Journal Pre-proof

EpidNews: Extracting, Exploring and Annotating News for Monitoring Animal Diseases

Rohan Goel, Sarah Valentin, Alexis Delaforge, Samiha Fadloun, Arnaud Sallaberry, Mathieu Roche, Pascal Poncelet

PII: S2590-1184(19)30061-9
DOI: https://doi.org/10.1016/j.cola.2019.100936
Reference: COLA 100936

To appear in: *Journal of Computer Languages*

Received date: 29 March 2019
Revised date: 5 November 2019
Accepted date: 11 November 2019

Please cite this article as: Rohan Goel, Sarah Valentin, Alexis Delaforge, Samiha Fadloun, Arnaud Sallaberry, Mathieu Roche, Pascal Poncelet, EpidNews: Extracting, Exploring and Annotating News for Monitoring Animal Diseases, *Journal of Computer Languages* (2019), doi: https://doi.org/10.1016/j.cola.2019.100936

EpidNews: Extracting, Exploring and Annotating News for Monitoring Animal Diseases

Rohan Goel[a,b], Sarah Valentin[c,d], Alexis Delaforge[b,e], Samiha Fadloun[b], Arnaud Sallaberry[b,e,*], Mathieu Roche[d], Pascal Poncelet[b]

[a] *BITS Pilani, Department of Computer Science, Pilani - 333031 (Rajasthan), India*
[b] *LIRMM - Univ. Montpellier - CNRS, 860 rue de St Priest, 34095 Montpellier, France*
[c] *CIRAD - ASTRE - Univ. Montpellier, Campus international de Baillarguet, 34398 Montpellier, France*
[d] *CIRAD - TETIS - Univ. Montpellier, 500 rue Jean-Franois Breton, 34000 Montpellier, France*
[e] *Univ. Paul-Valéry Montpellier 3, Route de Mende, 34199 Montpellier, France*

Abstract

In the recent years, there has been a massive increase in the amount of data published on the web about human and animal health events. Epidemiologists use this spatio-temporal information on a daily basis to detect and monitor disease outbreaks over time. While official sources such as the World Organization for Animal Health release formal outbreak notifications, unofficial sources such as online newspapers contain unstructured information with different levels of reliability. Manually retrieving the data from a website like Google News and then deriving sensible insights from the huge dataset takes a lot of time and effort. We present *EpidNews*, a new visual analytics tool that helps to visualize and explore epidemiological data for monitoring animal disease outbreaks. The tool uses several views depicting various levels of abstraction, which helps fulfill almost all the data analysis requirements of epidemiologists. *EpidNews* allows to visualize and compare data from both official and unofficial sources. We also present the use case of an epidemiology expert, wherein the expert assesses the usability and productivity of *EpidNews* by using the tool in her daily work.

*Corresponding author
Email address: arnaud.sallaberry@lirmm.fr (Arnaud Sallaberry)

A preliminary version of this paper was presented at the 11th International Symposium on Visual Information Communication and Interaction (VINCI 2018) [1].

Keywords: Visual analytics, Animal epidemiology, Spatio-temporal data

1. Introduction

The surveillance of disease outbreaks at an international level is a major challenge for both animal and human health. In this paper, we focus on animal diseases which have been least covered by previous work whereas it is of utmost importance: animals play a key role in human activities such as agriculture and poultry, and some animal diseases, called "zoonoses", have the potential to affect humans through the livestock-human interface [2].

A significant amount of effort in epidemiology research goes into detecting and monitoring disease outbreaks to prevent their spread. For this purpose, the epidemiologists need to collect and study large amounts of information on a daily basis.

Traditionally, epidemiologists working in the animal health area have relied upon official sources, such as the World Organization for Animal Health (OIE)[2]. The European Food Safety Authority (EFSA), for example, used the Animal Disease Notification System (ADNS) to highlight how the *African swine fever* virus which was initially reported in the Baltic countries in 2014, spread to the eastern parts of Poland, causing *wild boars* deaths in large numbers [3]. Outbreak information produced by official sources is shared into a well-defined format, i.e. with proper tagging of epidemiological entities like *diseases*, *hosts*, *symptoms*, *dates* and *locations*.

Over the past two decades, several platforms dedicated to automatic surveillance of electronic media have been developed to monitor potential health events worldwide [4]. Some of them are designed for the surveillance of a large range of threats [5], while others focus on animal health [6]. Data from such sources is not structured and therefore, the epidemiological entities (*diseases*, *hosts*, *symptoms*, *dates* and *locations*) are extracted by data mining techniques [7] - we refer

[2] http://www.oie.int/en/

to data obtained from such techniques as "unofficial".

Both types of sources are valuable for the epidemiologists. The official ones contain validated and reliable information in a structured format. The unofficial sources are usually publicly available. When they release information about a health event before its official confirmation, unofficial sources can provide early insights of new outbreaks [8]. However, as they contain unverified information, they are also prone to produce false alerts.

Exploring and analyzing the complex, heterogeneous and multi-sourced epidemiological data is a daunting task without the use of a dedicated tool. Visual analytics is a domain that focuses on designing visual interfaces to facilitate visual reasoning [9]. In our context, it could provide us a way to build a suitable tool for exploring and comparing the epidemiological data from official and unofficial sources efficiently. Unfortunately, existing visual analytics tools ([10, 11]) do not meet all the requirements of the epidemiologists.

In order to fill this gap, we present *EpidNews*, a new visual analytics tool for epidemiological data. It provides different views to: (1) handle different types of sources (official / unofficial) and different types of epidemiological entities (*diseases* / *hosts* / *symptoms*), (2) visualize geographic and temporal data, (3) observe data at different levels of aggregation, (4) annotate unofficial data. A demonstration video is available on YouTube[3].

In this paper, Section 2 gives an overview of the related work and Section 3 presents a list of requirements for the tool which were identified by computer scientists in collaboration with the French Epidemic Intelligence System, from the French Animal Health Surveillance Platform (ESA Platform)[4]. In Section 4, we depict the method to design both official and unofficial data for the visualization process. Section 5 gives a detailed description of *EpidNews* and in Section 6, we present a real world example of the analysis performed by an epidemiologist using *EpidNews*, highlighting the benefits of our approach. Finally, we conclude

[3] `https://youtu.be/iXjV4XRp6Rs`

[4] `https://www.plateforme-esa.fr/`

our work in Section 7.

2. Related Work

This section first discusses existing techniques in visualizing spatio-temporal data, followed by the various visualization approaches applied in epidemiology.

2.1. Visualizing spatio-temporal information

A piece of spatio-temporal data contains information about both space and time. In [12], Peuquet proposed a triad framework to describe how this can be exploited through three components: where (location), when (time) and what (information). The author also highlights their correlation in the following way:

- **when** + **where** → **what**: describes the information at a given location and time.
- **when** + **what** → **where**: shows the location of an information at a given time.
- **where** + **what** → **when**: gives the time of a particular information at a given location.

Several visualizations have been proposed to represent these components. Generally, the information (what) is combined with the other attributes where and when. For example, one of the first well-known representations combining location and information was made in the 19th century, by a British physician John Snow, to highlight a cholera outbreak in London [13]. He carried out a statistical analysis of the deaths which highlighted the spread of cholera via water distribution.

Aigner et al. [14][5] discusses about the time-oriented visualizations that combine time (when) and the information (what). Such an example is *EventRiver* [15], which represents events as bubbles positioned along the time axis.

[5] `https://vcg.informatik.uni-rostock.de/~ct/timeviz/timeviz.html`, (accessed on March 29, 2019)

Moreover, several visualizations ([16, 17, 18]) combine all the components (where, when and what). The work of Ferreira et al. [16] is one such example, where they propose an interactive visualization combining three different views: dot map, scatter plots and timeline.

2.2. Data visualization in epidemiology

In this subsection, we discuss about the monitoring systems in animal epidemiology. Although some visualization approaches have been proposed in the epidemiological field (e.g., [10, 11, 19, 20, 21, 22]), most of them focus on human diseases rather than on animal diseases. Moreover, none of them provides a complete visualization capturing the overall complexity of the data: various epidemiological entities -diseases, hosts and symptoms-, spatio-temporal dimensions, official and unofficial sources.

Existing approaches can track signs of emerging diseases and are based on a formalized process of epidemic intelligence[6] [23] focusing on *indicator-based* surveillance (e.g., *Empres-i*, *WAHIS*[7]) or on *event-based* surveillance (e.g., *HealthMap* [8]), *MedISys*[9]).

Indicator-based surveillance uses structured data collected through the official sources of traditional monitoring systems. For example, *Empres-i*[10] is a web application that visualizes the information with the help of a map and fetches the data from international animal health authorities: OIE, Food and Agriculture Organization of the United Nations (FAO), Government Ministries of Agriculture and Health, etc. Diseases are represented as glyphs on the map to highlight their type (e.g., circles for domestic animal diseases, and trian-

[6]Epidemic intelligence includes all activities related to early detection of health hazards, their verification, assessment, and investigation in order to recommend public health control measures.

[7]http://www.oie.int/en/animal-health-in-the-world/wahis-portal-animal-health-data/, (accessed on March 14, 2018)

[8]https://www.healthmap.org/en/, (accessed on March 14, 2018)

[9]http://medisys.newsbrief.eu, (accessed on March 14, 2018)

[10]http://empres-i.fao.org/eipws3g/, (accessed on March 14, 2018)

gles for wild animals diseases). The application also offers several interactions like choosing time/location by creating a rectangle on the map, along with the ability to save or select data, etc.

Event-based surveillance uses unstructured data collected from heterogeneous sources, such as the web, social networks, field experts, etc. For example, *HealthMap* [11] proposes an interactive map which uses word-processing algorithms from multiple sources like news feeds, multi-sources reporting systems (e.g. *ProMED-mail*), as well as official notifications. It represents diseases as circles, with noteworthiness of the event being highlighted by a color scale from yellow (minimum value) to purple (maximum value), and the size of the circle reflecting the level of alert. It depicts the disease information over time on a timeline and the geographic distribution as a heatmap. The tool also offers several interactions to filter data by location, time-slices, diseases, species, or sources.

The existing spatio-temporal information visualization tools represent the data using classical charts. For temporal information, representations like a static line chart in *HealthMap*, and a bar chart in *Empres-i* have been used. Furthermore, the spatial information is usually visualized on a static non-interactive map. Unfortunately, these basic visualization techniques do not properly present all the dimensions of the data. Also, it is important to note that the interactions between these spatial and temporal views are minimal, thereby limiting their usage and effectiveness. Moreover, although these tools use multiple data sources, they fail to separately identify and compare these sources in the visualizations (e.g. *HealthMap*). All of these existing approaches focus on visualization, but do not offer the possibility to interact with the data: the user cannot work on the dataset by labeling it or export an updated version.

3. Requirements

To develop *EpidNews*, a tool capable of handling different sources of data (official and unofficial) and the underlying spatio-temporal information, a list of

requirements were identified through a collaboration between computer scientists and three domain experts (i.e. veterinary epidemiologists). The involved experts were working within or closely to the French Epidemic Intelligence System (FEIS).

Hereafter, the term **outbreak** refers to a verified occurrence of an infectious animal disease, i.e. information obtained from official sources. By **signal**, we refer to an unverified epidemiological information, i.e. information obtained from unofficial sources.

The requirements identified are as follows:

[R0] Abstraction of official and unofficial data. Data from each type of sources needs to be in a proper format in order to be used as input for the visualization tool. This requirement is essential to ensure that the tool can efficiently handle various types of sources and allows the comparison between them. The process also needs to include the extraction of specific types of epidemiological entities, from both official and unofficial databases. By entity types, we mean the different categories of information commonly used to describe an outbreak (i.e. *disease*, *host*, *symptoms*, *location* and *date*). Unofficial data might contain additional information like: the name of the source website or the link to the original news report.

[R1] Exploring geographical data. This requirement refers to visualizing the data in the geographic space and generally, the most suitable visualization for this type is a map. The objective is to highlight and observe the different entity values of geographically tagged data in the same view. By entity values, we mean the different instances for each type of entity, e.g. *Avian influenza* is one possible value for the entity type *disease*. For instance, an epidemiologist wants to study the spread of the diseases *African swine fever* and *Bluetongue* in Europe and then compare it to the patterns found in other parts of the world.

[R2] Exploring temporal data. This is required by experts in order to reveal patterns, find peaks and track disease outbreaks over time. For example,

epidemiology experts want to observe the evolution of the disease *African swine fever* over the years 2015-2017, and check if the number of outbreaks or signals has increased, decreased or remained the same during this period. Epidemiologists are also keen to know if a disease spikes in a particular month or season of the year.

[R3] Summarizing and filtering data. Very often, the epidemiologists want to visualize a summary and aggregation of the data, in order to easily extract new insights from them. For instance, an expert might want to focus his/her research on only one disease (*African swine fever*), involving a particular host (*pig*) and to use only official sources. Therefore, he/she would like to filter away all the other diseases and the unofficial data from the analysis.

[R4] Producing data. It is of utmost importance to detect and label unofficial signals that do not match officials ones: they can be either early alerts or false/irrelevant signals. Experts also need to match unofficial signals and official reports that describe the same outbreak. Analyzing the signals from a raw dataset is time consuming, whereas visualizing the signals on the map makes the process easier. Using the spatio-temporal distance between an unofficial signal and a real outbreak helps to assess its relevance. Therefore, the tool should include features to filter signals using spatial and temporal parameters. Once an expert has evaluated a signal as irrelevant, he/she may want to record this information (for instance, to remove false signals from the original dataset). Therefore, the tool should allow the user to label each signal with its relevance and to export the updated version.

[R5] Interactivity and synchronization. Along with all the above requirements, it is of utmost importance to have synchronization between the different data representations, i.e. an activity in one, should reflect in the other views too. For instance, suppose that *V1* is a visualization that handles only the time information in data, and *V2* visualizes the location aspect. If a user changes the time period of analysis in *V1*, then in order to maintain synchronization,

V2 should visualize only the records falling in that time period. It can be very well noted that this requirement transverses all of R1, R2, and R3, R4. Also, the tool should engage the users by offering various interactions along with the visualizations: like dragging, zooming, ability to update the data, etc.

4. Data abstraction

Before creating a process to extract information from official and unofficial sources [**R0**], we list the epidemiological features extracted by the epidemiologists from both official and unofficial data sources, as well as the additional information needed to evaluate the reliability of the information.

- **Thematic entities**: This category includes the name of the *disease*, the *hosts* (animal species affected by the disease) and the *symptoms*.
- **Spatio-temporal entities**: These entities are used for the localization in space and time of the outbreak/the signal. The *location* has to be represented by a longitude and a latitude. The *date* is the temporal reference for each outbreak or signal in a standardized format, i.e. "YYYY-MM-DD".
- **Information about the source**: As our tool is designed to handle both official and unofficial data, it has to display the *source name*, a *textual content* as well as the *link* to the original content, and a *confidence index*, ranging from 0 to 1.

4.1. Official sources

Official sources store disease information in structured databases that are usually publicly available. Therefore, epidemiologists can easily download datasets in which each row corresponds to a unique outbreak, and columns correspond to the epidemiological entities. As the entities have consistent labels among official sources, the process simply consists in selecting the required ones from the dataset.

- **Thematic entities**: *disease* and *hosts* are directly extracted from the dataset. *Symptoms* are not provided by official data.

- **Spatio-temporal entities**: Several levels of spatial granularity are available, i.e. the continent, the country, the administrative division, etc. The most fine-grained information is given through the longitude and latitude, and corresponds to the specific location of the outbreak. As *date*, official sources provide two distinct references: the date of notification, i.e. the date on which the outbreak was officially communicated, and the date of occurrence, which is the real date on which the outbreak occurred.

- **Information about the source**: The *source name* is the name of the official source, and the *link* leads towards the official report. As *textual content*, the user can choose to aggregate additional epidemiological information provided by the source about the outbreak such as the number of cases, the mortality, etc. By default, the *confidence index* for official sources is set to 1 (the maximum value).

4.2. Unofficial sources

Unlike official sources, unofficial ones do not provide epidemiological entities in a structured format. Therefore, dedicated tools and data-mining methods are needed to extract and process the required information. For designing *EpidNews*, we relied on *PADI-web* (Platform for Automated Extraction of Animal Disease Information from the web), a system which tracks online news articles on animal disease outbreaks [6]. Figure 1 shows the three main steps of *PADI-web* pipeline.

PADI-web collects news articles from Google News through customized Really Simple Syndication (RSS) feeds. Each RSS feeds consists of an association of terms for a given disease (disease names terms, clinical signs or hosts). These combinations of terms are built with an integrated approach of automatic extraction of terms using text mining and domain experts [7]. As the second step, news articles are classified according to their relevance with the epidemiological

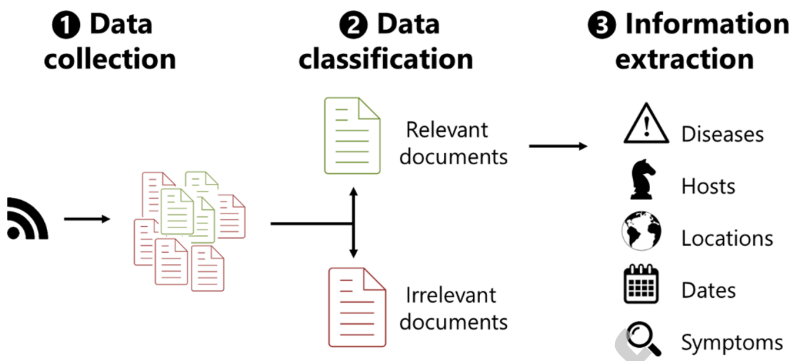

Figure 1: Overview of *PADI-web* pipeline.

field. This step allows filtering out irrelevant news articles (i.e. articles not related to an animal health event). The classification module relies on a Support Vector Machine (SVM) model trained on a dataset manually labeled by an epidemiologist expert.

The final step aims to convert raw textual data (news article content) into a structured dataset with epidemiological entities. This step is performed by the *PADI-web* information extraction module and applies only to the news articles classified as relevant at the previous step. The information extraction module relies on a combined method founded on rule-based systems and data mining techniques. A detailed description of the method is given in [6]. Briefly:

- **Thematic entities** are identified by matching with a list of *diseases* and *hosts* keywords. This list was created using text mining methods and validated by domain experts [7].

- **Spatio-temporal entities**: The *date* is set to the article publication date. *Locations* are identified by matching the text with location names from the gazetteer GeoNames [24]. The extraction module automatically calculates the accuracy with which the location is known and is related with the outbreak. A signal is generated for each location with an accuracy higher than 0.5.

- **Information about the source**: During the data collection step, *PADI-web* automatically extracts the *name of the source* (i.e. the website) and its *link* (i.e. the URL to the website). The *textual content* consists in a cleaned version of the original news content. The *confidence index* is set to the location accuracy.

5. EpidNews

In this section, we present *EpidNews* - a visual analytics tool for monitoring animal health events from multi-sources data. Fulfilling all the requirements, it is composed of several interactive views that allow the users to easily analyze, compare and label animal disease outbreaks and signals.

Figure 2 presents an overview of the tool. A map (Figure 2.a) shows spatial information [**R1**] and the two streamgraphs below it (Figure 2.b) visualize the temporal information [**R2**]. A sunburst (Figure 2.c) presents the relation between different entity types, namely *diseases*, *hosts* and *symptoms* [**R3**] in a hierarchical fashion. Each entity type contains a set of entities values: for example, the type *diseases* contains *African swine fever*, *Avian influenza*, *Bluetongue*, *Foot-and-mouth disease*, etc. A data manager (Figure 2.d) allows the user to select and filter entity types and values according to the requirements [**R3**] and [**R4**]. A toolbar (Figure 2.e) enhances the interactivity of all the views and allows uploading new data files. All these components are synchronized to reflect changes and interactions in each other [**R5**].

5.1. Map

The spatial information [**R1**] is represented using an interactive map (see Figure 2.a). We propose two kinds of geographical maps: (1) an icon map where each icon/glyph represents an outbreak or a signal which occurred at the corresponding location, and (2) a heatmap showing the distribution & density of the outbreaks/signals. The user can toggle between these two types by clicking the third or the fourth button on the toolbar (Figure 2.e). Also each map can be interacted with by zooming in/out and dragging to the required area [**R5**].

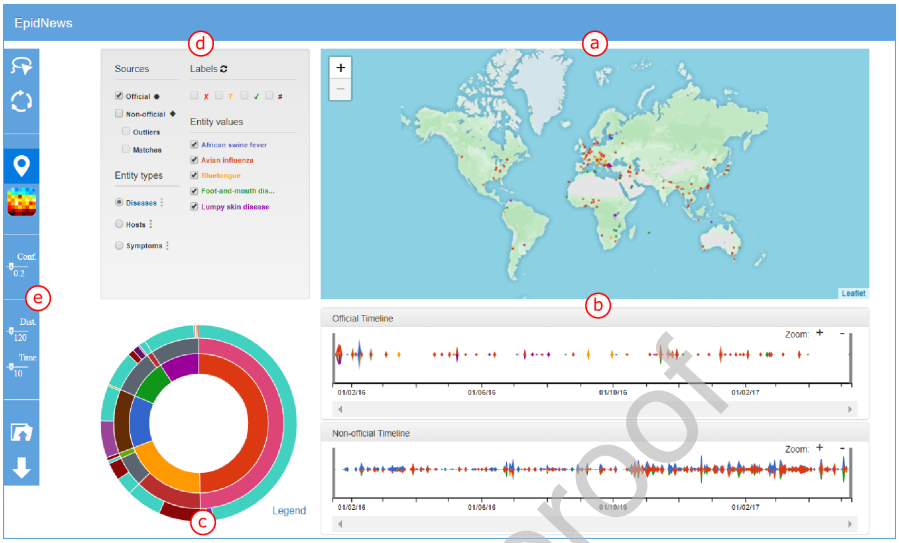

Figure 2: An overview of the tool *EpidNews*. **(a)** *Map* shows disease outbreaks and signals locations using a circle or diamond depending on source type. **(b)** *Streamgraphs* compare the temporal evolution of official and unofficial sources. **(c)** *Sunburst* presents relationships between *diseases*, *hosts* and *symptoms* in a hierarchical view. **(d)** *Data manager* allows manipulation of the data represented in the other views (sources, entities types and entities values). **(e)** *Toolbar* offers other interactive functionalities.

The heatmap (see Figure 3) represents the density of outbreaks/signals at a particular place. We use the viridis color scale following the study on quantitative colormaps performed by Liu and Heer [25].

The icon map depicts items with different icons/glyphs: circles for data from official sources (outbreaks) and diamonds for the unofficial ones (signals) [**R1**] (see Figure 2.a), thereby making their differentiation easier. Also, the color of glyphs represents and matches each entity value (Figure 2.d). For example, a red colored circle in Figure 2 represents an outbreak of *Avian influenza* extracted from an official source.

Moreover, in practice, many outbreaks or signals can occur at the same location which causes overlapping glyphs. The color opacity of the glyphs helps the user to estimate the number of items at a given point. For a precise view,

Figure 3: An example of a heatmap representing the density distribution of outbreaks using official sources.

the user can click on a glyph, which triggers a pop-up containing the list of all the items at that location (see Figure 4). Each item of this list mentions the following information: the shape and color of the glyph (signifying the source and entity value), the date of occurrence, the confidence index, the hyperlink to the source and the label icon for unofficial signals.

By default, the signals are labelled as *'not verified'*. The user can label each signal as *'relevant'* (the signal corresponds to an official outbreak), *'irrelevant'* (the signal does not correspond to an outbreak) or *'unknown'* (the signal cannot be verified) [**R4**]. To change the label of an item, the user simply clicks on the label icon until he/she reaches the right one.

To aid in the process of labeling the items and find the interesting ones, unofficial items that do not match officials ones can be identified by using sliders (Figure 2.e). Those sliders offer the possibility to change spatial and temporal parameters used to extract the items: each unofficial item that does not match an official one with a spatial and temporal distance less than the parameters is identified as an outlier [**R4**]. As the relevant spatio-temporal window can change for each disease, the sliders offer the flexibility needed to handle each case.

For both types of map, the user can filter [**R3**] a type of source or some

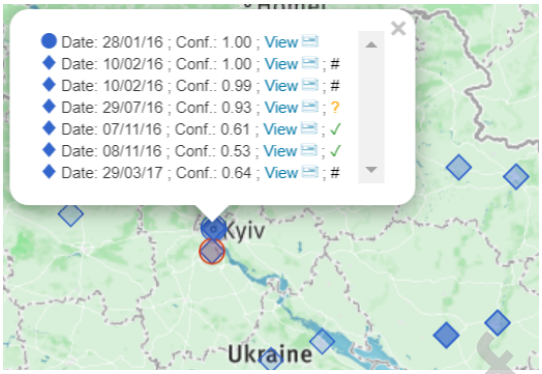

Figure 4: Outbreaks (circles) and signals (diamonds) news about *African swine fever* near Kyiv, Ukraine.

particular entities values by selecting the corresponding check-boxes in the "Sources" and "Entities" sections of the data manager (Figure 2.d). The user can also filter by label value by selecting the corresponding check-boxes, and visualize outliers/matched items by selecting the corresponding check-boxes. When the user changes the checked boxes, all the views are simultaneously updated [**R5**].

Finally, we discuss about the two last buttons available in the toolbar (Figure 2.e). The fifth one can be used to set the minimum confidence index, i.e. signals having location accuracy less than the selected value are removed from the map and from the other views [**R5**]. For instance, in Figure 2, all the signals displayed have a location confidence of at least 20% (0.2). The last button enables the users to upload/change the data for analysis.

5.2. Streamgraphs

We represent the time evolution [**R2**] of the number of outbreaks and signals involving the different entities using streamgraphs [26] (Figure 2.b). We use two streamgraphs: the top one contains data from official sources and the bottom one from unofficial ones. Markings on the x-axis of both the graphs are placed at same positions to facilitate comparison.

Entity colors are the same as in the data manager and icon-map (Figure 2.d). Here, again, the user can filter on entity type (*hosts* / *diseases* / *symptoms*) by using the checkboxes. Changing "Sources" has no impact here because the official and unofficial data are already visualized separately.

Since both the streamgraphs are synchronized **[R5]**, hovering the mouse over the streams shows additional information in both of them: the number of outbreaks/signals at the date defined by the x-position of the mouse along with the exact date. This helps in comparing the 2 different types sources over time. For example, in Figure 5, when a user hovers over the streamgraph of official data, labels are displayed in both of them, showing that there were 10 official outbreaks about *Avian influenza* on the 14th of November 2016 whereas there were only 4 unofficial signals on the same date. One can also observe that the stream hovered upon (here in red - matching the entity color) is also highlighted in both the streamgraphs.

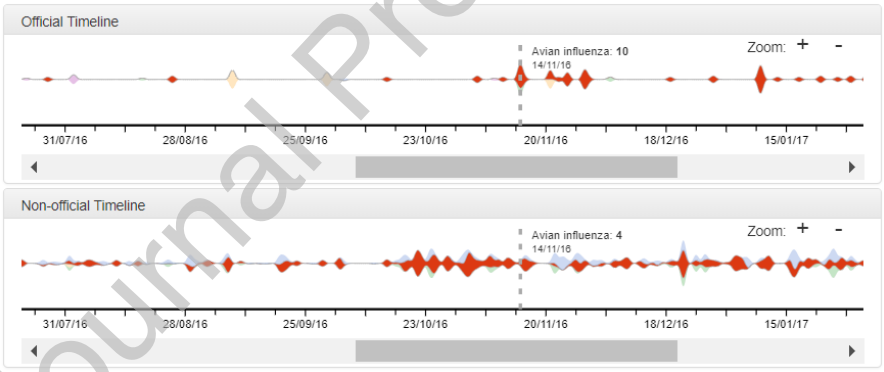

Figure 5: Representing and comparing temporal information on streamgraphs: number of outbreaks/signals about *Avian influenza* in official and unofficial sources between November 2016 and December 2016.

The user can modify the zoom level of a streamgraph by clicking on the dedicated buttons placed conveniently on the top right of each of them. On zooming, the y-axis is not modified but only the x-axis is extended or contracted. A scrollbar at the bottom enables the user to view the parts of the

streamgraph which can fall outside the view when zoomed in. Of course, both the scrollbars are synced with each other and also, changing the zoom level in either of them reflects in both graphs. Hence, the time period covered by both the streamgraphs always remains the same, thereby avoiding any possible mistake by the user during analysis.

The toolbar (Figure 2.e) provides another layer of interactivity between the streamgraphs and the map [**R5**]. The user can activate the lasso functionality by choosing the first option on the toolbar, and then select a custom shaped area on the map. The streamgraphs then show the date of outbreak/signal in this area using vertical lines, thereby allowing the analysis of spatial and temporal information simultaneously (Figure 6). Solid lines show signals which do not match with any official ones [**R4**], dashed lines show other signals. Also, the second tool on the toolbar resets the lasso selection.

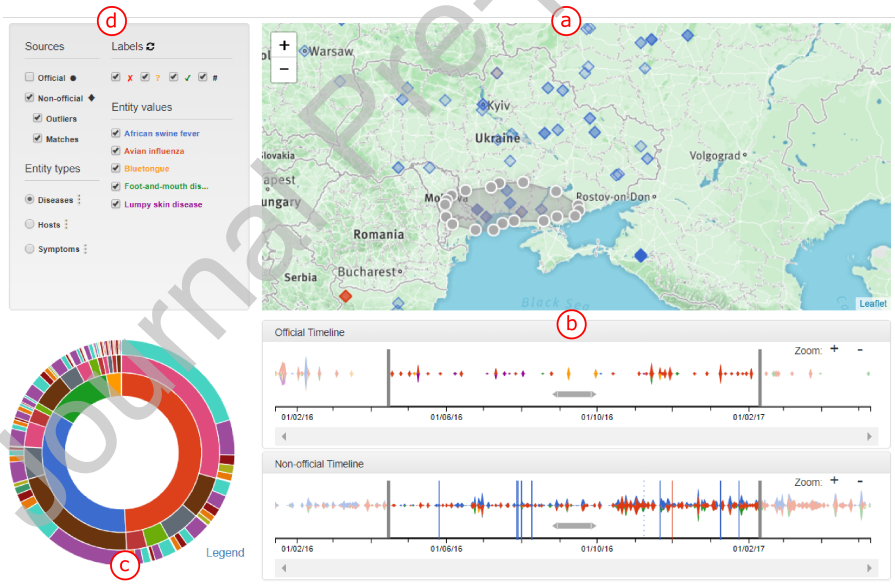

Figure 6: Lasso functionality activated, with an area hosting outliers and matches signals.

Finally, two dark gray vertical lines, initially positioned at the left and right ends of the streamgraphs (Figure 2.b) can be moved to select a specific time

period (see Figure 7.b) for analysis. The selected period (lying between those lines) is highlighted, and the data outside this time period is removed from the other views [**R5**].

5.3. Sunburst

EpidNews summarizes the data by visualizing the hierarchical relationships [**R3**] between the different types of entities, namely *diseases*, *hosts* and *symptoms* on a sunburst[11] [28]. The view includes three hierarchical levels which correspond to this three types of entity. For instance, in Figure 7.c, the innermost ring represents *hosts*, the middle ring represents the *symptoms* and the outermost ring represents the *diseases*. The rings are further divided into arcs, each of which represents an entity value of the corresponding type. The colors of entities match with the other views: map, streamgraphs and data manager. Users can observe the percentage of different combinations by hovering the mouse over these arcs (entities). For example, Figure 7.c shows that 30% of the signals/outbreaks are about *birds* (a *host*), *mortality* (a *symptom*) and *Avian influenza* (a *disease*). The user can also click on an entity to filter the data in the other views [**R3**]. We can also see in Figure 7 that, if a user clicks on the*Avian Influenza's* arc after hovering the mouse over the arcs for *birds*, *mortality*, *Avian Influenza*, the corresponding outbreak or signal will appear in the map and the streamgraphs, while the others will be removed from the map and faded in the streamgraphs [**R5**]. Also if a user clicks on *mortality* (i.e. an entity value of the middle ring), then the outbreak or signal combining *birds* (innermost ring) and *mortality* will be highlighted, without taking into account any constraints on the entity types of the outermost ring, i.e. *diseases*.

The users also have an option of changing the order of the levels in the hierarchy by intuitively dragging the entity types in the data manager [**R5**] one

[11] An alternative to sunburst would be an icicle plot [27], but we preferred the former as it can be seen as a hierarchical extension of a pie chart, popular diagram that people are used to read for assessing proportions.

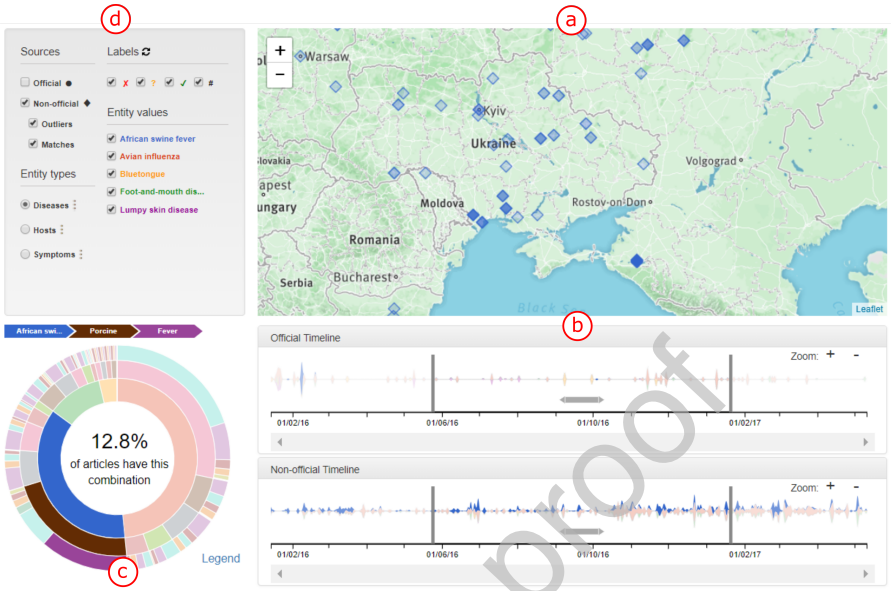

Figure 7: Synchronization between different views of *EpidNews*: (a) map, (b) streamgraphs and (c) sunburst.

below/above the other. For example, entity types in Figure 2.d and Figure 7.d have a different level order. The sunburst in Figure 2.c is ordered by *diseases* in the innermost ring, followed by *hosts* in the middle ring and finally *symptoms* in the outermost ring. This basically visualizes different entity values of *hosts* for various *diseases*, and also shows the *symptoms* observed in each *host*. Here it helps in identifying which disease affected which host, and what symptoms that particular host showed for that disease. On the other hand the sunburst in Figure 7.c is ordered by *hosts*, then *symptoms*, and finally *diseases*.

6. Use case: Epidemic intelligence for monitoring animal diseases

The use case aims to provide an example where one can use the tool *EpidNews* to effectively monitor the emergence and spread of animal diseases. This example helps highlight the usability and the ease that a visualization tool like *EpidNews* brings to the field of epidemiology, specifically in using multi-source

data. This use case was performed by one of the three epidemiology experts involved in the requirements. She decided to select *African swine fever* (*ASF*) as the disease to test this tool since her current research focuses on the same. For the study, the official data is obtained from the Empres-i[12] and ADNS[13] databases. The unofficial data is extracted by *PADI-web* as described in Section 4.2. For both types of sources, the expert extracted the signals and outbreaks reported between the dates 2016-02-10 and 2017-11-10. Depending on the host, official outbreaks have different epidemiological units: outbreak unit for domestic pigs is the pig holding (an outbreak corresponds to several individual cases), whereas for wild boars the reference unit is the animal (individual cases). The aspects tested in this use case are as follows:

- visualizing the official information (outbreaks) about the two distinct hosts (*wild boars* and *domestic pigs*) of the disease *African swine fever*,
- visualizing the different entity types (*diseases*, *hosts* and *symptoms*) extracted from unofficial sources and
- extracting new insights from unofficial data.

6.1. Task 1: Visualizing & analyzing official news

For the first task, the expert selected the relevant disease (*ASF*) and varied the streamgraph slider to observe data in different time periods. She used the heatmap (Figures 8.a, and 8.b) to aggregate the outbreaks, which distinguished the high density areas (in red) from the low density areas (in blue). This allowed the expert to understand the evolution of *ASF* outbreaks in Europe: how it spread from Eastern to Western Europe (in the second map, we notice less outbreaks in Ukraine and, the first occurrence of outbreaks in Czech Republic).

[12] http://empres-i.fao.org/eipws3g/, (accessed on March 14, 2018)

[13] https://ec.europa.eu/food/animals/animal-diseases/not-system_en, (accessed on March 14, 2018)

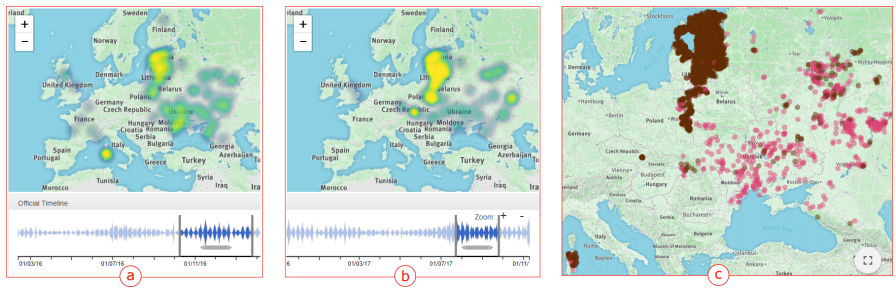

Figure 8: Distribution of *African swine fever* official outbreaks and its hosts using a heatmap and an icon-map. The heatmap represents the outbreaks (a) from 2016/10/10 to 2017/01/20 and (b) from 2017/08/01 to 2017/10/05. The icon-map (c) compares *ASF* outbreaks during the whole studied period with *wild boars* in brown, and *domestic pigs* in pink.

For comparing the spatio-temporal evolution between the two *ASF*'s hosts, the expert used the icon-map which presented *wild boars* in brown and *domestic pigs* in pink (Figure 8.c), making them easily distinguishable. The expert noticed that the *wild boar* cases were predominantly concentrated in the Baltic countries.

Observing locations of the official outbreaks, the expert noticed a potential geographical abnormality: although being constantly surrounded by *ASF* outbreaks, Belarus did not (officially) notify any *ASF* outbreak (Figure 8).

The lasso tool helped the expert in focusing on particular locations independent of countries' frontiers. The dates corresponding to the selected outbreak locations are highlighted with dashed lines in the streamgraph. Selecting a cluster of outbreaks in Western Russia, and then observing their corresponding dates that show up in the streamgraph, suggest that those *ASF* outbreaks were not only related in space but also in time (Figure 9.a.). On the contrary, outbreaks in Sardinia, Italy, regularly occurred during the whole period (Figure 9.b.).

As explained in Section 4.1, both date of occurrence and date of notification can be extracted from official datasets. The expert chose to compare outbreak streams using both the notification date (in blue) and the occurrence date (in

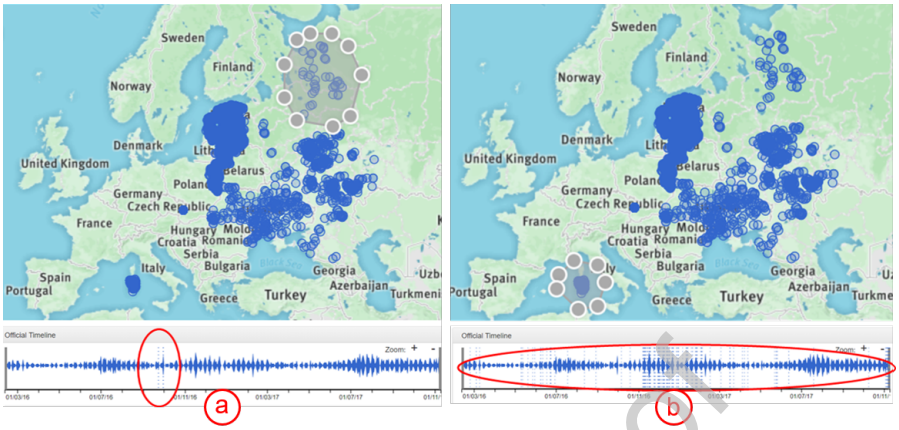

Figure 9: Streamgaph shows the highlighted dates using the lasso tool for the selected locations: (a) in Western Russia and (b) in Sardinia, Italy.

red) in Russia (Figure 10). The visualization of the streams shows a good overlap, but suggests that a cluster of outbreaks which occurred in mid September 2016 were reported almost three weeks later.

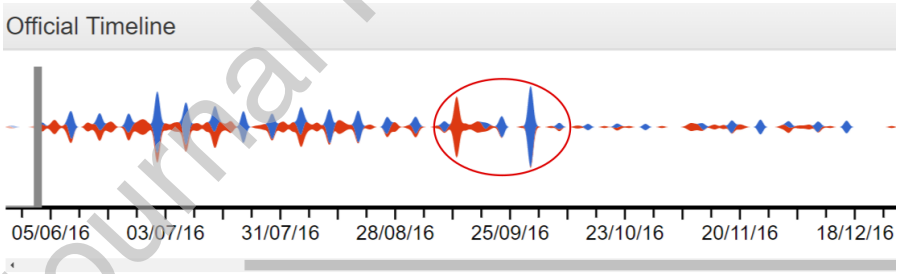

Figure 10: Outbreak occurrence date stream (red) & notification date stream (blue) for *ASF* outbreaks in Russia.

6.2. Task 2: Analyzing different entity types from unofficial signals

Some unofficial news reports can contain data which is not specific to the relevant disease: *ASF* (i.e. information about other *diseases* / *hosts* / *symptoms*).

For instance, almost 76.5 percent of the unofficial signals had the combination of *ASF* and either *domestic pigs* or *wild boars*, which are the *ASF* specific *hosts* (Figure 11.a) and the rest 23.5% had mentions of other diseases and host combinations.

After evaluating the proportion of each combination of *diseases*, *hosts* and *symptoms*, it can be noticed that *fever* and *mortality* are the most common *symptoms* associated with the *ASF* disease (Figure 11.b).

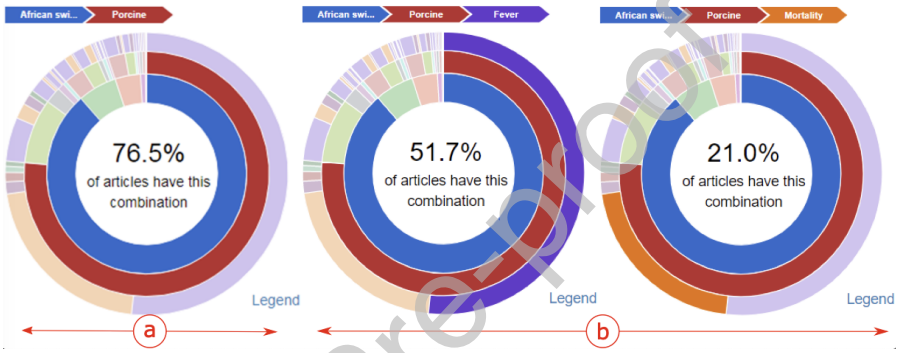

Figure 11: **(a)** Example of *disease/host* combinations in unofficial news sources. **(b)** Example of *disease/host/symptom* combinations in unofficial news sources.

6.3. Task 3: Further insights from the unofficial signals

The expert wanted to find and highlight useful information from unofficial sources, i.e. evaluate the relevance of the unofficial signals. During the studied period, 518 signals were extracted by *PADI-web* (Figure 12.a). By default for this view, all signals having a confidence index lower than 0.2 are hidden, and no filter is used to select either matches or outliers. Evaluating each signal directly from this view is a time-consuming task. To ease this, the expert adopted different strategies involving various features of *EpidNews*.

6.3.1. Subtask 3.1: Investigating outliers

Firstly, the expert wanted to study unofficial signals in an unusual geographic zone, i.e. signals outside a spatio-temporal window around official outbreaks.

These signals are likely to be either false signals, or early warnings. The expert first used the confidence index filter to select signals with confidence index greater than 0.7. Then, she used the sliders to select a maximum distance of 70 km[14] and a temporal window of 10 days[15], thereby displaying the outliers on the map (Figure 12.b).

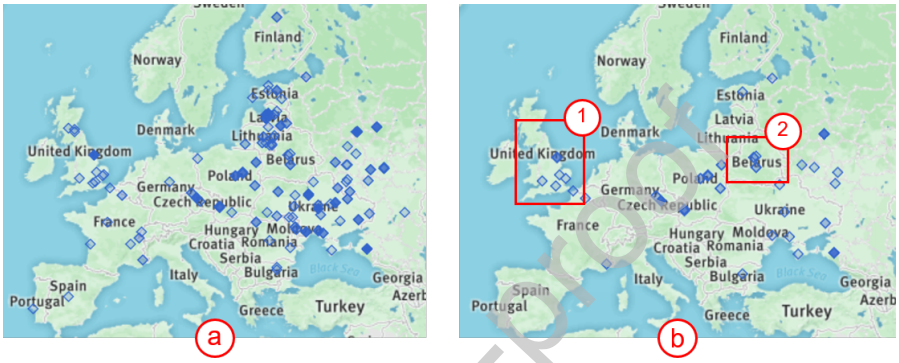

Figure 12: Unofficial signals for ASF disease **(a)** without filtering **(b)** using filtering tools (outliers with a confidence index >0.7). **(1)** and **(2)** are the two zones investigated by the expert.

The expert chose two zones with opposite status: United Kingdom (zone **(1)**), which is totally free from *ASF*, and Belarus (zone **(2)**), which did not officially notify any outbreak despite of being surrounded by affected countries.

The Figure 13 shows the workflow followed by the expert to evaluate the signals from Belarus. The expert clicked on the first glyph to show the list of signals and their details (Figure 13.a). By default, all the signals were labelled as *not verified*. For each signal, the expert displayed the news textual content to evaluate its relevance. It was found that all the signals were related to *ASF*,

[14]On the basis of the mean radius of clusters of African swine fever cases in wild boars in the Russian Federation, estimated to 74,1 km [29].

[15]On the basis of the maximal delay between detection dates and notification dates for African swine fever outbreaks in Eastern Europe, according to official data from Empres-i and ADNS.

but three of them were not related to an outbreak in Belarus: two news articles were about the pork ban imposed by Belarus due to the *ASF* outbreaks in the bordering countries and one was about the creation of a pig breeding center. The expert labelled all those three signals as *irrelevant*. The latest article was a relevant news report which referred to outbreaks not reported by official sources (Figure 13.b). As this signal could not be confirmed, but was a potential early alert, the expert labelled it as *unknown* (Figure 13.c).

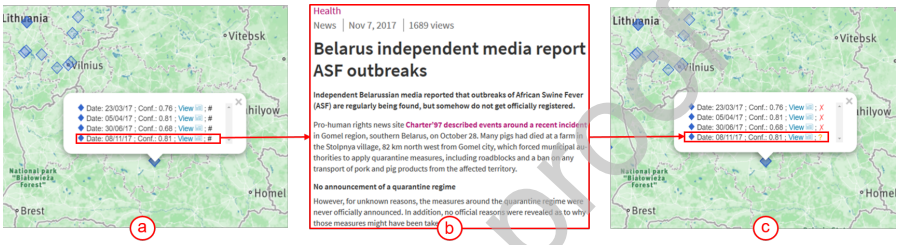

Figure 13: Workflow for evaluation of four unofficial signals for *ASF* in Belarus (a) visualisation of the signal details, (b) visualisation of source document content (view from the original website), (c) annotation of the signal.

The expert followed the same workflow to label signals from the United Kingdom. Contrary to signals from Belarus, some signals were not specific to the geographical zone on which they were displayed or not specific to *ASF*. For all those signals, the expert easily traced the source of error by viewing the article contents. Indeed, three signals were errors due to mentions of UK in news reports describing outbreak in eastern Europe. One signal was about the detection of antibiotic-resistant *E.coli* in a UK supermarket and contained a sum-up of on-going *ASF* outbreaks. The expert eventually labelled all the UK's signals as *irrelevant*.

6.3.2. Subtask 3.2: Linking unofficial and official data

The expert then evaluated signals in Ukraine. The context is different from the two previous zones, since the country is officially affected by *ASF* and a lot of outbreaks were reported. For this subtask, the expert wanted to link

unofficial information with the official one, in order to validate the relevance of unofficial signals. Then, contrary to the previous subtask, she selected the matches, i.e. the unofficial signals occurring inside a specific spatio-temporal window (using the same parameters as in Section 6.3).

Figure 14 shows the workflow followed by the expert, with the example of two signals occurring in Chernivestka region, published on June 09, 2016 and on June 10, 2016 (Figure 14.a). By reading the news reports contents, the expert noticed that both signals described the same outbreak, involving 700 pigs among which 50 had died. The expert labelled the signals as *unknown*. Then, she added the official outbreaks to the view in order to find the related one. As a lot of outbreaks were notified during the whole studied period, the view was overloaded by circles (Figure 14.b). The timeline allowed the expert to restrict the temporal window around June 2016, which filtered most of official outbreaks. She easily visualized an official outbreak closed to the unofficial signals (Figure 14.c). The expert read its content and found that it corresponded to the outbreak described by unofficial signals. Then, the expert changed the label from *unknown* to *relevant*.

6.3.3. Subtask 3.3: Annotation of the whole dataset

The expert used the combination of the workflows described in Section 6.3 to evaluate the 425 unofficial signals which had a confidence index greater than 0.7. During its evaluation, the expert labelled 213 signals as *irrelevant*, 170 as *relevant* and 42 as *unknown*. As the dataset could be saved/loaded, she could pause her work and resume later. The final labelled dataset allowed to filter the signals, to keep only the relevant ones for instance (Figure 15).

6.4. Concluding the use case

EpidNews proved to be of immense help in analyzing the spatio-temporal aspects of both official and unofficial data. The expert found it useful to visualize and easily interact with official data, for instance to locate clusters of outbreaks with a potential epidemiological link. The visualization also helped

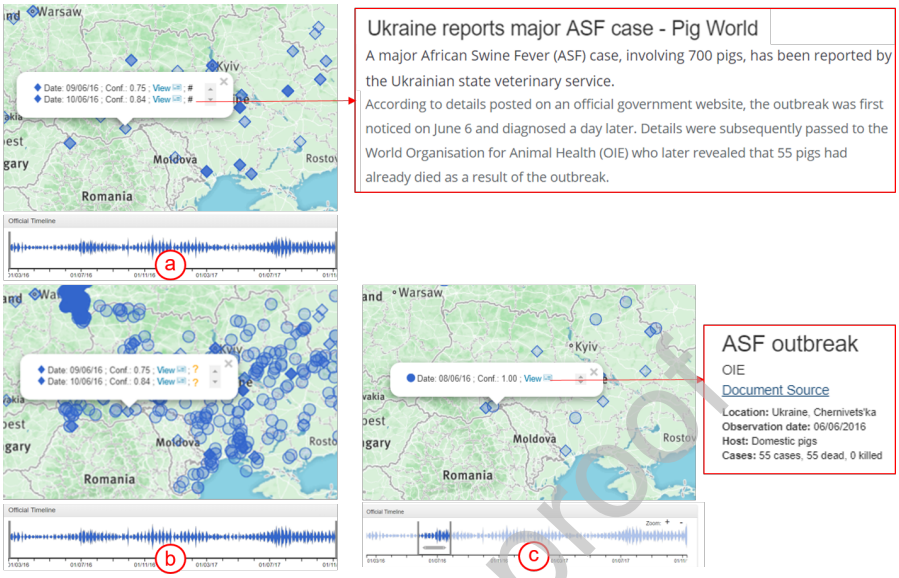

Figure 14: Workflow for evaluation of an unofficial signals for *ASF* in Romania (a) visualisation of the list of signals and their content, (b) visualisation adding the official outbreaks, (c) visualisation selecting a specific time window with the timeline. The texts displayed are the content of 2nd unofficial signal (above) and the content of the official outbreak (below).

detect bias in the dataset, like the potential absence of official notifications in some areas. Regarding unofficial data, it allowed to visualize the combination of three important entities in animal epidemiology: *diseases*, *hosts*, and *symptoms*. The tool was helpful in evaluating the relevance of unofficial signals, for instance in detecting geographical ambiguities and potential outbreaks which are not detected by official sources. The possibility to filter signals using their spatio-temporal distance with official data allowed to efficiently evaluate them. The expert found it very convenient to label the relevance of signals once evaluated and to export the processed dataset. Therefore, this tool should be used in the daily monitoring and analysis of both official and unofficial sources, thereby reducing the manual work of the epidemic intelligence team.

The expert in her evaluation mentioned that colors can sometimes be meaningful in the epidemiological context. For instance, *domestic pigs* are naturally

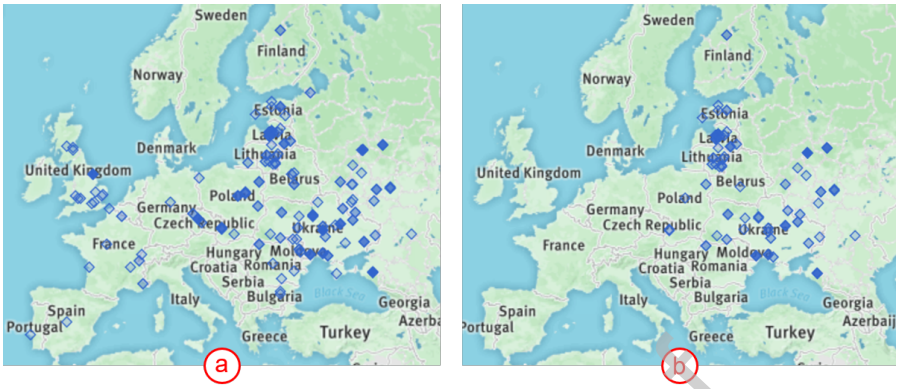

Figure 15: Visualisation of unofficial signals for *ASF* (a) before the annotation, (b) after the annotation, selecting only the relevant signals.

associated with the color pink and *wild boars* with brown. However, currently the tool ignores these color associations and uses the most visibly distinctive colors. In fact, automatically assigning a specific color to a term is complex because the choice partly relies on the socio-cultural context [30]. This limitation can be addressed by allowing the user to select or change the color for terms himself but also ensuring distinctive colors. The expert also suggested to cluster the glyphs of the icon-map to prevent cluttering issues when zoomed out beyond a certain level.

7. Conclusion and Future work

In this paper, we presented *EpidNews*, a new visual analytics tool for spatio-temporal data which can be used for monitoring animal diseases. It combines several views, map, streamgraphs and sunburst, that help the users to manage outbreak data from multiple sources and aggregate this information on different levels of abstraction. Thus, it allows the experts to easily analyze and track diseases outbreaks over different geographic regions and time intervals. Moreover, *EpidNews* was tested by an expert in epidemiology and the resulting use case is presented in the last section of the paper, which highlights the benefits of the

tool.

In future, we firstly plan to address the two issues raised towards the end of Section 6.4. Secondly, our tool will be extended in order to cover other application domains as well, like environmental threats, etc. Eventually, we also aim to include other data representations and offer a dynamic view for the spread of a disease.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the Ministry of Higher Education and Scientific Research of Algeria, the SONGES project (FEDER and Occitanie) and the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004. We thank all the expert teams & gatekeepers of this project. We thank the French Epidemic Intelligence System team for providing the data from ADNS database.

References

[1] R. Goel, S. Fadloun, S. Valentin, A. Sallaberry, M. Roche, P. Poncelet, EpidNews: An epidemiological news explorer for monitoring animal diseases, in: Proceedings of the 11th International Symposium on Visual Information Communication and Interaction (VINCI'18), 2018, pp. 1–8.

[2] S. Cleaveland, M. K. Laurenson, L. H. Taylor, Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence., Philosophical Transactions of the Royal Society B: Biological Sciences 356 (1411) (2001) 991–999.

[3] J. Cortiñas Abrahantes, A. Gogin, J. Richardson, A. Gervelmeyer, Epidemiological analyses on african swine fever in the baltic countries and poland, EFSA Journal 15 (3) (2017) 1–73.

[4] J. Choi, Y. Cho, E. Shim, H. Woo, Web-based infectious disease surveillance systems and public health perspectives: a systematic review, BMC Public Health 16 (1) (2016) 1238.

[5] 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, 2011.

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

[7] 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, Computers and Electronics in Agriculture 123 (2016) 104–115.

[8] E. H. Chan, T. F. Brewer, L. C. Madoff, M. P. Pollack, A. L. Sonricker, M. Keller, C. C. Freifeld, M. Blench, A. Mawudeku, J. S. Brownstein, Global capacity for emerging infectious disease detection, Proceedings of the National Academy of Sciences 107 (50) (2010) 21701–21706.

[9] K. A. Cook, J. J. Thomas, Illuminating the path: The research and development agenda for visual analytics, IEEE Computer Society Press, 2005.

[10] F. Claes, D. Kuznetsov, R. Liechti, S. Von Dobschuetz, B. Dinh Truong, A. Gleizes, D. Conversa, A. Colonna, E. Demaio, S. Ramazzotto, et al., The

EMPRES-i genetic module: a novel tool linking epidemiological outbreak information and genetic characteristics of influenza viruses, Database 2014.

[11] C. Freifeld, K. Mandl, B. Reis, J. Brownstein, Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports, Journal of Medical Informatics Association 15 (2) (2008) 150–157.

[12] D. J. Peuquet, It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems, Annals of the Association of American Geographers 84 (3) (1994) 441–461.

[13] M. Ward, G. Grinstein, D. Keim, Interactive data visualization: foundations, techniques, and applications, A K Peters, 2010.

[14] W. Aigner, S. Miksch, H. Schumann, C. Tominski, Visualization of time-oriented data, Springer Science & Business Media, 2011.

[15] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, D. Keim, Eventriver: Visually exploring text collections with temporal references, IEEE Transactions on Visualization and Computer Graphics 18 (1) (2012) 93–105.

[16] N. Ferreira, J. Poco, H. T. Vo, J. Freire, C. T. Silva, Visual exploration of big spatio-temporal urban data: a study of new york city taxi trips, IEEE Transactions on Visualization and Computer Graphics 19 (12) (2013) 2149–2158.

[17] I. Cho, W. Dou, D. X. Wang, E. Sauda, W. Ribarsky, Vairoma: A visual analytics system for making sense of places, times, and events in roman history, IEEE Transactions on Visualization and Computer Graphics 22 (1) (2016) 210–219.

[18] G. Sun, R. Liang, H. Qu, Y. Wu, Embedding spatio-temporal information into maps by route-zooming, IEEE Transactions on Visualization and Computer Graphics 23 (5) (2017) 1506–1519.

[19] H. Rosling, Z. Zhang, Health advocacy with gapminder animated statistics, Journal of epidemiology and global health 1 (1) (2011) 11–14.

[20] W. Van den Broeck, C. Gioannini, B. Gonçalves, M. Quaggiotto, V. Colizza, A. Vespignani, The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale, BioMed Central infectious diseases 11 (1) (2011) 37.

[21] C. Dunne, M. Muller, N. Perra, M. Martino, Vorograph: Visualization tools for epidemic analysis, in: Proceedings of the Conference on Human Factors in Computing Systems (CHI'15), 2015, pp. 255–258.

[22] S. Fadloun, A. Sallaberry, A. Mercier, E. Arsevska, M. Roche, P. Poncelet, EpidVis: a visual web querying tool for animal epidemiology surveillance, Information Visualization to appear.

[23] 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 Quarterly 92 (1) (2014) 7–33.

[24] D. Ahlers, Assessment of the accuracy of geonames gazetteer data, in: Proceedings of the 7th Workshop on Geographic Information Retrieval (GIR '13), 2013, pp. 74–81.

[25] Y. Liu, J. Heer, Somewhere over the rainbow: An empirical assessment of quantitative colormaps, in: Proceedings of the Conference on Human Factors in Computing Systems (CHI'18), 2018, pp. 598:1–598:12.

[26] L. Byron, M. Wattenberg, Stacked Graphs - Geometry & Aesthetics, IEEE Transactions on Visualization and Computer Graphics 14 (6) (2008) 1245–1252.

[27] J. B. Kruskal, J. M. Landwehr, Icicle plots: Better displays for hierarchical clustering, The American Statistician 37 (2) (1983) 162–168.

[28] R. O'Donnell, A. Dix, L. J. Ball, Exploring the pietree for representing numerical hierarchical data, in: Proceedings of the HCI'06 Conference on People and Computers XX, Springer, 2007, pp. 239–254.

[29] I. Iglesias, M. J. Muoz, F. Montes, A. Perez, A. Gogin, D. Kolbasov, A. de la Torre, Reproductive ratio for the local spread of African swine fever in wild boars in the Russian federation, Transboundary and Emerging Diseases 63 (6) (2014) e237–e245.

[30] C. Ware, Information Visualization: Perception for Design, Morgan Kaufmann Publishers Inc., 2000.