AfIA

Association française
pour l'Intelligence Artificielle

IC

Journées francophones d'Ingénierie des Connaissances

PFIA 2024

EpiStrat-Eval: outil d'évaluation des stratégies d'extraction d'informations spatiales pour la veille en épidémiologie

Y. Mahdoubi[1], S. Valentin[2,3], N. Idrissi[1], M. Roche[2,3]

[1]Faculté des Sciences et Techniques Béni Mellal, Maroc
[2]TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
[3]CIRAD, UMR TETIS, Montpellier F-34398, France

Résumé

Les articles de presse publiés en ligne jouent un rôle essentiel dans la détection précoce des foyers de maladies. Toutefois, l'extraction et le traitement de l'information spatiale à partir de ces données textuelles demeurent un défi majeur. C'est dans cette optique que l'application EpiStrat-Eval, en lien avec l'outil de surveillance PADI-web, offre une interface cartographique permettant aux épidémiologistes d'évaluer diverses stratégies d'extraction d'entités spatiales. Cette initiative contribue ainsi à l'amélioration continue des systèmes de surveillance.

Mots-clés

données textuelles, entités spatiales, évaluation

Abstract

Online news articles play a vital role in the early detection of disease outbreaks. However, extracting and processing spatial information from this textual data remains a significant challenge. To address this, EpiStrat-Eval, in collaboration with the PADI-web monitoring tool, offers a mapping interface that enables epidemiologists to evaluate various strategies for extracting spatial entities. This initiative contributes to the continuous improvement of monitoring systems.

Keywords

textual data, spatial information, evaluation

1 Introduction

1.1. Surveillance basée sur les événements et information spatiale

Ces dernières années, de nombreux travaux se sont intéressés à l'utilisation de données textuelles issues de sources informelles pour la surveillance des épidémies. Alors que les organisations telles que l'Organisation mondiale de la santé animale diffusent des notifications officielles, les sources informelles telles que les journaux en ligne fournissent des informations de niveaux de fiabilité variables pouvant s'avérer plus précoces [1].

Les outils de surveillance événementielle, tels que HealthMap [2] et PADI-web [3], permettent d'automatiser différentes étapes de la chaîne de traitement des données textuelles en y intégrant des approches basées sur l'Intelligence Artificielle et l'apprentissage automatique, de la collecte de sources jusqu'à la production d'informations épidémiologiques pouvant être analysées. L'identification précise des localisations joue un rôle important dans l'évaluation des risques épidémiologiques, permettant aux épidémiologistes de détecter une possible émergence ou d'appréhender la propagation d'une maladie. Cette spatialisation est indispensable, *in fine*, à la mise en place de mesures de lutte et de protection adaptées. Cependant, cette tâche se heurte à des défis méthodologiques propres à l'extraction automatique d'information spatiale à partir de données textuelles dans un contexte événementiel. D'une part, la détection des entités spatiales peut produire des erreurs en raison d'ambiguïtés fréquentes, telles que l'utilisation de noms de lieux génériques ou d'abréviations, la confusion entre des noms de personne et de lieux, etc. D'autre part, les articles de presse peuvent contenir des informations spatiales qui ne sont pas associées aux foyers de maladie. Si ces informations connexes sont trop nombreuses, le processus d'extraction génère une importante quantité d'information non-pertinente, augmentant la tâche de filtre par l'expert et posant le risque de générer de fausses alertes. Enfin, la tâche de *geocoding*, indispensable à la représentation cartographique des entités, ajoute une couche supplémentaire de complexité à ce processus d'identification en raison des nombreuses homonymies et ambiguïtés [4].

1.2. Motivation

Face à ces verrous, différentes stratégies d'extraction de l'information spatiale combinant heuristiques simples et approches fondées sur l'Intelligence Artificielle peuvent être intégrées dans les outils de veille. Notre objectif consiste à proposer une application permettant d'évaluer la performance de ces différentes stratégies. Nous nous adossons aux travaux dédiés à l'outil PADI-web, un outil automatique pour la veille en épidémiologie animale basé sur les articles publiés en ligne [3]. Nous proposons EpiStrat-Eval[1], une interface conviviale pour aider les épidémiologistes dans leur processus décisionnel lors de la validation de la localisation des foyers de maladie[2].

[1] https://github.com/ysfmh14/EpiStratEval.git
[2] Vidéo de démonstration

1.3. Travaux connexes: interfaces et évaluation

L'avènement des technologies numériques a révolutionné le domaine de la surveillance épidémiologique, en offrant de nouveaux outils pour collecter, analyser et visualiser les données de santé. De nombreuses plateformes interactives, principalement appliquées en santé humaine, permettent de visualiser et d'interagir avec des données spatio-temporelles, afin de faciliter la détection d'épidémies à partir de données officielles [6, 7]. Cependant, ces approches s'appuyant sur des bases de données officielles, la notion de validation spatiale n'y est pas intégrée. Dans le cadre de la surveillance des sources informelles, HealthMap [2] propose une carte interactive intégrant les diverses sources d'information. Chaque article est associé à une localisation (ville, région ou pays) et est représenté sur la carte via un cercle dont la taille et la couleur représentant le niveau d'alerte. Les utilisateurs enregistrés peuvent évaluer le niveau de risque de chaque alerte grâce à un score de 1 à 5. EpidNews [8] propose une interface utilisateur intuitive permettant d'explorer les données épidémiologiques des maladies animales issues à la fois des sources officielles et non officielles. L'objectif de l'outil est de mettre en évidence les alertes issues de sources non-officielles dans des régions non couvertes par les données officielles.

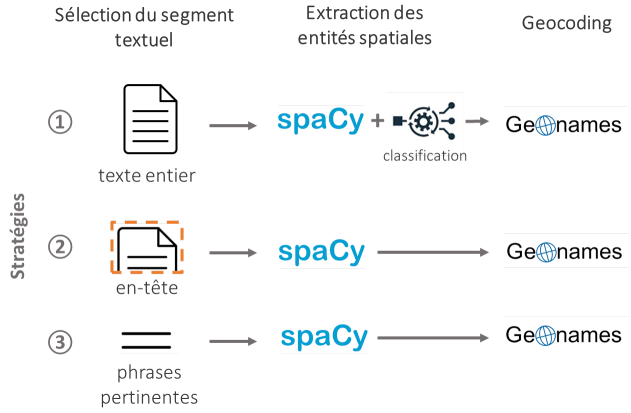
2 Données spatiales issues d'un système de veille

Dans cette section, nous présentons les différentes stratégies de spatialisation des foyers de maladie par PADI-web, et les étapes de leur préparation pour l'outil EpiStrat-Eval.

2.1. Stratégies d'extraction de PADI-web

L'interface de PADI-web permet d'extraire les informations spatiales relatives aux épidémies à partir d'articles publiés en utilisant six stratégies distinctes. Nous présentons dans la Figure 1 trois de ces stratégies, qui sont appliquées sur les articles automatiquement classés comme "Déclaration de foyer".

Figure 1. Stratégies d'extraction d'entités spatiales dans l'outil PADI-web.



La première étape consiste à préfiltrer le segment textuel à partir duquel vont être extraites les localisations : le texte entier (stratégie 1), le titre et les 300 premiers caractères (stratégie 2), les phrases classées comme pertinentes, c'est–à-dire relatives à un foyer récemment déclaré, grâce à une approche fondée sur l'apprentissage automatique détaillée dans [9] (stratégie 3). L'extraction des entités spatiales est ensuite réalisée en utilisant un modèle pré-entraîné de la librairie spaCy [10]. Dans la stratégie 1, une étape de sélection automatique des localisations pertinentes (associées à un foyer) via un module de classification est appliquée [5]. Le *geocoding*, qui permet de d'affecter des coordonnées géographiques à chaque entité spatiale, est réalisé grâce à la base de données géographiques GeoNames [11].

2.2. Préparation et nettoyage des données

Avant d'intégrer les données fournies par PADI-web dans notre application, nous effectuons plusieurs traitements pour obtenir un ensemble de données bien organisé, exempt de données erronées ou manquantes. Voici les traitements effectués :

(1) Suppression des lignes comportant des valeurs de latitude et de longitude nulles (ne pouvant pas être représentées sur la carte).
(2) Pour un même article, suppression des lignes représentant la même entité géographique, afin de réduire les redondances sur la carte.
(3) Ajout de la colonne *validation* afin d'enregistrer les validations des entités spatiales par les experts.

Nous intégrons également la colonne *location type* afin d'indiquer la hiérarchie de la localisation (i.e. continent, pays, région, ville), en utilisant le code et la classe géographique associés à l'identifiant GeoNames de chaque entité.

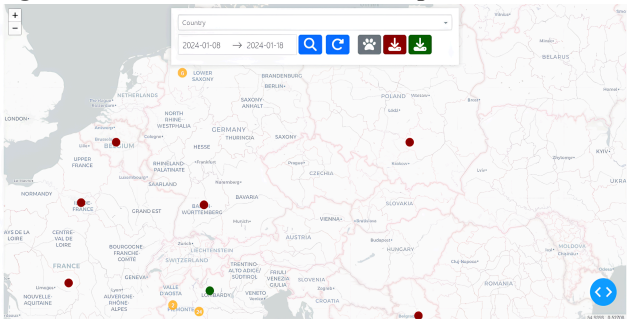
3 EpiStrat-Eval

Notre application permet de rendre les données extraites par PADI-web visuellement accessibles, afin d'accompagner la prise de décision des épidémiologistes. Cependant, l'objectif principal de notre application est de fournir aux experts la possibilité de comparer les différentes stratégies d'extraction d'informations spatiales, pour, *in fine*, identifier la ou les plus pertinentes.

3.1 Description générale de l'interface

L'interface principale de notre application, illustrée dans la Figure 2, se divise en deux composants majeurs. Le premier est une carte affichant les positions correspondant aux détections des maladies, tandis que la seconde est une barre de navigation contenant plusieurs composants. Parmi ces éléments, l'outil propose de filtrer les données en sélectionnant un pays et/ou en définissant une fenêtre temporelle. Cette barre offre également la possibilité d'afficher la liste des hôtes (espèces animales) qui sont extraits (bouton gris), ainsi que la possibilité de rafraîchir les données en cas de modification, ou encore de charger un fichier contenant des données déjà évaluées.

Figure 2. Visualisation de l'interface EpiStrat-Eval



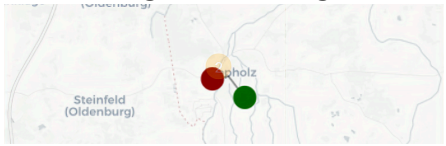
3.2 Représentation de l'information spatiale

Pour représenter les entités spatiales nous avons opté pour une représentation cartographique, dans laquelle les données sont symbolisées par des cercles de deux couleurs distinctes : les cercles rouges signalent les localisations de foyers extraites par la première stratégie, tandis que les cercles verts représentent les extractions de la deuxième stratégie. De plus, des cercles oranges sont utilisés pour condenser les informations lorsqu'un grand nombre de détections se situent dans une même zone. Cette zone, matérialisée en bleu, est établie en spécifiant une distance maximale entre les détections : toutes celles qui se trouvent à une distance égale ou inférieure les unes des autres sont regroupées dans une même région (Figure 3). Cette distance est un paramètre pouvant être adapté au contexte épidémiologique. Les cercles oranges offrent une représentation abstraite de la région en indiquant le nombre de détections présentes. Le détail des détections dans une région identifiée est affiché en cliquant sur la zone.

Figure 3. Représentation de détections multiples

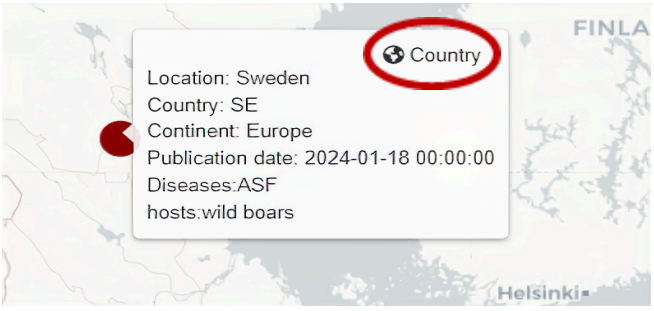
Les cercles orange mentionnés dans le paragraphe précédent sont également utilisés pour résoudre les problèmes de superposition, lorsqu'un foyer est détecté plusieurs fois dans la même localisation, soit par une seule stratégie, soit par les deux. Cela entraîne le chevauchement des cercles qui rend difficile la distinction des détections dans cette localisation. Les cercles orange permettent de clarifier cette situation en fournissant une représentation plus claire des détections multiples dans une même zone (Figure 4).

Figure 4. Visualisation d'une localisation identifiée par la stratégie 1 et la stratégie 2



Nous permettons également aux experts d'accéder à diverses informations concernant chaque détection en plaçant le curseur sur le cercle correspondant à la détection (Figure 5). Ces informations comprennent la valeur de la localisation dans le texte, le pays correspondant, la date de publication de l'article, et le nom de la maladie et le ou les hôtes extraits à partir de l'article. Le type de localisation (ville, région, pays ou continent) est également indiqué et permet de contextualiser la position d'une détection : si une région administrative est extraite, ses coordonnées géographiques associées vont être son centroïde.

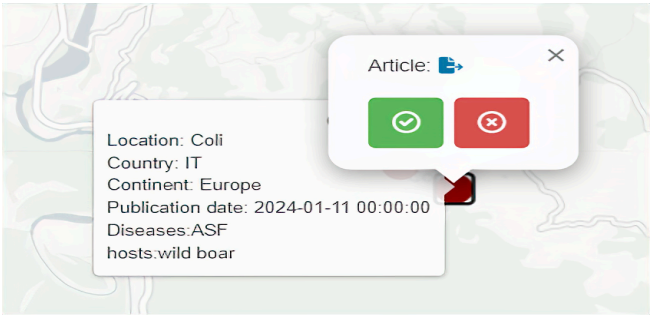
Figure 5. Visualisation des métadonnées associées à une détection



3.3 Evaluation

Nous allons à présent détailler l'aspect central de notre outil, permettant aux experts de valider ou non les localisations identifiées par chaque stratégie.

Figure 6. Validation de la pertinence d'une information spatiale



Pour pouvoir interpréter la pertinence d'une localisation, EpiStrat-Eval propose une redirection vers l'article dont l'entité spatiale a été extraite. L'expert peut ensuite valider ou non la pertinence de l'entité spatiale (Figure 6). Une fois ce choix effectué, un ticket est associé aux détections qui ont reçu une évaluation et le label (valide ou non valide) est ajouté à une colonne appelée "Validation". Les fichiers générés peuvent être téléchargés et ré-utilisés comme fichiers d'entrée pour reprendre une évaluation en cours ou modifier des valeurs existantes.

3.4. Outils

Les fonctionnalités d'EpiStrat-Eval reposent sur plusieurs outils. La bibliothèque Python *Dash* a été utilisée pour

concevoir des tableaux de bord réactifs. Cette approche permet aux composants de partager des informations entre eux grâce à l'utilisation de "callbacks" et permet d'intégrer des éléments interactifs. Nous avons également utilisé la bibliothèque Python *Folium* pour créer une carte interactive. Enfin, la bibliothèque Python *Flask* a été intégrée pour consommer des API, émises depuis un code JavaScript associé à la carte.

4 Cas d'étude

Nous avons réalisé un cas d'étude sur un échantillon d'articles extraits de l'outil PADI-web. L'évaluation a été réalisée par une épidémiologiste en santé animale. Nous avons sélectionné les articles de PADI-web classés comme des déclarations de foyers de maladies animales publiés entre les 22/02/2024 et le 26/02/2024 (31 articles). Nous avons comparé la stratégie 1 (extraction à partir de l'ensemble de l'article puis sélection automatique) avec la stratégie 2 (extraction limitée à l'en-tête de l'article, sans sélection automatique). Nous avons choisi de comparer ces deux stratégies car elles répondent à deux niveaux de complexité méthodologique : la stratégie 2 repose sur l'hypothèse que les entités spatiales associées à un événement sont présentes en début d'article, tandis que la 1ère stratégie évalue la pertinence de chaque entité spatiale quelle que soit leur position dans le texte. Avant filtrage, les deux stratégies génèrent 157 et 110 entités spatiales, respectivement. Après filtrage (section 2.2), 93 entités spatiales sont obtenues via la stratégie 1 *versus* 37 entités via la stratégie 2.

Pour être labellisées comme valides, les localisations doivent répondre à deux critères : (1) correspondre à une entité spatiale citée dans l'article et (2) correspondre à un foyer de maladie. L'outil ne permet pas d'évaluer l'étape de geoparsing (attribution de coordonnées spatiales), qui est une étape indépendante de l'identification de localisations pertinentes. Dans ce cas d'étude, nous avons donc considéré comme valides les localisations répondant aux deux critères précédents, même associées à des coordonnées spatiales incorrectes.

Tableau 1. Evaluation des stratégies d'extraction 1 et 2

		Stratégie 1	**Stratégie 2**
Articles	Localisés	93.5% (29/31)	61% (22/31)
	Non localisés	6.5% (2/31)	29% (9/31)
Localisations	Valides	75.3% (70/93)	94.6% (35/37)
	Non valides	24.7% (23/93)	5.4% (2/37)

La stratégie plus conservatrice (stratégie 2) permet d'obtenir une excellente précision (94.6% des entités extraites correspondent à des foyers mentionnés dans un article). Cependant, 29% des articles ne sont pas associés à une localisation (dans ces articles, aucune entité spatiale n'a été détectée par la librairie spaCy dans l'en-tête de l'article). La stratégie 1 génère davantage de localisations non valides, mais permet détecter un plus grand nombre absolu d'entités spatiales valides, et notamment de spatialiser la quasi-totalité des articles (93.5%). Les deux articles non spatialisés contiennent des entités spatiales bien détectées par GeoNames, mais qui n'ont pas été considérées comme liées à un foyer suite à la classification automatique. Dans ce cas d'étude, la stratégie 1 est à privilégier afin d'éviter un trop grand nombre de faux négatifs (localisations pertinentes associées à des foyers non détectés par la stratégie 1).

L'extension de ce cas d'étude à l'ensemble des stratégies d'extraction de PADI-web et en combinant l'évaluation de plusieurs experts, permettra d'évaluer de manière plus représentative les informations spatiales issues d'approches automatiques. Cette évaluation est indispensable à l'intégration de solutions fondées sur l'Intelligence Artificielle dans le cadre d'activités quotidiennes de veille épidémiologique.

5 Références

.[1] C. Paquet, D. Coulombier, R. Kaiser, and M. Ciotti. Epidemic intelligence : a new framework for strengthening disease surveillance in Europe. Eurosurveillance, 11(12) :5–6, December 2006.

[2] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein. HealthMap : Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. Journal of the American Medical Informatics Association, 15(2) :150–157, March 2008.

[3] S. Valentin, E. Arsevska, J. Rabatel, S. Falala, A. Mercier, R. Lancelot, and M. Roche. PADI-web 3.0 : A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance. One Health, 2021.

[4] M. Gritta, M. T. Pilehvar, and N. Collier. A pragmatic guide to geoparsing evaluation. Language Resources and Evaluation, 54(3) :683–712, September 2020.

[5] E. Arsevska, S. Valentin, J. Rabatel, J. de Goër de Hervé, S. Falala, R. Lancelot, and M. Roche. Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System. PLOS ONE, 13(8) :e0199960, August 2018.

[6] R. Arias-Carrasco, J. Giddaluru, L. E. Cardozo, F. Martins, V. Maracaja-Coutinho, and H. I. Nakaya. OUTBREAK : a user-friendly georeferencing online tool for disease surveillance. Biological Research, 54, 2021. Publisher : BMC.

[7] L. N. Carroll, A. P. Au, L. T. Detwiler, T.-c. Fu, I. S. Painter, and N. F. Abernethy. Visualization and analytics tools for infectious disease epidemiology : A systematic review. Journal of Biomedical Informatics, 51 :287–298, October 2014.

[8] R. Goel, S. Valentin, A. Delaforge, S. Fadloun, A. Sallaberry, M. Roche, and P. Poncelet. EpidNews : Extracting, exploring and annotating news for monitoring animal diseases. Journal of Computer Languages, 56 :100936, February 2020.

[9] S. Valentin, E. Arsevska, A. Vilain, V. De Waele, R. Lancelot, and M. Roche. Elaboration of a new framework for fine-grained epidemiological annotation. Scientific Data, 9(1) :655, October 2022.

[10] spaCy · Industrial-strength Natural Language Processing in Python.

[11] D. Ahlers. Assessment of the Accuracy of GeoNames Gazetteer Data. In Proceedings of the 7th Workshop on Geographic Information Retrieval, pages 74–81, New York, NY, USA, 2013. ACM.