PADI-web 3.0: A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance

Sarah Valentin, Elena Arsevska, Julien Rabatel, Sylvain Falala, Alizé Mercier, Renaud Lancelot, Mathieu Roche

One Health 13 (2021) 100357

Contents lists available at ScienceDirect

One Health

journal homepage: www.elsevier.com/locate/onehlt

PADI-web 3.0: A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance

Sarah Valentin [a,b,c], Elena Arsevska [a,b], Julien Rabatel [a], Sylvain Falala [a,b], Alizé Mercier [a,b], Renaud Lancelot [a,b], Mathieu Roche [a,c,*]

[a] CIRAD, UMR ASTRE / UMR TETIS, F-34398 Montpellier, France
[b] ASTRE, Univ Montpellier, CIRAD, INRAE, Montpellier, France
[c] TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France

ARTICLE INFO

Keywords:
Animal disease surveillance
Software
Text mining

ABSTRACT

PADI-web (Platform for Automated extraction of animal Disease Information from the web) is a biosurveillance system dedicated to monitoring online news sources for the detection of emerging animal infectious diseases. PADI-web has collected more than 380,000 news articles since 2016. Compared to other existing biosurveillance tools, PADI-web focuses specifically on animal health and has a fully automated pipeline based on machine-learning methods. This paper presents the new functionalities of PADI-web based on the integration of: (i) a new fine-grained classification system, (ii) automatic methods to extract terms and named entities with text-mining approaches, (iii) semantic resources for indexing keywords and (iv) a notification system for end-users. Compared to other biosurveillance tools, PADI-web, which is integrated in the French Platform for Animal Health Surveillance (ESA Platform), offers strong coverage of the animal sector, a multilingual approach, an automated information extraction module and a notification tool configurable according to end-user needs.

1. Introduction

Over the past decades, the number of outbreaks due to (re)emerging animal and human infectious diseases has been increasing in many parts of the world. In addition to the well-known role of human and animal mobility in the spread of pathogens, climate change and biodiversity loss are likely to exacerbate the global burden of these diseases [1,2]. National and international institutions are currently experiencing a global paradox, conciliating trade extension with the control of the risk to human and animal health.

In this context, the Epidemic Intelligence Service of the Center for Disease Control (CDC) is considered to be the earliest public health system dedicated to early warning [3]. It was created to enhance the surveillance and eradication of both infectious and noninfectious human diseases, such as poliomyelitis and leukemia, and also monitors for bioterrorism. The epidemic intelligence (EI) concept, as it is used today, was developed in the early 2000s. The French *Institut de Veille Sanitaire* (*Institute of Health Surveillance*) and the European Centre for Disease Prevention and Control (ECDC) proposed an EI framework to enhance disease surveillance in Europe in 2006 [4,5]. Eight years later, the World Health Organisation (WHO) published a comprehensive guide providing key definitions and detailing the implementation of early warning activities [6]. EI corresponds to a formalized surveillance process that encompasses '*all activities related to the early identification of potential health hazards that may represent a risk to health, and their verification, assessment and investigation*' [6]. It relies on two main channels of information: *indicator-based surveillance* (IBS) and *event-based surveillance* (EBS). Indicator-based surveillance is defined as '*the systematic collection, monitoring, analysis and interpretation of structured data* (i.e. *indicators)*' [6]. It corresponds to conventional surveillance of formal sources and is based on established case definitions. Event-based surveillance is defined by the WHO as '*the organized collection, monitoring, assessment and interpretation of mainly unstructured* ad hoc *information regarding health events or risks, which may represent an acute risk to human [or animal] health*' [6]. The definitions and concepts from both ECDC and WHO were elaborated for public health. However, they have been successfully transferred to other domains, such as plant health [7] and both terrestrial and aquatic animal health [8–10]. Both EBS and IBS can be formally represented as consecutive steps, corresponding to the flow of epidemiological information from its detection to its communication to the

* Corresponding author at: CIRAD, UMR TETIS, F-34398 Montpellier, France.
E-mail address: mathieu.roche@cirad.fr (M. Roche).

https://doi.org/10.1016/j.onehlt.2021.100357
Received 7 July 2021; Received in revised form 30 November 2021; Accepted 1 December 2021
Available online 3 December 2021

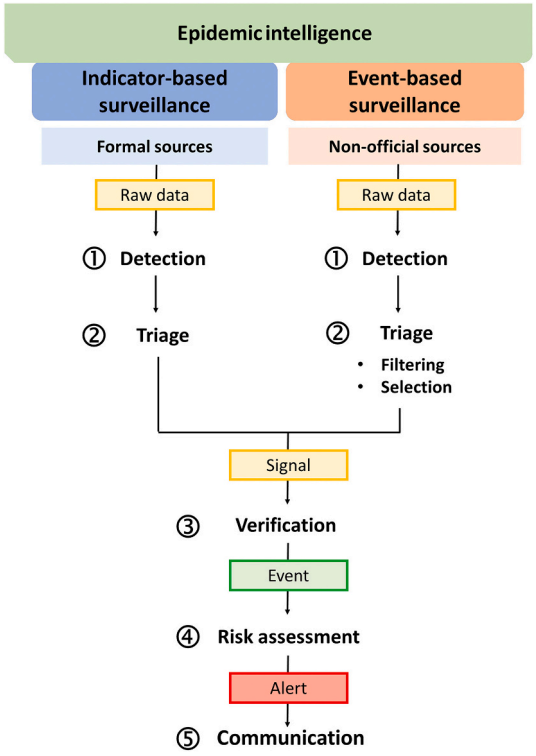


Fig. 1. Epidemic intelligence workflow (adapted from [6]).

relevant authorities [4,11,12] (see Fig. 1).

The first stages of the EI process consist of identifying and extracting relevant information from heterogeneous data. This process is based on 4 steps: (1) data detection, (2) data triage, (3) signal verification and (4) risk assessment and alert communication.

1. Data detection consists of defining modalities (e.g., format) through which raw data are detected and collected. The strategy differs according to the type of source. IBS usually relies on well-established notification procedures submitted by a country for a given disease according to international health regulations. EBS systems use specific queries implemented through RSS (Really Simple Syndication) feeds to detect news on an outbreak of a disease.
2. Data triage is an important step to avoid overwhelming the EI system with irrelevant data. Raw data is filtered (i.e. triaged) based on given selection criteria. In IBS, selection criteria could vary according to the institutional mandates, such as (i) its geographical coverage, e.g., worldwide (with the World Organisation for Animal Health (OIE)), regional (with ECDC), and (ii) its thematic mandate, e.g., animal health or public health. In EBS, data triage is based on (i) data filtering to remove duplicates and irrelevant data and (ii) data selection for sorting information according to the EBS system priority criteria. Data retained as relevant regarding early warning activities are referred to as 'signals'.
3. Signal verification aims to validate the truthfulness of a signal. This step is crucial in the EBS workflow since the data sources are informal. Once validated, a signal can become an 'event', i.e. a manifestation of the threat in a given affected host or population and at a given location and date [13]. This definition encompasses events from all possible known origins, such as infectious, zoonotic, food safety, chemical, radiological or nuclear, as well as pathogens of unknown origin in the EBS context.
4. The final stage deals with risk assessment and alert communication. Risk assessment aims at determining the level of risk related to a detected event. The event becomes an alert if the risk is considered significant to health. Alerts are communicated through channels adapted to relevant authorities (e.g., public health national networks, ministries of health, international organizations) or a larger network (e.g., end-users of EBS systems).

This paper presents new functionalities of the EBS system, PADI-web (Platform for Automated extraction of animal Disease Information from the web - https://padi-web.cirad.fr/) dedicated to the monitoring of online news sources for the detection of emerging/new animal infectious diseases. PADI-web has collected more than 380,000 news items since 2016. The first descriptions of PADI-web have been published elsewhere [8,14]. This paper presents new functionalities of PADI-web based on the integration of: (i) a new fine-grained classification system, (ii) automatic methods to extract terms and named entities with text-mining approaches, (iii) semantic resources for indexing keywords and (iv) an automatic notification system for end-users that used these new methods.

PADI-web was developed to meet the needs of the French Epidemic Intelligence System (i.e. FEIS) via online news monitoring. FEIS has been involved in the activities of the French Platform for Animal Health Surveillance (ESA Platform) since 2013. FEIS aims to identify, monitor and analyze reports of animal health hazards (including zoonotic diseases) threatening animal populations in France by monitoring official and unofficial information sources. PADI-web has been integrated into FEIS activities by ad hoc use depending on the epidemiological news. PADI-web successfully identified signals for current outbreaks of diseases that are notifiable to OIE, such as avian influenza (AI), African swine fever (ASF), foot-and-mouth disease (FMD), and bluetongue (BTV) [15], and human diseases such as COVID-19 [16]. Moreover, PADI-web is able to detect the first signals of emerging infectious disease outbreaks in a timely manner, as illustrated by the detection of primary FMD outbreaks in East Asia in 2016 [8]. PADI-web also provided alerts of ASF, FMD and BTV emergence within previously unaffected areas, for which we could not find any official confirmation [8].

The paper is structured as follows: in Section 2, we discuss related work; in Sections 3 and 4, we describe the proposed extensions of PADI-web and the results obtained; and in Section 5, we conclude the work.

2. Related work

In the context of EBS implementations, two important tasks must be considered. First, EBS systems must identify relevant texts (e.g., news, documents, sentences) related to epidemiological issues. The related work associated with this issue is detailed in Section 2.1. Second, epidemiological information (e.g., locations, symptoms, etc.) related to an event have to be extracted in these relevant texts. This issue is presented in Section 2.2.

2.1. *Identification of relevant texts related to the epidemiological domain*

Most EBS systems (e.g., MediSys, HealthMap, GPHIN, Argus, AquaticHealth.net, PADI-web) [8]) involve binary classification, i.e., news articles identified as relevant or irrelevant. Interestingly, there is no formal definition of relevance. This is a significant limitation in comparing EBS performances. In addition, the lack of shared gold standards and annotated resources hampers knowledge and experience sharing. Most commonly, there are two types of classification methods: (i) PADI-web 1.0 (first version [8]), MediSys and AquaticHealth use a keyword-based approach, (ii) GPHIN, Argus and HealthMap rely on machine learning-based classifiers.

2.1.1. *Keyword-based classification*

PADI-web 1.0 [8] categorized collected news articles by using a list of 32 outbreak-related keywords. News articles are classified as relevant if they contain in the text (title and body) one of the keywords related to an outbreak event (e.g. 'outbreak' 'cases' 'spread') [17]. MediSys classification relies on a more sophisticated approach involving Boolean

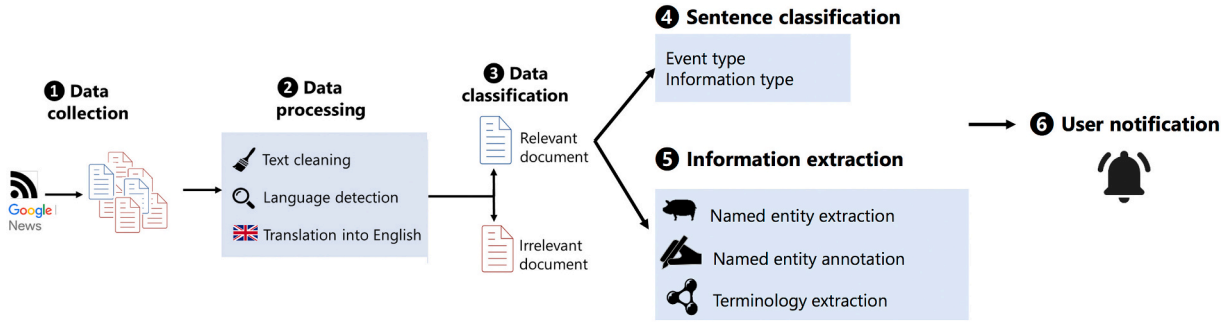


Fig. 2. PADI-web 3.0 pipeline.

combinations and keyword weightings. A document is considered relevant if it matches one of a predefined set of alerts [18,19]. Two types of signals (i.e. single and combination) are implemented. A single signal consists of attributing positive and negative weights to relevant and irrelevant keywords. An article is kept if the sum of the keyword weights it contains is above a given threshold. A combination signal is based on keywords combined by Boolean expressions (i.e. 'AND' and 'AND NOT'). Documents are selected if they contain at least two relevant keywords and do not include any irrelevant keywords. News articles of AquaticHealth.net are tagged by the users if they contain specific key terms usually the scientific names for diseases. This strategy is based on the assumption that the authors using correct scientific terminology are more likely to disseminate relevant information.

2.1.2. Machine learning-based classification

Other systems (i.e. HealthMap, BioCaster, Argus, GPHIN) rely on supervised machine learning classifiers, namely, three Bayesian algorithms and support vector machines (SVMs). The classifiers are trained on manually labeled data and automatically learn rules to label unclassified news articles. HealthMap uses a Bayesian machine learning algorithm. GPHIN automatically computes a relevance score for each retrieved report. This score corresponds to the confidence estimate of the SVM classifier [20]. Relevant articles in Argus were identified by keyword matching (with a set of concepts and keywords relevant to infectious disease surveillance) combined with Bayesian software tools. Experts further evaluate the automatic classification with GPHIN, HealthMap and Argus systems. In GPHIN, articles with a high relevance score are published immediately, while the system discards low-scoring reports automatically. Analysts triage the remaining medium-relevance reports. Analysts also review automatically discarded articles to verify that relevant information has not been erroneously filtered out by the automated system [20]. BioCaster classification is totally automated. A naive Bayes classifier was trained on a gold standard corpus. Each labeled article was manually assigned to the following classes: alert, publish, check and reject. However, a binary classification was implemented with the alert, publish and check classes being merged into a single category (i.e. relevant). In its current version, PADI-web integrates a supervised classifier [14]. Our paper proposes an original fine-grained classification based on sentence classification to highlight new epidemiological information, as described in Section 3.

Epidemiological information related to an event must be extracted into relevant texts (e.g., articles, sentences). This issue is discussed in the following subsection.

2.2. Epidemiological information extraction

Information extraction (IE) aims at locating specific pieces of information in textual data [21]. Entity extraction, also called named entity recognition (NER), is an IE subtask that seeks to locate and classify textual elements into predefined categories: (i) locations (e.g., 'Lagos','China'), (ii) temporal expressions (e.g., 'last month', 'July 28, 1990'), (iii) organizations (e.g., 'Ministry of Health'), (iv) person names, (v) quantities (e.g., '2'), and so on. This list of predefined categories can be extended to include domain-specific entities (thematic entities). In the animal health domain, we deal with (i) disease names (e.g., 'avian influenza', 'AI'), (ii) causal agents (e.g.,'H5N1 virus'), (iii) animal species (e.g., 'chicken'), (iv) symptoms (e.g., 'appetite loss'), and so on. In the context of online news, it is important to distinguish geographic entity extraction and resolution from identifying event-related locations. Geographic entity extraction and resolution aim at correctly extracting and identifying all locations from a text. Two types of approaches are used and combined to extract entities from texts: (1) *dictionary-based* approaches and (2) *classifier-based* approaches.

The *dictionary-based* approach involves matching terms from a document with a list of keywords. Some dictionaries can have an ontological structure rather than a simple list of terms. Ontologies aim at modeling the relations between entities [22]. In the health domain, an ontology can represent the causality relationships between a disease and a pathogen [23]. Geographical dictionaries are usually called gazetteers.

Dictionary and ontologies need regular updates to include new terms. This requires time-consuming manual work. Note that terms can be ambiguous, e.g., 'May' can refer either to a date, a location or a person's name. In the sentence 'The virus can be transmitted between pigs by their body fluids', the term 'body fluids' refers to the transmission route, yet in another context it may relate only to an anatomy concept.

In location extraction, this level of ambiguity is referred to as geo/non-geo ambiguity [24]. To overcome the rigidity of the dictionary-based approach, another approach consists of considering NER as a classification task, where the type of entity is the label to assign. Conditional random fields (CRFs) are among the most prominent classifiers used for NER [25] at the core of well-established pretrained NER tools, including StanfordNER [26] and NLTK [27]. This approach is designed for sequential data; CRFs predict the probability of an output sequence according to an input sequence [28]. The classification approach is particularly suitable for misspelt locations or short texts in terms of length, such as tweets, for which gazetteer lookups suffer from low precision due to irrelevant matches [29].

While *classifier-based* approaches achieve good results, they are limited to the given categories used for the training steps. Recent tools allow users to add new types of entities to NER algorithms by training the model on annotated datasets, such as the neural network-based NER algorithm from the SpaCy package [30]. Locations are also prone to another level of ambiguity that occurs when several distinct places have the same name, i.e. the *geo/geo ambiguity*, or *referent ambiguity*. Several methods are described in the literature to address these kinds of spatial ambiguities [24,31,32].

3. Material and methods

The PADI-web pipeline involves six steps ranging from online news collection to the extraction of epidemiological features: (1) data

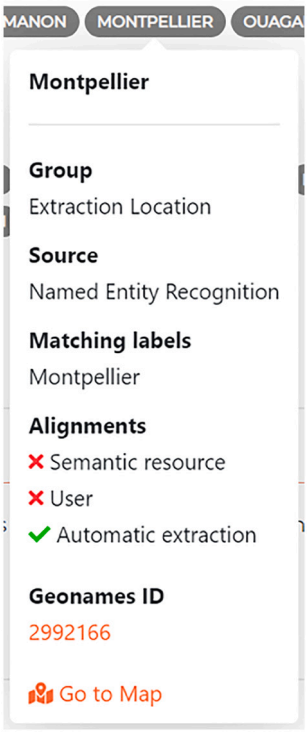


Fig. 3. Spatial Entity 'Montpellier' recognized by the automatic extraction (i.e. *location* with SpaCy) and associated with the Geonames ID used for the geo-tagging task.

- Disease-based RSS feeds consist of disease names (e.g., bluetongue OR BTV) and target seven animal diseases.
- Symptom-based RSS feeds include clinical signs and hosts (e.g., fever AND pigs) without any disease names. These feeds enable the detection of diseases that are yet to be monitored by PADI-web, as well as unknown diseases [33].

RSS feeds are implemented in 16 languages (e.g. English, Chinese, Arabic, Italian, French, Russian, Turkish, etc.).

3.2. Data processing

To avoid duplicates, PADI-web checks if collected articles already exist in the database based on their URL. The webpages of the news articles that were not duplicates are visited to retrieve their content.

The *BeautifulSoup*[1] and *readability-lxml*[2] Python libraries are used to collect the content of webpages and remove irrelevant elements (e.g., pictures, hyperlinks, etc.) [34]. Used as a piece of the preprocessing task of the PADI-web pipeline, these libraries allow us to isolate and work on the cleaned content of web documents from the web, i.e. only the body, titles, etc.

All news articles that are not in English are translated using the Translator API of the Microsoft Azure system [35].

3.3. Data classification

The classification step allows the selection of relevant news among all the news collected. A relevant news article is a news article that is related to a disease event. Relevant news includes the description of a current outbreak as well as its socioeconomic impacts, preparedness,

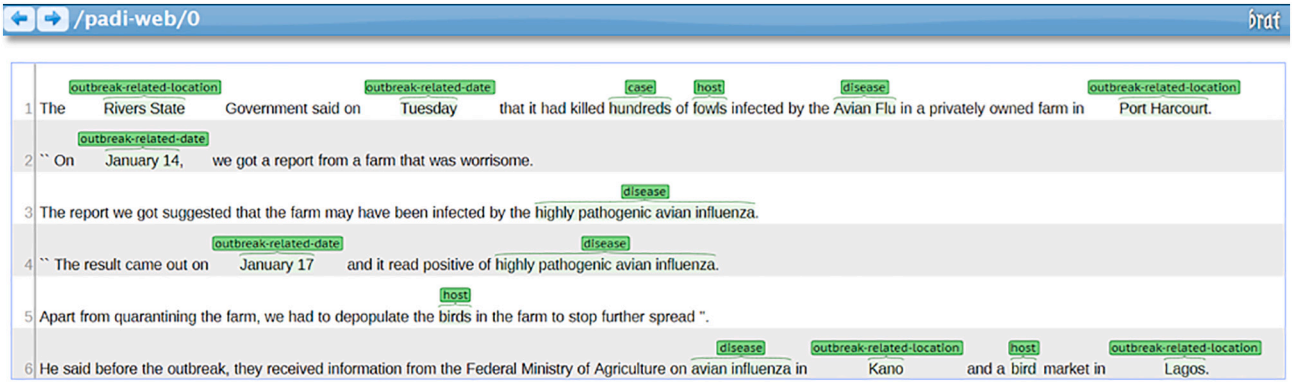


Fig. 4. Brat annotation integrated into PADI-web 3.0.

collection, (2) data processing, (3) data classification, (4) sentence classification, (5) information extraction and (6) user notification. All these steps are summarized in the following subsections, and they were previously detailed in [14]. The extensions proposed into PADI-web 3.0 (i.e. sentence classification, named entity annotation, terminology extraction, and notification) are described in subsections 3.4, 3.5 and 3.6. The different steps of the pipeline are shown in Fig. 2.

3.1. Data collection

PADI-web retrieves articles daily from the news aggregator Google News through customized RSS feeds. An RSS feed is a combination of terms (disease names, symptoms or hosts). These terms have been identified by an approach combining text mining and domain experts. The RSS feeds are of two types:

prevention and control measures, etc. The classification module is based on a supervised machine learning approach described in [14].

3.4. Sentence classification

For classification tasks, annotation is usually at the document level. The labels are often related to the news' relevance to filter out the irrelevant ones [14,36–38]. Other classification frames assign a broad thematic label to the news, such as 'outbreak-related' or 'socioeconomic' [39]. PADI-web already includes a classification module for texts (i.e. news). Into PADI-web 3.0, the sentence classification feature extends it to automatically classify all the sentences of a new text. This enables

[1] https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[2] https://pypi.org/project/readability-lxml/

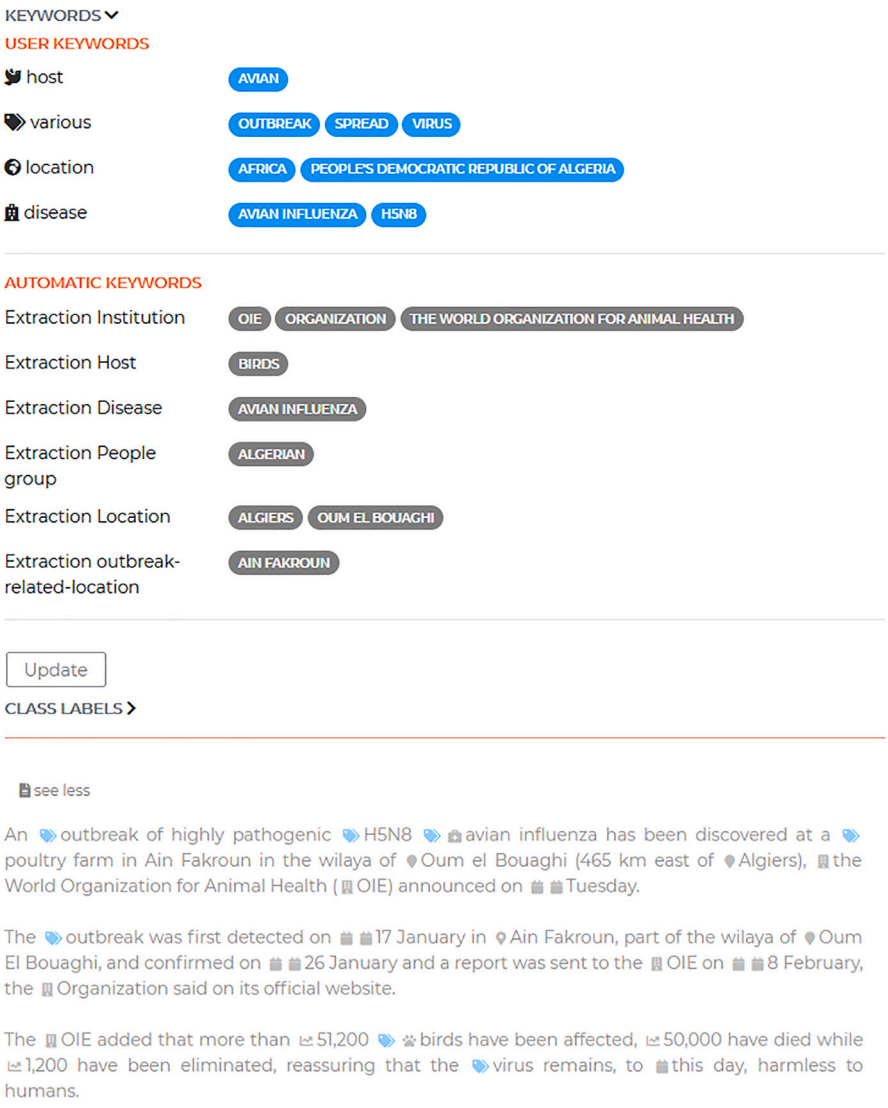


Fig. 5. Organisation of keyword lists in PADI-web 3.0.

highlighting fine-grained epidemiological information (e.g., risk events, preventive and control measures, etc.). Each new text acquired by PADI-web is segmented into sentences (with SpaCy[3]). These sentences are then automatically processed by a classifier that contains a specific classification model. The sentences are represented with a classical vector space model (i.e. bag-of-words).

To learn the models and automatically classify sentences, we need to provide labeled data (i.e. training data). For the proposed fine-grained classification, a dedicated corpus has been built [40]. This one is summarized below. A detailed version of the guidelines and the labeled corpus are publicly available in a Dataverse repository [40]. For each sentence, two labeling approaches are proposed: event and information types. The *Event type* label aims to differentiate sentences referring to the current/recent outbreak (*Current event* and *Risk event*) from sentences referring to old outbreaks (*Old event*) or general information (*General*). Sentences that do not contain any epidemiological information are considered irrelevant (*Irrelevant*). The *information type* level describes the sentence epidemiological topic. As an epidemiological topic, we include the notification of a suspected or confirmed event, the description of a disease in an area (*Descriptive epidemiology* and *Distribution*), preventive or control measures against a disease outbreak (*Preventive and control measures*), an event's economic and/or political impacts (*Economic and political consequences*), its suspected or confirmed transmission mode (*Transmission pathway*), the expression of concern and/or facts about risk factors (*Concern and risk factors*) and general information about the epidemiology of a pathogen or a disease (*General epidemiology*).

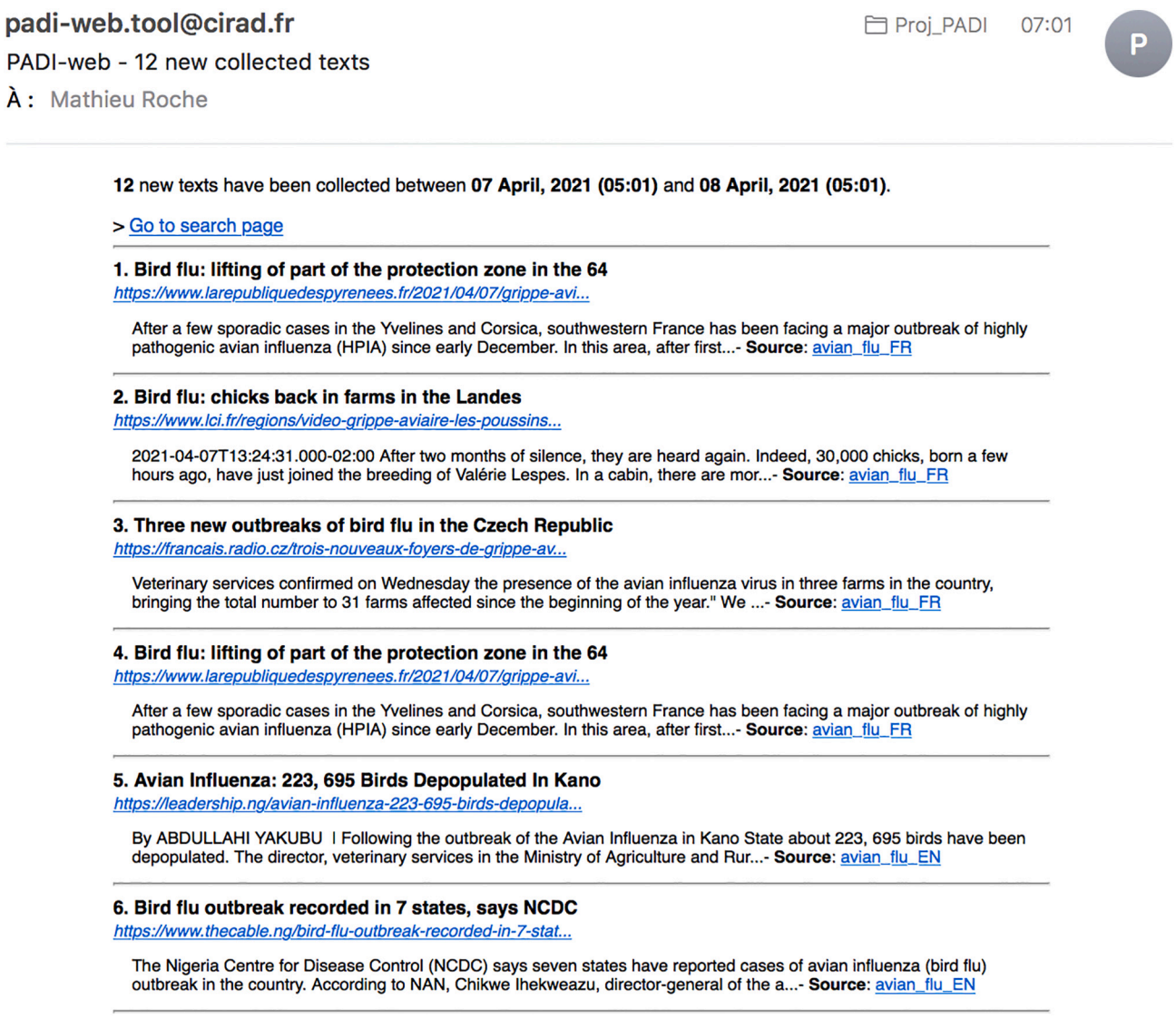
3.5. Information extraction

The final step aims to extract epidemiological entities from the relevant news content. In PADI-web 3.0, the previous information extraction method founded on rule-based systems and data mining techniques has been entirely replaced by a named entity recognition process detailed below.

3.5.1. Named entity recognition

As described in Section 2.2, several tools exist for named entity recognition (NER). SpaCy has been integrated into PADI-web 3.0. SpaCy already includes powerful NER models that allow recognizing general named entities in texts using several languages. With PADI-web, we can

[3] https://spacy.io/


padi-web.tool@cirad.fr

Proj_PADI 07:01

PADI-web - 12 new collected texts

À : Mathieu Roche

12 new texts have been collected between **07 April, 2021 (05:01)** and **08 April, 2021 (05:01)**.

> Go to search page

1. Bird flu: lifting of part of the protection zone in the 64
https://www.larepubliquedespyrenees.fr/2021/04/07/grippe-avi...

After a few sporadic cases in the Yvelines and Corsica, southwestern France has been facing a major outbreak of highly pathogenic avian influenza (HPIA) since early December. In this area, after first...- **Source**: avian_flu_FR

2. Bird flu: chicks back in farms in the Landes
https://www.lci.fr/regions/video-grippe-aviaire-les-poussins...

2021-04-07T13:24:31.000-02:00 After two months of silence, they are heard again. Indeed, 30,000 chicks, born a few hours ago, have just joined the breeding of Valérie Lespes. In a cabin, there are mor...- **Source**: avian_flu_FR

3. Three new outbreaks of bird flu in the Czech Republic
https://francais.radio.cz/trois-nouveaux-foyers-de-grippe-av...

Veterinary services confirmed on Wednesday the presence of the avian influenza virus in three farms in the country, bringing the total number to 31 farms affected since the beginning of the year." We ...- **Source**: avian_flu_FR

4. Bird flu: lifting of part of the protection zone in the 64
https://www.larepubliquedespyrenees.fr/2021/04/07/grippe-avi...

After a few sporadic cases in the Yvelines and Corsica, southwestern France has been facing a major outbreak of highly pathogenic avian influenza (HPIA) since early December. In this area, after first...- **Source**: avian_flu_FR

5. Avian Influenza: 223, 695 Birds Depopulated In Kano
https://leadership.ng/avian-influenza-223-695-birds-depopula...

By ABDULLAHI YAKUBU | Following the outbreak of the Avian Influenza in Kano State about 223, 695 birds have been depopulated. The director, veterinary services in the Ministry of Agriculture and Rur...- **Source**: avian_flu_EN

6. Bird flu outbreak recorded in 7 states, says NCDC
https://www.thecable.ng/bird-flu-outbreak-recorded-in-7-stat...

The Nigeria Centre for Disease Control (NCDC) says seven states have reported cases of avian influenza (bird flu) outbreak in the country. According to NAN, Chikwe Ihekweazu, director-general of the a...- **Source**: avian_flu_EN

Fig. 6. Extract of a notification received the 8th of April 2021 related to avian influenza disease - List of new articles collected with French (FR) (automatically translated in English) and English (EN) feeds.

use a classical model to identify well-known named entities such as locations and organizations. Moreover, specific models for entity recognition for the animal disease surveillance domain (e.g., host, etc.) could be used (see Section 2.2). We have used a labeled dataset [41] to learn and integrate a domain-specific model, which is able to detect host and disease names, as well as numbers of cases related to an outbreak. Both types of entities (i.e. general and specific) are then recognized with PADI-web 3.0. For location names, regular calls to the Geonames gazetteer API [42] aim to associate each recognized location name with a Geonames entity ID (see Fig. 3).

3.5.2. *Named entity annotation*

To annotate textual data, Brat[4] was integrated into PADI-web 3.0. Brat is a powerful annotation tool that contains all the needed functionalities to prepare and annotate a corpus (see Fig. 4). Moreover, PADI-web includes the possibility to export texts from the PADI-web database and convert them into the Brat format. PADI-web also includes the option to use the Brat annotated corpora to update an existing model and update it with new examples.

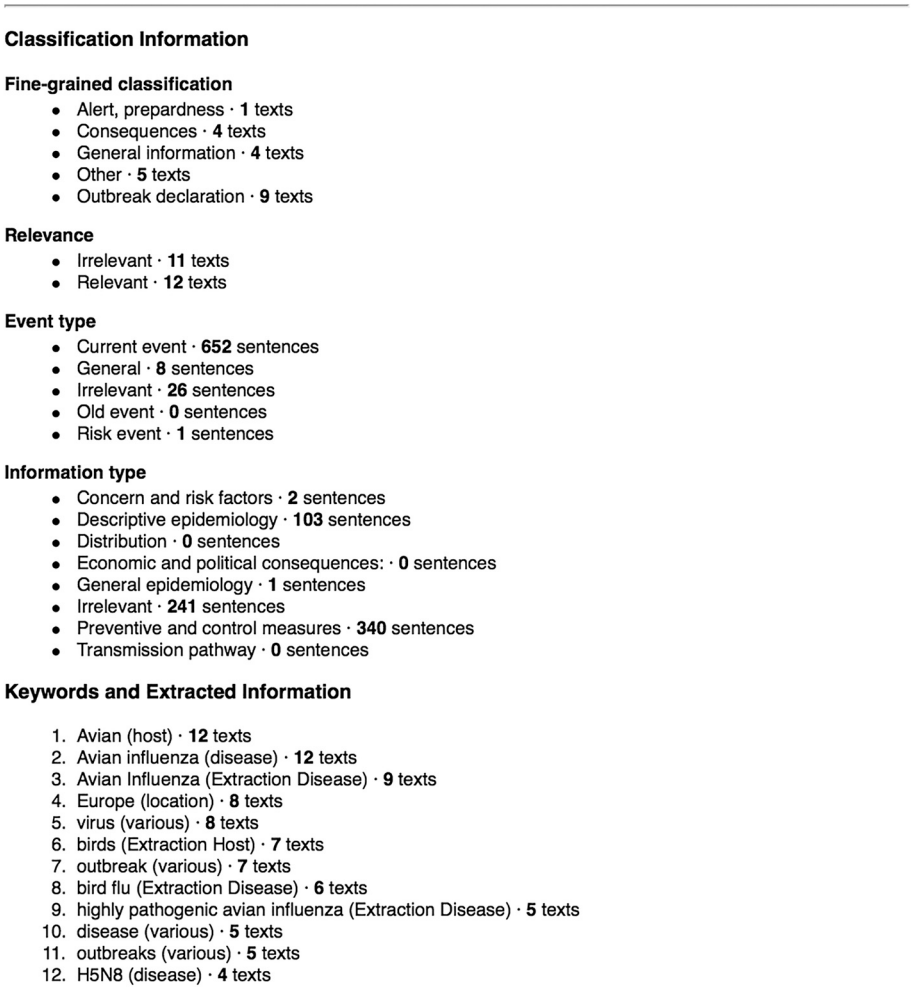
3.5.3. *Terminology extraction and semantic resource*

Terminology extraction is an important task in the natural language processing (NLP) domain. This enables the extraction of relevant and discriminative terms in textual data. Into PADI-web 3.0, we integrated a tool called BioTex [43] to highlight terms extracted from the textual data (i.e. news) of PADI-web. We use BioTex because it was initially built for the medical domain. BioTex combines linguistic and statistical information adapted to biomedical areas. To select the appropriate terms, BioTex uses two principles: (i) a combination of statistical methods, e.g., term frequency-inverse document frequency (TF-IDF), Okapi BM25 and C-value measures [43]; and (ii) use of a list of syntactic structures of the terms that have been learned with relevant sources in the medical domain, e.g., MeSH (Medical Subject Heading). The terms extracted with BioTex can be either words (e.g., 'pig') or multiword terms (e.g., 'wild pig', 'domestic pig').

To index data dealing with the agriculture domain, the AGROVOC[5] thesaurus has been integrated into PADI-web 3.0. AGROVOC is the largest linked open dataset dealing with the agriculture domain. AGROVOC is a thesaurus that contains 38,780 concepts and 808,000

[4] https://brat.nlplab.org/

[5] http://www.fao.org/agrovoc/



Classification Information

Fine-grained classification

- Alert, prepardness · **1** texts
- Consequences · **4** texts
- General information · **4** texts
- Other · **5** texts
- Outbreak declaration · **9** texts

Relevance

- Irrelevant · **11** texts
- Relevant · **12** texts

Event type

- Current event · **652** sentences
- General · **8** sentences
- Irrelevant · **26** sentences
- Old event · **0** sentences
- Risk event · **1** sentences

Information type

- Concern and risk factors · **2** sentences
- Descriptive epidemiology · **103** sentences
- Distribution · **0** sentences
- Economic and political consequences: · **0** sentences
- General epidemiology · **1** sentences
- Irrelevant · **241** sentences
- Preventive and control measures · **340** sentences
- Transmission pathway · **0** sentences

Keywords and Extracted Information

1. Avian (host) · **12** texts
2. Avian influenza (disease) · **12** texts
3. Avian Influenza (Extraction Disease) · **9** texts
4. Europe (location) · **8** texts
5. virus (various) · **8** texts
6. birds (Extraction Host) · **7** texts
7. outbreak (various) · **7** texts
8. bird flu (Extraction Disease) · **6** texts
9. highly pathogenic avian influenza (Extraction Disease) · **5** texts
10. disease (various) · **5** texts
11. outbreaks (various) · **5** texts
12. H5N8 (disease) · **4** texts

Fig. 7. Extract of a notification received the 8th of April 2021 related to avian influenza disease - Information about (i) classification and (ii) keywords extracted. First, news classification is proposed with 2 types of classifications (i.e. fine-grained and relevance classifications). Second, the sentence classification described in subsection 3.3 is notified (i.e. event and information types).

Table 1
Performances of MLP for Event type classification.

	Precision	Recall	F-measure
Current event ($n = 799$)	0.74	0.98	0.81
Risk event ($n = 105$)	0.39	0.29	0.33
Old event ($n = 44$)	0.33	0.09	0.14
General ($n = 136$)	0.79	0.58	0.67
Irrelevant ($n = 160$)	0.69	0.41	0.52
Weighted average	0.72 (±0.02)	0.70 (±0.02)	0.69 (±0.02)

terms.

Each keyword source is in one of 3 different source types that PADI-web 3.0 handles: from semantic resources (i.e., AGROVOC), from users (e.g., epidemiological-related keywords provided by a user), and from automatic processes (i.e., SpaCy and BioTex). Keywords are organized according to their source type and have a different color depending on it. In the text body, icons standing for keyword matches follow the same color code as in lists of keywords (see Fig. 5).

Table 2
Performances of MLP for Information type classification.

	Precision	Recall	F-measure
Descriptive epidemiology ($n = 401$)	0.70	0.78	0.73
Distribution ($n = 27$)	0.67	0.15	0.24
Preventive and control measures ($n = 309$)	0.57	0.75	0.65
Concern and risk factors ($n = 110$)	0.53	0.35	0.42
Transmission pathway ($n = 69$)	0.56	0.28	0.37
Economic and political consequences ($n = 58$)	0.68	0.26	0.38
General epidemiology ($n = 109$)	0.83	0.70	0.76
Weighted average	0.66 (±0.03)	0.66 (±0.04)	0.66 (±0.03)

3.5.4. *Linking*

Finally, a keyword alignment mechanism is integrated into PADI-web 3.0 for entity and keyword recognition with elements of the PADI-web database. This keyword alignment mechanism is based on the Levenshtein distance [44,45]: two keywords are aligned if their distance is below a user-specified threshold. The distance measure is a

normalized Levenshtein distance (normalized between 0 and 1 by dividing it by the length of the longest string). The Levenshtein distance between two strings is given by the minimum number of operations needed to transform one source string into the target string. An operation is an insertion, deletion, or substitution of a single character.

The following subsection presents a notification system that uses the new functionalities presented in Subsections 3.4 and 3.5.

3.6. User notifications

Registered users can subscribe to notification emails in PADI-web 3.0 to receive regular emails summarizing some basic information about the recently collected texts. Notification emails aim to provide a PADI-web user with a compact list of texts collected recently. It is also possible to filter articles by specifying some text sources (e.g., RSS feeds) in the users' preferences. They are sent daily or weekly (according to the users' preferences) and contain three types of information:

- The list of new texts that have been collected during the period covered by the notification. The maximum amount of texts listed in an email is thirty. A link to the PADI-web search page is provided for a complete inspection (see Fig. 6).
- Information about the classification of new texts covered by the notification email. For each classification task, the number of new texts that have been assigned to each classification label is provided (see Fig. 7).
- Information about keywords and extracted information from the new texts. The 30 most frequent keywords found in the texts are listed with their respective frequency, i.e., the number of texts where each entity or keyword occurs (see Fig. 7).

Daily notification emails are sent every morning starting from 7 am (according to the PADI-web server time). Weekly notifications are sent every Monday morning.

4. Results

This section presents experimental results obtained with sentence classification. We evaluated a global classification model able to correctly identify both the event type and information type of an unlabeled sentence (see Section 3.4). We used an annotated corpus that contains 1244 sentences (from 87 news articles). We applied classical NLP techniques to transform the sentences into numerical vectors (punctuation removal, conversion to lowercase, splitting into tokens and Term Frequency - Inverse Document Frequency weighting) [46,47]. We compared three classifiers that are widely used for text classification, i.e. Naive Bayes (NB), support vector machines (SVMs) and multilayer perceptrons (MLPs). We estimated the performances of the trained models using precision, recall, F-measure and 5-cross-validation methods. Precision corresponds to the number of relevant sentences of a given class over the number of sentences attributed to this class. Recall is the number of relevant sentences of a given class over the real number of relevant sentences associated with this class. The F-measure is the harmonic mean of precision and recall. We presented the results using an MLP classifier in Tables 1 and 2. Classification scores for some classes are better (e.g., *Current event*, *Descriptive epidemiology*) than other ones (e.g., *Old event*, *Distribution*). We could explain this situation with the unbalanced datasets used. Moreover, during the annotation phase, many instances of dedicated classes (e.g., general epidemiology) involved the same sentence structure (e.g., 'The virus causes a hemorrhagic fever with high mortality rates in pigs') that favors machine learning approaches. MLP and SVM achieved comparatively equal performances and outperformed the NB classifier [48]. These behaviors were identical for both event type and information type classifications [48].

Into PADI-web 3.0, a selection of model families is trained on the current dataset (random forests with various parameters, linear support vector classification, neural networks, Gaussian-based models, K-nearest neighbors, etc.), using a 5-fold cross-validation scheme. The model obtaining the highest mean accuracy score[6] is used to classify each new retrieved sentence. For article-level classification tasks, a model is trained and built automatically every night using existing user classification labels. Currently, the trained classification models integrated into PADI-web 3.0 reach a mean accuracy score of 0.94 for the article-level *Relevance* task using a random forest classifier. For the sentence-level *Event type* and *Information type* classification tasks, the accuracy is 0.66 and 0.49, respectively, with a random forest classifier in both cases. The preprocessing tasks applied in our experiments to optimize the results summarized in this paper and detailed in [48] have to be integrated into PADI-web 3.0.

5. Discussions and conclusion

This paper presented the extraction of epidemiological event information for animal disease surveillance using new functionalities of PADI-web. These functionalities are based on semantic information and fine-grained information integrated into machine-learning approaches.

[49] noted that health agencies have been reluctant to incorporate outputs from biosurveillance EBS tools into their systems because many technical issues had not yet been addressed. In the development of PADI-web 3.0 and its new functionalities, experts were solicited at different levels: identification of users' operational needs, annotation guideline creation, corpus annotation, qualitative evaluation of method outputs and regular feedback loops concerning the developments' methods and outputs. Health experts are inclined to integrate automatic processes beyond event-based surveillance outputs, as they directly support the decision-making process [50]. The links and collaborations between informatics and epidemiology should therefore be strengthened and promoted.

Authorship statement

All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufffciently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript. Furthermore, each author certiffes that this material or similar material as not been and will not be submitted to or published in any other publication before its appearance in *One Health*.

All persons who have made substantial contributions to the work reported in the manuscript (e.g., technical help, writing and editing assistance, general support) are named in the Acknowledgements.

Acknowledgements

This work was funded by the French General Directorate for Food (DGAL), the French Agricultural Research Centre for International Development (CIRAD) and the SONGES Project (FEDER and Occitanie). This work was also supported by the French National Research Agency under the Investments for the Future Program, referred to as ANR-16-CONV-0004.

We thank Valérie De Waele and Aline Vilain (Moriskin project coordinated by Sciensano (Institute of Public and Animal Health) - Belgian Federal Public Service Health) for the annotation task of the corpus cited in this paper. This study was partially funded by EU grant 874850 MOOD and is catalogued as MOOD014. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission.

[6] Ratio of number of correct predictions to the total number of elements to classify.

References

[1] F. Keesing, L.K. Belden, P. Daszak, A. Dobson, C.D. Harvell, R.D. Holt, P. Hudson, A. Jolles, K.E. Jones, C.E. Mitchell, S.S. Myers, T. Bogich, R.S. Ostfeld, Impacts of biodiversity on the emergence and transmission of infectious diseases, Nature 468 (7324) (2010) 647–652, number: 7324 Publisher: Nature Publishing Group.

[2] R.S. Ostfeld, Biodiversity loss and the rise of zoonotic pathogens, Clin. Microbiol. Infect. 15 (2009) 40–43, https://doi.org/10.1111/j.1469-0691.2008.02691.x. URL, http://www.sciencedirect.com/science/article/pii/S1198743X14604122.

[3] A.D. Langmuir, The epidemic intelligence Service of the Center for Disease Control, Public Health Rep. 95 (5) (1980) 470–477. URL, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1422746/.

[4] R. Kaiser, D. Coulombier, M. Baldari, D. Morgan, C. Paquet, What is epidemic intelligence, and how is it being improved in Europe? Weekly Releases (1997–2007) 11 (5) (2006) 2892, https://doi.org/10.2807/esw.11.05.02892-en. URL, https://www.eurosurveillance.org/content/10.2807/esw.11.05.02892-en.

[5] C. Paquet, D. Coulombier, R. Kaiser, M. Ciotti, Epidemic intelligence: a new framework for strengthening disease surveillance in Europe, Eurosurveillance 11 (2006) 5–6, https://doi.org/10.2807/esm.11.12.00665-en. URL. https://www.eurosurveillance.org/content/10.2807/esm.11.12.00665-en.

[6] WHO, Early Detection, Assessment and Response to Acute Public Health Events: Implementation of Early Warning and Response with a Focus on Event-Based Surveillance, Interim Version Edition, WHO Press, Geneva: The Organization, 2014. URL, https://apps.who.int/iris/bitstream/handle/10665/112667/WHO_HSE_GCR_LYO_2014.4_eng.pdf;jsessionid=750EA62E5F92D8D975244F320CB3D67B?sequence=1.

[7] O. Alomar, A. Batlle, J.M. Brunetti, R. García, R. Gil, T. Granollers, S. Jiménez, A. Laviña, C. Reverté, J. Riudavets, J. Virgili-Gomà, Development and testing of the media monitoring tool MedISys for the monitoring, early identification and reporting of existing and emerging plant health threats, EFSA Supporting Publications 13 (12) (2016). URL, https://efsa.onlinelibrary.wiley.com/doi/abs/10.2903/sp.efsa.2016.EN-1118.

[8] E. Arsevska, S. Valentin, J. Rabatel, J. de Goër, S. de Hervé, R. Falala, M. Roche Lancelot, Web monitoring of emerging animal infectious diseases integrated in the French animal health epidemic intelligence system, PLoS One 13 (8) (2018), e0199960, https://doi.org/10.1371/journal.pone.0199960. URL, https://dx.plos.org/10.1371/journal.pone.0199960.

[9] A. Lyon, A. Mooney, G. Grossel, Using AquaticHealth.net to detect emerging trends in aquatic animal health, Agriculture 3 (2) (2013) 299–309, https://doi.org/10.3390/agriculture3020299. URL, http://www.mdpi.com/2077-0472/3/2/299.

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

[11] P. Barboza, L. Vaillant, A. Mawudeku, N.P. Nelson, D.M. Hartley, L.C. Madoff, J. P. Linge, N. Collier, J.S. Brownstein, R. Yangarber, P. Astagneau, On behalf of the early alerting, reporting project of the Global Health security initiative, evaluation of epidemic intelligence systems integrated in the early alerting and reporting project for the detection of A/H5N1 influenza events, PLoS One 8 (3) (2013), e57252, https://doi.org/10.1371/journal.pone.0057252. URL, http://dx.plos.org/10.1371/journal.pone.0057252.

[12] B. Rotureau, P. Barboza, A. Tarantola, C. Paquet, International epidemic intelligence at the Institut de Veille Sanitaire, France, Emerg. Infect. Dis. 13 (2007) 1590–1592, https://doi.org/10.3201/eid1310.070522. URL, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2851537/.

[13] M.G. Baker, A.M. Forsyth, The new international health regulations: a revolutionary change in global health security, The New Zealand Med. J. 120 (1267) (2007) U2872.

[14] S. Valentin, E. Arsevska, S. Falala, J. de Goër, R. Lancelot, A. Mercier, J. Rabatel, M. Roche, PADI-web: a multilingual event-based surveillance system for monitoring animal infectious diseases, Comput. Electron. Agric. 169 (2020) 105163, https://doi.org/10.1016/j.compag.2019.105163. URL, http://www.sciencedirect.com/science/article/pii/S0168169919310646.

[15] S. Valentin, E. Arsevska, A. Mercier, S. Falala, J. Rabatel, R. Lancelot, M. Roche, Padi-web: An event-based surveillance system for detecting, classifying and processing online news, in: Z. Vetulani, P. Paroubek, M. Kubis (Eds.), Human Language Technology. Challenges for Computer Science and Linguistics, Springer International Publishing, Cham, 2020, pp. 87–101.

[16] S. Valentin, A. Mercier, R. Lancelot, M. Roche, E. Arsevska, Monitoring online media reports for early detection of unknown diseases: insight from a retrospective study of COVID-19 emergence, Transbound. Emerg. Dis. 68 (3) (2021) 981–986, https://doi.org/10.1111/tbed.13738.

[17] E. Arsevska, S. Falala, J. De Goer, R. Lancelot, J. Rabatel, M. Roche, PADI-web: platform for automated extraction of animal disease information from the web, in: Proceedings of LTC - Language and Technology Conference, 2017, pp. 241–245.

[18] J. Mantero, J. Belyaeva, J. Linge, European Commission, Joint Research Centre, Institute for the Protection and the Security of the Citizen, How to Maximise Event-Based Surveillance Web-Systems: the Example of ECDC/JRC Collaboration to Improve the Performance of MedISys, Publications Office, Luxembourg, 2011, https://doi.org/10.2788/69804 oCLC: 870614547. URL.

[19] R. Steinberger, F. Fuart, E. Goot, C. Best, P. Etter, R. Yangarber, Text mining from the web for medical intelligence, in: Mining Massive Data Sets for Security, IOS Press, 2008. URL, https://www.researchgate.net/profile/Erik_Van_der_Goot/publication/252032768_Text_Mining_from_the_Web_for_Medical_Intelligence/links/54e46a9d0cf2dbf6069671a0.pdf.

[20] D. Carter, M. Stojanovic, P. Hachey, K. Fournier, S. Rodier, Y. Wang, B. de Bruijn, Global Public Health Surveillance using Media Reports: Redesigning GPHIN, arXiv e-prints, 2020. arXiv:2004.04596_eprint: 2004.04596.

[21] R.J. Mooney, R. Bunescu, Mining knowledge from text using information extraction, ACM SIGKDD 7 (1) (2005) 3–10. URL, https://doi.org/10.1145/1089815.1089817.

[22] N. Guarino, D. Oberle, S. Staab, What is an ontology? in: S. Staab, R. Studer (Eds.), Handbook on Ontologies, International Handbooks on Information Systems Springer, Berlin, Heidelberg, 2009, pp. 1–17. URL, https://doi.org/10.1007/978-3-540-92673-3_0.

[23] H. Chanlekha, A. Kawazoe, N. Collier, A framework for enhancing spatial and temporal granularity in report-based health surveillance systems, BMC Med. Informat. Dec. Making 10 (1) (2010) 1.

[24] E. Amitay, N. Har'El, R. Sivan, A. Soffer, Web-a-where: geotagging web content, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '04, Association for Computing Machinery, Sheffield, United Kingdom, 2004, pp. 273–280. URL, https://doi.org/10.1145/1008992.1009040.

[25] J. Lafferty, A. McCallum, F.C.N. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proc. 18th International Conf. on Machine Learning, 2001, pp. 282–289.

[26] C.D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S.J. Bethard, D. McClosky, The Stanford CoreNLP natural language processing toolkit, Association for Computational Linguistics (ACL) System Demonstrations, 2014, pp. 55–60. URL, http://www.aclweb.org/anthology/P/P14/P14-5010.

[27] S. Bird, E. Loper, NLTK: the natural language toolkit, in: Proceedings of the ACL Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 214–217. URL, https://www.aclweb.org/anthology/P04-3031.

[28] S. Song, N. Zhang, H. Huang, Named entity recognition based on conditional random fields, Clust. Comput. 22 (2019) 1–12, https://doi.org/10.1007/s10586-017-1146-3.

[29] D. Inkpen, J. Liu, A. Farzindar, F. Kazemi, D. Ghazi, Location detection and disambiguation from twitter messages, J. Intell. Inf. Syst. 49 (2) (2017) 237–253, https://doi.org/10.1007/s10844-017-0458-3. URL.

[30] M. Honnibal, I. Montani, spaCy 2: natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, in: To appear, 2017.

[31] H. Li, K.R. Srihari, C. Niu, W. Li, Info Xtract location normalization: A hybrid approach to geographic references in information extraction, in: Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, 2003, pp. 39–44. URL, https://www.aclweb.org/anthology/W03-0106.

[32] B. Martins, H. Manguinhas, J. Borbinha, Extracting and exploring the geo-temporal semantics of textual resources, in: IEEE International Conference on Semantic Computing, 2008, pp. 1–9, https://doi.org/10.1109/ICSC.2008.86.

[33] E. Arsevska, M. Roche, P. Hendrikx, D. Chavernac, S. Falala, R. Lancelot, B. Dufour, Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web, Comput. Electron. Agric. 123 (2016) 104–115, https://doi.org/10.1016/j.compag.2016.02.010. https://linkinghub.elsevier.com/retrieve/pii/S0168169916300473.

[34] L. Richardson, Beautiful soup documentation, in: Doc April, 2007.

[35] M. Research, Customized neural machine translation with Microsoft Translator, library Catalog. www.microsoft.com, May 2018. URL, https://www.microsoft.com/en-us/research/blog/customized-neural-machine-translation-microsoft-translator/.

[36] M. Conway, S. Doan, A. Kawazoe, N. Collier, Classifying disease outbreak reports using N-grams and semantic features, Int. J. Med. Inform. 78 (2009) e47–e58.

[37] S. Doan, A. Kawazoe, N. Collier, The role of roles in classifying annotated biomedical text, in: Biological, Translational, and Clinical Language Processing, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 17–24. URL, https://www.aclweb.org/anthology/W07-1003.

[38] M. Torii, L. Yin, T. Nguyen, C.T. Mazumdar, H. Liu, D.M. Hartley, N.P. Nelson, An exploratory study of a text classification framework for internet-based surveillance of emerging epidemics, Int. J. Med. Inform. 80 (1) (2011) 56–66, https://doi.org/10.1016/j.ijmedinf.2010.10.015. URL, http://linkinghub.elsevier.com/retrieve/pii/S1386505610002030.

[39] Y. Zhang, Y. Dang, H. Chen, M. Thurmond, C. Larson, Automatic online news monitoring and classification for syndromic surveillance, Decis. Support. Syst. 47 (4) (2009) 508–517, https://doi.org/10.1016/j.dss.2009.04.016. URL, https://linkinghub.elsevier.com/retrieve/pii/S0167923609001109.

[40] S. Valentin, V. De Waele, A. Vilain, E. Arsevska, R. Lancelot, M. Roche, Annotation of epidemiological information in animal disease-related news articles: guidelines and manually labelled corpus, in: Dataset, CIRAD Dataverse, 2019, https://doi.org/10.18167/DVN1/YGAKNB.

[41] J. Rabatel, E. Arsevska, M. Roche, PADI-web corpus: labeled textual data in animal health domain, Data in Brief 22 (2019) 643–646, https://doi.org/10.1016/j.dib.2018.12.063. URL, https://linkinghub.elsevier.com/retrieve/pii/S2352340918316032.

[42] D. Ahlers, Assessment of the accuracy of geonames gazetteer data, in: Proceedings of the 7th Workshop on Geographic Information Retrieval, ACM, New York, NY, USA, 2013, pp. 74–81.

[43] J.A. Lossio-Ventura, C. Jonquet, M. Roche, M. Teisseire, Biomedical term extraction: overview and a new methodology, Informat. Ret. J. 19 (1) (2016) 59–99. URL, https://doi.org/10.1007/s10791-015-9262-2.

[44] V. Levenshtein, Binary Codes Capable of Correcting Deletions, Insertions and Reversals, Soviet Physics Doklady 10, 1966, p. 707.

[45] D. Lin, An information-theoretic definition of similarity, in: Proceedings of 15th International Conf. on Machine Learning, Morgan Kaufmann, 1998, pp. 296–304. URL, citeseer.ist.psu.edu/95071.html.

[46] A.K. Uysal, S. Gunal, The impact of preprocessing on text classification, Inf. Process. Manag. 50 (1) (2014) 104–112, https://doi.org/10.1016/j.ipm.2013.08.006. URL, http://www.sciencedirect.com/science/article/pii/S0306457313000964.

[47] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Inf. Process. Manag. 24 (5) (1988) 513–523, https://doi.org/10.1016/0306-4573(88)90021-0. URL, http://www.sciencedirect.com/science/article/pii/0306457388900210.

[48] S. Valentin, Extraction and Combination of Epidemiological Information from Informal Sources for Animal Infectious Diseases Surveillance, Ph.D. thesis, University of Montpellier, 2020.

[49] E. Velasco, T. Agheneza, K. Denecke, G. Kirchner, T. Eckmanns, Social media and internet-based data in global systems for public health surveillance: a systematic review, The Milbank Quart. 92 (1) (2014) 7–33. URL, https://doi.org/10.1111/1468-0009.12038/full.

[50] M. Cui, R. Bai, Z. Lu, X. Li, U. Aickelin, P. Ge, Regular expression based medical text classification using constructive heuristic approach, IEEE Access 7 (2019) 147892–147904, conference Name: IEEE Access, https://doi.org/10.1109/ACCESS.2019.2946622.