



HAL
open science

An evaluation framework for comparing epidemic intelligence systems

Nejat Arinik, Roberto Interdonato, Mathieu Roche, Maguelonne Teisseire

► **To cite this version:**

Nejat Arinik, Roberto Interdonato, Mathieu Roche, Maguelonne Teisseire. An evaluation framework for comparing epidemic intelligence systems. *IEEE Access*, 2023, 11, pp.31880-31901. 10.1109/ACCESS.2023.3262462 . hal-04049825

HAL Id: hal-04049825

<https://hal.science/hal-04049825v1>

Submitted on 30 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

An evaluation framework for comparing epidemic intelligence systems

Nejat Arinik, Roberto Interdonato, Mathieu Roche & Maguelonne Teisseire

March 30, 2023

Abstract

In the context of Epidemic Intelligence, many Event-Based Surveillance (EBS) systems have been proposed in the literature to promote the early identification and characterization of potential health threats from online sources of any nature. Each EBS system has its own surveillance definitions and priorities, therefore this makes the task of selecting the most appropriate EBS system for a given situation a challenge for end-users. In this work, we propose a new evaluation framework to address this issue. It first transforms the raw input epidemiological event data into a set of normalized events with multi-granularity, then conducts a descriptive retrospective analysis based on four evaluation objectives: spatial, temporal, thematic and source analysis. We illustrate its relevance by applying it to an Avian Influenza dataset collected by a selection of EBS systems, and show how our framework allows identifying their strengths and drawbacks in terms of epidemic surveillance.

Keywords: Epidemic Intelligence, Event-Based Surveillance, Retrospective Analysis.

Cite as: N. Arinik, R. Interdonato, M. Roche & M. Teisseire. An evaluation framework for comparing epidemic intelligence systems. *IEEE Access*, 2023. DOI: [10.1109/ACCESS.2023.3262462](https://doi.org/10.1109/ACCESS.2023.3262462)

1 Introduction

At least 60% of infectious human diseases originated in animals¹. The emergence and spread of any animal infectious disease, such as Avian Influenza, has serious consequences for animal health and a substantial socio-economic impact for agriculture. For instance, the 2021–2022 season have experienced the largest observed highly pathogenic avian influenza (HPAI) cases in Europe so far, with a total of 2,467 outbreaks in poultry, 3,573 HPAI events in wild birds, and 48 million birds culled in the affected establishments². Due to this highly contagious nature, it is critical to monitor new and emergent infectious animal diseases. To this aim, epidemic intelligence has been used to remedy this public health issue.

Traditionally, a public health surveillance system has long used Indicator-Based Surveillance (IBS) for a global epidemic monitoring approach, the well-known ones being the World Organisation for Animal Health (WOAH)³ and the Food and Agriculture Organization of the United Nations (FAO)⁴. This type of surveillance consists in collecting structured and verified official health threats, hereafter referred to as epidemiological events (or events for short), through routine national surveillance systems and public health authorities. However, IBS typically undergoes some reporting delay in the detection of these data, as it relies only on laboratory confirmed animal cases. To improve this timeliness issue, several Event-Based Surveillance (EBS) systems have been proposed with the aim of promoting the early identification and characterization of potential epidemiological events from online sources of any nature, including online news outlets and social media, thanks to the recent developments in internet and digital technologies [78]. Recently, several EBS platforms have shown their effectiveness by detecting the first signals of emerging infectious disease outbreaks in a timely manner and providing alerts within previously unaffected areas (e.g. [67]).

In the literature, there exist two categories of EBS systems by their functioning nature: 1) moderated (i.e. human-curated) and 2) automated. The first type of systems are human-curated ones that rely on pure manual data collection and analysis. The data can be provided by official or unofficial data sources, but in any case their accuracy is manually assessed by moderators. The Program for Monitoring

¹www.cdc.gov/onehealth/basics/zoonotic-diseases.html

²www.ecdc.europa.eu/en/news-events/2021-2022-data-show-largest-avian-flu-epidemic-europe-ever

³www.woah.org

⁴www.fao.org

Emerging Diseases (ProMED) is such an example of a moderated system [11]. The second type of systems differs from the first one in that it includes in some or all of their pipelines automated text-mining based steps for data collection and processing. Furthermore, automated systems are also categorized into semi- and fully-automated systems. The main difference between them is that the former includes a dedicated team of curators to assess and verify the outputs, whereas the latter does not. An example of semi-automated system is the Canadian Public Health Agency Global Public Health Intelligence Network (GPHIN) [46]. Likewise, fully-automated systems include BioCaster [14, 45], HealthMap [22], MediSys [39], PADI-web [70], DANIEL [35], Sentinel [79] and EpiTweetr [19].

Each EBS system has its own priorities (e.g. geography, disease) and surveillance definitions (e.g. collected epidemiological information), so there is no such candidate as a *best* EBS system, that would fit all situations. However, due to the profusion of available EBS systems, selecting the most appropriate one(s) for an effective surveillance system of a given situation is a challenge for end-users. Some existing works try to compare them according to the guideline of the Centers for Disease Control and Prevention (CDC) [6, 63], but they either focus only on few evaluation aspects or require human resources for manual assessment, which brings some cost to practitioners. Furthermore, there exist many other studies conducting a retrospective analysis using surveillance dataset, without any objective of comparison. These works deal with additional evaluation points that are not considered in the CDC's guideline, which would bring valuable additional information for evaluation purposes.

In this work, we propose a new automatic evaluation framework to solve all these issues. It is based on four evaluation objectives: 1) spatial analysis (how the events are geographically distributed), 2) temporal analysis (how the events evolve over time and what temporal aspects characterize it), 3) thematic entity analysis (what thematic entities are extracted from the events and how they are related to spatio-temporal analysis) and 4) news outlet analysis (what news sources play key role in epidemiological information dissemination). For each aspect, we compare the obtained results with a reference gold standard database, along with an appropriate visualization for end-users. All these analyses aim to highlight the strengths and drawbacks of the considered EBS systems in terms of epidemic surveillance. We illustrate its relevance by applying it to a selection of EBS systems. Our main contribution is essentially threefold. First, we propose a generic evaluation framework, which is not tied to any specific disease, geographical region, or surveillance definition, so it can be applied to any situation, as long as we have access to a gold standard database. Second, we model the studied epidemiological events with multi-granularity in order to better understand the spatial and temporal evolution of disease events, as well as their thematic characterization. Third, we take into consideration in our framework the fact that there exist some gaps between EBS systems in disease detection and collection, an issue so-called *reporting bias* [28, 29].

The rest of the article is organized as follows. First, in Section 2, we review the literature on EBS systems, focusing on different evaluation strategies. Next, in Section 3, we introduce our evaluation framework designed to study and compare EBS systems and their outputs. We put it into practice on a selection of EBS systems in Section 4 and discuss these results in Section 5. Finally, we review our main findings in Section 6, and identify some perspectives for our work.

2 Related Work

In this section, we review the existing evaluation strategies for EBS systems. The performance assessment of these systems are traditionally performed according to the CDC's guideline, which aims at understanding the internal and external performances of EBS systems [6, 63]. Nevertheless, most of these evaluation metrics are more in line with an end-user perspective, which require human resource for manual assessment.

On the other hand, there exist many studies which conduct a retrospective analysis using surveillance dataset, i.e. the output of an IBS/EBS system, without performing any comparative study. These works deal with additional evaluation points that are not considered in the traditional evaluation methods [5, 6]. In particular, we are interested in those works performing a descriptive analysis, rather than predictive analysis, which is in line with our work. For this reason, we widen the scope of our review with these works.

In the following, we overview the existing works in four parts: 1) Spatial (Section 2.1), 2) temporal (Section 2.2), 3) thematic (Section 2.3) and 4) source (Section 2.4) dimensions. Note that although the surveillance data is naturally spatio-temporal, we review each dimension separately for the sake of clarity.

2.1 Spatial dimension

The spatial dimension is the most studied dimension in the existing works. We summarize these works in two aspects: 1) geographic coverage and 2) hotspot analysis. The most widespread evaluation analysis is the assessment of geographic coverage of the surveillance data, and it is often time-invariant. This geographic coverage is calculated for either the whole world [10, 25, 29, 42] or some particular regions/countries [49]. Hence, this analysis allows showing to what degree the locations (e.g. countries) are covered by the data at hand. Moreover, it can be used to manually identify the events appearing in an unusual geographic zone in the context of early warning detection [25].

Another analysis for the spatial dimension is hotspot analysis. The hotspots are the areas, where a substantial number of events are concentrated over time. The task of identifying the hotspots is also referred to as *outbreak detection* in the literature of Epidemic Intelligence. Such hotspots are usually found through three different approaches. The first one is the exponentially weighted average method [22] by assigning large values to more recent alerts coming from multiple sources through the decay parameter of the exponential weight. The second one is the spatial auto-correlation analysis, which statistically identify the hotspots [23, 43, 53, 72]. The most used technics are the Moran's I [47] and the Getis-Ord G_i^* [24]. The last approach is spatio-temporal clustering analysis, which aims at determining regions where the number of events is significantly higher than expected. Space-time scan statistics [74] and ST-DBSCAN [9] are two such well-known clustering methods.

In our work, we only include the geographic coverage-based assessment. This is because we want to evaluate the epidemiological information collected by EBS systems at fine-grained level. This is only possible at event level, rather than at outbreak level. Nonetheless, as in hotspot analysis, we take the temporal aspect into account by adapting the traditional calculation of geographic coverage.

2.2 Temporal dimension

The temporal dimension is another important aspect in a retrospective analysis of surveillance data. Since our goal is to perform a descriptive analysis, in the following we focus only on it with two aspects: qualitative vs. quantitative assessments.

There are two main approaches for the qualitative assessments to describe the temporal evolution of the events. The most widespread approach is trend analysis to capture underlying temporal features in time-series event data. This includes methods that can identify discriminatory information about a particular time-series data (e.g., shapelets [26]), those that look for temporally frequent sub-sequences that occur in a majority of time-series (e.g., temporal patterns [48]), and those that investigate on seasonal [44] and periodic [31] effects. The second approach is to identify anomalous cases in a time-series [62]. These anomalous cases can represent either the locations having a significantly high infection cases (e.g. outbreaks) [79] or the locations that have remarkably different infection history than neighbor locations (e.g. potential early signals) [25].

Another evaluation analysis is through the quantitative assessments. The most widespread analysis is using the concept *timeliness* in order to evaluate how timely the events are detected by an EBS system [2, 4, 6, 52, 63, 71]. If an EBS system reports the events in a timely manner, this would allow public authorities to mitigate potentially dangerous situations as soon as possible. Another approach for quantitative assessments aims at evaluating how two time series data are in a similar trend. This evaluation is usually done with the correlation analysis between the daily or weekly event time series derived from IBS/EBS systems using Pearson's correlation coefficients [7]. The final approach relies on the concept of *transmissibility*. It is used to quantify how easily a disease can spread through a population, i.e. how rapidly an outbreak is growing or declining. It can be measured by estimating the basic [16], effective or time-varying reproduction numbers [8, 15, 50].

Regarding the connection with our work, we include a qualitative assessment based on frequent temporal events in time-series data, and we perform it with an appropriate frequent spatio-temporal pattern mining method. Moreover, we include a quantitative assessment based on timeliness. However, we do not include the other works presented above for the following reasons. First, EBS systems rely on unofficial data sources, therefore false alerts might be introduced in the data. This requires to handle it with a specific method, which is out of scope in this work. Second, each EBS system collects epidemiological data of different size, and their differences can be substantial. In which case, calculating the correlation coefficient of two time series data, each associated with a different source, can be biased towards the most populated source. Finally, estimating the transmissibility and the reproduction number are disease-dependent. This requires to develop a different model for each disease, which is also out of scope in this work.

2.3 Thematic dimension

The thematic dimension is not always well elaborated in the existing surveillance systems. This is probably because the collected events are characterized by spatio-temporal attributes in practice, rather than their thematic attributes (e.g. disease and host) [69]. Therefore, to the best of our knowledge, there exist only a handful of works for evaluating the thematic dimension of the existing EBS systems.

All the existing works in the literature are interested only in the ranking of thematic entities. This ranking can be obtained with the frequency [25], a statistical measure (e.g. F-measure [59, 70], chi-square [77]) and a constraint based objective (e.g. temporal periodicity [38]). In all these works, there are two factors, which directly affect the ranking results. The first one is related to the multidimensionality nature of the elements, for which the ranking is computed. In case of two or more dimensions, this corresponds to the identification of co-occurrences in the same events. The second factor is related to the normalization of thematic entities, i.e. how they are individually expressed for comparison purposes. This normalization step consists in transforming a raw text into one of well-defined taxonomy classes, which results in hierarchical information. In the literature, most of the works focus only on one fixed [45] or a few [69, 70] hierarchical levels.

In this work, we also include the assessment based on the ranking of thematic entities. We use the combination of all mentioned approaches: frequent pattern mining, F-measure [59, 70] and temporal periodicity [38]. Moreover, the thematic elements are normalized with multi-granularity. Finally, we take temporal periodicity into account in the ranking results.

2.4 Source dimension

All EBS systems partially or completely rely on various online news and press agencies, news outlets for short, for ensuring their monitoring of emerging infectious diseases across the world. Nevertheless, there are not enough studies that characterize and assess the news sources involved in EBS systems. The existing works study these sources at two different levels: news aggregator and news outlet levels.

On the one hand, the first level aims to assess the degree to which news aggregators contribute to the news collected by EBS systems. Lyon *et al.* [42] show based on the main EBS systems that the most contribution is provided by Google News, then to a lesser extent ProMED, MeltWater and Baidu. On the other hand, the second level focuses on how countries are covered by the news outlets at hand. [1, 63] show that international news outlets do not capture well news infection events occurring in some less-developed regions, which results in a reporting bias. In which case, local news outlets performs better, because these events are mostly reported in local television or recorded in local print media in local or regional languages. Finally, the news aggregators and news outlets are inherently related and dependent to each other, if an EBS system collects its news data from news aggregators. The authors of [68] analyze this aspect with a network analysis approach by describing how outbreak-related information disseminates from a news outlet to a news aggregator.

In this work, we only analyze the publishing sources at news outlet level. This is because not all EBS systems rely on multiple news aggregators (e.g. PADI-web). Unlike the existing works, we rely on a ranking based assessment of news outlets with two different objectives: importance and timely detection.

3 Evaluation Framework

In this section, we describe the framework that we propose to evaluate and compare a number of EBS systems based on the epidemiological data that they collect. Our goal here is to highlight the strengths and drawbacks of the considered EBS systems in terms of epidemic surveillance. Put differently, we want to know what we lose when we monitor a number of high-threat diseases with a single EBS platform, while there might be some different epidemiological information captured by other EBS systems.

To this aim, we propose a two-step pipeline approach, which is illustrated in Figure 1. The input of the pipeline is a set of unnormalized events, accompanied by the associated news documents. Since each EBS system can collect and extract epidemiological information from online sources in a different way, the first step is to extract the normalized events from the input. We detail this step in the Appendix (Section B), for space matters. Then, the second step consists in performing a retrospective analysis of these events with four objectives: 1) spatial, 2) temporal, 3) thematic and 4) source dimensions. Each dimension allows answering a question that naturally arises in our analysis, and it is implemented through a well-known existing tool deemed appropriate for this purpose. Our methodological contribution is found in the combination of these tools. In the rest of this section, we describe the different steps of our framework in detail.

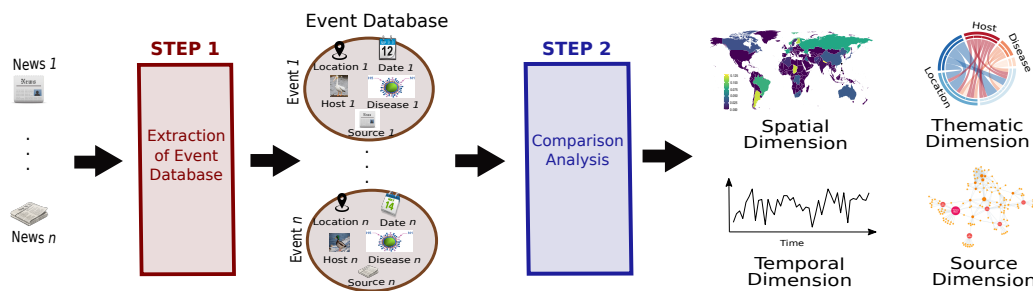


Figure 1: Workflow of our evaluation framework.

Location	Date	Disease	Host	Source
Skelmersdale	31-03-2021	H7N9 serotype	captive bird	Liverpool Echo

Table 1: The corresponding event for the text reported from Liverpool Echo⁵: "A highly pathogenic strain of bird flu (H7N9) has been detected in two captive birds of prey at a private property in Skelmersdale on March 31."

In the following, we first describe in Section 3.1. how we define an event in our context, as well as event related definitions and notations. Second, we explain how to identify common events across IBS/EBS platforms in an automatic manner, a task that we call *event matching* (Section 3.2). Then, we evaluate an EBS system based on the spatial (Section 3.3), temporal (Section 3.4), thematic (Section 3.5) and source (Section 3.6) dimensions. Except the last dimension, these evaluations are always performed with respect to a gold standard database, which is supposed to contain all events confirmed and notified by national and supranational authorities. Moreover, we take into consideration the fact that there exist some gaps between EBS systems in disease detection and collection, an issue so-called *reporting bias* [28, 29]. To do so, our evaluation relies on either ranking results (see Section C.1 in the Appendix) or the concept of representativeness (e.g. Section 3.3).

3.1 Definitions and Notations Related to Events

An event definition depends on the application at hand, and there is no unified standard. In the literature, an event is minimally defined as a disease-location pair, and associated with an infection time (or time period) [35, 70]. Although this minimal definition shows to what degree the locations are covered by the data at hand over time for a particular disease, the other works expand this definition with 1) the news outlets by which news documents are found [42, 70] and 2) thematic information (e.g. disease serotype, hosts, symptoms) [14, 45, 46, 61, 70, 79]. Although the extracted thematic information can be very rich, depending on a system, relating thematic entities to the events can be challenging. This is because there can be multiple events in the same news document, even in the same sentence. Based on these previous event definitions, we define an event throughout this work as the detection of the virus for a specific host at a specific date and in a specific location. Moreover, we also consider the fact that an event is reported by a news outlet. We illustrate in Table 1 how we define an event from the following text: "A highly pathogenic strain of bird flu (H7N9) has been detected in two captive birds of prey at a private property in Skelmersdale on March 31."

Next, we introduce the definitions and notations related to events. An event database \mathcal{E} is a finite set of events collected by an EBS system. Let $\mathcal{D} = \{D_1, \dots, D_n\}$ be a set of dimensions to define the events in \mathcal{E} . Every event is expressed as a tuple $e = (d_1, \dots, d_n)$, where we call d_i event attribute for every $i = 1, \dots, n$. Concretely, in this work, the set \mathcal{D} contains five dimensions, and it is defined as $\{D_Z, D_T, D_D, D_H, D_S\}$. The dimension D_Z is the location, where disease events have been occurred, and they are expressed as polygons (e.g. country or city polygons). The dimension D_T is the notification date of events and it is a totally ordered domain. Moreover, the dimension D_D is the disease which infects a number of hosts. The dimension D_H is the host who have caught the viruses of a specific disease. Finally, the dimension D_S is the news outlet publishing a given epidemiological event.

Each dimension D_i is associated with a domain of (discrete) values, denoted by $dom(D_i)$. Given an event database \mathcal{E} over \mathcal{D} , for every $i = 1, \dots, n$, we denote by $Dom_{\mathcal{E}}(D_i)$ (or simply $Dom(D_i)$ if \mathcal{E} is clear from the context), the active domain of D_i in \mathcal{E} , which corresponds to the set of all values

⁵www.liverpoolecho.co.uk/news/liverpool-news/highly-pathogenic-bird-flu-detected-20309813

Level	Location (D_Z)	Date (D_T)	Disease (D_D)	Host (D_H)	Source (D_S)
0	ALL_Z	ALL_T	ALL_D	ALL_H	ALL_S
1	Europe	2021	avian flu	bird	Liverpool Echo
2	United Kingdom	03-2021	highly patho.	captive bird	
3	England	week 13	H7N9		
4	Lancashire	31-03-2021			
5	West Lancashire				
6	Skelmersdale				

Table 2: Hierarchical event representation for the event illustrated in Table 1.

Time Interval	Country
week 1	France, Italy, China, India
week 2	France, Italy, Spain, China, India, Nepal
week 4	France, Spain, Portugal, India, Nepal
week 6	Spain, Portugal, India
week 7	Spain, Portugal, India
week 8	Portugal, India, Pakistan
week 10	India, Pakistan
week 11	Italy, India, Pakistan

Table 3: Illustrative example of $\mathcal{E}_{l_T}^{l_Z}$, where $l_Z = \text{country}$ and $l_T = \text{week}$. The first column describe the time intervals, for which epidemiological events occur, and the second column indicates the countries, in which epidemiological events occur. Note that $\mathcal{E}_{l_T}^{l_Z}$ contains only the country information, even though more specific spatial information can be present in the data.

of $Dom(D_i)$ occurring in \mathcal{E} . In this work, we consider only values in active domains. Moreover, we assume that each dimension $D_i \in \mathcal{D}$ is associated with a hierarchy, denoted by H_i , in order to consider different granularity levels of domain values. Every hierarchy H_i is a tree whose nodes are elements of $Dom(D_i)$ and whose root is ALL_i . For instance, for the spatial dimension D_Z , ALL_Z corresponds to the whole world containing all existing locations. We illustrate in Table 2 the hierarchical representation of the event from Table 1.

Moreover, we sometimes need to fix the spatial and temporal scales (i.e. hierarchical levels) of an event database \mathcal{E} . This operation amounts to discretize the dimensions D_Z and D_T over a set Z of geographic zones and a set T of time intervals, depending on the spatial and temporal scales. We denote this fixed scaled event database by $\mathcal{E}_{H_T \sim l_T}^{H_Z \sim l_Z}$ ($\mathcal{E}_{l_T}^{l_Z}$ for short), where l_Z (resp. l_T) represents a spatial (resp. temporal) scale in H_Z (resp. H_T). When this fixed scaled database is ordered by time, then it is defined as $\mathcal{E}_{l_T}^{l_Z} = \{(t_1, X_1), (t_2, X_2), \dots, (t_{|T|}, X_{|T|})\}$, where $|T|$ represents the size of database, $X_j \subseteq Z$ is a set of spatial entities and t_j represents a time interval for which $X_j \in Z$ occurs in $\mathcal{E}_{l_T}^{l_Z}$. Note that if an event does not have precise information with respect to l_Z and l_T (e.g. an event occurring in France, while $l_Z = \text{city}$), we do not include it in $\mathcal{E}_{l_T}^{l_Z}$. We illustrate a fixed scale event database $\mathcal{E}_{l_T}^{l_Z}$ with an example in Table 3. This example relies on a toy fictional event database, in which we fix the spatial and temporal scales to *country* and *week*, respectively. Each row in Table 3 includes the countries reporting at least one epidemiological event for a given weekly time interval. For instance, we observe the first disease cases in France, Italy, China and India during the first week. Then, in the second week the viruses spread over neighbor countries, which are Spain, India and Nepal. Finally, it is also possible to restrain all literal values of a dimension D_i , i.e. $Dom_{\mathcal{E}}(D_i)$, with a fixed spatial or temporal scale. For a given spatial (resp. temporal) scale, we denote it as $Dom_{\mathcal{E}}(D_i, l_Z)$ (resp. $Dom_{\mathcal{E}}(D_i, l_T)$).

3.2 Event matching

In this section, our goal is to identify common events between two event databases in an automatic manner, which is not a trivial task. We propose here an approximation scheme by modeling this task as an assignment problem, also known as maximum weighted bipartite matching problem, as already done in the literature (e.g. [57]). In the end, we obtain a set of "putatively" associated events between two event databases.

Let \mathcal{E}_1 (resp. \mathcal{E}_2) be two event databases associated with IBS or EBS systems, containing $N_{\mathcal{E}_1}$ and $N_{\mathcal{E}_2}$ events, respectively. Also, we assume $N_{\mathcal{E}_1} \leq N_{\mathcal{E}_2}$ without loss of generality. Moreover, let S be

the $N_{\mathcal{E}_1} \times N_{\mathcal{E}_2}$ similarity matrix of \mathcal{E}_1 and \mathcal{E}_2 . The term S_{ij} , with $1 \leq i \leq N_{\mathcal{E}_1}$ and $1 \leq j \leq N_{\mathcal{E}_2}$, represents the similarity score between events e_i and e_j and it is calculated with Equation 8 (Section A in the Appendix). Then, we look for a bijection $f : \{1, 2, \dots, N_{\mathcal{E}_1}\} \rightarrow \{1, 2, \dots, N_{\mathcal{E}_2}\}$ such that the objective is to maximize the similarity between \mathcal{E}_1 and \mathcal{E}_2 , as defined in Equation 1.

$$\text{Max} \sum_{i=1}^{N_{\mathcal{E}_1}} S_{if(i)}. \quad (1)$$

Since this problem can be modelled as an assignment or a maximum weighted bipartite matching problem, it can be solved in various ways. One of them is through the well-known Hungarian algorithm, whose complexity is $O(n^3)$ [33].

Finally, in the solution of the assignment problem, some events might be assigned to other events with negative or weak positive similarity scores. Therefore, we perform a post-processing by removing the assignment results, whose similarity scores are lower than some threshold value.

3.3 Spatial Dimension

Our evaluation strategy for the spatial dimension relies on the concept *representativeness*. Barboza et al. [6] define this concept as the ability of describing accurately the distribution of events in terms of place, time and host. Particularly, geographic representativeness constitutes an important aspect in Epidemic Intelligence. For this reason, we propose to compare the spatial dimension of the events collected by an EBS system through geographic representativeness by taking the temporal aspect into account. We call it *spatio-temporal representativeness*, and it allows measuring how well the event database \mathcal{E} of an EBS system represents geographic zones (e.g. country, regions) in terms of the events found in a gold standard database $\mathcal{E}_{\mathcal{R}}$, for a given time period. In the end, the obtained results enable us to know to what degree geographic zones are represented by \mathcal{E} .

In the definition of the spatio-temporal representativeness, we say that an EBS system represents well a specific geographic zone for a given time interval, if it finds at least one event in $\mathcal{E}_{\mathcal{R}}$. For this reason, its calculation requires fixing the spatial and temporal scales of the events in \mathcal{E} (resp. $\mathcal{E}_{\mathcal{R}}$) with l_Z and l_T , i.e. $\mathcal{E}_{l_T}^{l_Z}$ (resp. $\mathcal{E}_{\mathcal{R}l_T}^{l_Z}$). Since there can be some reporting delay between the events of \mathcal{E} and $\mathcal{E}_{\mathcal{R}}$, we also consider in this calculation the previous (resp. next) time interval in order not to penalize an EBS system. For a given geographic zone, we perform this calculation for all the time intervals, and then we take their average to obtain a final score. This score is in the range $[0, 1]$, where the score of 0 (resp. 1) indicates that $\mathcal{E}_{l_T}^{l_Z}$ never (resp. always) finds an event in $\mathcal{E}_{\mathcal{R}l_T}^{l_Z}$ for a given geographic zone.

For space matters, we explain in the Appendix how we calculate the spatio-temporal representativeness score of an event database \mathcal{E} with respect to a gold standard database $\mathcal{E}_{\mathcal{R}}$ (Section C.2).

3.4 Temporal Dimension

For the temporal dimension, we include two evaluation assessments. The first one is a quantitative assessment based on the concept *timeliness* (Section 3.4.1). The second one is a qualitative assessment related to the consistent periodic behavior of the events (Section 3.4.2).

3.4.1 Timeliness

We start with the first comparison, which is based on the concept *timeliness* [2, 4, 6, 52, 63]. Barboza et al. [6] define this concept as the ability of identifying disease events in a timeframe enabling utilization of the information by decision makers to mitigate potentially dangerous situations as soon as possible.

In the literature, timeliness is measured as the time difference between the publication date of an event in an EBS system and that of the same event in a gold-standard database. Nevertheless, we model it with an exponential decay function in order to obtain a normalized score, as proposed in [13, 40, 41]. Its calculation for an event database \mathcal{E} of an EBS system is performed with respect to a gold standard database $\mathcal{E}_{\mathcal{R}}$. This requires to know the binding of the events between \mathcal{E} and $\mathcal{E}_{\mathcal{R}}$, which is unknown in advance. To estimate such a binding we rely on the method described in Section 3.2. In this method, for a given event $e \in \mathcal{E}$, we define a bijective function $f(e, \mathcal{E})$, which returns the putatively associated event $e' \in \mathcal{E}_{\mathcal{R}}$ with $e \neq e'$. Then, when we repeat it for each event in \mathcal{E} , and we obtain the set $\tilde{\mathcal{E}}$ of events with $\tilde{\mathcal{E}} \subseteq \mathcal{E}$, which represents a subset of events having the correspondence with the events in $\mathcal{E}_{\mathcal{R}}$. Note that not all events in \mathcal{E} has a binding in $\mathcal{E}_{\mathcal{R}}$. In the end, the obtained score is in the range $[0, 1]$, where the score of 0 (resp. 1) indicates that an EBS system is never (resp. always) timely in the detection of the events in $\mathcal{E}_{\mathcal{R}}$.

Country	Spatially Close Countries
France	Italy, Spain, Switzerland, Germany, Belgium, Luxembourg
Italy	France, Switzerland, Germany, Slovenia, Austria, Liechtenstein
Spain	France, Portugal
Portugal	Spain
China	Mongolia, Russia, Italy, Afghanistan, Tajikistan, Kyrgyzstan, Nepal, India, North Korea
India	China, Pakistan, Nepal, Bhutan, Bangladesh
Nepal	China, India
Pakistan	Iran, Afghanistan, China, India

Table 4: Spatial closeness between the considered countries used in the example illustrated in Table 3. For the sake of clarity, we consider only neighbor countries sharing a border as spatially close.

For space matters, we explain in the Appendix how we calculate the timeliness score of an event database \mathcal{E} with respect to a gold standard database $\mathcal{E}^{\mathcal{R}}$ (Section C.3.1).

3.4.2 Full and Partial Periodicity

Another interesting temporal dimension analysis is to check if there are any periodically occurred events (e.g., at least once every n weeks), which are geographically close to each other. For instance, we know from the literature that some Avian Influenza events can occur seasonally due to migratory birds, or it can become endemic due to its persistence in some regions. Therefore, it can be useful to characterize the cyclic behavior of the epidemiological events by taking into account the spatial information. We perform this task by identifying periodic-frequent spatial patterns from the field of spatio-temporal frequent pattern mining [31, 56]. Next, we first introduce the necessary definitions and concepts.

As in Section 3.3, in the following, we also fix the spatial and temporal scales of the events in \mathcal{E} (resp. $\mathcal{E}_{\mathcal{R}}$) with l_Z and l_T . Therefore, we investigate on the temporal aspects discussed above through $\mathcal{E}_{l_T}^{l_Z}$ (resp. $\mathcal{E}_{\mathcal{R}l_T}^{l_Z}$). Let Z (resp. T) represent all spatial (resp. temporal) entities with respect to a spatial (resp. temporal) scale l_Z (resp. l_T) in $\mathcal{E}_{l_T}^{l_Z}$, i.e. $Z = \text{Dom}_{\mathcal{E}}(D_Z, l_Z)$ (resp. $T = \text{Dom}_{\mathcal{E}}(D_T, l_T)$). Each element in $\mathcal{E}_{l_T}^{l_Z}$ is called transaction. Moreover, in each transaction, we call *pattern* a set X of spatial entities, with $X \subseteq Z$. If X contains k spatial entities, then it is called a k -pattern. A pattern X is called *spatial*, if the maximum distance between any two of its spatial entities is no more than the user-specified distance α . That is, X is a spatial pattern if $\max(\text{Dist}(z_p, z_q) | \forall z_p, z_q \in X) \leq \alpha$.

Furthermore, the number of transactions containing a spatial pattern X in $\mathcal{E}_{l_T}^{l_Z}$ is called the support of X , and denoted as $\text{sup}(X)$. If this support is large, then one can naturally ask how recurrent X is in $\mathcal{E}_{l_T}^{l_Z}$. Let t_i^X and t_j^X be two consecutive time intervals at which X appears in $\mathcal{E}_{l_T}^{l_Z}$. The time difference between t_i^X and t_j^X is defined as an inter-arrival time of X , and defined as $t_j^X - t_i^X$. Let T_ι^X be the set of all inter-arrival times of X in $\mathcal{E}_{l_T}^{l_Z}$. The recurrence of a spatial pattern X is considered *full periodic* (*periodic* for short), if any value in the set T_ι^X is never no more than the user-specified maximum inter-arrival time ι . The cardinality of T_ι^X in $\mathcal{E}_{l_T}^{l_Z}$ constitutes the period-support of X , denoted as $\text{psup}(X)$. In other words, X periodically appears $\text{psup}(X)$ times within the data, and at least once every ι time intervals. Our aim in this section is to find all spatial patterns that periodically appear in $\mathcal{E}_{l_T}^{l_Z}$. We call them periodic spatial patterns.

We illustrate all these concepts with the same example illustrated in Table 3. On top of that, since we are interested in spatial patterns, Table 4 depicts the spatial neighborhood of the countries with respect to the parameter α . Overall, some countries (e.g. India) face against a long infection period, whereas the others (e.g. France) succeed in stopping quickly the propagation of the viruses. Regarding the frequency of the spatial patterns from Table 3, we have $\text{sup}(\text{India}) = 8$ (the most frequent), whereas we have $\text{sup}(\text{Nepal}) = 2$ and $\text{sup}(\text{China}) = 2$ (the least frequent). Moreover, *India* is the only periodic spatial pattern (with $\text{psup}(\text{India}) = 7$) when $\iota = 2$. Nevertheless, when we set $\iota = 4$, in this case, the periodic spatial patterns are *India* (with $\text{psup}(\text{India}) = 7$), *Portugal* (with $\text{psup}(\text{Portugal}) = 3$), *Spain* (with $\text{psup}(\text{Spain}) = 3$) and *Portugal-Spain* (with $\text{psup}(\text{Portugal-Spain}) = 3$). Note that the patterns *India-Portugal* and *India-Spain* are also periodic (with $\text{psup}(\text{India-Portugal}) = 3$ and $\text{psup}(\text{India-Spain}) = 3$), but they do not fulfill the requirement of spatial closeness (see Table 4).

For some cases, the periodicity condition can be too strict. For instance, in Table 3 France (resp. Pakistan) appears in the first (resp. last) 3 transactions, which is also valuable information. To weaken this strict definition, we also consider the partial periodicity condition. In this weaker condition, it is sufficient for a spatial pattern to periodically appear only in *some* transactions of $\mathcal{E}_{l_T}^{l_Z}$. Concretely, a spatial pattern X is said to be a *partial* periodic spatial pattern if its period-support $\text{psup}(X)$ is no less

Pattern	Periodic Support
India	7
Portugal	3
Spain	3
Portugal-Spain	2
Pakistan	2
India-Pakistan	2
France	2

Table 5: All partial periodic spatial patterns obtained with $\iota = 2$ and $\rho = 2$ based on the example illustrated in Table 3. The spatial closeness between the considered countries is defined in Table 4.

than the user-specified minimum period-support ρ . For instance, if we take the same example illustrated in Table 3, all partial periodic spatial patterns for $\iota = 2$ and $\rho = 2$ are illustrated in Table 5. It is worth noticing that the parameter ι has a positive effect on the generation of partial periodic spatial patterns, while ρ has a negative effect on the number of patterns being generated from the database. Moreover, note that the input parameters ι and ρ can be both expressed in percentage or in count, respectively. For instance, when we set $\rho = 1.0$ (resp. $\rho < 1.0$), this amounts to generate full (resp. partial) periodic spatial patterns.

In this work, given spatial and temporal scales l_Z and l_T , we discover all full and partial periodic spatial patterns in $\mathcal{E}_{l_T}^{l_Z}$ with respect to the input parameters ι , ρ and α through the method *ST-ECLAT* (ST for short), proposed in [31]. Particularly, we are interested in two use cases for obtaining these patterns. First, we want to know what spatial entities (e.g. countries) have consistently epidemiological events throughout the year. We call the obtained results *continuous* periodic patterns. Second, we want to know what spatial entities (e.g. countries) have a seasonal effect and are exposed to disease events only for some period of time every year. We call the obtained results *seasonal* (or *yearly*) periodic patterns.

For space matters, we explain in the Appendix how we quantitatively evaluate the performance of \mathcal{E} in terms of its ability to detect these continuous and seasonal periodic patterns with respect to a gold standard database $\mathcal{E}^{\mathcal{R}}$ (Section C.3.2).

3.5 Thematic Dimension

In this section, we aim to evaluate EBS systems in terms of thematic entities they extract from the events. In other words, we want to know whether the dimensions D_D and D_H in an event database \mathcal{E} provides fine- or coarse-grained information. Ideally, we expect \mathcal{E} to provide very detailed information, as in a gold standard database $\mathcal{E}^{\mathcal{R}}$. Note that this aspect is related to one of the relevant characteristics of an EBS system in the CDC's guideline, so-called *completeness* [6].

In our evaluation, we want to discover the rich data relations between spatial, temporal and thematic entities with two use cases. In the first use case, we totally omit the temporal aspect, and we propose to find out what thematic entities characterize most a spatial entity. For instance, if we take the same example illustrated in Table 1, we might want to know where the specific Avian Influenza serotype H7N9 is more prevalent. Our second use case is the temporal version of the first one [38], in which we are interested in the periodic aspects, as in Section 3.4.2. For instance, when several Avian Influenza events with particular spatial and thematic characteristics repeat themselves at regular intervals in the data, this would indicate an ongoing spreading pattern with specific characteristics. In this work, we propose to perform these two use cases within a single evaluation scheme through the identification of frequent patterns, as in Section 3.4.2.

As opposed to Section 3.4.2, there are several key differences in this section, because we consider the fact that each event in \mathcal{E} can be described with different hierarchical event attributes. First, we do not fix any temporal, spatial or thematic scale on \mathcal{E} , and we ensure that each transaction in \mathcal{E} corresponds to a single event. Second, a transaction in \mathcal{E} does not simply consist of atomic spatial entities, it is rather represented by a tuple $Y = (d_Z, d_D, d_H)$, with $d_Z \in \text{Dom}_Z(\mathcal{E})$, $d_D \in \text{Dom}_D(\mathcal{E})$ and $d_H \in \text{Dom}_H(\mathcal{E})$. In this context, we call this tuple Y *multidimensional pattern* [55]. Therefore, with the multidimensionality of the patterns, \mathcal{E} is defined as $\mathcal{E} = \{(t_1, Y_1), (t_2, Y_2), \dots, (t_{|N_{\mathcal{E}}|}, Y_{|N_{\mathcal{E}}|})\}$, where $|N_{\mathcal{E}}|$ represents the size of \mathcal{E} , Y_j is a multidimensional pattern and t_j represents the timestamp of Y_j . Third, we adapt \mathcal{E} to include various hierarchical information of the event attributes. To do so, we modify \mathcal{E} by adding all ancestors in the associated hierarchy of every multidimensional pattern. In the end, each transaction consists of the original multidimensional pattern and its variants with all ancestors in the associated hierarchy. We denote this modified event database as \mathcal{E}^+ . We illustrate how

we obtain \mathcal{E}^+ from \mathcal{E} with an example in Table 6.

Time	Multidimensional pattern
t_1	(Paris, AI, bird) , (Île de France, AI, bird), (France, AI, bird)
t_2	(Italy, AI, wild bird) , (Italy, AI, bird)
t_3	(Spain, H5N1, wild bird) , (Spain, HPAL, wild bird), (Spain, AI, wild bird), (Spain, H5N1, bird), (Spain, HPAL, bird), (Spain, AI, bird)

Table 6: Illustrative example of the modified version \mathcal{E}^+ of a subset of an event database \mathcal{E} with three events. The first column indicates the timestamp of the events, and the second column describes the multidimensional patterns. Note that the second column contains all ancestors in the associated hierarchy of every multidimensional pattern. For the sake of clarity, we show in bold the elements in \mathcal{E} , before obtaining its modified version \mathcal{E}^+ .

In this work, we perform the two use cases discussed above by discovering frequent multidimensional patterns, accompanied by the partial periodicity condition. When the temporal aspect is omitted, we simply calculate the support $\text{sup}(Y)$ of each multidimensional pattern Y in \mathcal{E}^+ , as such pattern always corresponds to a single tuple (d_Z, d_D, d_H) in this context. We call them *static* multidimensional patterns. When we take the partial periodicity condition into account, this amounts to find partial periodic multidimensional patterns, as it ensures that two multidimensional patterns appear in the same time interval in \mathcal{E}^+ . We call them *temporal* multidimensional patterns. In practice, we use the ST algorithm described in Section 3.4.2 to generate these *static* and *temporal* multidimensional patterns. The flexibility of ST is that when we set a very large inter-arrival time value to ι , this allows us to omit the partial periodicity condition.

For space matters, we explain in the Appendix how we quantitatively evaluate the performance of an event database \mathcal{E} of an EBS system in terms of its ability to detect these static and temporal multidimensional patterns with respect to a gold standard database $\mathcal{E}^{\mathcal{R}}$ (Section C.4).

3.6 Source Dimension

Finally, the last part of our evaluation framework is regarding online news and press agencies, that we call short news outlets or news sources, involved in the propagation of epidemiological information on the web.

All EBS systems rely partially or completely on various online news outlets for ensuring their monitoring of emerging infectious diseases across the world. Nevertheless, there are not enough studies that characterize and assess the news sources involved in EBS systems. For instance, Schwind *et al.* [63] point out that local news outlets are more likely to report ongoing epidemiological events than international media sources do. For this reason, we aim to identify and characterize important news outlets, and we propose in this section an evaluation scheme for the news outlets involved in the propagation of epidemiological information on the web. Our evaluation scheme consists of two different objectives. Our first objective is that we want to identify important news outlets for information dissemination. In our second objective, we are interested in the ability of news reporting in timely manner. In other words, we want to rank news outlets, which publish epidemiological events as fast as possible. In the following, we propose to perform these tasks through network analysis. Note that some EBS systems are designed to collect epidemiological data from both official and unofficial data sources. In order to have fairer evaluation across EBS systems, we do a preprocessing step by eliminating the official data sources and keeping only unofficial ones within the data.

3.6.1 Identification of Important News Outlets

We say that a news outlet needs to fulfill two conditions in order to be considered as important. First, it reports epidemiological information that are reported by both local and international news media. In other words, if someone follows the news reported by an important news outlet, it means she receives sufficiently necessary epidemiological information for her country and nearby. Second, it also reports the events that are reported by important news outlets. We perform the task of identification of important news outlets through network analysis. We design our approach in two steps. First, we extract the news outlet network $G_{\mathcal{E}}$ from an event database \mathcal{E} , where nodes represent news outlets and edges describe the relations for node pairs. We do this process on the whole or a subset of data for each considered EBS platform. Then, we apply a well-suited centrality measure to rank the news outlets by their importance score. Next, we describe how we process these steps.

Traditionally, most of the existing works in the literature extract a news outlet network, when citation information (i.e. what sources cite what other sources) between news outlets is available. Nevertheless,

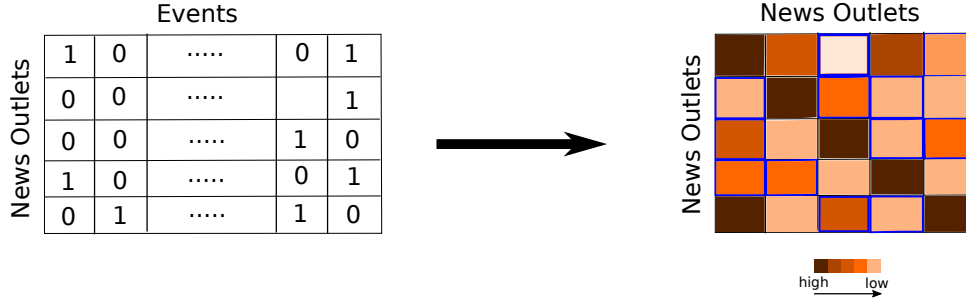


Figure 2: Illustration of the network construction steps. In the first step, we build an initial matrix, in which we indicate for each event if the news outlets report it or not. In the second step, we convert the first matrix into a second one to encode to what degree the news outlets report the same events.

this information is hardly available in the data collected by EBS systems. For this reason, we propose to use a co-occurrence fraction counting method known from the field of scientometrics [37], as also used in [12]. The construction of $G_{\mathcal{E}}$ is as follows and illustrated in Figure 2. Let us say an EBS system monitor $|P|$ news outlets for $|E|$ distinct events. First, we construct a $|P| \times |E|$ matrix A , where the rows represent news outlets and the columns represents the distinct events detected by an EBS system. Each element of matrix A is defined as in Equation 2.

$$a_{ie} = \begin{cases} 1, & \text{if news outlet } i \text{ reports event } e, \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Next, we transform the matrix A into another $|P| \times |P|$ symmetric matrix B to measure how frequent two news outlets report the same events. Each element b_{ij} of matrix B is defined as in Equation 3.

$$b_{ij} = \sum_{e=1}^E \frac{a_{ie}a_{je}}{a_{ie}^2} \quad (3)$$

In the end, we obtain a score of 1 (resp. 0), when two news outlets always (resp. never) report the same events, or a score in $[0, 1]$ otherwise.

Finally, in the second step of our approach, we apply a centrality measure over $G_{\mathcal{E}}$ to identify important news outlets. A centrality measure aims to rank the vertices of a network by assigning them a score. The more central a vertex is, the larger score it has. In the literature, there is a large number of centrality measures, each having a particular objective. In this work, we propose to use the PageRank centrality for $G_{\mathcal{E}}$, as it is more suitable to our definition of important news outlets. In the end, we want to see how similar important news outlets are among multiple EBS systems. If it is very similar, this would indicate that they rely mostly on the same news sources. In the rest of the work, we denote the first k most important news outlets from the PageRank centrality result by $\text{PageRank}(G_{\mathcal{E}}, k)$.

3.6.2 Timely Detection

One specific criteria that one may want to optimize in event detection is to minimize detection time (i.e. capturing an epidemiological event as soon as possible). Our second objective is related to this timeliness capability of the news outlets. We want to identify the news outlets, which are timely in event detection, and not those detecting as many events. To do so, we follow the work of Leskovec *et al.* [36]. Their method first creates the news outlet network $G_{\mathcal{E}}$ from \mathcal{E} , then finds through their method CELF a set A of news outlets, which minimizes detection time, while covering all the event set E .

First, we extract our news outlet network $G_{\mathcal{E}}$ as follows. Let us suppose that the set S_e of news outlets reports through their news documents the same event e in an event database \mathcal{E} . For every event $e \in \mathcal{E}$, we create a path structure P_e , that we call *cascade*, such that a news outlet in S_e sequentially join the cascade P_e by linking to other news outlets in S_e , whereby the edges obey time order and the weights of directed edges represent the time difference between two news documents. When we repeat this process for each event $e \in \mathcal{E}$, this gives us a network in the end. We illustrate this network creation for several events with an example in Figure 3.

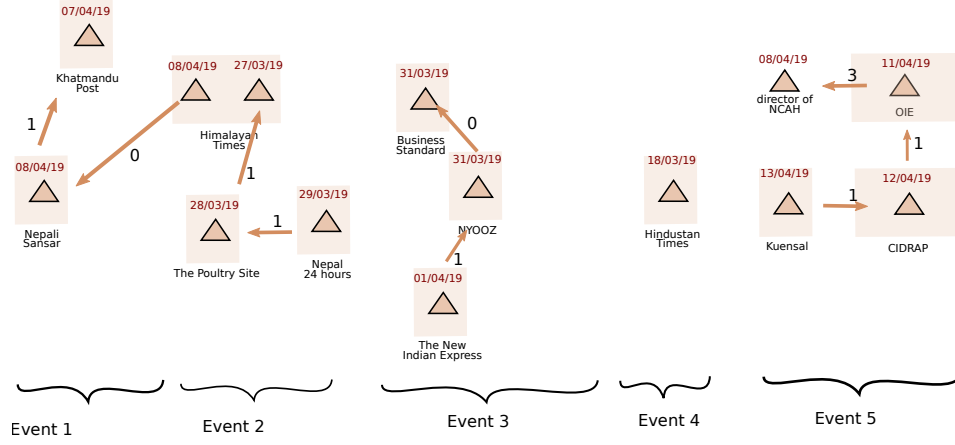


Figure 3: Illustration of the event cascades for five events. The nodes represent news outlets publishing news documents, and the edges between them obey time order and the edge weights represent their time difference. For instance, the cascade for the first event starts at the news document published by *Khatmandu Post*, and then the edges are sequentially created by adding other news documents in S_e linking to it.

Then, we apply the CELF algorithm to $G_{\mathcal{E}}$ in order to identify a set A of timely news outlets. This algorithm starts with the empty set $A_0 = \emptyset$ and iteratively adds in step k the news outlet s_k maximizing the marginal gain as in Equation 4.

$$s_k = \operatorname{argmax}_{s \in P \setminus A_{k-1}} R(A_{k-1} \cup \{s\}) - R(A_{k-1}). \quad (4)$$

The algorithm stops, once it has selected $k = |A|$ elements. The marginal gain is expressed for a subset A of news outlets in terms of the function $R(A)$, which is used as a penalty reduction function. It is defined as in Equation 5.

$$R(A) = \sum_{e \in E} P(e) (T_{max} - \min_{s \in A} T(e, s)), \quad (5)$$

where T_{max} is time horizon, P is a (given) probability distribution over the events and $T(e, s)$ represents the time delay in days, until news outlet s participates in the event path P_e . Note that $T(e, s)$ equals T_{max} , if s does not report event e . Moreover, in our context each event has uniform probability, therefore we omit $P(e)$ from the definition of $R(A)$. In the rest of the work, we denote the first k most timely news outlets obtained from the CELF method for $G_{\mathcal{E}}$ by $\text{CELF}(G_{\mathcal{E}}, k)$.

For space matters, we explain in the Appendix how we quantitatively evaluate the performance of an event database \mathcal{E} in terms of its ability to detect the important and timely news outlets (Section C.5).

4 Experimental Setup

In this section, we define our experimental setup in order to illustrate how to use our framework and interpret its results. We first present the selected EBS/IBS systems, to which we apply our framework (Section 4.1). Then, we describe the input event data, as well as its processing (Section 4.2). The results are presented afterwards, in Section 5.

4.1 Selected EBS systems

We show the relevancy of our framework on two well-known EBS systems PADI-web [67] and ProMED [11]. Moreover, we use the reference gold standard database Empres-i from the World Organisation for Animal Health (WOAH) and the Food and Agriculture Organization of the United Nations (FAO) [18] to evaluate the performances of PADI-web and ProMED. Our choice of the Empres-i database is based on the fact that it is a well-populated official database for the main animal diseases, such as Avian Influenza and African Swine Fever [3, 20]. ProMED collects and organizes its disease events

through 50 subject matter expert moderators from 34 countries⁶, who provide written commentary, giving the reader the necessary historical context and/or clinical background to understand the importance of the information being reported. ProMED also supply references to previous reported events and to the scientific literature for the sake of completeness. In principle, ProMED first searches for an official source (e.g. WOA report) if it is available at the same time that an online news document is available. For this reason, ProMED relies on both official and unofficial sources for event detection. On the contrary, PADI-web is an automated surveillance system, which automatically collects online news documents with customized queries using the Google News aggregator, translates all non-English documents into English, classifies the documents, and extracts epidemiological information (diseases, dates, symptoms, hosts and locations) from the relevant news documents. PADI-web is currently integrated in the French Platform for Animal Health Surveillance (ESA Platform) [3].

We choose ProMED and PADI-web in our experiments for several reasons. First, each one belongs to a different EBS category: moderated vs. automated. Second, both EBS systems have a well-established surveillance system, since they are operational for a long time. They are currently collaborating with and used by national and supranational health authorities. Third, they are open-access tools. Finally, several works have assessed PADI-web and ProMED, separately [63, 66] or together [2, 3]. We base our discussion in Section 5 on these previous results, when possible.

4.2 Event Data and Processing

The event datasets we use concern the Avian Influenza (AI) cases affecting bird species from 2019 to 2021. These AI cases can be high pathogenic AI (HPAI) or low pathogenic AI (LPAI). They are retrieved from PADI-web, ProMED and Empres-i, respectively. Regarding the PADI-web data, we rely only on those considered as relevant from PADI-web’s automatic process. We chose a three-year study period (2019-2021) to sufficiently capture the space-time epidemiological characteristics of the AI events around the world. In order not to penalize ProMED we keep all its data provided by official and unofficial data sources, although PADI-web relies only on online news outlets. Nevertheless, for the sake of completeness, we compare both systems in Table 8 and Figure 8 by discarding the ProMED data provided by official data sources (i.e. WOA reports). Furthermore, it is also worth noticing that automated EBS systems, such as PADI-web, might report false event information due to their automated location detection and extraction strategies. Evaluating the rate of reporting false events for such systems is not the scope of this work.

We process the collected raw event datasets by transforming them into normalized event databases, as explained in the Appendix (Section B). During this processing, we deal with the different event definitions that PADI-web and ProMED have, which do not exactly match the one proposed in Section 3.1, as the definition of an event can be different from one EBS system to another. Regarding ProMED, we only extract the information regarding news outlets from the raw news documents. Regarding PADI-web, the events are essentially disease-location pairs. Moreover, PADI-web extracts event-related thematic information for each collected news document, without relating them to any event. For this reason, we complete the minimally defined events with the extracted thematic entities, as detailed in Section B.1. In the end, we obtain three normalized event databases for PADI-web, ProMED and Empres-i, denoted by \mathcal{E}_{PW} , \mathcal{E}_{PM} and \mathcal{E}_{EI} , respectively. These normalized data are not publicly available due to third party restrictions, nevertheless, they are available on request⁷.

5 Results

We now assess, compare and discuss the performances of the considered EBS systems when applied to our framework on the Avian Influenza event databases \mathcal{E}_{PW} , \mathcal{E}_{PM} and \mathcal{E}_{EI} . In our experiments, the number of events by year for those event databases are shown in Table 7. In total, there are 1515, 786 and 5229 events for \mathcal{E}_{PW} , \mathcal{E}_{PM} and \mathcal{E}_{EI} , respectively. We present the results in line with our four evaluation objectives: 1) spatial (Section 5.1), 2) temporal (Section 5.2), 3) thematic (Section 5.3) and 4) source (Section 5.4) dimensions. Our source code is publicly available⁸.

5.1 Spatial Dimension

We evaluate through spatio-temporal representativeness how accurate EBS systems describe the distribution of events found in the gold standard Empres-i database in terms of place and time. We calculate

⁶<https://promedmail.org/team>

⁷<https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.57745/Y3XR0X>

⁸github.com/arinik9/compebs

EBS/IBS system	Number of events in 2019	Number of events in 2020	Number of events in 2021	Total
PADI-web	116	436	963	1515
ProMED	28	245	513	786
Empres-i	267	1539	3423	5229

Table 7: Event statistics for PADI-web, ProMED and Empres-i in the Avian Influenza dataset.

	PADI-web covers	PADI-web never covers
ProMED covers	Ghana, Philippines, Saudi Arabia, Senegal, Mali, Hong Kong, Ivory Coast, Albania, Algeria, Netherlands, Niger, Mauritania, Botswana, Ireland, Australia (15 countries in total) ^γ	Chile, Bosnia and Herzegovina, Lithuania, Serbia, Bhutan, Lesotho, Namibia, Pakistan, Greece, Canada, Slovenia, Egypt, Afghanistan, Portugal (14 countries in total)
ProMED never covers	- (0 country in total)	Mexico, Greenland, Dominican Republic (3 countries in total)

Table 8: Countries covered by PADI-web and ProMED according to the spatio-temporal representativeness scores.

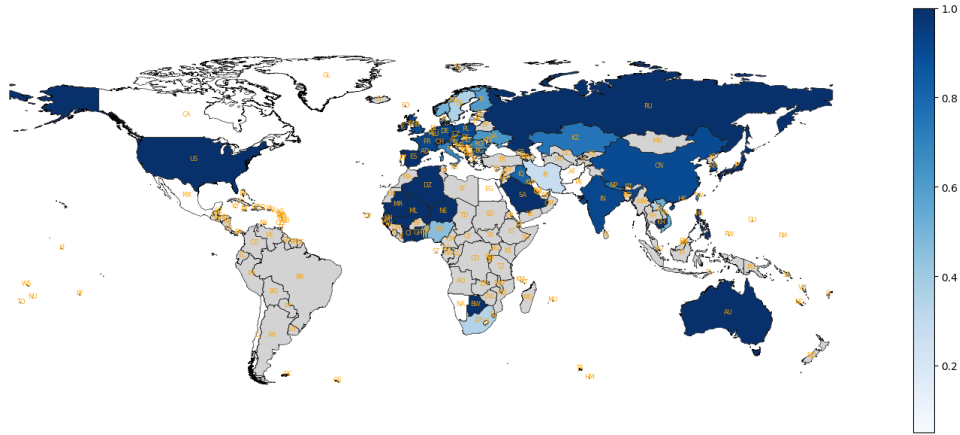
^γ Here, we focus only on the countries, where the spatio-temporal representativeness score is the maximum value of 1 for both PADI-web and ProMED. If we focus on the countries covered both PADI-web and ProMED instead, there are 58 countries in common.

the spatio-temporal representativeness scores for each region and country with monthly time intervals. In the following, we discuss only these scores at country scale and leave those at region scale in the Appendix, due to lack of space. We plot the scores at country scale in Subfigures 4a and 4b (see Figures 9 and 10 in the Appendix for region scale). In these plots, countries without an Empres-i event are indicated in gray, and the degree to which an EBS system covers the events occurring in a country is shown with different blue scales, where the large (resp. small) values are indicated in dark (resp. light) blue. When an EBS system never finds an event in \mathcal{E}_{EI} , it is shown in white.

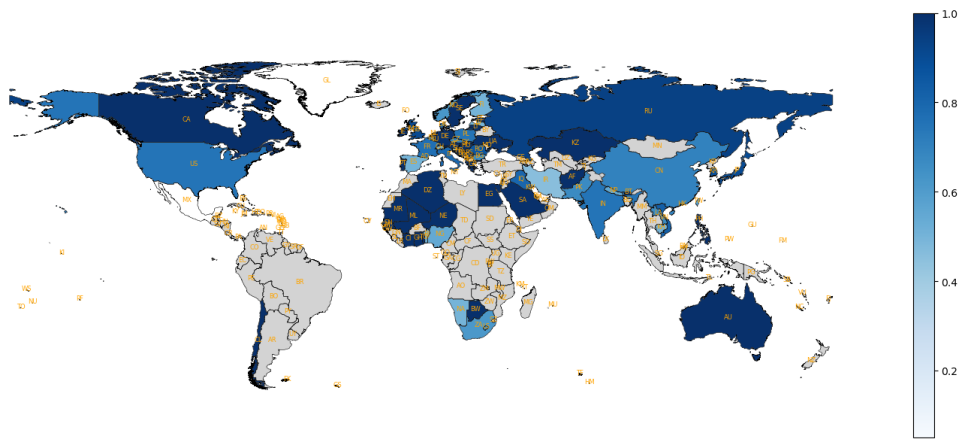
We see from these plots that both PADI-web and ProMED report the events from a large number of countries, but they represent well only some of them. Moreover, although some countries (e.g. Ireland, Australia) are equally represented by PADI-web and ProMED, there are some discrepancies in the spatial focus of these EBS systems. For instance, PADI-web (resp. ProMED) better covers the USA, Spain, France, India and China (resp. South Africa, Vietnam, Kazakhstan, Ukraine and Sweden) than ProMED (resp. PADI-web). Furthermore, PADI-web never reports an event from Canada, Portugal and Afghanistan that ProMED covers well. All these similarities and differences at country scale are summarized in Table 8.

Finally, we also plot the spatio-temporal representativeness score differences between PADI-web and ProMED in Subfigure 4c (See Figure 11 for region scale) to ease their comparison. In this figure, it is colored in blue (resp. red) when ProMED (resp. PADI-web) gives better spatio-temporal representativeness score for a country and in yellow in case of non-zero equality. We see from the figure that ProMED gives better scores than PADI-web does for the overwhelming majority of countries. The average spatio-temporal representativeness score over all these countries also confirms this superiority (0.59 vs. 0.80, see Section C.2 in the Appendix for the calculation details). Note that when we consider the country and region scales together in the calculation of this score, it still confirms the superiority (0.40 vs. 0.55), although the score is lower than that at country scale. This decrease is mostly because of the difficulty of geocoding task in event normalization, i.e. accurately assigning geographic coordinates to spatial entities, due to the ambiguity among place names [51].

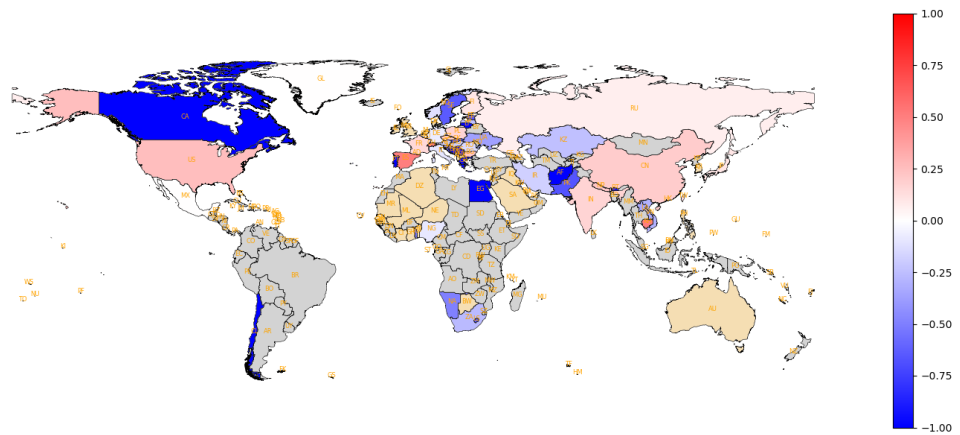
In summary, both PADI-web and ProMED report the Avian Influenza events for a large number of countries. This is mostly due to the fact that Avian Influenza (with African Swine Fever) is one of the animal disease cases reported well by both systems [2], and that the number of detected Avian Influenza events increases each year (see Table 7). Nevertheless, there are some discrepancies in the spatial focus of these EBS systems. These discrepancies are also consistent with the previous works. For instance, in [2], the official Avian Influenza events lie mostly in Central America, Africa (mostly



(a) Spatio-temporal representativeness scores at country scale for PADI-web with respect to the results of Empres-i.



(b) Spatio-temporal representativeness scores at country scale for ProMED with respect to the results of Empres-i.



(c) Spatio-temporal representativeness score differences between PADI-web (Figure 4a) and ProMED (Figure 4b) at country scale.

Figure 4: Spatio-temporal Representativeness scores at country scale for PADI-web and ProMED with respect to the results of Empres-i. In (a) and (b), the degree to which an EBS system covers the Empres-i events occurring in a country is shown with different blue scales, and it is shown in white when an EBS system never finds an event in \mathcal{E}_{EI} . In (c), the score differences between (a) and (b) are shown. It is colored in blue (resp. red) when ProMED (resp. PADI-web) gives better spatio-temporal representativeness scores for a country and in yellow in case of non-zero equality. In all these plots, countries without an Empres-i event are indicated in gray.

	PADI-web covers	PADI-web never covers
ProMED covers	Ghana, Senegal, Mali, Ivory Coast, Albania, Algeria, Mauritania, Botswana, Australia (9 countries in total) ^γ	Chile, Lithuania, Serbia, Namibia, Greece, Canada, Egypt, Afghanistan, Portugal (9 countries in total)
ProMED never covers	Iran, Croatia, Switzerland, Laos, Philippines, Saudi Arabia, Hong Kong, Austria, Italy, Luxembourg, Spain (11 countries in total)	Mexico, Greenland, Bosnia and Herzegovina, Bhutan, Dominican Republic, Lesotho, Pakistan, Slovenia (8 countries in total)

Table 9: Countries covered by PADI-web and ProMED according to the spatio-temporal representativeness scores, when we discard the ProMED data provided by official data sources (i.e. WOA reports).

^γ Here, we focus only on the countries, where the spatio-temporal representativeness score is the maximum value of 1 for both PADI-web and ProMED.

	Number of putatively associated events
PADI-web and Empres-i	422
ProMED and Empres-i	469
PADI-web and ProMED	450

Table 10: Event matching statistics in the Avian Influenza dataset.

Egypt, Nigeria and South Africa), Middle East and Asia. Although both PADI-web and ProMED cover these areas, the degree to which they report the events are different in [2]. Indeed, PADI-web covers more countries in Asia (particularly in China and India) than ProMED does, a point also highlighted in a ProMED’s publication [11]. Similarly, ProMED better covers Africa, Eastern Europe and Middle East than PADI-web does.

The most important factor that determines the events the EBS systems find is inevitably related to the online news outlets [42]. We also investigate on this aspect in Section 5.4. PADI-web relies only on the news aggregator Google News, whereas ProMED cooperates with 50 human moderators and curators from all around the world. Although these moderators rely on both the WOA reports and online news outlets, the latter plays a substantial role for ProMED (338 out of 786 Avian Influenza events in our experiments). For the sake of completeness, we also compare in Table 9 these systems in terms of spatio-temporal representativeness scores by discarding the ProMED data provided by official data sources (i.e. WOA reports).

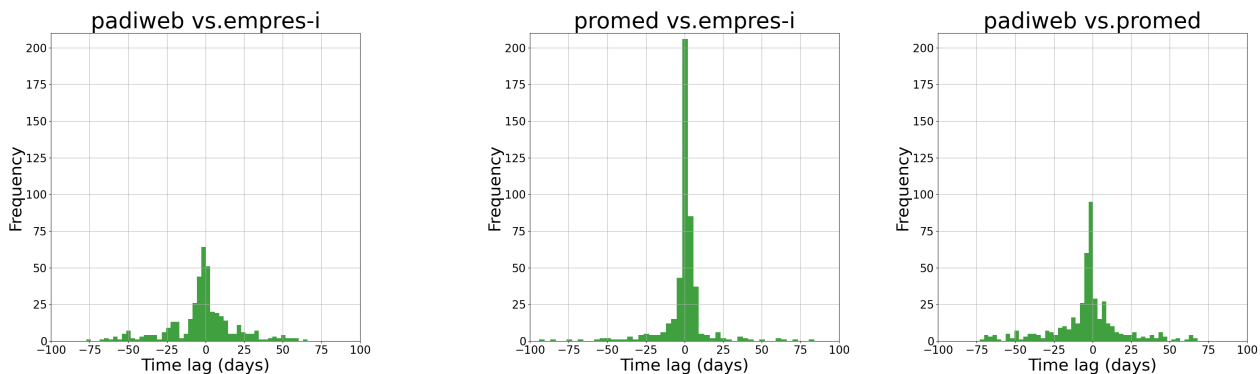
5.2 Temporal Dimension

We present the results of the temporal dimension in two parts: Timeliness (Section 5.2.1) and periodicity (Section 5.2.2) analyses.

5.2.1 Timeliness analysis

We study how timely the EBS systems PADI-web and ProMED are compared to the Empres-i events, as well as the assessment of timeliness between them. Note that we perform this assessment based on the identification of the putatively associated events between a pair of event databases, as explained in Section 3.2. The obtained statistics regarding these putatively associated events are shown in Table 10.

In Figure 5, we plot the time lag values for each pair of systems. In these plots, when the first (resp. second) system reports an event earlier than the second (resp. first) one, then this results in a negative (resp. positive) value. Moreover, we summarize the statistics in terms of timeliness in Table 11. We can see from Figure 5 and Table 11 that although both PADI-web and ProMED can be timely depending on events, PADI-web is more timely with respect to the Empres-i events compared to ProMED (49% vs. 28%). Indeed, the timeliness scores also confirm this superiority (0.18 vs. 0.12, see Section C.3.1 in the Appendix for the calculation details). Furthermore, when we look at the same events detected by PADI-web and ProMED (see the third row in Table 11), PADI-web is also more timely (60% vs. 33%). The delay for ProMED is mostly related to its events collected from the official data sources (i.e. the WOA reports). For instance, when we discard the ProMED data provided by official data sources (i.e. WOA reports), the timely detection rate of ProMED increases from 28% to 44%. Nevertheless,



(a) Time lags between PADI-web and Empres-i.

(b) Time lags between ProMED and Empres-i. (c) Time lags between PADI-web and ProMED.

Figure 5: Time lags of the putatively associated events between PADI-web and Empres-i (a), between ProMED and Empres-i, between PADI-web and ProMED. In these plots, when the first (resp. second) EBS system reports an event earlier than the second (resp. first) one, then this results in a negative (resp. positive) value.

	Number of events if the first system is in advance	Number of events if the first system is better and 30 days in advance	Number of events if the second system is in advance	Number of events if the second system is better and 30 days in advance	Average delay in days for the first system	Average delay in days for the second system	Timeliness score for the first system	Timeliness score for the second system
PADI-web vs. EMPRES-i (422 events)	209 (49%)	39	185	29	14.8	16.37	0.18	0.21
ProMED vs. EMPRES-i (469 events)	134 (28%)	16	293	12	6.00	12.76	0.12	0.09
PADI-web vs. ProMED (450 events)	273 (60%)	47	150 (33%)	30	17.08	15.04	0.15	0.23

Table 11: Timeliness summary statistics for the comparisons between PADI-web and Empres-i, between ProMED and Empres-i, between PADI-web and ProMED.

when we focus only on the events, where PADI-web and ProMED are late, i.e. those reported after the Empres-i events, the average delay in days for ProMED is better compared to PADI-web (6.00 vs 14.8, here the less is better). This fact is also due to the official WOHAI reports, as the publication dates of the WOHAI-based ProMED events and those of the associated Empres-i events are very close, which reduces the average value.

Overall, PADI-web detects the 209 putatively associated events (49%) before their publication in Empres-i, and those 39 of them are 30 days in advance. Likewise, ProMED detects the 134 putatively associated events (28%) before Empres-i, and those 16 of them are 30 days in advance. Moreover, PADI-web (resp. ProMED) detects the 273 (resp. 150) putatively associated events before ProMED (resp. PADI-web), and those 47 (resp. 30) of them are 30 days in advance. Finally, our results for time lags and timeliness are also consistent with the previous works [2, 66]. Indeed, the performances of PADI-web and ProMED are also comparable in these works, and PADI-web is slightly more timely than ProMED.

5.2.2 Periodicity Analysis

We now study how accurate EBS systems detect full or partial periodic continuous and seasonal patterns with different temporal scales based on the evolution of the epidemiological events. We compare the obtained results based on the Empres-i dataset to see to what extent PADI-web and ProMED can capture similar patterns.

To ease our discussion, we visualize the evolution of the events occurring in some countries of interest with fine-grained temporal scale from 2019 to 2021 in Figure 6. We describe it generically here, for matters of convenience. This figure is in a form of heatmap matrix. The columns represent distinct events provided by PADI-web (in pink), ProMED (in gray) and Empres-i (in yellow). The rows correspond to the bi-weeks of 2019, 2020 and 2021. Each cell of the matrix indicates the absence or presence of at least one event for a given time period and country. Only the cells in brown indicate the presence of events. Finally, the columns (i.e. the events) are regrouped by country, as indicated on the

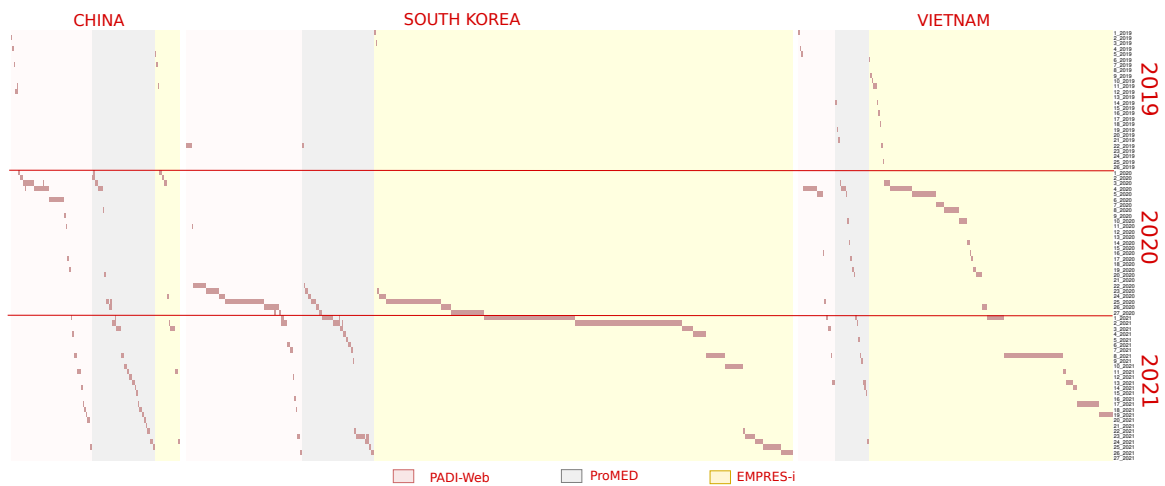


Figure 6: Evolution of epidemiological events for China, South Korea and Vietnam from 2019 to 2021 with fine-grained temporal scale. The columns represent distinct epidemiological events provided by PADI-web (in pink), ProMED (in gray) and Empres-i (in yellow). The rows correspond to the bi-weeks of 2019, 2020 and 2021. Each cell of the matrix indicates the absence or presence of at least one event for a given time period and country. Only the cells in red indicate the presence of events. Finally, the columns (i.e. the events) are regrouped by country, as indicated on the top part of the plot.

top part of the plot.

We start with the full and partial periodic seasonal patterns. We obtain the results from PADI-web, ProMED and Empres-i by discretizing the time (resp. spatial) dimension into monthly intervals (resp. country zones) and by applying the ST method with the parameters $\iota = 12$, $\varrho = \{0.5, 1.0\}$ and $\alpha = 1000km$. Recall that in our dataset the full periodicity, with $\varrho = 1.0$, amounts to be the events occurring every year for the same time period from 2019 to 2021. Regarding the partial periodicity, with $\varrho = 0.5$, a seasonal pattern is valid in our dataset, when the events occur during two consequent years between 2019 and 2021 for the same time period. In the comparison, we need to take into account the fact that the detection time for the same events can differ up to 30 days in average for PADI-web and ProMED with respect to the Empres-i events (see Figure 5). Therefore, it is reasonable to observe some time delay in the results.

Table 12 shows the full and partial seasonal periodic frequent patterns of six countries for each EBS/IBS system to see when and where epidemiological events repeatedly occur every year from 2019 to 2021. These countries are China, South Korea, Vietnam, India, United Kingdom and France. In Table 12, the Empres-i patterns detected by both PADI-web and ProMED are in orange, and it is colored in red (resp. blue) when only PADI-web (resp. ProMED) finds them. We see from the table that the results of PADI-web and ProMED are not very inline with the Empres-i seasonal patterns, and we summarize the comparison in two points.

First, we observe the full periodic seasonal patterns in the Empres-i data for some countries, and we expect PADI-web and ProMED to detect them. These countries are China, Vietnam, India, United Kingdom, Taiwan, South Africa, Bulgaria, Japan and Denmark (see Table 12 and Figure 6 for some of them). PADI-web captures its full periodic seasonal patterns for the considered four countries of Table 12, plus for Taiwan and Japan. In these patterns, PADI-web accurately detects only the pattern for the United Kingdom. Nonetheless, there are some discrepancies for the other countries. Indeed, it underrepresents (resp. overrepresents) Vietnam (resp. China and India). For instance, PADI-web overrepresents the events occurring in China, because particularly since 2020, with the rise of Covid-19 cases, media sources make much news about China and coronavirus. For instance, one of the news titles is "*Chinese authorities say viral pneumonia outbreak is not SARS, MERS or bird flu*"⁹. Overall, the average evaluation scores for PADI-web and ProMED based on the seasonal full periodic frequent patterns are 0.49 and 0.17, respectively (see Section C.3.2 in the Appendix for the calculation details).

Second, we see many more seasonal patterns for partial periodicity. This indicates that a country witnesses an event during two consecutive years from 2019 to 2021. On the one hand, ProMED is interestingly able to capture almost accurately the patterns for China, South Korea, India, United Kingdom and France. But, it largely underrepresents the patterns for Vietnam. On the other hand,

⁹www.reuters.com/article/us-china-pneumonia-idUSKBN1Z40G3

EBS system	country	full periodicity (2019-2021)												partial periodicity (2019-2021)											
		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
PADI-web	China																								
PADI-web	South Korea																								
PADI-web	Vietnam																								
PADI-web	India																								
PADI-web	United Kingdom																								
PADI-web	France																								
ProMED	China																								
ProMED	South Korea																								
ProMED	Vietnam																								
ProMED	India																								
ProMED	United Kingdom																								
ProMED	France																								
Empres-i	China																								
Empres-i	South Korea																								
Empres-i	Vietnam																								
Empres-i	India																								
Empres-i	United Kingdom																								
Empres-i	France																								

Table 12: Full and partial seasonal (or yearly) periodic patterns for PADI-web, ProMED and Empres-i. For the sake of comparison, we only show them only for five countries: China, South Korea, Vietnam, India, United Kingdom and France. The Empres-i patterns detected by both PADI-web and ProMED are colored in orange, and it is colored in blue (resp. red) when only ProMED (resp. PADI-web) finds them.

Rank	PADI-web (month)	PADI-web (week)	ProMED (month)	ProMED (week)	EMPRES-i (month)	EMPRES-i (week)
1	China (16)	India (41)	Russia (15)	India (23)	Taiwan (30)	Taiwan (66)
2	India (15)	France (30)	China (14)	Russia (23)	South Africa (23)	Germany (36)
3	United Kingdom (14)	United Kingdom (25)	India (14)	South Korea (23)	Vietnam (21)	South Africa (29)
4	France (13)	Germany (15)	Germany (12)	Germany (22)	Russia (14)	Russia (29)
5	Germany (12)	Japan (13)	Netherlands (8)	China (21)	Sweden (13)	Sweden (23)
6	Japan (11)	South Korea (13)	Vietnam (8)	Japan (17)	Germany (12)	Poland (23)
7	Germany-United Kingdom (10)	N/A	France (7)	United Kingdom (11)	Netherlands (11)	United Kingdom (22)
8	Netherlands (9)	N/A	South Korea (7)	N/A	Belgium (10)	France (21)
9	France-Germany (9)	N/A	Sweden (7)	N/A	France (10)	Germany-United Kingdom (20)
10	Taiwan (8)	N/A	Denmark (5)	N/A	Denmark (9)	South Korea (20)

Table 13: Most 10 frequent partial weekly and monthly periodic patterns for PADI-web, ProMED and Empres-i. The periodic support values for these patterns are indicated in parenthesis. These results are produced with the parameter values $\iota = 2$, $\varrho = 0.1$ and $\alpha = 1000km$. N/A indicates that there is no available entry. The Empres-i patterns detected by both PADI-web and ProMED are colored in orange, and it is colored in blue (resp. red) when only ProMED (resp. PADI-web) finds them.

PADI-web captures two seasonal patterns for Vietnam, but it still underrepresents it. Moreover, it also overrepresents the seasonal patterns for the United Kingdom, China and India. Overall, the average evaluation scores for PADI-web and ProMED based on the partial periodic seasonal frequent patterns are 0.51 and 0.85, respectively. When we consider the partial and full periodic frequent seasonal patterns together, we obtain the average scores of 0.50 and 0.51 for PADI-web and ProMED, respectively.

Next, we pass to continuous periodic patterns, i.e. the epidemiological events occurring consistently throughout the year, by applying the ST method with $\iota = 2$, $\varrho = 0.1$ and $\alpha = 1000km$. Some of our results are shown in Table 13 and Figure 6. Table 13 shows the most 10 frequent partial weekly and monthly periodic continuous patterns at country scale for each EBS/IBS system from 2019 to 2021. We expect an EBS system to provide a similar ranking as in Empres-i. In Table 13, the Empres-i partial continuous patterns detected by both PADI-web and ProMED are in orange, and it is colored in blue (resp. red) when only PADI-web (resp. ProMED) finds them.

Overall, we observe some differences across the results, and we interpret them in two parts. We first analyze the monthly partial periodic continuous patterns. On the one hand, we see that the ranking of ProMED at monthly scale is much more in line with Empres-i compared to PADI-web, as it captures seven monthly Empres-i patterns (out of 10). On the other hand, PADI-web detects only few partial continuous patterns. Furthermore, both PADI-web and ProMED are able to detect the same three monthly Empres-i patterns, related to the recurrent events occurring in Germany, Netherlands and France. Finally, only PADI-web (resp. ProMED) is able to detect the patterns for Taiwan (resp. Russia, Vietnam, Sweden and Denmark). Overall, when we consider the whole ranking results, we obtain the evaluation scores of 0.86 and 0.91 for PADI-web and ProMED, respectively.

Now, we pass to the weekly partial periodic continuous patterns. The identification of these patterns

is harder compared to the monthly patterns, because this amounts to seek the events occurring at least once every two weeks. Indeed, PADI-web (resp. ProMED) can only find six (resp. seven) partial weekly periodic patterns in total. Moreover, we see that the obtained patterns are slightly different compared to the monthly patterns, hence they give another temporal vision of the EBS systems. For instance, South Korea is not that frequent at monthly scale in the Empres-i data, but it is one of the most 10 frequent weekly partial periodic patterns (see also Figure 6). Of the detected patterns by PADI-web and ProMED, four of them are also found in Empres-i's result. This shows that both systems have a comparable performance at weekly scale. Overall, when we consider the weekly and monthly partial patterns together, we obtain the average evaluation scores of 0.69 and 0.75 for PADI-web and ProMED, respectively (see Section C.3.2 in the Appendix for the calculation details).

To conclude this part, identifying the full and partial weekly, monthly and seasonal periodic patterns gives a different analysis perspective to assess the performances of the EBS systems. Overall, both PADI-web and ProMED have comparable results. Nevertheless, there are some substantial differences between them. When it comes to the seasonal patterns presented in Table 12, on the one hand, ProMED finds less partial periodic seasonal patterns compared to PADI-web, but most of them are found in the Empres-i's result. On the other hand, PADI-web finds more seasonal patterns, but they are not as accurate as the patterns provided by ProMED. This is probably because several of them might be either false alerts, i.e. suspected cases being not confirmed by the national authorities, or directly erroneous due to the automatic processing framework of PADI-web. Consequently, it overrepresents some countries. When it comes to the weekly and monthly continuous patterns presented in Table 13, the obtained results are barely in line with the Empres-i's results. Nonetheless, ProMED performs slightly better, since it correctly finds more monthly partial patterns. Overall, we obtain the final periodicity scores of 0.59 and 0.63 by combining both continuous and seasonal periodicity aspects for PADI-web and ProMED, respectively. Finally, our results are also partially confirmed by the previous works [7]. In [7], the authors measure the correlation between the weekly event time series derived from ProMED, HealthMap¹⁰ and an official source WHO data using Pearson's correlation coefficient. They find out that the results derived from ProMED and HealthMap are moderately correlated with the ones reported by World Health Organization (WHO) on West African Ebola, and that there exist some substantial differences between them, particularly at the peak of the epidemics.

5.3 Thematic Dimension

In this section, we study how detailed EBS systems provide the thematic information encoded in their data. We want to know how similar the frequent multidimensional patterns across EBS systems are. We compare the results based on the Empres-i dataset to see to what extent PADI-web and ProMED can capture similar patterns. To ease our discussion, we also visualize the relations between spatial and thematic entities with a *chord diagram* in Figure 12, which is found in the Appendix for space matters.

Table 14 shows the most 13 frequent static and temporal multidimensional patterns for PADI-web, ProMED and Empres-i. We obtain these patterns with the parameters $\iota \in \{10, 30, \infty\}$ and $\varrho = 1$ (in count). Note that the use of ∞ represents a very large value for eliminating the periodicity aspect from the method ST. This amounts to obtain the static version of frequent multidimensional patterns, without any temporal aspect. We describe Table 14 generically here, for matters of convenience. Given a specific spatial scale, each row corresponds to a spatial entity and these entities are regrouped by a specific system. The columns represent the existing host entities in a specific hierarchical level, and they are regrouped by the existing disease entities in a specific hierarchical level. For instance, in Table 14 we stick to the second level of hierarchy for spatial, disease and host entities (see Table 18 in the Appendix for thematic taxonomy). Each cell can encode four different information. First, we display a dash character, when a specific multidimensional pattern, be static or temporal, is not frequent. Second, the statistics of a given multidimensional pattern is expressed in the format $x|y$. The first value x corresponds to the static condition, and represents its frequency number (i.e. support) without considering the partial periodicity constraint. The second value corresponds to the temporal condition, and represents its periodic support with respect to the parameters ι and ϱ . Third, we also highlight with different gray scales to what degree a multidimensional pattern is partially periodic in the data at hand. For instance, in Table 14, we consider two different ι values, which are 10 and 30 days. The results for the former (resp. latter) are indicated in dark (resp. light) gray. Finally, we show the Empres-i multidimensional patterns detected by both PADI-web and ProMED in orange, and they are colored in blue (resp. red) when only PADI-web (resp. ProMED) finds them.

We can summarize Table 14 in five points. First, as expected, Empres-i provides only fine-grained disease information. Interestingly, the data collected by ProMED is also fine-grained, whereas PADI-

¹⁰similar to PADI-web, as they both are automated systems.

	country	HPAI			LPAI			AI unknown		
		Domestic	Wild	Bird un-known	Domestic	Wild	Bird un-known	Domestic	Wild	Bird un-known
PADI-web	Japan	-	-	-	-	-	-	32 31	-	12 10
	United Kingdom	8 5	9 8	-	-	-	-	-	13 10	24 19
	South Korea	-	-	8 7	-	-	-	10 7	9 7	20 18
	China	11 10	-	9 7	-	-	-	-	-	-
	India	-	-	-	-	-	-	9 6	-	-
ProMED	Germany	6 2	12 9	-	-	-	-	-	-	-
	Vietnam	-	-	8 3	-	-	-	-	-	-
	Japan	7 6	-	-	-	-	-	-	-	-
	Ireland	5 3	-	-	-	-	-	-	-	-
	India	4 2	-	3 1	-	-	-	-	-	-
	China	3 2	-	4 1	-	-	-	-	-	-
	United Kingdom	-	3 1	-	-	-	-	-	-	-
	South Korea	-	3 2	-	-	-	-	-	-	-
	United States	-	-	-	-	-	-	3 2	-	-
	Hungary	3 1	-	-	-	-	-	-	-	-
EMPRES-i	Hungary	-	-	268 267	-	-	-	-	-	-
	Germany	-	267 265	41 38	-	-	-	-	-	-
	United Kingdom	-	102 101	-	-	-	-	-	-	-
	Denmark	-	96 95	-	-	-	-	-	-	-
	Vietnam	-	-	65 56	-	-	-	-	-	-
	Netherlands	-	57 55	-	-	-	-	-	-	-
	Taiwan	-	-	-	45 37	-	35 -	-	-	-
	Japan	-	-	44 43	-	-	-	-	-	-
	Russia	-	-	41 36	-	-	-	-	-	-
	Poland	-	-	40 33	-	-	-	-	-	-
	South Korea	-	37 35	- 33	-	-	-	-	-	-
	France	-	-	- 32	-	-	-	-	-	-

Table 14: Most 13 frequent static and temporal multidimensional patterns. To ease our discussion, these patterns are only at country and solely concern the events occurring in 2020. When a multidimensional pattern is static, it is shown in white. When it is partial periodic with 10 days (resp. 30 days), it is shown in dark (resp. light) gray. The Empres-i patterns detected by both PADI-web and ProMED are colored in orange, and it is colored in blue (resp. red) when only ProMED (resp. PADI-web) finds them.

web provides mostly coarse-grained disease information (see also Figure 12 in the Appendix). Second, the overwhelming majority of the frequent multidimensional patterns provided by PADI-web, ProMED and Empres-i concern the HPAI cases. This fact highlights how national and international authorities prioritize the surveillance of HPAI cases, since it is highly contagious among birds, and can be deadly, especially for domestic poultry. Third, apart the unknown bird category, most of the frequent multidimensional patterns of Empres-i (resp. PADI-web and ProMED) concerns wild (resp. domestic) birds. On this point, we can say that the distribution of host categories are not very balanced, with a slight dominance for unknown bird category. Fourth, the overwhelming majority of the frequent static multidimensional patterns for all systems are strongly partial periodic, with $\iota = 10$ days. Finally, when we compare the frequent multidimensional patterns across EBS systems, we observe that both PADI-web and ProMED detect few Empres-i patterns (four and two patterns for ProMED and PADI-web, respectively). This fact shows how the thematic data collected by PADI-web and ProMED can be different with respect to Empres-i. Interestingly, although PADI-web is currently collaborating with the French Platform for Animal Health Surveillance (see Section 4.1), France is not in the first 13 frequent patterns for PADI-web. Overall, when we consider all the static and temporal frequent multidimensional patterns together, we obtain the ranking scores of 0.64 and 0.63 for PADI-web and ProMED, respectively (see Section C.4 in the Appendix for the calculation details).

5.4 Source dimension

Finally, we assess how important and timely the news outlets involved in EBS systems for information dissemination. Next, we rank them in terms of these two objectives and see if PADI-web and ProMED obtain similar results. Overall, it is worth noticing that PADI-web (resp. ProMED) includes 480 (resp. 189) distinct news outlets. Only 63 of them are in common between PADI-web and ProMED. On top of the news outlets, ProMED also mostly relies on official reports from WOA (472 events out of 786). For this reason, we also include WOA in our analysis for ProMED. Nevertheless, we solely discuss the

rank	Asia		Europe		World	
	PADI-web	ProMED	PADI-web	ProMED	PADI-web	ProMED
1	hindustantimes (IND)	WOAH	reuters	WOAH	reuters	WOAH
2	indiatimes (IND)	yna.co.kr (KOR)	heraldscotland (GBR)	rossaprimavera.ru (RUS)	hindustantimes (IND)	rossaprimavera.ru (RUS)
3	businessworld.in (IND)	outbreaknewstoday	thepoultrysite	tatar-inform.ru (RUS)	thepoultrysite	yna.co.kr (KOR)
4	reuters	nippon (JPN)	francetvinfo (FRA)	bbc (GBR)	indiatimes (IND)	reuters
5	middleeastmonitor	newindianexpress (IND)	outbreaknewstoday	reuters	businessworld.in (IND)	nippon (JPN)
6	thepoultrysite	tuoitrenews.vn (VNM)	phys	foodingredientsfirst	heraldscotland (GBR)	niknews.mk.ua (UKR)
7	nippon (JPN)	niknews.mk.ua (UKR)	agriculture.com	cheshire-live.co.uk (GBR)	outbreaknewstoday	outbreaknewstoday
8	oneindia (IND)	outlookindia (IND)	dgwgo (GBR)	kazakh-zerno (KAZ)	nippon (JPN)	bbc
9	newindianexpress (IND)	thebeijinger (CHN)	republicain-lorrain.fr (FRA)	nltimes.nl (NLD)	middleeastmonitor	tass.ru (RUS)
10	indianexpress (IND)	russian.news.cn (CHN)	farminguk (GBR)	nv.ua (UKR)	agriculture.com	tatar-inform (RUS)

Table 15: PageRank centrality results for PADI-web, ProMED and Empres-i. The first (resp. second) part of the table corresponds to the results based on the events occurring only in Asia (resp. Europe). In the last part, the results are produced from the whole dataset.

rank	Asia		Europe		World	
	PADI-web	ProMED	PADI-web	ProMED	PADI-web	ProMED
1	indiatimes (IND)	WOAH	thepoultrysite	WOAH	indiatimes (IND)	WOAH
2	hindustantimes (IND)	outbreaknewstoday	francebleu (FRA)	vetandlife.ru (RUS)	hindustantimes (IND)	outbreaknewstoday
3	yna.co.kr (KOR)	yna.co.kr (KOR)	fwi.co.uk (GBR)	regnum.ru (RU)	thepoultrysite	vetandlife.ru (RUS)
4	thepoultrysite	koreaherald	farminguk (GBR)	tass.ru (RUS)	yna.co.kr (KOR)	yna.co.kr (KOR)
5	newindianexpress (IND)	indiatimes (IND)	reuters	rossaprimavera.ru (RUS)	reuters	reuters
6	thehindu (IND)	reuters	albaniandailynews (ALB)	life.ru (RUS)	outbreaknewstoday	koreaherald (KOR)
7	indianexpress (IND)	nippon (JPN)	heraldscotland (SCO)	bbc (GBR)	wattagnet	regnum.ru (RUS)
8	nippon (JPN)	vetandlife.ru (RUS)	wattagnet	khaleejtimes (ARE)	francebleu (FRA)	tass.ru (RUS)
9	reuters	poultrymed	20minutes.fr (FRA)	farminguk (GBR)	newindianexpress (IND)	indiatimes (IND)
10	kashmirobserver (IND)	newindianexpress (IND)	phys	aphascience.blog.gov.uk (GBR)	thehindu (IND)	rossaprimavera.ru (RUS)

Table 16: Timely detection results for PADI-web, ProMED and Empres-i. The first (resp. second) part of the table corresponds to the results based on the events occurring only in Asia (resp. Europe). In the last part, the results are produced from the whole dataset.

performances of the news outlets for a fair comparison.

First, we compare PADI-web and ProMED in terms of their important news outlets for information dissemination, obtained with the PageRank algorithm. Table 15 shows the first 10 (resp. 9) news outlets having the largest PageRank scores for PADI-web (resp. ProMED) based on the events occurring in Asia, Europe and in the whole world, respectively. We can summarize the results in two points. First, overall, the results show that the most important news outlets for both sources are almost completely different. PADI-web relies mostly on the Indian news outlets for the events in Asia and French and British/Scottish ones for Europe, whereas spatially more diverse news outlets are in ProMED’s results, with a slight prevalence for the Russian (and nearby countries such as Ukraine and Kazakhstan) and British/Scottish news outlets. Consequently, PADI-web and ProMED have only three news outlets in common. Finally, on top of national news outlets, several international ones, such as *Reuters* and *Outbreak News Today*, can also take an important role for information dissemination for PADI-web and ProMED. Nevertheless, their rankings can be very different. For instance, Reuters is the first news outlets for the world-wide events, whereas it is at 10th place for ProMED. This also confirms us how different the news collection strategies of PADI-web and ProMED are.

Next, we pass to the results of timely news outlets involved in PADI-web and ProMED, obtained with the method CELF by limiting the output size to the first 30 news outlets. Similar to the previous results, Table 16 shows only the first 10 (resp. 9) timely news outlets for PADI-web (resp. ProMED) based on the events occurring in Asia, Europe and in the whole world, respectively. Compared to the previous results in Table 15, we see here that PADI-web and ProMED share more common news outlets in terms of timely detection. Furthermore, we observe that the most important news outlets in terms of PageRank score are not necessarily timely in event detection. In other words, we observe

Dimension	Description	Qualitative Evaluation		Qualitative Evaluation		Quantitative Eval.	
		PADI-web (+)	PADI-web (-)	ProMED (+)	ProMED (-)	PADI-web	ProMED
Spatial	Spatio-temporal representativeness	<ul style="list-style-type: none"> Event reporting for 56 countries (Table 8). Well covering 21 countries. 	<ul style="list-style-type: none"> Absence of events from 14 countries that ProMED covers. Less reporting capacity for Eastern Europe and Middle East. 	<ul style="list-style-type: none"> Event reporting for 70 countries (Table 8). Well covering 36 countries. Thanks to the WOAH reports, covering the countries better than PADI-web. 	<ul style="list-style-type: none"> Absence of events from 3 countries that Empres-i reports. Less reporting capacity for Western Europe, Asia and the USA. 	0.40	0.55
Temporal	Timeliness	<ul style="list-style-type: none"> Timely detection performance: 49% (Table 11). 	<ul style="list-style-type: none"> 14.8 days delay in average (by focusing solely on late events). 	<ul style="list-style-type: none"> Timely detection performance: 28% (Table 11). 	<ul style="list-style-type: none"> 6 days delay in average (by focusing solely on late events). 	0.18	0.12
	Periodicity	<ul style="list-style-type: none"> Detecting more accurate full periodic seasonal patterns than ProMED does (Table 12). Detecting weekly and montly partial periodic patterns for 2 countries that ProMED does not (Table 13). 	<ul style="list-style-type: none"> Detecting few weekly partial continuous patterns. Detecting less accurate monthly partial periodic continuous patterns than ProMED does. 	<ul style="list-style-type: none"> Detecting more accurate partial periodic seasonal and monthly partial periodic continuous patterns than PADI-web does (Table 12). Detecting weekly and montly partial periodic patterns for 4 countries that ProMED does not (Table 13). 	<ul style="list-style-type: none"> Detecting less accurate full periodic seasonal patterns than PADI-web does. Detecting few weekly partial continuous patterns. 	0.59	0.63
Thematic	Static and temporal multidimensional patterns	<ul style="list-style-type: none"> Detecting 1 frequent multidimensional pattern that ProMED does not (Table 14). Balanced distribution of domestic and wild bird cases. Strongly partial periodicity for its frequent multidimensional patterns. 	<ul style="list-style-type: none"> Providing less detailed disease and host information than ProMED does. Absence of LPAI cases. 	<ul style="list-style-type: none"> Providing more detailed disease and host information than PADI-web does. Detecting 3 frequent multidimensional patterns that PADI-web does not (Table 14). Strongly partial periodicity for its frequent multidimensional patterns. 	<ul style="list-style-type: none"> Providing events concerning mostly domestic birds, i.e. less balanced distribution. Absence of LPAI cases. 	0.64	0.63
Source	Important and timely news outlets	<ul style="list-style-type: none"> Relying on more timely important news outlets than ProMED Tables 15 and 16. Using different national and international news outlets than ProMED. 	<ul style="list-style-type: none"> Relying mostly on the Indian, French and British/Scottish news outlets. Relying less timely important news outlets than PADI-web. 	<ul style="list-style-type: none"> Using different national and international news outlets than PADI-web. Relying on WOAH, which is important and timely Tables 15 and 16. 	<ul style="list-style-type: none"> Relying mostly on the Russian (also, nearby countries such as Kazakhstan) and British news outlets. Relying on less timely important news outlets than PADI-web. 	0.96	0.89

Table 17: Summary of findings regarding all evaluation results between PADI-web and ProMED.

some inconsistency issues in the results obtained by the methods PageRank and CELF for PADI-web and ProMED. Recall that this consistency assessment allows us verifying whether news outlets playing a key role in epidemiological information dissemination are also timely in event detection. For instance, regarding the events occurring in Europe for PADI-web, the French news outlets *France Bleu* and *20 Minutes* appear only in Table 16, whereas we observe two other French news outlets in Table 15. Nevertheless, this kind of inconsistencies seems not to appear much in the whole data, i.e. world scale. Indeed, the rank evaluation scores also confirm this last point, as we obtain the ranking scores of 0.96 and 0.89 for PADI-web and ProMED, respectively (see Section C.5 in the Appendix for the calculation details).

To conclude this part, we show that PADI-web and ProMED rely mostly on different important and timely news outlets. This is mostly because the Indian, French and British/Scottish news outlets take an important role for PADI-web, whereas these are mostly the Russian (and nearby countries such as Ukraine and Kazakhstan) and British/Scottish news outlets for ProMED. Moreover, both EBS systems also rely on the same international news outlets (e.g. Reuters), nevertheless these news outlets do not contribute to both EBS systems in the same manner. All these results suggest for these EBS systems to include more spatially more diverse news outlets for a greater geographic coverage (e.g. Baidu for Chinese news). Finally, we present a summary of findings in Table 17 and Figure 7 based on all obtained evaluation results from this section and the previous ones. We see that PADI-web and ProMED seem to be complementary. PADI-web (resp. ProMED) performs slightly better for the timeliness and source (resp. spatial and periodicity) dimensions, and they have a comparable performance for the thematic dimension. For the sake of completeness, we also compare in Figure 8 these EBS systems in terms of the presented five dimensions by discarding the ProMED data provided by official data sources (i.e. WOAH reports).

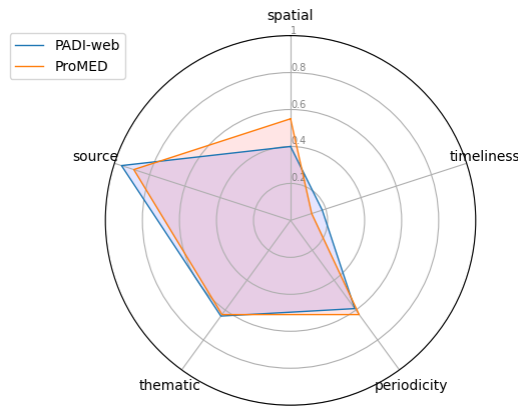


Figure 7: Radar chart summarizing all quantitative evaluation results between PADI-web and ProMED.

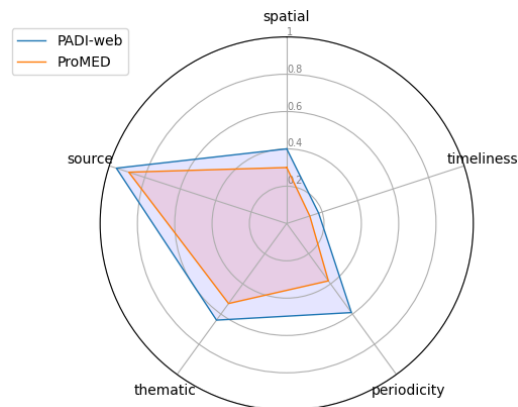


Figure 8: Radar chart summarizing all quantitative evaluation results between PADI-web and ProMED, when we discard the ProMED data provided by official data sources (i.e. WOH reports).

6 Conclusion

In this article, we have presented a new evaluation framework to identify the strengths and drawbacks of EBS systems in terms of epidemic surveillance. This evaluation is very valuable from the epidemiological standpoint, since it allows end-users to select the most appropriate EBS system(s) for an effective surveillance of a particular situation. We want not only to compare EBS systems, but also to produce results that the end user can easily interpret. For this purpose, we proposed a two-step framework based on our review of the literature. It first transforms the raw input event data into a set of normalized distinct events, then conducts a descriptive retrospective analysis of these events with four objectives: spatial, temporal, thematic and source analysis. We illustrated its relevance by applying it to an Avian Influenza dataset collected by PADI-web, ProMED and Empres-i. We showed that our framework allows identifying the strengths and drawbacks of the considered EBS systems. For some of our evaluation aspects, our results confirm the findings already published in the literature. For others, the systematic nature of our approach uncovers new findings for the considered EBS systems.

Our work could be extended in several ways. First, our method can be applied systematically to other EBS systems and other animal diseases, for the sake of completeness. This would give a better overview of the capabilities of the existing systems. Second, the source dimension can be better evaluated in Section 3.6, if we can obtain an appropriate gold standard dataset. This would imply to conduct an extensive work like in [76], but tailored to Epidemic Intelligence. Third, our evaluation framework focuses only on a descriptive retrospective analysis. Nevertheless, it would be also valuable to extend this work with a predictive analysis to see to what extent the existing EBS data can give an insight on the short or long term future using past event information, accompanied by domain-specific data (e.g. animal mobility models, environmental data). Some examples are risk mapping [17, 29, 54, 64] and epidemic forecasting with sparse data [30]. Finally, due to the generic nature of our evaluation framework, it can be also applied to other spatio-temporal systems with similar properties (e.g. natural

disaster surveillance systems), so this could constitute another perspective.

Acknowledgments. This study was partially funded by EU grant 874850 MOOD and is catalogued as MOOD063. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission. We thank the ProMED-mail staff for sharing with us their dataset, as well as for their constructive comments to improve the content of this paper. ProMED is a program of the International Society for Infectious Diseases (ISID).

References

- [1] T. T. Ao, M. Rahman, F. Haque, A. Chakraborty, M. J. Hossain, S. Haider, A. S.M. Alamgir, J. Sobel, S. P. Luby, and E. S. Gurley. “Low-Cost National Media-Based Surveillance System for Public Health Events, Bangladesh”. In: *Emerging Infectious Diseases* 22.4 (2016), pp. 720–722. DOI: [10.3201/eid2204.150330](https://doi.org/10.3201/eid2204.150330).
- [2] E. Arsevska, M. Roche, P. Hendriks, D. Chavernac, S. Falala, R. Lancelot, and B. Dufour. “Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web”. In: *Computers and Electronics in Agriculture* 123 (2016), pp. 104–115. DOI: [10.1016/j.compag.2016.02.010](https://doi.org/10.1016/j.compag.2016.02.010).
- [3] E. Arsevska, S. Valentin, J. Rabatel, J. de Goër de Hervé, S. Falala, R. Lancelot, and M. Roche. “Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System”. In: *PLOS ONE* 13.8 (2018), pp. 1–25. DOI: [10.1371/journal.pone.0199960](https://doi.org/10.1371/journal.pone.0199960).
- [4] C. Y. Bahk, D. A. Scales, S. R. Mekaru, J. S. Brownstein, and C. C. Freifeld. “Comparing timeliness, content, and disease severity of formal and informal source outbreak reporting”. In: *BMC Infectious Diseases* 15.1 (2015). DOI: [10.1186/s12879-015-0885-0](https://doi.org/10.1186/s12879-015-0885-0).
- [5] P. Barboza, L. Vaillant, Y. Le Strat, D. M. Hartley, N. P. Nelson, A. Mawudeku, L. C. Madoff, J. P. Linge, N. Collier, J. S. Brownstein, and P. Astagneau. “Factors Influencing Performance of Internet-Based Biosurveillance Systems Used in Epidemic Intelligence for Early Detection of Infectious Diseases Outbreaks”. In: *PLoS ONE* 9.3 (2014). Ed. by Vishnu Chaturvedi, e90536. DOI: [10.1371/journal.pone.0090536](https://doi.org/10.1371/journal.pone.0090536).
- [6] 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, and P. Astagneau. “Evaluation of Epidemic Intelligence Systems Integrated in the Early Alerting and Reporting Project for the Detection of A/H5N1 Influenza Events”. In: *PLoS ONE* 8.3 (2013). Ed. by Hiroshi Nishiura, e57252. DOI: [10.1371/journal.pone.0057252](https://doi.org/10.1371/journal.pone.0057252).
- [7] S. Bhatia, B. Lassmann, E. Cohn, A. N. Desai, M. Carrion, M. U. G. Kraemer, M. Herringer, J. Brownstein, L. Madoff, A. Cori, and P. Nouvellet. “Using digital surveillance tools for near real-time mapping of the risk of infectious disease spread”. In: *npj Digital Medicine* 4.1 (2021). DOI: [10.1038/s41746-021-00442-3](https://doi.org/10.1038/s41746-021-00442-3).
- [8] S. Bhatia, J. Wardle, R. K. Nash, P. Nouvellet, and A. Cori. “A generic method and software to estimate the transmission advantage of pathogen variants in real-time : SARS-CoV-2 as a case-study”. In: *medRxiv* (2021). DOI: [10.1101/2021.11.26.21266899](https://doi.org/10.1101/2021.11.26.21266899).
- [9] D. Birant and A. Kut. “ST-DBSCAN: An algorithm for clustering spatial-temporal data”. In: *Data & Knowledge Engineering* 60.1 (2007), pp. 208–221. DOI: [10.1016/j.datak.2006.01.013](https://doi.org/10.1016/j.datak.2006.01.013).
- [10] M. N. K. Boulos and E. M. Geraghty. “Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fight against outbreaks and epidemics”. In: *International Journal of Health Geographics* 19.1 (2020). DOI: [10.1186/s12942-020-00202-8](https://doi.org/10.1186/s12942-020-00202-8).
- [11] M. Carrion and L. C. Madoff. “ProMED-mail: 22 years of digital surveillance of emerging infectious diseases”. In: *International Health* 9.3 (2017), pp. 177–183. DOI: [10.1093/inthealth/ihx014](https://doi.org/10.1093/inthealth/ihx014).
- [12] J. Chołoniowski, J. Sienkiewicz, G. Leban, and J. A. Hołyst. “Modeling of temporal fluctuation scaling in online news network with independent cascade model”. In: *Physica A: Statistical Mechanics and its Applications* 523 (2019), pp. 129–144. DOI: [10.1016/j.physa.2019.02.035](https://doi.org/10.1016/j.physa.2019.02.035).
- [13] F. Claes, D. Kuznetsov, R. Liechti, S. Von Dobschuetz, B. D. Truong, A. Gleizes, D. Conversa, A. Colonna, E. Demaio, S. Ramazzotto, F. Larfaoui, J. Pinto, P. Le Mercier, I. Xenarios, and G. Dauphin. “The EMPRES-i genetic module: a novel tool linking epidemiological outbreak information and genetic characteristics of influenza viruses”. In: *Database* 2014.0 (2014), bau008–bau008. DOI: [10.1093/database/bau008](https://doi.org/10.1093/database/bau008).
- [14] N. Collier, S. Doan, A. Kawazoe, R. M. Goodwin, M. Conway, Y. Tateno, Q.-H. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, M. Shigematsu, and K. Taniguchi. “BioCaster: detecting public health rumors with a Web-based text mining system”. In: *Bioinformatics* 24.24 (2008), pp. 2940–2941.

- [15] A. Cori, N. M. Ferguson, C. Fraser, and S. Cauchemez. "A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics". In: *American Journal of Epidemiology* 178.9 (2013), pp. 1505–1512. DOI: [10.1093/aje/kwt133](https://doi.org/10.1093/aje/kwt133).
- [16] S. Dharmaratne, S. Sudaraka, I. Abeyagunawardena, K. Manchanayake, M. Kothalawala, and W. Gunathunga. "Estimation of the basic reproduction number (R0) for the novel coronavirus disease in Sri Lanka". In: *Virology Journal* 17.1 (2020). DOI: [10.1186/s12985-020-01411-0](https://doi.org/10.1186/s12985-020-01411-0).
- [17] B. Durand, A. Tran, G. Balança, and V. Chevalier. "Geographic variations of the bird-borne structural risk of West Nile virus circulation in Europe". In: *PLOS ONE* 12.10 (2017). Ed. by Massimo Ciccozzi, e0185962. DOI: [10.1371/journal.pone.0185962](https://doi.org/10.1371/journal.pone.0185962).
- [18] EMPRES-i. *empres-i.fao.org*. Sept. 2013.
- [19] L. Espinosa, A. Wijermans, F. Orchard, M. Höhle, T. Czernichow, P. Coletti, L. Hermans, C. Faes, E. Kissling, and T. Mollet. "Epi tweeter: Early warning of public health threats using Twitter data". In: *Eurosurveillance* 27.39 (2022). DOI: [10.2807/1560-7917.es.2022.27.39.2200177](https://doi.org/10.2807/1560-7917.es.2022.27.39.2200177).
- [20] Matthew L. Farnsworth, Christopher Hamilton-West, Stephanie Fitchett, Scott H. Newman, Stephane de La Rocque, Lorenzo De Simone, Juan Lubroth, and Julio Pinto. "Comparing national and global data collection systems for reporting, outbreaks of H5N1 HPAI". In: *Preventive Veterinary Medicine* 95.3-4 (2010), pp. 175–185. DOI: [10.1016/j.prevetmed.2010.03.012](https://doi.org/10.1016/j.prevetmed.2010.03.012).
- [21] S. Federhen. "The NCBI Taxonomy database". In: *Nucleic Acids Research* 40.D1 (2011), pp. D136–D143. DOI: [10.1093/nar/gkr1178](https://doi.org/10.1093/nar/gkr1178).
- [22] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein. "HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports". In: *Journal of the American Medical Informatics Association* 15.2 (2008), pp. 150–157. DOI: [10.1197/jamia.m2544](https://doi.org/10.1197/jamia.m2544).
- [23] M. Gehlen, M. R. C. Nicola, E. R. D. Costa, V. K. Cabral, E. L. L. de Quadros, C. O. Chaves, R. A. Lahm, A. D. R. Nicoletta, M. L. R. Rossetti, and D. R. Silva. "Geospatial intelligence and health analytics: Its application and utility in a city with high tuberculosis incidence in Brazil". In: *Journal of Infection and Public Health* 12.5 (2019), pp. 681–689. DOI: [10.1016/j.jiph.2019.03.012](https://doi.org/10.1016/j.jiph.2019.03.012).
- [24] A. Getis and J. K. Ord. "The Analysis of Spatial Association by Use of Distance Statistics". In: *Geographical Analysis* 24.3 (2010), pp. 189–206. DOI: [10.1111/j.1538-4632.1992.tb00261.x](https://doi.org/10.1111/j.1538-4632.1992.tb00261.x).
- [25] R. Goel, Valentin S., Delaforge A., S. Fadloun, A. Sallaberry, M. Roche, and P. Poncet. "EpidNews: Extracting, exploring and annotating news for monitoring animal diseases". In: *Journal of Computer Languages* 56 (2020), p. 100936. DOI: [10.1016/j.col.2019.100936](https://doi.org/10.1016/j.col.2019.100936).
- [26] J. Hills, J. Lines, E. Baranauskas, J. Mapp, and A. Bagnall. "Classification of time series by shapelet transformation". In: *Data Mining and Knowledge Discovery* 28.4 (2013), pp. 851–881. DOI: [10.1007/s10618-013-0322-1](https://doi.org/10.1007/s10618-013-0322-1).
- [27] M. Honnibal and I. Montani. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". To appear. 2017.
- [28] D.-C. Huang, J.-F. Wang, J.-X. Huang, D. Z. Sui, H.-Y. Zhang, M.-G. Hu, and C.-D. Xu. "Towards Identifying and Reducing the Bias of Disease Information Extracted from Search Engine Data". In: *PLOS Computational Biology* 12.6 (2016). Ed. by Marcel Salathé, e1004876. DOI: [10.1371/journal.pcbi.1004876](https://doi.org/10.1371/journal.pcbi.1004876).
- [29] K. E. Jones, N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman, and P. Daszak. "Global trends in emerging infectious diseases". In: *Nature* 451.7181 (2008), pp. 990–993. DOI: [10.1038/nature06536](https://doi.org/10.1038/nature06536).
- [30] S. Kandula, T. Yamana, S. Pei, W. Yang, H. Morita, and J. Shaman. "Evaluation of mechanistic and statistical methods in forecasting influenza-like illness". In: *Journal of The Royal Society Interface* 15.144 (2018), p. 20180174. DOI: [10.1098/rsif.2018.0174](https://doi.org/10.1098/rsif.2018.0174).
- [31] R. U. Kiran, C. Saideep, K. Zettsu, M. Toyoda, M. Kitsuregawa, and P. K. Reddy. "Discovering Partial Periodic Spatial Patterns in Spatiotemporal Databases". In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019. DOI: [10.1109/bigdata47090.2019.9005693](https://doi.org/10.1109/bigdata47090.2019.9005693).
- [32] Kazuaki Kishida. "Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments". In: *NII Technical Reports* 2005.14 (2005), pp. 1–19. ISSN: 1346-5597.
- [33] H. W. Kuhn. "The Hungarian method for the assignment problem". In: *Naval Research Logistics Quarterly* 2.1-2 (1955), pp. 83–97. DOI: [10.1002/nav.3800020109](https://doi.org/10.1002/nav.3800020109).
- [34] G. Lejeune, A. Doucet, R. Yangarber, and N. Lucas. "Filtering news for epidemic surveillance: towards processing more languages with fewer resources". In: *CLIA: Fourth International Workshop On Cross Lingual Information Access* (2010), pp. 3–10.
- [35] Gaël Lejeune, Romain Brixtel, Antoine Doucet, and Nadine Lucas. "Multilingual event extraction for epidemic detection". In: *Artificial Intelligence in Medicine* 65.2 (2015), pp. 131–143. DOI: [10.1016/j.artmed.2015.06.005](https://doi.org/10.1016/j.artmed.2015.06.005).

- [36] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. “Cost-effective outbreak detection in networks”. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07*. KDD '07. ACM Press, 2007, pp. 420–429. DOI: [10.1145/1281192.1281239](https://doi.org/10.1145/1281192.1281239).
- [37] L. Leydesdorff and H. W. Park. “Full and fractional counting in bibliometric networks”. In: *Journal of Informetrics* 11.1 (2017), pp. 117–120. DOI: [10.1016/j.joi.2016.11.007](https://doi.org/10.1016/j.joi.2016.11.007).
- [38] Y. Li, Z. Lou, Y. Shi, and J. Han. “Temporal motifs in heterogeneous information networks”. In: *MLG Workshop@ KDD* (2018).
- [39] J. P. Linge, R. Steinberger, F. Fuat, S. Bucci, J. Belyaeva, M. Gemo, D. Al-Khudhairy, R. Yangarber, and E. van der Goot. “MedlSys”. In: *Advanced ICTs for Disaster Management and Threat Detection*. Ed. by E. Asimakopoulou and N. Bessis. IGI Global, 2010, pp. 131–142. DOI: [10.4018/978-1-61520-987-3.ch009](https://doi.org/10.4018/978-1-61520-987-3.ch009).
- [40] Y. Liu, H. Peng, J. Li, Y. Song, and X. Li. “Event detection and evolution in multi-lingual social streams”. In: *Frontiers of Computer Science* 14.5 (2020), p. 145612.
- [41] Z. Lu, W. Yu, R. Zhang, J. Li, and H. Wei. “Discovering Event Evolution Chain in Microblog”. In: *IEEE 17th International Conference on High Performance Computing and Communications*. 2015, pp. 635–640. DOI: [10.1109/HPC-CSS-ICCESS.2015.81](https://doi.org/10.1109/HPC-CSS-ICCESS.2015.81).
- [42] A. Lyon, M. Nunn, G. Gossel, and M. Burgman. “Comparison of Web-Based Biosecurity Intelligence Systems: BioCaster, EpiSPIDER and HealthMap”. In: *Transboundary and Emerging Diseases* 59.3 (2011), pp. 223–232. DOI: [10.1111/j.1865-1682.2011.01258.x](https://doi.org/10.1111/j.1865-1682.2011.01258.x).
- [43] Y. Mao, N. Zhang, B. Zhu, J. Liu, and R. He. “A descriptive analysis of the Spatio-temporal distribution of intestinal infectious diseases in China”. In: *BMC Infectious Diseases* 19.1 (2019). DOI: [10.1186/s12879-019-4400-x](https://doi.org/10.1186/s12879-019-4400-x).
- [44] Y. Matsubara, Y. Sakurai, W. G. van Panhuis, and C. Faloutsos. “FUNNEL: Automatic Mining of Spatially Coevolving Epidemics”. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014. DOI: [10.1145/2623330.2623624](https://doi.org/10.1145/2623330.2623624).
- [45] Z. Meng, A. Okhmatovskaia, M. Polleri, Y. Shen, G. Powell, Z. Fu, I. Ganser, M. Zhang, N. B. King, D. Buckeridge, and N. Collier. “BioCaster in 2021: automatic disease outbreaks detection from global news media”. In: *Bioinformatics* 38.18 (2022). Ed. by Zhiyong Lu, pp. 4446–4448. DOI: [10.1093/bioinformatics/btac497](https://doi.org/10.1093/bioinformatics/btac497).
- [46] N. M. M'ikanatha, R. Lynfield, K. G. Julian, C. A. Van Beneden, and H. de Valk. “Infectious disease surveillance: a cornerstone for prevention and control”. In: *Infectious Disease Surveillance*. John Wiley & Sons Ltd, 2013. Chap. 1, pp. 1–20. DOI: [10.1002/9781118543504.ch1](https://doi.org/10.1002/9781118543504.ch1).
- [47] P. A. P. Moran. “Notes on Continuous Stochastic Phenomena”. In: *Biometrika* 37.1/2 (1950), pp. 17–23. DOI: [10.2307/2332142](https://doi.org/10.2307/2332142).
- [48] A. Mueen. “Time series motif discovery: dimensions and applications”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4.2 (2014), pp. 152–159. DOI: [10.1002/widm.1119](https://doi.org/10.1002/widm.1119).
- [49] S. N. Mukhi, E. Andres, B. Demianyk, B. Gammon, and H. Kloetze. “KIWI: A technology for public health event monitoring and early warning detection”. In: *Online Journal of Public Health Informatics* 8.3 (2016). DOI: [10.5210/ojphi.v8i3.6937](https://doi.org/10.5210/ojphi.v8i3.6937).
- [50] R. K. Nash, P. Nouvellet, and A. Cori. “Real-time estimation of the epidemic reproduction number: Scoping review of the applications and challenges”. In: *PLOS Digital Health* 1.6 (2022). Ed. by Michele Tizzoni, e0000052. DOI: [10.1371/journal.pdig.0000052](https://doi.org/10.1371/journal.pdig.0000052).
- [51] V. Ng, E. E. Rees, J. Niu, and Zaghlool A. “Application of natural language processing algorithms for extracting information from news articles in event-based surveillance”. In: *Canada Communicable Disease Report* 46.6 (2020), pp. 186–191. DOI: [10.14745/ccdr.v46i06a06](https://doi.org/10.14745/ccdr.v46i06a06).
- [52] S. V. de Oliveira and R. N. Angerami. “Timeliness in the notification of spotted fever in Brazil: Evaluating compulsory reporting strategies and digital disease detection”. In: *International Journal of Infectious Diseases* 72 (2018), pp. 16–18. DOI: [10.1016/j.ijid.2018.04.4317](https://doi.org/10.1016/j.ijid.2018.04.4317).
- [53] R. Pakzad, I. Pakzad, S. Safiri, M. R. Shirzadi, M. Mohammadpour, A. Behroozi, Mark J. M. Sullman, and A. Janati. “Spatiotemporal analysis of brucellosis incidence in Iran from 2011 to 2014 using GIS”. In: *International Journal of Infectious Diseases* 67 (2018), pp. 129–136. DOI: [10.1016/j.ijid.2017.10.017](https://doi.org/10.1016/j.ijid.2017.10.017).
- [54] M. C. Paul, F. L. Goutard, F. Roulleau, D. Holl, W. Thanapongtharm, F. L. Roger, and A. Tran. “Quantitative assessment of a spatial multicriteria model for highly pathogenic avian influenza H5N1 in Thailand, and application in Cambodia”. In: *Scientific Reports* 6.1 (2016). DOI: [10.1038/srep31096](https://doi.org/10.1038/srep31096).
- [55] H. Pinto, J. Han, J. Pei, K. Wang, Q. Chen, and U. Dayal. “Multi-dimensional sequential pattern mining”. In: *Proceedings of the tenth international conference on Information and knowledge management - CIKM'01*. ACM Press, 2001. DOI: [10.1145/502585.502600](https://doi.org/10.1145/502585.502600).

- [56] M. Plantevit, A. Laurent, D. Laurent, M. Teisseire, and Y. W. Choong. “Mining multidimensional and multilevel sequential patterns”. In: *ACM Transactions on Knowledge Discovery from Data* 4.1 (2010), pp. 1–37. DOI: [10.1145/1644873.1644877](https://doi.org/10.1145/1644873.1644877).
- [57] L. Ramshaw and R. E. Tarjan. *On Minimum-Cost Assignments in Unbalanced Bipartite Graphs*. Tech. rep. HPL-2012-40R1. Palo Alto, CA: HP Labs, 2012.
- [58] Hal Roberts, Rahul Bhargava, Linas Valiukas, Dennis Jen, Momin M. Malik, Cindy Sherman Bishop, Aashka Dave, Emily Ndulue, Anushka Shah, Justin Clark, Jasmin Rubinowitz, Alexis Hope, Bruce Etling, Robert Faris, Catherine D’Ignazio, Fernando Bermejo, Yochai Benkler, and Ethan Zuckerman. *Snapshot data for: Media Cloud: Massive Open Source Collection of Global News on the Open Web*. Version V1. 2021. DOI: [10.7910/DVN/4F7TZW](https://doi.org/10.7910/DVN/4F7TZW). URL: <https://doi.org/10.7910/DVN/4F7TZW>.
- [59] M. Roche, J. Azé, Y. Kodratoff, and M. Sebag. “Learning Interestingness Measures in Terminology Extraction. A ROC-based approach”. In: *ROCAI* (2004), pp. 81–88.
- [60] M. A. Rodriguez and M. J. Egenhofer. “Determining semantic similarity among entity classes from different ontologies”. In: *IEEE Transactions on Knowledge and Data Engineering* 15.2 (2003), pp. 442–456. DOI: [10.1109/tkde.2003.1185844](https://doi.org/10.1109/tkde.2003.1185844).
- [61] Agnès Rortais, Jenya Belyaeva, Monica Gemo, Erik Van der Goot, and Jens P Linge. “MedISys: An early-warning system for the detection of (re-) emerging food-and feed-borne hazards”. In: *Food Research International* 43.5 (2010), pp. 1553–1556.
- [62] D. Scales, A. Zelenev, and J. S. Brownstein. “Quantifying the effect of media limitations on outbreak data in a global online web-crawling epidemic intelligence system, 2008–2011”. In: *Emerging Health Threats Journal* 6.1 (2013), p. 21621. DOI: [10.3402/ehj.v6i0.21621](https://doi.org/10.3402/ehj.v6i0.21621).
- [63] J. S. Schwind, D. J. Wolking, J. S. Brownstein, J. A. K. Mazet, and W. A. Smith. “Evaluation of Local Media Surveillance for Improved Disease Recognition and Monitoring in Global Hotspot Regions”. In: *PLoS ONE* 9.10 (2014). Ed. by Michelle L. Gatton, e110236. DOI: [10.1371/journal.pone.0110236](https://doi.org/10.1371/journal.pone.0110236).
- [64] K. B. Stevens, M. Gilbert, and D. U. Pfeiffer. “Modeling habitat suitability for occurrence of highly pathogenic avian influenza virus H5N1 in domestic poultry in Asia: A spatial multicriteria decision analysis approach”. In: *Spatial and Spatio-temporal Epidemiology* 4 (2013), pp. 1–14. DOI: [10.1016/j.sste.2012.11.002](https://doi.org/10.1016/j.sste.2012.11.002).
- [65] J. Strötgen and M. Gertz. “HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions”. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. SemEval ’10. Association for Computational Linguistics, 2010, 321–324.
- [66] F.-J. Tsai, E. Tseng, C.-C. Chan, H. Tamashiro, S. Motamed, and A. C. Rougemont. “Is the reporting timeliness gap for avian flu and H1N1 outbreaks in global health surveillance systems associated with country transparency?” In: *Globalization and Health* 9.1 (2013), p. 14. DOI: [10.1186/1744-8603-9-14](https://doi.org/10.1186/1744-8603-9-14).
- [67] S. Valentin, E. Arsevska, J. Rabatel, S. Falala, A. Mercier, R. Lancelot, and M. Roche. “PADI-web 3.0: A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance”. In: *One Health* 13 (2021), p. 100357. DOI: [10.1016/j.onehlt.2021.100357](https://doi.org/10.1016/j.onehlt.2021.100357).
- [68] S. Valentin, B. Boudoua, K. Sewalk, N. Arınık, M. Roche, R. Lancelot, and E. Arsevska. “Dissemination of information in event-based surveillance, a case study of Avian Influenza (submitted)”. In: *PLOS ONE* (2022).
- [69] S. Valentin, R. Lancelot, and M. Roche. “Fusion of spatiotemporal and thematic features of textual data for animal disease surveillance”. In: *Information Processing in Agriculture* (2022). DOI: [10.1016/j.inpa.2022.03.004](https://doi.org/10.1016/j.inpa.2022.03.004).
- [70] S. Valentin, R. Lancelot, and M. Roche. “Identifying associations between epidemiological entities in news data for animal disease surveillance”. In: *Artificial Intelligence in Agriculture* 5 (2021), pp. 163–174. DOI: [10.1016/j.aiia.2021.07.003](https://doi.org/10.1016/j.aiia.2021.07.003).
- [71] S. Valentin, A. Mercier, R. Lancelot, M. Roche, and E. Arsevska. “Monitoring online media reports for early detection of unknown diseases: Insight from a retrospective study of COVID-19 emergence”. In: *Transboundary and Emerging Diseases* 68.3 (2020), pp. 981–986. DOI: [10.1111/tbed.13738](https://doi.org/10.1111/tbed.13738).
- [72] T. Wang, X. Wang, P. Tie, Y. Bai, Y. Zheng, Changfu Yan, Z. Chai, J. Chen, H. Rao, L. Zeng, L. Chen, and L. Qiu. “Spatio-temporal cluster and distribution of human brucellosis in Shanxi Province of China between 2011 and 2016”. In: *Scientific Reports* 8.1 (2018). DOI: [10.1038/s41598-018-34975-7](https://doi.org/10.1038/s41598-018-34975-7).
- [73] W. Xiang and B. Wang. “A Survey of Event Extraction From Text”. In: *IEEE Access* 7 (2019), pp. 173111–173137. DOI: [10.1109/access.2019.2956831](https://doi.org/10.1109/access.2019.2956831).
- [74] M. Xue, Z. Huang, Y. Hu, J. Du, M. Gao, R. Pan, Y. Mo, J. Zhong, and Z. Huang. “Monitoring European data with prospective space–time scan statistics: predicting and evaluating emerging clusters of COVID-19 in European countries”. In: *BMC Public Health* 22.1 (2022), p. 2183. DOI: [10.1186/s12889-022-14298-z](https://doi.org/10.1186/s12889-022-14298-z).

- [75] R. Yan, Y. Li, Y. Zhang, and X. Li. “Event Recognition from News Webpages through Latent Ingredients Extraction”. In: *Information Retrieval Technology*. Ed. by P.-J. Cheng, M.-Y. Kan, W. Lam, and P. Nakov. Springer Berlin Heidelberg, 2010, pp. 490–501. DOI: [10.1007/978-3-642-17187-1_47](https://doi.org/10.1007/978-3-642-17187-1_47).
- [76] J. Ye and S. Skiena. “MediaRank: Computational Ranking of Online News Sources”. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Association for Computing Machinery, 2019, 2469–2477. DOI: [10.1145/3292500.3330709](https://doi.org/10.1145/3292500.3330709).
- [77] E. Yom-Tov, D. Borsa, I. J. Cox, and R. A. McKendry. “Detecting Disease Outbreaks in Mass Gatherings Using Internet Data”. In: *Journal of Medical Internet Research* 16.6 (2014), e154. DOI: [10.2196/jmir.3156](https://doi.org/10.2196/jmir.3156).
- [78] D. Zeng, Z. Cao, and D. B. Neill. “Artificial intelligence-enabled public health surveillance—from local detection to global epidemic monitoring and control”. In: *Artificial Intelligence in Medicine*. Ed. by Lei Xing, Maryellen L. Giger, and James K. Min. Academic Press, 2021. Chap. 22, pp. 437–453. DOI: [10.1016/b978-0-12-821259-2.00022-3](https://doi.org/10.1016/b978-0-12-821259-2.00022-3).
- [79] O. Şerban, N. Thapen, B. Maginnis, C. Hankin, and V. Foot. “Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification”. In: *Information Processing & Management* 56.3 (2019), pp. 1166–1184. DOI: <https://doi.org/10.1016/j.ipm.2018.04.011>.

Appendix A Event similarity

Throughout this work, we need to assess the similarity of two events. In this section, we explain how we perform this similarity calculation in the presence of hierarchical data. The main idea is that we consider two events to be similar, if 1) their event attributes, with $D_i \in \{D_Z, D_D, D_H\}$, are identical or hierarchically linked in H_i , and 2) their event dates are very close. In the similarity assessment, we first calculate the similarity for each event attribute, and then we sum up the obtained values in order to get the final score. Note that we do not take the dimension D_S into account in this calculation.

Before introducing this calculation, let us define some additional notations and definitions. First, for convenience, we denote the depth (or level) of a node v in the hierarchy H_i by $l_i(v)$. Likewise, the depth of the closest common ancestor of v_1 and v_2 is denoted by $\bar{l}_i(v_1, v_2)$. In H_i , a path connecting two nodes from the root (i.e. the most general level) to the leaves (i.e. the most specific level) (resp. from the leaves to the root) gives the specialization relation (resp. the generalization relation). Given an element $x \in \text{Dom}(D_i)$ and the associated hierarchy H_i , we denote by x^\uparrow (respectively x^\downarrow) the set containing x along with all generalizations (respectively specializations) of x with respect to H_i that belong to $\text{Dom}(D_i)$. Based on this, we also define the specificity relation, which allows comparing the attributes of two events being at different hierarchical levels. Namely, for a pair of event attributes d_i and d'_i , with $i \in \{Z, T, D, H\}$, and associated with $e = (d_Z, d_T, d_D, d_H, d_S)$ and $e' = (d'_Z, d'_T, d'_D, d'_H, d'_S)$, event attribute d_i is said to be more specific than d'_i , denoted by $d_i \preceq d'_i$, if $d'_i \in d_i^\uparrow$. For instance, in Table 2 *H7N9* is more specific than *avian influenza* in D_D , i.e. $H7N9 \preceq \text{avian influenza}$.

Then, let e and e' be two events, for which we calculate the similarity. For the dimension $D_i \in \{D_Z, D_D, D_H\}$, we use an ontology-based semantic similarity measure proposed in [60], but tailored for our purposes. In this measure, we consider that two literal values x and y of a dimension $D_i \in \{D_Z, D_D, D_H\}$, with $x, y \in \text{Dom}(D_i)$, are similar, if they are identical or hierarchically linked in H_i , i.e. $y \preceq x \vee x \preceq y$. Otherwise, we assign a large negative value for a penalization. Concretely, we calculate the similarity between x and y in Equation 6 as

$$\text{sim}_i(x, y) = \lambda_{x,y}^i \frac{2\bar{l}_i(x, y)}{l_i(x) + l_i(y)} - (1 - \lambda_{x,y}^i)\sigma_i, \quad (6)$$

where the variable $\lambda_{x,y}^i$ takes a binary value, where 1 means x and y are identical or hierarchically linked in H_i , and 0 otherwise. Finally, the variable σ_i is a large penalization factor, used when $\lambda = 0$. Note that the obtained score without the penalization factor is in the range $[0, 1]$. But, in the end, the score is not normalized from one side, i.e. $[-\sigma_i, 1]$.

Furthermore, we handle the temporal dimension D_T in a different manner. We calculate the similarity between two dates in Equation 7 as

$$\text{sim}_T(t_1, t_2) = 1 - \frac{|t_2 - t_1|}{L}, \quad (7)$$

where $|t_1 - t_2|$ represents temporal distance in days and L corresponds to a time delay which controls how the temporal distance between two dates are small enough. In this work, we set L to 21 (i.e. 3 weeks). Finally, we compute the final similarity score by summing the individual similarity scores.

$$\text{sim}(e, e') = \sum_{\substack{i \in \{Z, T, D, H\} \\ d_i \in e, d'_i \in e'}} \text{sim}_i(d_i, d'_i), \quad (8)$$

Appendix B Extraction of Event Database

In this section, we detail how we produce an event database \mathcal{E} from the events collected by an EBS/IBS system. First, we need to clarify our terminology, as we use the term *event* in two different contexts. We call the events being usually extracted at news document-level *document events*. Since information about events can be scattered over different news documents in the corpus, this implies that several copies of a specific event can co-exist in the end. The fusion of these copies can produce fully-fledged event description and it is simply called *corpus event*.

The input of our process is a set of document events, accompanied by their associated news documents. As mentioned in the main manuscript that the definition of an event can be different from one system to another, the minimalist one being a disease-location pair. For this reason, when the event definition of a system does not match the one proposed in Section 3.1, we process the associated news documents in order to complete the missing information (Section B.1). Then, we normalize the document events that can be used for evaluation and comparison purposes. In this context, the normalization

of event attributes is an important step, since it allows transforming a raw text into one of well-defined taxonomy classes (Section B.2). Finally, since the information about events can be scattered over different news documents, the last step consists in aggregating/fusing all the event information in order to produce fully-fledged event descriptions, i.e. cross-document information fusion (Section B.3). In the end of our process, we produce a set of normalized corpus events. To simplify our processing, we suppose that the input data are collected for a particular disease and that each news document contains at least one event.

B.1 Document Event Completion

In this section, when the event definition of an EBS system does not match the one proposed in Section 3.1, we process the associated news documents in order to complete the missing information, if possible, assuming that the position of each extracted event is known in its associated news document. In the following, we explain this process in two parts: 1) Thematic and source entity extraction (Section B.1.1) and 2) document event completion (Section B.1.2). Note that these tasks are different from the event extraction methods [70, 73], which are not the scope of this work.

B.1.1 Thematic and Source Entity Extraction

What we propose is a classical preprocessing for information extraction in EBS systems through the application of natural language processing (NLP) [51]. In the first step, we extract thematic and source-related information from the header, title and raw content of each news document. We first start with the header part. We retrieve the publication date and the news outlet publishing the news document at hand.

Then, we extract disease and host entity information from the title and the raw content of a news document. We consider a title is a special sentence of the document at hand, which summarizes the content with a few keywords. For this reason, we detail the extraction of event-related keywords together for the title and the raw content. First, the raw textual content is preprocessed by sentence splitting, tokenization and lemmatization using SpaCy [27]. We perform our entity extraction for each sentence, as well as the title. In each sentence, we extract by building the lexicon of host and disease names. We extract thematic entities with them by keyword search [25, 70].

B.1.2 Event Completion

In the event completion task, we can find multiple events in the same news document. Indeed, it is possible that a sentence can contain multiple events, or events can be located in the different part of the document. In the following, we follow the work of [75] for the event completion task. To simplify our processing, we assume that the position of each extracted event is known in its associated news document.

As in [75], we treat each sentence containing all essential event attributes as head sentence. To identify the head sentences we need to locate (host, time, host, disease) entities in the same sentence. The event completion starts from a head sentence and continues up to k next sentences. Although a head sentence needs to contain all essential information, its subsequent sentences can contain more detailed thematic information about an event. At the end of the process, for each news document, we have a set of document events defined as in Section 3.1.

B.2 Event Normalization

This task consists in normalizing the attributes of each event in an event database, assuming that the events are defined as in Section 3.1. As mentioned before, a normalization task aims at transforming a raw text into one of well-defined taxonomy classes, which results in hierarchical information.

We start with the normalization of spatial entities, also called *geocoding*. This task consists in assigning geographic coordinates to spatial entities. This can be a challenging task due to the ambiguity among place names [51]. For instance, Avignon is a regional county municipality in Quebec/Canada, but also a city in France. In this work, we perform the geocoding task with the gazetteer GeoNames¹¹. For a given query of spatial entity, GeoNames outputs a ranked list of most appropriate geographic coordinates associated with the input text. Although the first result is usually the desired/correct geographic coordinates, we still need to apply a disambiguation technic to resolve the place ambiguity issue. This disambiguation is easily solved, if we know the associated country. Indeed, EBS systems

¹¹www.geonames.org/

Entity type (Level 0)	Level 1	Level 2	Level 3
Host	human	male female	– –
	mammal	arctoidea	seal, etc.
		canidae	fox, dog, etc.
		equidae	horse, donkey, mule, etc.
	avian	camelidae	lama, alpaca, etc.
		wild	eagle, falcon, gull, hawk, black swan, crow, etc.
mosquito	domestic	chicken, turkey, fowl, ostrich, house crow, mute swan, etc.	
	culex	culex pipiens, culex tarsalis, etc.	
	aedes	aedes aegypti, aedes vexans	
Disease	Avian Infl.	low pathogenic	other serotypes
		high pathogenic	H5N1, H7N9, H5N6, H5N8
	West Nile	–	–

Table 18: Adjusted taxonomy classes associated with the disease and host dimensions for Avian Influenza and West Nile Virus.

often provide the country information of a spatial entity in an event. Otherwise, we employ the following heuristic strategy.

First, we find the corresponding news document, and we extract the country information of the spatial entities mentioned in the title, because we consider that a title is a special sentence of the document at hand, which summarizes the content with a few main keywords. If there is not any spatial entity in the title, we look for a nationality mention (e.g. French farm). In the worst case, we use the country information associated with the news outlet (e.g. France for Le Monde¹²), thanks to a publicly available dataset [58]. The rationale behind this processing is according to pragmatic rules when one reads a news document of a given country [34]. Often, a city name is explicitly mentioned without citing the associated country name. Nevertheless, if the news document is about another country, it is more likely to be mentioned so that the reader has fewer chances of misunderstanding. We extract spatial entities from the title using SpaCy [27] and a lookup table. This lookup table is used to locate the spatial entities which are not detected by SpaCy. It can be constructed with a gazetteer and extracts spatial entities with keyword search. We also extract nationalities using a second lookup table. Then, we finalize the geocoding task from the ranked list provided by GeoNames by selecting the first result whose associated country is the same as the one extracted from the title.

In terms of the temporal dimension D_T , the temporal expressions are usually in a well-structured format. The only task we need to do is to correctly estimate this format [65]. For instance, a date can be written in *MM-DD-YYYY* or *DD-MM-YYYY*. We do this estimation by checking the whole data. Finally, we normalize the temporal expressions according to the TIMEX3 annotation standard.

On what regards the thematic dimensions D_D and D_H , we normalize these disease and host names against the NCBI Taxonomy database [21], using a manually composed table of species name synonyms. Moreover, we tailor some of these NCBI taxonomies to diseases of interest. This is because the NCBI database is supposed to be generic enough in order to be used for any application domain. Nevertheless, for a particular area of expertise (e.g. Epidemic Intelligence), the taxonomy classes would be better constructed. For instance, categorizing the bird types as *wild* and *domestic*, or the Avian Influenza serotypes as *high* and *low pathogenic* based on their ability to cause mortality in birds, is a valuable information for epidemiological standpoint. We illustrate some of the adjusted taxonomy classes for thematic and disease entities in Table 18.

¹²www.lemonde.fr

B.3 Corpus Event Construction

The process of extracting corpus events requires identifying the news documents reporting the same document events. To do so, we apply an overlapping clustering method. The idea of our method is that when an event has less precise information, then it is more likely to belong to different clusters. In the beginning of our method, every document event constitutes its own cluster. We start the method iterations with the events with most detail spatial entity information, and continue with one granularity level lower in each iteration. In each iteration and for each event of this iteration, we find a set of similar events of the same or higher granularity level. We assess the similarity between events, as described in Section A. This score is the sum of four scores reflecting the similarities of geographical locations, infection dates, disease names and host species by taking their hierarchical levels into account. When a set of similar events are found for a given event, then we put them in the same cluster. In each cluster update, we perform an information fusion step based on a voting approach. Namely, for each event attribute except date, we take the most frequent value. For the temporal attribute, we take the oldest infection date.

Appendix C Quantitative Evaluation of EBS Systems

In this section, we explain how we quantitatively evaluate the performance of an event database \mathcal{E} of an EBS system with respect to a gold standard database \mathcal{E}_R . In this evaluation, the results usually take the form of a ranked list. For this reason, we first need to define a ranking evaluation measure to evaluate the quality of the ranked lists with respect to those produced by a gold standard database (Section C.1). Then, we pass to the quantitative evaluations for the spatial (Section C.2), temporal (Section C.3), thematic (Section C.4) and source (Section C.5) dimensions.

Before presenting the quantitative evaluations, we need to define the selection operator σ for the dimensions D_Z and D_T as $\sigma_{\text{Dom}(D_T) \in T}^{\text{Dom}(D_Z) \in Z}(\mathcal{E})$ ($\sigma_T^Z(\mathcal{E})$ for short), for two sets Z and T of literal values defined for D_Z and D_T . This operator is used to select a subset of the input data \mathcal{E} . For the sake of simplicity, when the selection operator is not applied on D_Z or D_T , we simply omit it from the notation σ . For instance, $\sigma^{\{France\}}(\mathcal{E}_{week}^{country})$ represents a set of weekly events at country level occurring only in France. Finally, the number of remaining events after the selection is calculated by taking the cardinality, i.e. $|\sigma^{\{France\}}(\mathcal{E}_{week}^{country})|$.

C.1 Evaluation with Ranking

Throughout this framework, we compare the results produced by the event database \mathcal{E} of an EBS system with those from a gold standard database \mathcal{E}_R , assuming that \mathcal{E}_R is available to use. As we will see in the following sections, these results usually take the form of a ranked list. For instance, such a list might represent the periodic patterns ordered by their frequency in an event database \mathcal{E} . Furthermore, since \mathcal{E} has a limited data compared to a gold standard database \mathcal{E}_R , these two ranked lists are usually of different size, and might contain disjoint values. For this reason, we need a specific ranking evaluation measure to evaluate the quality of the ranked list produced by an EBS system, called *candidate list* and denoted by $L_{\mathcal{E}}$, with respect to that produced by a gold standard database, called *reference list* and denoted by $L_{\mathcal{E}_R}$. In this work, we use the normalized F-measure $F(L_{\mathcal{E}}, L_{\mathcal{E}_R})$ [32, 69, 70], which is the combination of normalized precision $P(L_{\mathcal{E}}, L_{\mathcal{E}_R})$ and recall $R(L_{\mathcal{E}}, L_{\mathcal{E}_R})$. These last two are based on the difference between the sum of ranks of the candidate and reference lists.

The calculation of the normalized precision and recall between $L_{\mathcal{E}}$ and $L_{\mathcal{E}_R}$, of sizes $N_{L_{\mathcal{E}}}$ and $N_{L_{\mathcal{E}_R}}$, requires extracting a list of common values found in both lists, that we call *relevant list* of size N_{Λ} . Concretely, we calculate $F(L_{\mathcal{E}}, L_{\mathcal{E}_R})$ between this relevant list and the initial reference list. Finally, we define a bijection $r : \{1, 2, \dots, N_{\Lambda}\} \rightarrow \{1, 2, \dots, N_{L_{\mathcal{E}}}\}$, such that it gives the rank of the i^{th} element of the relevant list in the reference list. Based on this, the *normalized recall* is defined as in Equation 9.

$$R(L_{\mathcal{E}}, L_{\mathcal{E}_R}) = 1 - \frac{\sum_{i=1}^{N_{\Lambda}} r(i) - \sum_{i=1}^{N_{\Lambda}} i}{N_{\Lambda} N_{L_{\mathcal{E}}} + N_{\Lambda}^2}. \quad (9)$$

When we transform the index i representing each position of elements into $\log(i)$, it is called *normalized precision*, which is calculated as in Equation 10.

$$P(L_{\mathcal{E}}, L_{\mathcal{E}_R}) = 1 - \frac{\sum_{i=1}^{N_{\Lambda}} \log(r(i)) - \sum_{i=1}^{N_{\Lambda}} \log(i)}{\log(C(N_{L_{\mathcal{E}}}, N_{\Lambda}))}, \quad (10)$$

where $C(N_{L_{\mathcal{E}}}, N_{\Lambda}) = \frac{N_{L_{\mathcal{E}}}!}{N_{\Lambda}!(N_{L_{\mathcal{E}}} - N_{\Lambda})!}$. Finally, $F(L_{\mathcal{E}}, L_{\mathcal{E}_R})$ is the harmonic mean of $P(L_{\mathcal{E}}, L_{\mathcal{E}_R})$ and $R(L_{\mathcal{E}}, L_{\mathcal{E}_R})$, which is calculated as

$$F(L_{\mathcal{E}}, L_{\mathcal{E}_R}) = \frac{2R(L_{\mathcal{E}}, L_{\mathcal{E}_R})P(L_{\mathcal{E}}, L_{\mathcal{E}_R})}{R(L_{\mathcal{E}}, L_{\mathcal{E}_R}) + P(L_{\mathcal{E}}, L_{\mathcal{E}_R})}. \quad (11)$$

C.2 Quantitative Evaluation for Spatial Dimension

In this section, we describe how we compute the spatio-temporal representativeness score of an event database \mathcal{E} with respect to a gold standard database $\mathcal{E}^{\mathcal{R}}$. The definition of the spatio-temporal representativeness requires fixing the spatial and temporal scales, denoted by l_Z and l_T , respectively. Depending on these scales, we first discretize the spatial and temporal dimensions over a set Z of geographic zones and a set T of time intervals. Then, for a particular disease, we say that an event database \mathcal{E} of an EBS system represents well a specific geographic zone z for a given time interval t , if it finds at least one event in $\mathcal{E}^{\mathcal{R}}$, which is defined in Equation 12 as

$$\mathbb{1}_{l_T, t}^{l_Z, z}(\mathcal{E}) = \begin{cases} 1, & |\sigma_{\{t\}}^{\{z\}}(\mathcal{E}_{l_T}^{l_Z})| > 0 \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

The calculation of the spatio-temporal representativeness scores is done with respect to a gold standard database \mathcal{E}_R . Let Z be the set of spatial entities found in \mathcal{E}_R with respect to the spatial scale l_Z , i.e. $Z = \text{Dom}_{\mathcal{E}_R}(D_Z, l_Z)$. Also, let T a set of time intervals with respect to the temporal scale l_T , i.e. $T = \text{Dom}_{\mathcal{E}_R}(D_T, l_T)$. For a given geographic zone $z \in Z$, its spatio-temporal representativeness score for \mathcal{E} with respect to \mathcal{E}_R is calculated as

$$\Phi_{l_T}^{l_Z, z}(\mathcal{E}, \mathcal{E}_R) = 1 - \frac{1}{|T|} \sum_{t \in T} \max \left(0, \mathbb{1}_{l_T, t}^{l_Z, z}(\mathcal{E}_R) - \max(\mathbb{1}_{l_T, t-1}^{l_Z, z}(\mathcal{E}), \mathbb{1}_{l_T, t}^{l_Z, z}(\mathcal{E}), \mathbb{1}_{l_T, t+1}^{l_Z, z}(\mathcal{E})) \right). \quad (13)$$

In Equation 13, the term after the subtraction operator calculates an error score, which is similar to the mean absolute error. The difference is that when an EBS system detects an event, which is not detected in \mathcal{E}_R , is not considered as an error. Furthermore, since there can be some reporting delay between the events of \mathcal{E} and \mathcal{E}_R , we also consider in Equation 13 the previous (resp. next) time interval $t-1$ (resp. $t+1$) in order not to penalize \mathcal{E} . The obtained score is in the range $[0, 1]$, where the score of 0 (resp. 1) indicates that an EBS system never (resp. always) finds an event in $\mathcal{E}^{\mathcal{R}}$ for a given geographic zone. Finally, we obtain a single spatio-temporal representativeness score for \mathcal{E} by taking the average of the obtained scores over all geographic zones Z , as in Equation 14.

$$\Phi_{l_T}^{l_Z}(\mathcal{E}, \mathcal{E}_R) = \frac{1}{|Z|} \sum_{z \in Z} \Phi_{l_T}^{l_Z, z}(\mathcal{E}, \mathcal{E}_R). \quad (14)$$

This score is tied to particular l_Z and l_T scales. If we consider several different scale values, then we obtain an average value as in Equation 15, i.e.

$$\Phi_{L_Z, L_T}^{L_Z}(\mathcal{E}, \mathcal{E}_R) = \frac{1}{|L_Z||L_T|} \sum_{l_Z \in L_Z} \sum_{l_T \in L_T} \Phi_{l_T}^{l_Z}(\mathcal{E}, \mathcal{E}_R), \quad (15)$$

where L_Z and L_T are two sets of values for spatial and temporal scales, respectively.

C.3 Quantitative Evaluation for Temporal Dimension

In this section, we describe how we compute the timeliness (Section C.3.1) and periodicity (Section C.3.2) scores of an event database \mathcal{E} with respect to a gold standard database $\mathcal{E}^{\mathcal{R}}$.

C.3.1 Quantitative Evaluation for Timeliness

In this section, we describe how to compute the timeliness score between an event database \mathcal{E} and a gold standard database $\mathcal{E}^{\mathcal{R}}$. First, we calculate in Equation 16 the temporal distance between an event $e \in \mathcal{E}_R$ and its corresponding one $e' \in \mathcal{E}$ as

$$\delta(e, e') = \begin{cases} t(e') - t(e), & t(e') \geq t(e) \\ 0, & t(e') < t(e), \end{cases} \quad (16)$$

where the function $t(e)$ gives the date of a given event e . Note that in Equation 16 when an EBS system finds an event before the corresponding report date in \mathcal{E}_R , we do not reward or penalize it, and put a score of 0. Then, we calculate the timeliness score for \mathcal{E} with respect to \mathcal{E}_R as

$$\Psi(\mathcal{E}, \mathcal{E}_R) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} 1 - \exp^{-\frac{\delta(f(e, \mathcal{E}_R), e)}{L}}, \quad (17)$$

where the variable L in Equation 17 indicates the maximum expected reporting delay, which controls the temporal similarity of two events. In this work, we set L to 21 days.

C.3.2 Quantitative Evaluation for Periodicity

In this section, we explain how we quantitatively evaluate the performance of an event database \mathcal{E} of an EBS system in terms of its ability to detect the continuous and seasonal periodic patterns with respect to a gold standard database \mathcal{E}^R . In the following, we first evaluate an event database \mathcal{E} for each pattern type separately, then we take an average of the obtained scores.

We first start with the evaluation of \mathcal{E} in terms of its ability to detect *continuous* full and partial periodic patterns with respect to a gold standard database \mathcal{E}^R . Let l_Z and l_T be spatial and temporal scales for \mathcal{E} and \mathcal{E}^R . Moreover, let I and P be also two sets of parameter values for ι and ρ in ST. On top of that, the parameter value α controls the spatial closeness between the spatial entities in ST. Then, we propose to calculate the evaluation score $\tilde{F}_{l_T}^{l_Z}(\mathcal{E}, \mathcal{E}_R, I, P, \alpha)$, which corresponds to the average score of F , as defined in Section C.1, over the combination of all the considered parameters. It is calculated in Equation 18 as

$$\tilde{F}_{l_T}^{l_Z}(\mathcal{E}, \mathcal{E}_R, I, P, \alpha) = \frac{1}{|I||P|} \sum_{\iota \in I} \sum_{\rho \in P} F(\text{ST}(\mathcal{E}_{l_T}^{l_Z}, \iota, \rho, \alpha), \text{ST}(\mathcal{E}_R^{l_Z}, \iota, \rho, \alpha)). \quad (18)$$

As we see, Equation 18 is calculated for particular spatial and temporal scales. We want to make our evaluation more robust by considering several scale values for dimensions D_Z and D_T . To consider this aspect, we define two sets L_Z and L_T of values for spatial and temporal scales, respectively. Then, we compute the final average score for continuous full and partial periodic patterns in Equation 19 as

$$\tilde{F}_{L_T}^{L_Z}(\mathcal{E}, \mathcal{E}_R, I, P, \alpha) = \frac{1}{|L_Z||L_T|} \sum_{l_Z \in L_Z} \sum_{l_T \in L_T} \tilde{F}_{l_T}^{l_Z}(\mathcal{E}, \mathcal{E}_R, I, P, \alpha). \quad (19)$$

Second, we propose to perform the evaluation of \mathcal{E} in terms of its ability to detect *seasonal* full and partial periodic patterns with respect to a gold standard database \mathcal{E}^R . We propose to do it, similar to the spatio-temporal representativeness evaluation defined in Equation 14. To detect such a seasonal effect within the data, we need to apply the method ST to a subset of \mathcal{E} , where the events occur only within a particular time interval over several years, and we repeat this process for all the subsets. For instance, a subset of data can represent all the events occurring in every January, whatever the year is, and we repeat similar selections for the other months. To do so, in Equation 20 we first need to define if a spatial pattern x is in the output of $\text{ST}(\sigma_t(\mathcal{E}_{l_T}^{l_Z}), \iota, \rho, \alpha)$ for a given time interval t and the input parameters ι, ρ, α , i.e.

$$\mathbb{1}_{l_T, t}^{l_Z}(\mathcal{E}, x, \iota, \rho, \alpha) = \begin{cases} 1, & x \in \text{ST}(\sigma_t(\mathcal{E}_{l_T}^{l_Z}), \iota, \rho, \alpha) \\ 0, & \text{otherwise.} \end{cases} \quad (20)$$

Then, let X be the set of all spatial patterns found in the reference ranking result of $\text{ST}(\sigma_t(\mathcal{E}_R^{l_Z}, \iota, \rho, \alpha))$ with respect to l_Z, l_T, ι, ρ and α . Based on this, we compute in Equation 21 to what degree the set X is discovered in \mathcal{E} through ST for a particular time interval t .

$$\tilde{\Phi}_{l_T, t}^{l_Z}(\mathcal{E}, \mathcal{E}_R, \iota, \rho, \alpha) = \frac{1}{|X|} \sum_{x \in X} \max \left(\mathbb{1}_{l_T, t-1}^{l_Z}(\mathcal{E}, x, \iota, \rho, \alpha), \mathbb{1}_{l_T, t}^{l_Z}(\mathcal{E}, x, \iota, \rho, \alpha), \mathbb{1}_{l_T, t+1}^{l_Z}(\mathcal{E}, x, \iota, \rho, \alpha) \right), \quad (21)$$

Then, let T be all the literal values of a given temporal scale l_T in \mathcal{E}_R , i.e. $T = \text{Dom}_{\mathcal{E}_R}(D_T, l_T)$. We apply Equation 22 to obtain an average score with respect to T, I and P for seasonal full and partial periodic patterns as

$$\tilde{\Phi}_{l_T}^{l_Z}(\mathcal{E}, \mathcal{E}_R, I, P, \alpha) = \frac{1}{|T||I||P|} \sum_{t \in T} \sum_{\iota \in I} \sum_{\varrho \in P} \tilde{\Phi}_{l_T, t}^{l_Z}(\mathcal{E}, \mathcal{E}_R, \iota, \varrho, \alpha). \quad (22)$$

Similar to Equation 19, we compute the final average score for seasonal full and partial periodic patterns in Equation 23 by considering several values for spatial and temporal scales to make our evaluation more robust.

$$\tilde{\Phi}_{L_T}^{L_Z}(\mathcal{E}, \mathcal{E}_R, I, P, \alpha) = \frac{1}{|L_Z||L_T|} \tilde{\Phi}_{l_T}^{l_Z}(\mathcal{E}, \mathcal{E}_R, I, P, \alpha). \quad (23)$$

Finally, we calculate in Equation 24 the final periodicity score $\tilde{\Gamma}(\mathcal{E}, \mathcal{E}_R)$ for both continuous and seasonal periodic patterns as

$$\tilde{\Gamma}(\mathcal{E}, \mathcal{E}_R) = \frac{\tilde{F}_{L_T}^{L_Z}(\mathcal{E}, \mathcal{E}_R, I, P, \alpha) + \tilde{\Phi}_{l_T}^{l_Z}(\mathcal{E}, \mathcal{E}_R, I, P, \alpha)}{2}. \quad (24)$$

C.4 Quantitative evaluation for thematic dimension

In this section, we explain how we evaluate the modified version \mathcal{E}^+ of an event database \mathcal{E} in terms of its ability to detect *static* and *temporal* multidimensional patterns with respect to the modified version \mathcal{E}_R^+ of a gold standard database.

Let I and P be the parameter values for ST. Note that the set I of values allows handling both the static and temporal cases, as a very large inter-arrival time value for ι amounts to omit the partial periodicity condition in ST. Then, we propose to calculate the evaluation score $\Omega(\mathcal{E}^+, \mathcal{E}_R^+, I, P, \alpha)$, which corresponds to the average score of F , as defined in Section C.1, over the combination of all the considered parameters. It is calculated in Equation 25 as

$$\Omega(\mathcal{E}^+, \mathcal{E}_R^+, I, P) = \frac{1}{|I||P|} \sum_{\iota \in I} \sum_{\varrho \in P} F(\text{ST}(\mathcal{E}^+, \iota, \varrho), \text{ST}(\mathcal{E}_R^+, \iota, \varrho)). \quad (25)$$

C.5 Quantitative evaluation for source dimension

Despite the existence of several prominent works aiming at ranking a large number of news outlets in terms of various objectives (e.g. news quality, popularity [76]), to the best of our knowledge, none of them is specialized in Epidemic Intelligence. This means that, unlike the previous studied dimensions, we are not able to compare the obtained results with a gold standard database.

For this reason, we propose to evaluate an EBS system for the source dimension from a different perspective. Instead of evaluating the quality of the results, we rather assess the consistency of the results obtained in Sections 3.6.1 and 3.6.2 one another. The rationale is that for an effective surveillance system, we expect news outlets playing a key role in epidemiological information dissemination (as detected in Section 3.6.1) to be those which are timely in event detection (as in Section 3.6.2). Otherwise, this inconsistency amounts to decrease the main strength of an EBS system, which is its timeliness.

Concretely, given an event database \mathcal{E} we consider the ranked list produced for a timely detection objective (Section 3.6.2) as a reference list, and compare it with that obtained in Section 3.6.1 by applying Equation 11, as shown in Equation 26.

$$\hat{F}(\mathcal{E}) = F(\text{PageRank}(G_{\mathcal{E}}, k), \text{CELF}(G_{\mathcal{E}}, k)) \quad (26)$$

Appendix D Additional Results

D.1 Additional Results for Spatio-temporal Representativeness

In this section, we complete our discussion in Section 5.1 with the spatio-temporal representativeness scores obtained at region scale with monthly time intervals (Equation 13). We plot the scores at region scale in Figures 9 and 10. In these plots, regions without an Empres-i event are indicated in gray, and the degree to which an EBS system covers the events occurring in a region is shown with different blue scales, where the largest values are in dark blue and the smallest ones in light blue. When an EBS system never finds an event in \mathcal{E}_{EI} , it is shown in white.

Moreover, we also plot the score differences in Figure 11 to ease the comparison between PADI-web and ProMED. In this figure, it is colored in blue (resp. red) when ProMED (resp. PADI-web) gives better spatio-temporal representativeness score for a country and in yellow in case of non-zero equality.

D.2 Additional Results for Thematic Dimension

To ease our discussion in Section 5.3, in this section, we visualize in Figure 12 the relation between spatial and thematic entities with a *chord diagram*. The advantage of this diagram is that we combine two aspects: hierarchy and temporal information. In each diagram, the connections between spatial and thematic entities are shown in the center of the circle. Each entity is represented by a fragment on the outer part of the circular layout. These entities are organized in a hierarchical way. In the most outer part of the circular layout, we show through a histogram how a given entity is involved in the events over time. In the histogram, the x -axis represents the weekly time intervals and the y -axis indicates the number of occurrences of a given entity in the events. Note that the same entities are colored in the same way across all the plots in order to ease the comparison.

We can compare the considered EBS systems in three aspects in these plots. The first aspect concerns the height of the segments, since it indicates the amount of information provided by a given EBS system. By looking at these segments, we can easily observe the frequent entities for each EBS system. For instance, the most frequent country for ProMED is India, whereas India is only the third most frequent country for PADI-web, after South Korea and Japan. Nevertheless, both EBS systems do not detect most of the events occurring in Taiwan. The second aspect concerns the links between entities. For instance, in ProMED we have more information regarding the details of Avian Influenza, compared to PADI-web. Nevertheless, both EBS systems do not detect most of the LPAI events. Finally, the third aspect is related to the temporal information, illustrated in the most outer layer of the circle. For instance, we see from Subfigure 12a that the Empres-i LPAI events occur throughout the year, whereas the HPAI ones occur only in the beginning and the end of the year.

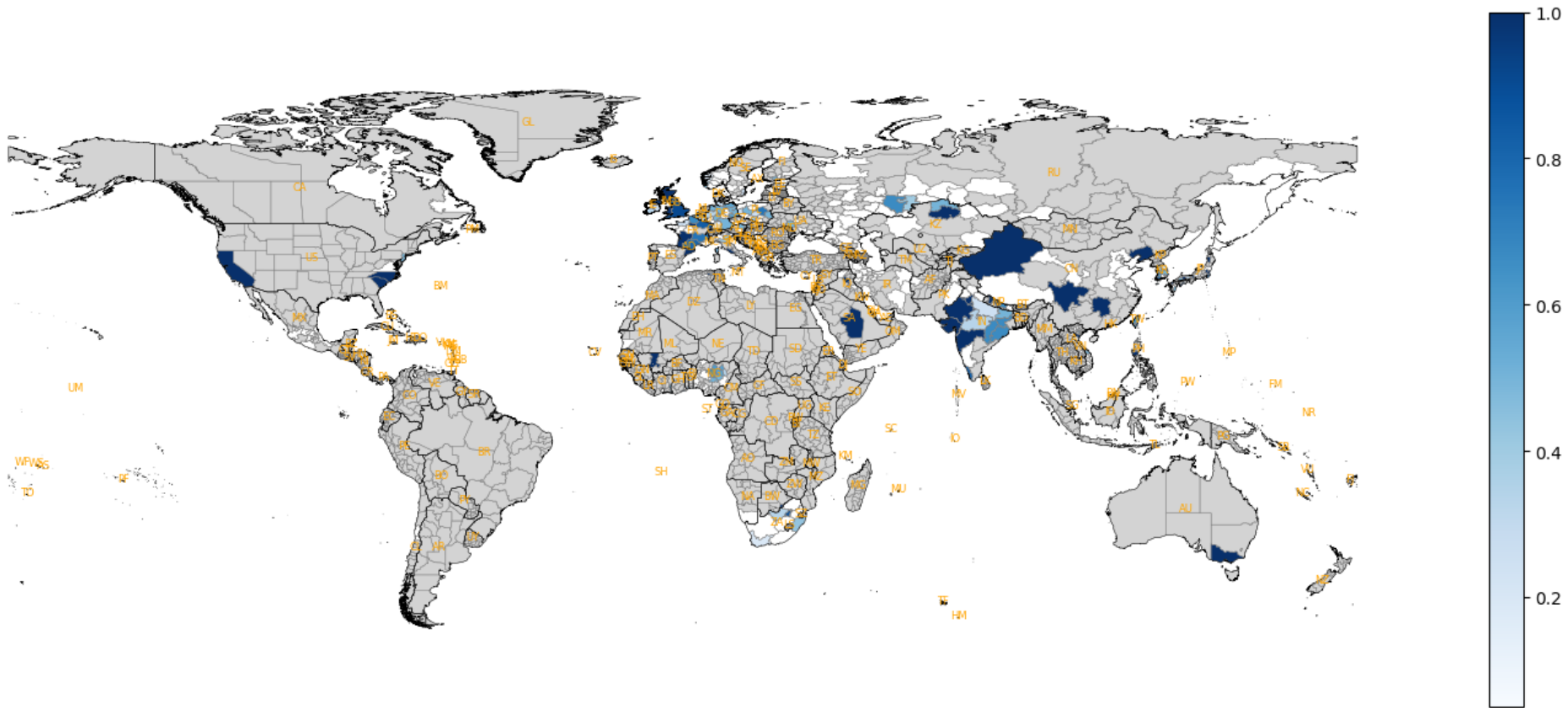


Figure 9: Spatio-temporal Representativeness scores at region scale for PADI-web with respect to the results of Empres-i. The degree to which an EBS system covers the Empres-i events occurring in a country is shown with different blue scales, and it is shown in white when an EBS system never finds an event in \mathcal{E}_{EI} . Moreover, countries without an Empres-i event are indicated in gray.

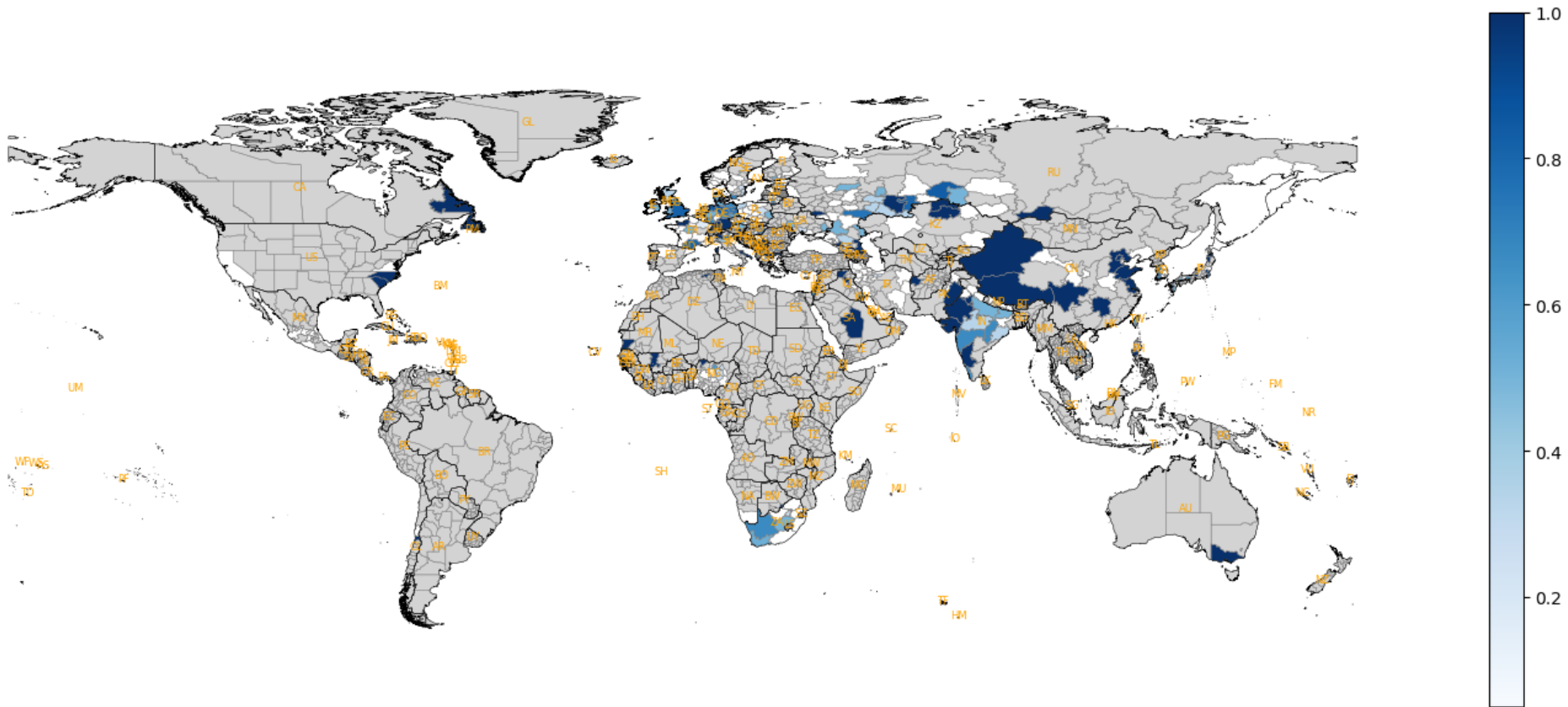


Figure 10: Spatio-temporal Representativeness scores at country scale for ProMED with respect to the results of Empres-i. The degree to which an EBS system covers the Empres-i events occurring in a country is shown with different blue scales, and it is shown in white when an EBS system never finds an event in \mathcal{E}_{EI} . Moreover, countries without an Empres-i event are indicated in gray.

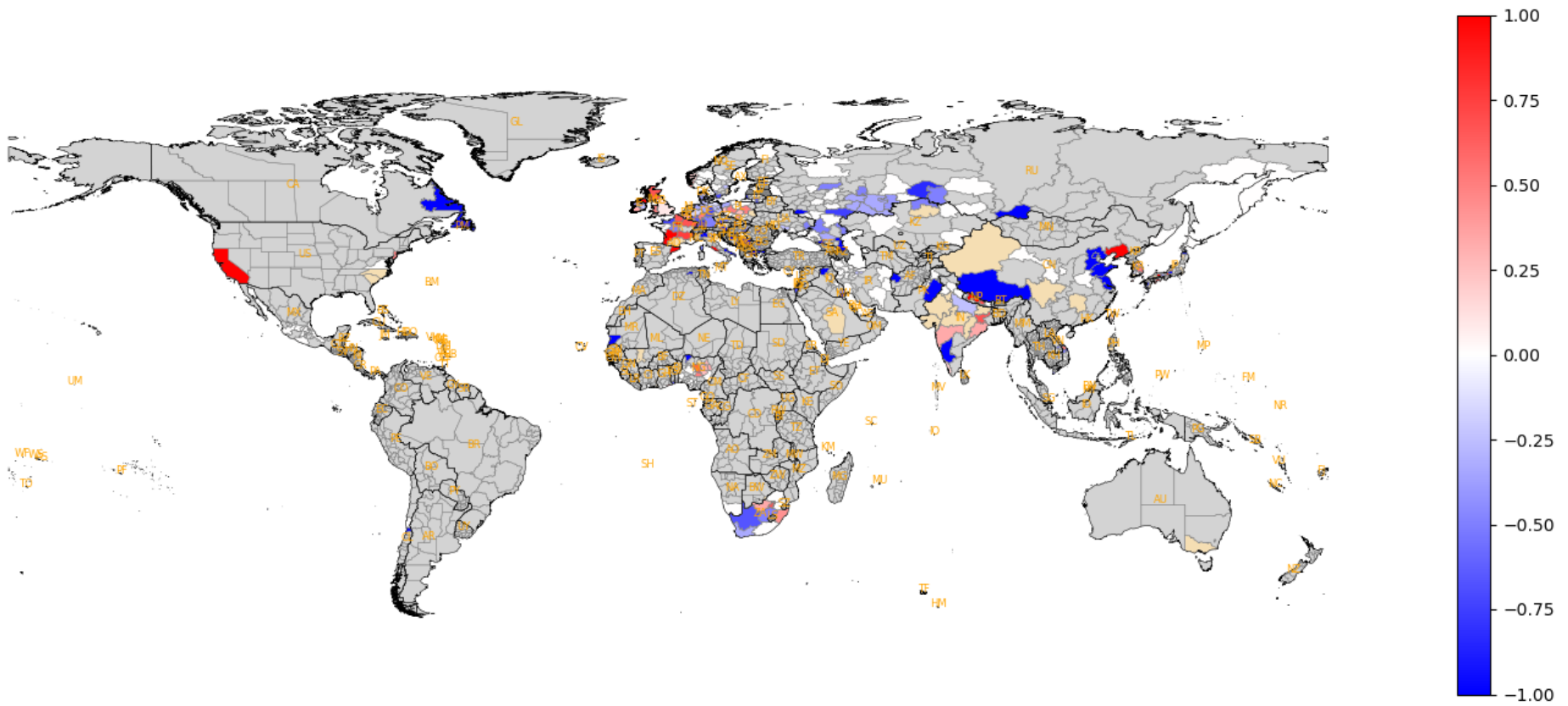


Figure 11: Spatio-temporal Representativeness score differences between PADI-web (Figure 9) and ProMED (Figure 10) at region scale. The regions are colored in blue (resp. red) when ProMED (resp. PADI-web) gives better spatio-temporal representativeness scores for a region and in yellow in case of non-zero equality. Moreover, countries without an Empres-i event are indicated in gray.

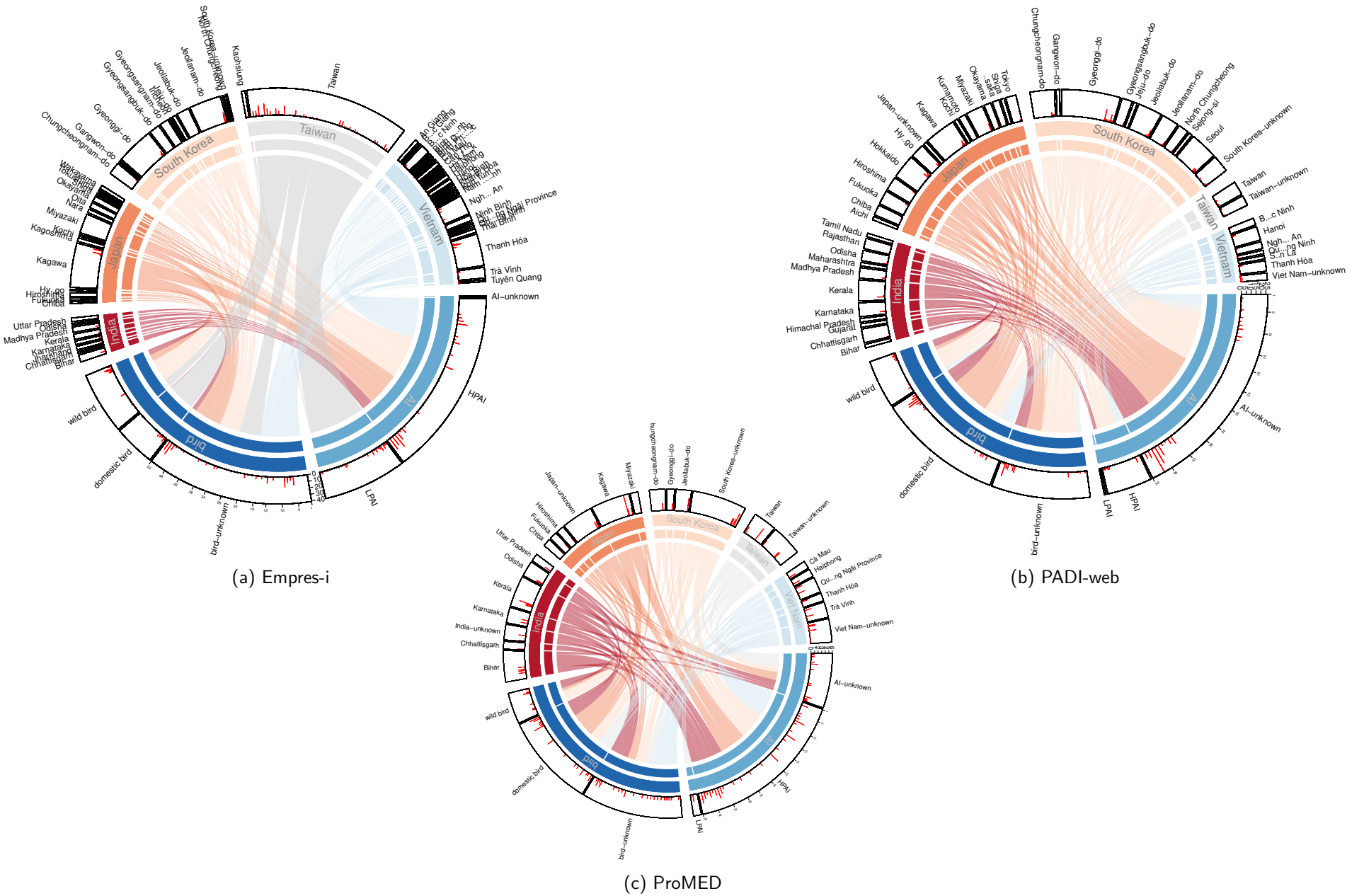


Figure 12: Visualization of the relations between spatial and thematic entities with a chord diagram for Empres-i (a), PADI-web (b) and ProMED (c). In these diagrams, we focus only on the Avian Influenza events occurring in Asian countries during 2020.