cases, outbreaks	ThePigSite
	African swine fever	Porcine	Fever, Haemorrhagic	disease, spread, outbreak	GlobalMeatNews
	African swine fever	Porcine	Fever	outbreak, cases, case, disease, virus	Times LIVE
	African swine fever	Porcine	Fever	outbreak, cases, case, disease, virus	Times LIVE
	African swine fever	Porcine	Fever	outbreaks, cases, virus	ThePigSite
2016-05-26,2016-05-25	African swine fever	Porcine	Fever, Mortality	cases, virus	Interfax
2016-05-26,2016-05-25	African swine fever	Porcine	Fever, Mortality	cases, virus	Interfax
2016-06-06,2016-06-07	African swine fever	Porcine	Fever	case, outbreak, disease	Pig World
	African swine fever	Porcine	Fever	cases, disease, bacteria, virus, outbreaks	AllAfrica

Bibliographie

Adebayo, S. (2013). « Evolving epidemic intelligence : Towards improved health events detection over social media streams ». Mém.de mast. Univ. Saint-Andrews.

Ahn, D. (2006). « The stages of event extraction ». In : *Proceedings of the Workshop on Annotating and Reasoning about Time and Events ARTE 06*, p. 1–8.

Akobeng, A. K. (2007). « Understanding diagnostic tests 3 : Receiver operating characteristic curves. » In : *Acta paediatrica* 96.5, p. 644–647.

Alexandrov, T., D. Stefanov, P. Kamenov, A. Miteva, S. Khomenko, K. Sumption, H. Meyer-Gerbaulet et K. Depner (2013). « Surveillance of foot-and-mouth disease (FMD) in susceptible wildlife and domestic ungulates in Southeast of Bulgaria following a FMD case in wild boar ». In : *Veterinary Microbiology* 166.1-2, p. 84–90.

Alomar, O., A. Batlle, J. Brunetti, R. García, R. Gil, A. Granollers, S. Jiménez, A. Laviña, J. P. Linge, M. Pautasso, C. Reverté, J. Riudavets, A. Rortais, G. Stancanelli, S. Volani et S. Vos (2015). « Development and testing of the media monitoring tool MedISys for early identification and reporting of existing and emerging plant health threats ». In : *EPPO Bulletin* 45.2, p. 288–293.

Amato-Gauci, A. et A. Ammon (2008). « The surveillance of communicable diseases in the European Union–a long-term strategy (2008-2013). » In : *Euro surveillance* 13.26, p. 3.

Ammon, A. et P. Makela (2010). « Integrated data collection on zoonoses in the European Union, from animals to humans, and the analyses of the data ». In : *International Journal of Food Microbiology* 139.SUPPL. 1, S43–S47.

Amrine, D. E., B. J. White, R. L. Larson, A. D.E., W. B.J. et L. R.L. (2014). « Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for bovine respiratory disease ». In : *Computers and Electronics in Agriculture* 105, p. 9–19.

ANSES (2012). *Risques d'introduction et de diffusion d'agents pathogènes exotiques en France métropolitaine et propositions de mesures pour réduire ces risques*. Rapp. tech. Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail, p. 239.

Anyamba, A., J.-P. Chretien, J. Small, C. J. Tucker, P. B. Formenty, J. H. Richardson, S. C. Britch, D. C. Schnabel, R. L. Erickson et K. J. Linthicum (2009). « Prediction of a Rift Valley fever outbreak. » In : *Proceedings of the National Academy of Sciences of the United States of America* 106.3, p. 955–9.

Arsevska, E., J. Hellal, S. Mejri, S. Hammami, P. Marianneau, D. Calavas et V. Hénaux (2015). « Identifying Areas Suitable for the Occurrence of Rift Valley Fever in North Africa : Implications for Surveillance ». In : *Transboundary and Emerging Diseases* 63.6, p. 658–674.

Arsevska, E., T. Balenghien, C. Garros, R. Lancelot, S. Zientara, C. Sailleau et E. Bréard (2014a). « Fièvre catarrhale ovine en Europe en 2014 : épizootie dans les Balkans, progression de la circulation en Italie et en Espagne ». In : *Bulletin épidémiologique* 69, p. 16–18.

Arsevska, E., D. Calavas, M. Dominguez, H. Guis, P. Hendrikx, R. Lancelot, B. Peiffer et J.-B. Perrin (2014b). « Fièvre catarrhale ovine en Sardaigne – un point épidémiologique pour les années 2012 et 2013 ». In : *Bulletin épidémiologique* 56, p. 13–14.

Arsevska, E., M. Dominguez, B. Peiffer, J.-B. Perrin, C. Marcé, P. Hendrikx, F. Etore, C. Collignon, R. Lancelot, T. Lefrançois et D. Calavas (2014c). « Développement d'une veille sanitaire internationale en santé animale dans le cadre de la Plateforme ESA ». In : *Bulletin épidémiologique* 60, p. 30.

Arsevska, E., A. Mercier, A. Bronner, D. Calavas, J. Cauchard, P. Caufur, S. Falala, M. Hamon, P. Hendrikx, R. Lancelot, S. Rauterau et C. Tisseuil (2016a). « Dermatose nodulaire contagieuse des bovins : état des connaissances et situation épidémiologique dans les Balkans au 31 juillet 2016 ». In : *Bulletin épidémiologique* 74, p. 25–29.

Arsevska, E., M. Roche, S. Falala, R. Lancelot, D. Chavernac, P. Hendrikx et B. Dufour (2016b). « Monitoring Disease Outbreak Events on the Web Using Text-mining Approach and Domain Expert Knowledge ». In : *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Sous la dir. de N. Calzolari et al. Portoroz, Slovenia : European Language Resources Association (ELRA).

Arsevska, E., M. Roche, P. Hendrikx, D. Chavernac, S. Falala, R. Lancelot et B. Dufour (2016c). « Identification of associations between clinical signs and hosts to monitor the web for detection of animal disease outbreaks ». In : *International Journal of Agricultural and Environmental Information Systems* 7.3, p. 1–20.

Arsevska, E., M. Roche, P. Hendrikx, D. Chavernac, S. Falala, R. Lancelot et B. Dufour (2016d). « Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web ». In : *Computers and Electronics in Agriculture* 123, p. 104–115.

Assunção, R. et T. Correa (2009). « Surveillance to detect emerging space-time clusters ». In : *Computational Statistics and Data Analysis* 53.8, p. 2817–2830.

Azé, J. (2003). « Extraction de Connaissances dans des Données Numériques et Textuelles ». Thèse de doct. Université Paris Sud - Paris XI.

Bahk, C. Y., D. A. Scales, S. R. Mekaru, J. S. Brownstein et C. C. Freifeld (2015). « Comparing timeliness, content, and disease severity of formal and informal source outbreak reporting. » In : *BMC infectious diseases* 15.1, p. 135.

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

Barboza, P., L. Vaillant, A. Mawudeku, N. P. Nelson, D. M. Hartley, L. C. Madoff, J. P. Linge, N. Collier, J. S. Brownstein, R. Yangarber et P. Astagneau (2013). « Evaluation of Epidemic Intelligence Systems Integrated in the Early Alerting and Reporting Project for the Detection of A/H5N1 Influenza Events ». In : *PLoS ONE* 8.3. Sous la dir. de H. Nishiura, e57252.

Barboza, P., L. Vaillant, Y. L. Strat, D. M. Hartley, N. P. Nelson, A. Mawudeku, L. C. Madoff, J. P. Linge, N. Collier, J. S. Brownstein et P. Astagneau (2014). « Factors influencing performance of internet-based biosurveillance systems used in epidemic intelligence for early detection of infectious diseases outbreaks ». In : *PLoS ONE* 9.3. Sous la dir. de V. Chaturvedi, e90536.

Ben Jebara, K. (2007). « WAHIS and the role of the OIE's reference laboratories and collaborating centres. » In : *Developments in biologicals* 128, p. 69–72.

Ben Jebara, K. (2010). « The OIE World Animal Health Information System : the role of OIE Reference Laboratories and Collaborating Centres in disease reporting. » In : *Revue scientifique et technique* 29.3, p. 451–458.

Bohigas, P. A., F. Santos-O'Connor, D. Coulombier, F. Santos-O&apos ;connor, D. Coulombier, F. Santos-O'Connor et D. Coulombier (2009). « Epidemic intelligence and travel-related diseases : ECDC experience and further developments ». In : *Clinical Microbiology and Infection* 15.8, p. 734–739.

Brownstein, J. S., C. C. Freifeld et L. C. Madoff (2009). « Digital Disease Detection — Harnessing the Web for Public Health Surveillance ». In : *New England Journal of Medicine* 360.21. P. 2153–2157.

Brownstein, J. S., C. C. Freifeld, B. Y. Reis et K. D. Mandl (2008). « Surveillance Sans Frontieres : Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project ». In : *PLoS Med* 5.7, p. 1–6.

Bruno, P., D. Delamarre et P. L. Beux (2002). « Indexation de textes médicaux par extraction de concepts, et ses utilisations ». In : *6th International Conference on the Statistical Analysis of Textual Data (JADT'2002)*. T. 2, p. 617–628.

Brunton, L. A., R. Nicholson, A. Ashton, N. Alexander, W. Wint, G. Enticott, K. Ward, J. M. Broughan et A. V. Goodchild (2015). « A novel approach to mapping and calculating

the rate of spread of endemic bovine tuberculosis in England and Wales ». In : *Spatial and Spatio-temporal Epidemiology* 13, p. 41–50.

Buza, T. M., S. W. Jack, H. Kirunda, M. L. Khaitsa, M. L. Lawrence, S. Pruett et D. G. Peterson (2015). « ERAIZDA : A model for holistic annotation of animal infectious and zoonotic diseases ». In : *Database* 2015.1.

Cakici, B., K. Hebing, M. Grünewald, P. Saretok et A. Hulth (2010). « CASE : a framework for computer supported outbreak detection. » In : *BMC medical informatics and decision making* 10, p. 14.

Chang et C. Manning (2013). « SUTIME : Evaluation in TempEval-3 ». In : *Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. T. 2. Second Joint Conference on Lexical and Computational Semantics (*SEM). Association for Computational Linguistics. Atlanta, Georgia : Association for Computational Linguistics, p. 78–82.

Chanlekha, H. et N. Collier (2010). « Analysis of syntactic and semantic features for fine-grained event-spatial understanding in outbreak news reports ». In : *Journal of biomedical semantics* 1.1, p. 1.

Chiang, J. H. et H. C. Yu (2005). « Literature extraction of protein functions using sentence pattern mining ». In : *IEEE Transactions on Knowledge and Data Engineering* 17.8, p. 1088–1098.

Choi, B. C. K. (2012). « The Past, Present, and Future of Public Health Surveillance ». In : *Scientifica* 2012.Table 1, p. 1–26.

Church, K. W., B. Laboratories, M. Hill et P. Hanks (1990). « Word Association Norms,Mutual Information , and Lexicography ». In : *Computational linguistics* 16.1, p. 22–29.

Cilibrasi, R. et P. M. B. Vitanyi (2007). « The Google Similarity Distance ». In : *arXiv :cs/0412098* 19.3, p. 370–383.

Clements, A. C. A., D. U. Pfeiffer, M. J. Otte, K. Morteo et L. Chen (2002). « A global livestock production and health atlas (GLiPHA) for interactive presentation, integration and analysis of livestock data ». In : *Preventive Veterinary Medicine*. T. 56. 1, p. 19–32.

Collier, N. (2010). « WhatLs unusual in online disease outbreak news ? » In : *Journal of biomedical semantics* 1.1, p. 2.

Collier, N. (2012). « Uncovering text mining : a survey of current work on web-based epidemic intelligence. » In : *Global public health* 7.7, p. 731–49.

Collier, N. et S. Doan (2012). « GENI-DB : A database of global events for epidemic intelligence ». In : *Bioinformatics* 28.8, p. 1186–1188.

Collier, N., S. Doan, A. Kawazoe, R. M. Goodwin, M. Conway, Y. Tateno, Q. H. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, M. Shigematsu et K. Taniguchi (2008). « BioCaster :

Detecting public health rumors with a Web-based text mining system ». In : *Bioinformatics* 24.24, p. 2940–2941.

Conraths, F. J., M. Peters et M. Beer (2013a). « Schmallenberg virus, a novel orthobunyavirus infection in ruminants in Europe : potential global impact and preventive measures. » In : *New Zealand veterinary journal* 61.2, p. 63–7.

Conraths, F. J., D. Kämer, K. Teske, B. Hoffmann, T. C. Mettenleiter et M. Beer (2013b). « Reemerging Schmallenberg virus infections, Germany, 2012. » In : *Emerging infectious diseases* 19.3, p. 513–514.

Conway, M., S. Doan, A. Kawazoe et N. Collier (2009). « Classifying disease outbreak reports using n-grams and semantic features ». In : *International Journal of Medical Informatics* 78.12, e47–e58.

Conway, M., A. Kawazoe, H. Chanlekha et N. Collier (2010). « Developing a disease outbreak event corpus ». In : *Journal of Medical Internet Research* 12.3, e43.

Copeland, J., G. Rainisch, J. Tokars, H. Burkom, N. Grady et R. English (2007). « Syndromic prediction power : comparing covariates and baselines ». In : *Advances in Disease Surveillance* 2, p. 46.

Cowen, P., T. Garland, M. E. Hugh-Jones, A. Shimshony, S. Handysides, D. Kaye, L. C. Madoff, M. P. Pollack et J. Woodall (2006). « Evaluation of ProMED-mail as an electronic early warning system for emerging animal diseases : 1996 to 2004 ». In : *Journal of the American Veterinary Medical Association* 229.7, p. 1090–1099.

Cui, X., N. Yang, Z. Wang, C. Hu, W. Zhu, H. Li, Y. Ji et C. Liu (2015). « Chinese social media analysis for disease surveillance ». In : *Personal and Ubiquitous Computing* 19.7, p. 1125–1132.

Dalkey, N. et O. Helmer (1963). « An experimental application of the Delphi method to the use of experts ». In : *Management Science* 9.3, p. 458–467.

Dawson, P. M., M. Werkman, E. Brooks-Pollock, M. J. Tildesley, S. Figure, E. The et F. Tns (2015). « Epidemic predictions in an imperfect world : modelling disease spread with partial data. » In : *Proceedings. Biological sciences / The Royal Society* 282.1808, p. 27.

Delichère, M. et D. Memmi (2002). « Analyse Factorielle Neuronale pour Documents Textuels ». In : *TALN*. Sous la dir. d'INRIA. 1. INRIA, Le Chesnay, France, p. 24–27.

Dibie, J., S. Dervaux, E. Doriot, L. Ibanescu et C. Pénicaud (2016). « [MS]^2O - A Multiscale and Multi-step Ontology for Transformation Processes : Application to Micro-Organisms ». In : *Graph-Based Representation and Reasoning - 22nd International Conference on Conceptual Structures, ICCS 2016, Annecy, France, July 5-7, 2016, Proceedings*, p. 163–176.

Doan, S., N. Collier, H. Xu, H. D. Pham et M. P. Tu (2012). « Recognition of medication information from discharge summaries using ensembles of classifiers. » In : *BMC medical informatics and decision making* 12.1, p. 36.

Doan, S., A. Kawazoe, M. Conway et N. Collier (2009). « Towards role-based filtering of disease outbreak reports ». In : *Journal of Biomedical Informatics* 42.5, p. 773–780.

Economopoulou, A., P. Kinross, D. Domanovic et D. Coulombier (2014). « Infectious diseases prioritisation for event-based surveillance at the European Union level for the 2012 Olympic and Paralympic games ». In : *Eurosurveillance* 19.15, p. 20770.

Elrahman, S. M. A. et A. Abraham (2013). « A Review of Class Imbalance Problem ». In : *Network and Innovative Computing* 1.2013, p. 332–340.

European Council (2016). *Council Directive 82/894/EEC on the notification of animal diseases within the Community*.

Falagas, M. E., E. I. Pitsouni, G. A. Malietzis et G. Pappas (2008). « Comparison of PubMed, Scopus, Web of Science, and Google Scholar : strengths and weaknesses. » In : *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 22.2, p. 338–42.

Ferreira, J. D., D. Paolotti, F. M. Couto et M. J. Silva (2012). « On the usefulness of ontologies in epidemiology research and practice. » In : *Journal of epidemiology and community health* 67.5, p. 385–8.

Formenty, P., C. Roth, F. Gonzalez-Martin, T. Grein, M. Ryan, P. Drury, M. K. Kindhauser et G. Rodier (2006). « Les pathogénes émergents, la veille internationale et le réglement sanitaire international (2005) ». In : *Médecine et Maladies Infectieuses* 36.1, p. 9–15.

Frantzi, K., S. Ananiadou et H. Mima (2000). « Automatic recognition of multi-word terms : The C-value/NC-value method ». In : *International Journal on Digital Libraries* 3.2, p. 115–130.

Freifeld, C. C., K. D. Mandl, B. Y. Reis et J. S. Brownstein (2008). « HealthMap : Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports ». In : *Journal of the American Medical Informatics Association* 15.2, p. 150–157.

Furrer, L., S. Küker, J. Berezowski, H. Posthaus, F. Vial et F. Rinaldi (2015). « Constructing a syndromic terminology resource for veterinary text mining ». In : *Proceedings of the 11th International Conference on Terminology and Artificial Intelligence*. Sous la dir. de T. Poibeau et P. Faber. T. 1495, p. 61–70.

Gilbert, M. et A. Liebhold (2010). « Comparing methods for measuring the rate of spread of invading populations ». In : *Ecography* 33.5, p. 809–817.

Glanville, W. A. de, L. Vial, S. Costard, B. Wieland et D. U. Pfeiffer (2014). « Spatial multi-criteria decision analysis to predict suitability for African swine fever endemicity in Africa. » In : *BMC veterinary research* 10, p. 9.

Godfrey, E. R. et S. E. Randolph (2011). « Economic downturn results in tickborne disease upsurge ». In : *Parasites Vectors* 53.2, p. 84.

Guglielmetti, P., D. Coulombier, G. Thinus, F. Van Loock et S. Schreck (2006). « The early warning and response system for communicable diseases in the EU : an overview from 1999 to 2005. » In : *Euro surveillance* 11.12, p. 215–220.

Gustafson, L. L., D. H. Gustafson, M. C. Antognoli et M. D. Remmenga (2013). « Integrating expert judgment in veterinary epidemiology : Example guidance for disease freedom surveillance ». In : *Preventive Veterinary Medicine* 109.1-2, p. 1–9.

Hallgren, K. A. (2012). « Computing Inter-Rater Reliability for Observational Data : An Overview and Tutorial ». In : *Tutorials in quantitative methods for psychology* 8.1, p. 23–34.

Hanley, A. et J. McNeil (1982). « The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve ». In : *Radiology* 143.1, p. 29–36.

Hartley, D., N. Nelson, R. Walters, R. Arthur, R. Yangarber, L. Madoff, J. Linge, A. Mawudeku, N. Collier, J. Brownstein, G. Thinus et N. Lightfoot (2010). « The landscape of international event-based biosurveillance ». In : *Emerg Health Threats Journal* 3.e3.

Hausberg, M. C., A. Hergert, C. Kröger, M. Bullinger, M. Rose et S. Andreas (2012). « Enhancing medical students' communication skills : development and evaluation of an undergraduate training program ». In : *BMC Medical Education* 12.1, p. 16.

Hay, S. I., K. E. Battle, D. M. Pigott, D. L. Smith, C. L. Moyes, S. Bhatt, J. S. Brownstein, N. Collier, M. F. Myers, D. B. George et P. W. Gething (2013a). « Global mapping of infectious disease. » In : *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 368.1614, p. 20120250.

Hay, S. I., D. B. George, C. L. Moyes et J. S. Brownstein (2013b). « Big Data Opportunities for Global Infectious Disease Surveillance ». In : *PLoS Medicine* 10.4, e1001413.

He, H. et E. A. Garcia (2009). « Learning from imbalanced data ». In : *IEEE Transactions on Knowledge and Data Engineering* 21.9, p. 1263–1284.

Heredia-Langner, A., L. R. Rodriguez, A. Lin et J. B. Webster (2015). « Selecting a Classification Ensemble and Detecting Process Drift in an Evolving Data Stream ». In : *Proceedings of the International Conference on Data Mining (DMIN'2015)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering et Applied Computing (WorldComp), p. 31.

Hutwagner, L., T. Browne, G. M. Seeman et A. T. Fleischauer (2005). « Comparing Aberration Detection Methods with Simulated Data ». In : *Emerging Infectious Diseases* 11.2, p. 314–316.

Hutwagner, L., W. Thompson, G. M. Seeman et T. Treadwell (2003). « The bioterrorism preparedness and response Early Aberration Reporting System (EARS) ». In : *Journal of Urban Health* 80.2 Suppl 1, p. i89–i96.

Jackson, M. L., A. Baer, I. Painter et J. Duchin (2007). « A simulation study comparing aberration detection algorithms for syndromic surveillance. » In : *BMC medical informatics and decision making* 7.1, p. 6.

Japkowicz, N. et S. Stephen (2002). « The class imbalance problem : A systematic study ». In : *Intelligent Data Analysis* 6.5, p. 429–449.

Jimeno, A., E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga et D. Rebholz-Schuhmann (2008). « Assessment of disease named entity recognition on a corpus of annotated sentences. » In : *BMC bioinformatics* 9 Suppl 3.Suppl 3, S3.

Jindal, P. et D. Roth (2013). « Extraction of events and temporal expressions from clinical narratives ». In : *Journal of Biomedical Informatics* 46.SUPPL. S13–S19.

Jones, K. E., N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman et P. Daszak (2008). « Global trends in emerging infectious diseases. » In : *Nature* 451.7181, p. 990–3.

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

Keller, M., C. C. Freifeld et J. S. Brownstein (2009b). « Automated vocabulary discovery for geo-parsing online epidemic intelligence ». In : *BMC Bioinformatics* 10.1, p. 385.

Klopotek, M. A., J. Koronacki, M. Marciniak, A. Mykowiecka et S. Wierzchon (2013). « Language Processing and Intelligent Information Systems – 20th International Conference, IIS 2013, Warsaw, Poland, June 17-18, 2013. Proceedings ». In : *Language Processing and Intelligent Information Systems – 20th International Conference, IIS 2013, Warsaw, Poland, June 17-18, 2013. Proceedings*. T. 7912. Springer, p. 281.

Kulldorff, M. et N. Nagarwalla (1995). « Spatial disease clusters : detection and inference. » In : *Statistics in medicine* 14.8, p. 799–810.

Kulldorff, M. (2001). « Prospective time periodic geographical disease surveillance using a scan statistic ». In : *Journal of the Royal Statistical Society Series a - Statistics in Society* 164.1, p. 61–72.

Kulldorff, M., R. Heffernan, J. Hartman, R. Assunçao et F. Mostashari (2005). « A space-time permutation scan statistic for disease outbreak detection ». In : *PLoS Medicine* 2.3. Sous la dir. de S. M. Blower, p. 0216–0224.

Lallich, S. et O. Teytaud (2004). « Evaluation et validation de l'intérêt des règles d'association ». In : *Revue des Nouvelles Technologies de l'Information (RNTI-E-1) : Mesures de qualité pour la fouille de données* 1.2, p. 193–218.

Langmuir, A. D. (1980). « The Epidemic Intelligence Service of the Centers for Disease Control ». In : *Public Health Rep* 95.5, p. 470–477.

Laroche, A., P. Drouin et G. Bernier-Colborne (2011). « Étude de l'influence de la taille du corpus de référence sur l'extraction terminologique automatique contrastive ». In : *9th International Conference on Terminology and Artificial Intelligence (TIA 2011)*. Paris, France, p. 66–72.

Le Potier, M., E. Arsevska et C. Marcé (2015). « Persistance de la peste porcine africaine en Europe de l'Est ». In : *Bulletin épidémiologique*, p. 28–29.

LeCun, Y., Y. Bengio et G. Hinton (2015). « Deep learning ». In : *Nature* 521.7553, p. 436–444.

Lee, K., A. Agrawal et A. Choudhary (2015). « Mining Social Media Streams to Improve Public Health Allergy Surveillance ». In : *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15*. ACM Press, p. 815–822.

Lightner, D. V. (2012). « Global transboundry disease politics : The OIE perspective ». In : *Journal of Invertebrate Pathology*. Diseases in Aquatic Crustaceans : Problems and Solutions for Global Food Security 110.2, p. 184–187.

Linge, J. P., R. Steinberger, T. P. Weber, R. Yangarber, E. van der Goot, D. H. Al Khudhairy et N. I. Stilianakis (2009). « Internet surveillance systems for early alerting of health threats. » In : *Euro Surveillance* 14.13, p. 2.

Lombardo, J. S. et D. L. Buckeridge (2006). *Disease Surveillance : A Public Health Informatics Approach*. John Wiley & Sons, p. 1–458.

Lossio-Ventura, J. A., C. Jonquet, M. Roche et M. Teisseire (2014). « Towards a mixed approach to extract biomedical terms from text corpus ». In : *Knowledge Discovery in Bioinformatics* 4.1, p. 15.

Lossio-Ventura, J. A., C. Jonquet, M. Roche et M. Teisseire (2016). « Biomedical term extraction : overview and a new methodology ». In : *Information Retrieval* 19.1-2, p. 59–99.

Lunardon, N., G. Menardi et N. Torelli (2014). « ROSE : A Package for Binary Imbalanced Learning ». In : *The R Journal* 6, p. 79–89.

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

Lyon, A., A. Mooney et G. Grossel (2013b). « Agriculture | Free Full-Text | Using AquaticHealth.net to Detect Emerging Trends in Aquatic Animal Health ». In : *Agriculture* 3.2, p. 299–309.

MAAF (2013). *Ministère de l'agriculture, de l'agroalimentaire et de la forêt. Arrêté du 29 juillet 2013 relatif à la définition des dangers sanitaires de première et deuxième catégorie pour les espèces animales*.

MacRae, J., T. Love, M. G. Baker, A. Dowell, M. Carnachan, M. Stubbe et L. McBain (2015). « Identifying influenza-like illness presentation from unstructured general practice clinical narrative using a text classifier rule-based expert system versus a clinical expert. » In : *BMC medical informatics and decision making* 15.1, p. 78.

Madoff, L. C. (2004). « ProMED-Mail : An Early Warning System for Emerging Diseases ». In : *Clinical Infectious Diseases* 39.2, p. 227–232.

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

Martin, V., L. D. Simone et J. Lubroth (2007). « Geographic information systems applied to the international surveillance and control of transboundary animal diseases , a focus on highly pathogenic avian influenza ». In : *Veterinaria Italiana* 43.3, p. 437–450.

Martinez, D., M. R. Ananda-Rajah, H. Suominen, M. A. Slavin, K. A. Thursky et L. Cavedon (2015). « Automatic detection of patients with invasive fungal disease from free-text computed tomography (CT) scans ». In : *Journal of Biomedical Informatics* 53, p. 251–260.

Menardi, G. et N. Torelli (2014). « Training and assessing classification rules with imbalanced data ». In : *Data Mining and Knowledge Discovery* 28.1, p. 92–122.

Morgan, K. L., I. G. Handel, V. N. Tanya, S. M. Hamman, C. Nfon, I. E. Bergman, V. Malirat, K. J. Sorensen et B. M. d. C. Bronsvoort (2014). « Accuracy of herdsmen reporting versus serologic testing for estimating foot-and-mouth disease prevalence ». In : *Emerging Infectious Diseases* 20.12, p. 2048–2054.

Munzert, S. (2015). *Automated Data Collection with R.A Practical Guide to Web Scraping and Text Mining*. Wiley. United Kingdom.

Mykhalovskiy, E. et L. Weir (2006). « Canadian Contribution to Global Public Health Intelligence Network and Early Warning Outbreak Detection ». In : *Canadian Journal of Public Health* 97.1, p. 42–44.

Nazar, R., J. Vivaldi et T. Cabre (2008). « A Suite to Compile and Analyze an LSP Corpus ». In : *Sixth International Conference on Language Resources and Evaluation (LREC, 2008)*.

Marrakech, Morocco : European Language Resources Association (ELRA), p. 1164–1169.

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

Nelson, N., L. Yang, A. R. A. Reilly, J. E. Hardin et D. M. Hartley (2012). « Event-based internet biosurveillance : relation to epidemiological observation ». In : *Emerg Themes* 9.1, p. 4.

Noto, K., M. H. S. Jr et C. Elkan (2008). « Learning to Find Relevant Biological Articles Without Negative Training Examples ». In : Lecture Notes in Computer Science. Sous la dir. de W. Wobcke et M. Zhang, p. 202–213.

Novak, B., D. Mladenič et M. Grobelnik (2006). « Text Classification with Active Learning ». In : *From Data and Information Analysis to Knowledge Engineering : Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation e.V. University of Magdeburg, March 9–11, 2005*. Sous la dir. de M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger et W. Gaul. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 398–405.

Okoli, C. et S. Pawlowski (2004). « The Delphi Method as a Research Tool : An Example , Design Considerations and Applications 1 Introduction 2 Overview of the Delphi method ». In : *Information & Management* 42.1, p. 15–29.

Paquet, C., D. Coulombier, R. Kaiser et M. Ciotti (2006). « Epidemic intelligence : a new framework for strengthening disease surveillance in Europe. » In : *Euro surveillance* 11.12, p. 212–214.

Pazienza, M. T., A. Stellato, A. G. Tudorache, A. Turbati et F. Vagnoni (2012). « An architecture for data and knowledge acquisition for the Semantic Web : the AGROVOC use case ». In : Springer, p. 426–433.

Perrin, J. B. (2013). « Modélisation de la mortalité bovine dans un objectif de surveillance épidémiologique ». Thèse de doct. Universite Claude Bernard, Lyon, France.

Pioz, M., H. Guis, D. Calavas, B. Durand, D. Abrial et C. Ducrot (2011). « Estimating front-wave velocity of infectious diseases : A simple, efficient method applied to bluetongue ». In : *Veterinary Research* 42.1, p. 60.

Piskorski, J. et R. Yangarber (2013). « Information Extraction : Past, Present and Future ». In : sous la dir. de T. Poibeau, H. Saggion, J. Piskorski et R. Yangarber. Theory and Applications of Natural Language Processing. Springer Berlin Heidelberg, p. 23–49.

Poibeau, T. (2003). *Extraction automatique d'information : du texte brut au Web sémantique*. Lavoisier, p. 238.

Roche, M. (2004). « Intégration de la construction de la terminologie de domaines spécialisées dans un processus global de fouille de textes ». Thèse de doct. Université de Paris-Sud. Faculté des Sciences d'Orsay (Essonne).

Roche, M., S. Fortuno, J. A. Lossio-Ventura, A. Akli, S. Belkebir, T. Lounis et S. Toure (2015). « Extraction automatique des mots-clés à partir de publications scientifiques pour l'indexation et l'ouverture des données en agronomie ». In : *Cah Agric* 24.5, p. 313–320.

Roche, M. et V. Prince (2010). « A web-mining approach to disambiguate biomedical acronym expansions ». In : *Informatica (Ljubljana)* 34.2, p. 243–253.

Rodeia, S. P. (2008). « EFSA assessment of the risk of introducing foot and mouth disease into the EU and the reduction of this risk through interventions in infected countries : A review and follow-up : Editorial ». In : *Transboundary and Emerging Diseases* 55.1, p. 3–4.

Rodriguez, L. L., T. A. Bunch, M. Fraire et Z. N. Llewellyn (2000). « Re-emergence of vesicular stomatitis in the western United States is associated with distinct viral genetic lineages. » In : *Virology* 271.1, p. 171–81.

Rortais, A., J. Belyaeva, M. Gemo, E. van der Goot et J. P. Linge (2010). « MedISys : An early-warning system for the detection of (re-)emerging food- and feed-borne hazards ». In : *Food Research International* 43.5, p. 1553–1556.

Sadatsafavi, M., N. Shahidi, F. Marra, M. J. FitzGerald, K. R. Elwood, N. Guo et C. A. Marra (2010). « A statistical method was used for the meta-analysis of tests for latent TB in the absence of a gold standard, combining random-effect and latent-class methods to estimate test accuracy ». In : *Journal of Clinical Epidemiology* 63.3, p. 257–269.

Saeed, A., S. Kanwal, M. Arshad, M. Ali, R. S. Shaikh et M. Abubakar (2015). « Foot-and-mouth disease : overview of motives of disease spread and efficacy of available vaccines. » In : *Journal of animal science and technology* 57.1, p. 10.

Salton, G. (1983). *Introduction to Modern Information Retrieval*. New York : Mcgraw-Hill College, p. 440.

Sánchez-Vizcaíno, J. M., L. Mur et B. Martínez-López (2013). « African swine fever (ASF) : Five years around Europe ». In : *Veterinary Microbiology* 165.1–2. P. 45–50.

Saneifar, H., P. Poncelet et M. Roche (2015). « From Terminology Extraction to Terminology Validation : An Approach Adapted to Log Files ». In : *Journal of Universal Computer Science* 21.4, p. 604–635.

Santamaria, S. L., M. Fallon, J. M. Green, S. Schulz et J. R. Wilcke (2012). « Developing the Animals in Context Ontology ». In : *3rd International Conference on Biomedical Ontology*. T. 897.

Santamaria, S. L. et K. L. Zimmerman (2011). « Uses of informatics to solve real world problems in veterinary medicine. » In : *Journal of veterinary medical education* 38.2, p. 103–9.

Santos, J. C. et S. Matos (2014). « Analysing Twitter and web queries for flu trend prediction. » In : *Theoretical biology & medical modelling* 11 Suppl 1.Suppl 1, S6.

Schmid, H. (1994). « Probabilistic Part-of-Speech Tagging Using Decision Trees ». In : *New Methods in Language Processing* 4, p. 44–49.

Schölkopf, B., C. J. C. Burges et A. J. Smola (1999). *Advances in Kernel Methods : Support Vector Learning*. MIT Press, p. 400.

Schriml, L. M., C. Arze, S. Nadendla, A. Ganapathy, V. Felix, A. Mahurkar, K. Phillippy, A. Gussman, S. Angiuoli, E. Ghedin, O. White et N. Hall (2009). « GeMInA, Genomic Metadata for Infectious Agents, a geospatial surveillance pathogen database ». In : *Nucleic Acids Research* 38.SUPPL.1, p. D754–D764.

Semenza, J. C., B. Sudre, J. Miniota, M. Rossi, W. Hu, D. Kossowsky, J. E. Suk, W. Van Bortel et K. Khan (2014). « International Dispersal of Dengue through Air Travel : Importation Risk for Europe ». In : *PLoS Neglected Tropical Diseases* 8.12.

Serrano, L., M. Bouzid, T. Charnois, S. Brunessaux et B. Grilheres (2013). « Extraction et agrégation automatique d'évenements pour la veille en sources ouvertes : du texte à la connaissance ». In : *24èmes Journées francophones d'Ingénierie des Connaissances*. Lille, France.

Sherman, D. M. (2007). *Tending Animals in the Global Village : A Guide to International Veterinary Medicine*. T. 27. John Wiley & Sons, p. 495.

Sherman, D. M. (2010). « A global veterinary medical perspective on the concept of one health : focus on livestock. » In : *ILAR journal / National Research Council, Institute of Laboratory Animal Resources* 51.3, p. 281–287.

Smith-Akin, K. A., C. F. Bearden, S. T. Pittenger et E. V. Bernstam (2007). « Toward a veterinary informatics research agenda : An analysis of the PubMed-indexed literature ». In : *International Journal of Medical Informatics* 76.4, p. 306–312.

Soderland, S. (1999). « Learning Information Extraction Rules for Semi-Structured and Free Text ». In : *Machine Learning* 34.1, p. 233–272.

Song, M., H. Yu et W.-S. Han (2015). « Developing a hybrid dictionary-based bio-entity recognition technique ». In : *BMC medical informatics and decision making* 15.1, p. 1.

Steinberger, R., F. Fuart, E. V. D. Goot et C. Best (2008). « Text Mining from the Web for Medical Intelligence ». In : *Health (San Francisco)*. NATO Science for Peace and Security Series - D : Information and Communication Security 19, p. 295–310.

Stevens, G., B. McCluskey, A. King, E. O'Hearn et G. Mayr (2015). « Review of the 2012 epizootic hemorrhagic disease outbreak in domestic ruminants in the United States ». In : *PLoS ONE* 10.8, e0133359.

Strötgen, J. et M. Gertz (2010). « HeidelTime : High quality rule-based extraction and normalization of temporal expressions ». In : *Proceedings of the 5th International Workshop on Semantic Evaluation* July, p. 321–324.

Strötgen, J. et M. Gertz (2015). « A Baseline Temporal Tagger for all Languages ». In : *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*. September, p. 541–547.

Tang, J., M. Hong, D. Zhang, B. Liang et J. Li (2008). « Information Extraction : Methodologies and applications ». In : *Emerging Technologies of Text Mining : Techniques and Applications*, p. 1–33.

Tisseuil, C., A. Gryspeirt, R. Lancelot, M. Pioz, A. Liebhold et M. Gilbert (2016). « Evaluating methods to quantify spatial variation in the velocity of biological invasions ». In : *Ecography* 39.5, p. 409–418.

Tolle, K. M. et H. Chen (2000). « Comparing Noun Phrasing Techniques for Use with Medical Digital Library Tools CancerLit Concept Space - MeSH & NPIhdex a a 12 a 13 a 14 a 25 a 26 a ». In : *Journal of the American Society for Information Science* 20.X, p. 352–370.

Torii, M., L. Yin, T. Nguyen, C. T. Mazumdar, H. Liu, D. M. Hartley et N. P. Nelson (2011). « An exploratory study of a text classification framework for Internet-based surveillance of emerging epidemics ». In : *International Journal of Medical Informatics* 80.1, p. 56–66.

Touhami, R., P. Buche, J. Dibie et L. Ibanescu (2015). « Ontology evolution for experimental data in food ». In : *Communications in Computer and Information Science*. Sous la dir. d'E. Garoufallou, R. J. Hartley et P. Gaitanou. T. 544. Communications in Computer and Information Science. Springer International Publishing, p. 393–404.

Tsai, F.-J., E. Tseng, C.-C. Chan, H. Tamashiro, S. Motamed et A. C. Rougemont (2013). « Is the reporting timeliness gap for avian flu and H1N1 outbreaks in global health surveillance systems associated with country transparency ? » In : *Globalization and health* 9, p. 14.

Tuarob, S., C. S. Tucker, M. Salathe et N. Ram (2014). « An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages ». In : *Journal of Biomedical Informatics* 49, p. 255–268.

Turney, P. D. (2001). « Mining the Web for synonyms : PMI-IR versus LSA on TOEFL ». In : *Proceedings of the 12th European Conference on Machine Learning (ECML-2001), Freiburg,*

Germany. Lecture Notes in Computer Science 2167. Sous la dir. de L. Raedt et P. Flach, p. 491–502.

Uzaman, N. et J. F. Allen (2010). « TRIPS and TRIOS System for TempEval-2 : Extracting Temporal Information from Text ». In : *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval '10)*. Los Angeles, California : Association for Computational Linguistics, p. 276–283.

Vallat, B., A. Thiermann, K. B. Jebara et A. Dehove (2013). « Notification of animal and human diseases : the global legal basis OIE notification system ». In : *Revue scientifique et technique* 32.2, p. 331–335.

Velardi, P., G. Stilo, A. E. Tozzi et F. Gesualdo (2014). « Twitter mining for fine-grained syndromic surveillance ». In : *Artificial Intelligence in Medicine* 61.3, p. 153–163.

Velasco, E., T. Agheneza, K. Denecke, G. Kirchner et T. Eckmanns (2014). *Social media and internet-based data in global systems for public health surveillance : A systematic review*.

Viera, A. J. et J. M. Garrett (2005). « Understanding interobserver agreement : The kappa statistic ». In : *Family Medicine* 37.5, p. 360–363.

Vinot, R., N. Grabar et M. Valette (2003). « Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'Internet ». In : *TALN*, p. 275–284.

Vivaldi, J., L. Màrquez et H. Rodríguez (2001). *Improving Term Extraction by System Combination using Boosting*. Sous la dir. de L. De Raedt et P. Flach. T. 2167. Lecture Notes in Computer Science. Springer Berlin Heidelberg, p. 515–526. 640 p.

Volkova, S., D. Caragea et W. H. Hsu… (2010). « Animal disease event recognition and classification ». In : *Proceedings of the Second International Workshop on Web Science and Information Exchange in the Medical Web (MedEx 2010)*. New York, NY, USA : ACM, p. 51–61.

Waidyanatha, N., A. Dubrawski, G. M. et G. Gow (2011). « Affordable System for Rapid Detection and Mitigation of Emerging Diseases ». In : *International Journal of E-Health and Medical Communications* 2.1, p. 73–90.

Wei, C. P. et Y. H. Lee (2004). « Event detection from online news documents for supporting environmental scanning ». In : *Decision Support Systems* 36.4, p. 385–401.

Wernike, K., B. Hoffmann et M. Beer (2013). « Schmallenberg virus ». In : *Developments in Biologicals* 135, p. 175–182.

Wilson, A. J. et P. S. Mellor (2009). « Bluetongue in Europe : past, present and future. » In : *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 364.1530, p. 2669–2681.

Witten, I. H. et E. Frank (2005). *Data Mining : Practical Machine Learning Tools and Techniques*. 2 edition. T. 2 edition. Amsterdam ; Boston, MA : Morgan Kaufmann, p. 560.

Wood, S. (2007). « Opening data to the world : Why health numbers matter ». In : *Bulletin of the World Health Organization* 85.10, p. 736.

Wood, S. (2003). « Thin plate regression splines ». In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 65.1, p. 95–114.

Woodall, J. P. (2001). « Global surveillance of emerging diseases : the ProMED-mail perspective. » In : *Cadernos de saude publica* 17, p. 147–154.

Yan, L., R. Dodier, M. C. Mozer et R. Wolniewicz (2003). « Optimizing Classifier Performance via an Approximation to the Wilcoxon-Mann-Whitney Statistic ». In : *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*. Sous la dir. de T. Fawcett et N. Mishra. T. 20. 2. Washington, DC : AAAI Press, Menlo Park, California, p. 848.

Ye, Y., F. R. Tsui, M. Wagner, J. U. Espino et Q. Li (2014). « Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers ». In : *Journal of the American Medical Informatics Association* 21.5, p. 815–823.

Yu, H. et D. Kaufman (2006). *A cognitive evaluation of four online search engines for answering definitional questions posed by physicians.* T. 12. World Scientific, p. 328–339. 524 p.

Zeldenrust, M. E., J. C. Rahamat-Langendoen, M. J. Postma et J. A. van Vliet (2008). « The value of ProMED-mail for the Early Warning Committee in the Netherlands : more specific approach recommended. » In : *Euro surveillance* 13.6.

Zeller, H., L. Marrama, B. Sudre, W. Van Bortel et E. Warns-Petit (2013). « Mosquito-borne disease surveillance by the European Centre for Disease Prevention and Control. » In : *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 19.8, p. 693–8.

Zesch, T. et I. Gurevych (2010). « The More the Better ? Assessing the Influence of Wikipedia ' s Growth on Semantic Relatedness Measures ». In : *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Sous la dir. de N. C. et al. Valletta, Malta : European Language Resources Association (ELRA), p. 1374–1380.

Zhang, Y. et B. Liu (2007). *Semantic Text Classification of Emergent Disease Reports*. PKDD 2007. Berlin, Heidelberg : Springer-Verlag, p. 629–637. XXIV p.

Zhang, Y., Y. Dang, H. Chen, M. Thurmond et C. Larson (2009). « Automatic online news monitoring and classification for syndromic surveillance ». In : *Decision Support Systems* 47.4, p. 508–517.

Zou, K. H., A. J. O'Malley et L. Mauri (2007). « Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models ». In : *Circulation* 115.5, p. 654–657.

Zuccon, G., S. Khanna, A. Nguyen, J. Boyle, M. Hamlet et M. Cameron (2015). « Automatic detection of tweets reporting cases of influenza like illnesses in Australia. » In : *Health information science and systems* 3.Suppl 1 HISA Big Data in Biomedicine and Healthcare 2013 Con, S4.

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique

Titre : Élaboration d'une méthode semi-automatique pour l'identification et le traitement des signaux d'émergence pour la veille internationale sur les maladies animales infectieuses

Mots clefs : intelligence épidemiologique, fouille de textes, données textuelles non-structurées, santé animale, Web

Résumé : La veille en santé animale, notamment la détection précoce de l'émergence d'agents pathogènes exotiques et émergents à l'échelle mondiale, est l'un des moyens de lutte contre l'introduction de ces agents pathogènes en France.
Récemment, il y a eu une réelle prise de conscience par les autorités sanitaires de l'utilité de l'information non-structurée concernant les maladies infectieuses publiée sur le Web.
C'est dans ce contexte que nous proposons une approche de veille basé sur une méthode de fouille de textes pour la détection, collecte, catégorisation et extraction de l'information sanitaire à partir des donnés textuelles non structurées (articles médias) publiées sur le Web.
Notre méthode est générique. Toutefois, pour l'élaborer, nous l'appliquons à cinq maladies animales infectieuses exotiques : la peste porcine africaine, la fièvre aphteuse, la fièvre catarrhale ovine, la maladie du virus Schmallenberg et l'influenza aviaire. Nous démontrons que des techniques de fouille de textes, complétées par les connaissances d'experts du domaine, sont la fondation d'une veille sanitaire du Web à la fois efficace et réactive pour détecter des émergences de maladies exotiques au niveau international.
Notre outil (PADI-web) peut servir le dispositif de veille sanitaire internationale en France. Il facilitera la détection précoce de signaux de dangers sanitaires émergents dans l'information publiée sur le Web.

Title : Elaboration of a semi-automatic method for identification and analysis of signals of emergence of animal infectious diseases at international level

Keywords : epidemic intelligence, text mining, non-structured text data, animal health, Web

Abstract : Monitoring animal health worldwide, especially the early detection of outbreaks of emerging and exotic pathogens, is one of the means of preventing the introduction of infectious diseases in France.
Recently, there is an increasing awareness among health authorities for the use of unstructured information published on the Web for epidemic intelligence purposes.
In this manuscript we present a text mining approach, which detects, collects, classifies and extracts information from non-structured textual data available in the media reports on the Web. Our approach is generic; however, it was elaborated using five exotic animal infectious diseases: african swine fever, foot-and-mouth disease, bluetongue, Schmallenberg and avian influenza.
We show that the text mining techniques, supplemented by the knowledge of domain experts, are the foundation of an efficient and reactive system for monitoring disease emergence from non-structured text published on the Web.
Our tool (PADI-web) can serve the French epidemic intelligence team for international monitoring of animal health. It can facilitate the early detection of events related to emerging health hazards identified from media reports on the Web.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France