Added-value of automatic multilingual text analysis for epidemic surveillance

Gaël Lejeune, Romain Brixtel, Charlotte Lecluze,
Antoine Doucet, Nadine Lucas
firstname.lastname@unicaen.fr

Normandy University – UNICAEN, GREYC CNRS UMR-6072
Boulevard du Maréchal Juin, CS 14032 Caen Cedex France

Abstract. The early detection of disease outbursts is an important objective of epidemic surveillance. The web news are one of the information bases for detecting epidemic events as soon as possible, but to analyze tens of thousands articles published daily is costly. Recently, automatic systems have been devoted to epidemiological surveillance. The main issue for these systems is to process more languages at a limited cost. However, existing systems mainly process major languages (English, French, Russian, Spanish...). Thus, when the first news reporting a disease is in a minor language, the timeliness of event detection is worsened. In this paper, we test an automatic style-based method, designed to fill the gaps of existing automatic systems. It is parsimonious in resources and specially designed for multilingual issues. The events detected by the human-moderated ProMED mail between November 2011 and January 2012 are used as a reference dataset and compared to events detected in 17 languages by the system DAnIEL2 from web articles of this time-window. We show how being able to process press articles in languages less-spoken allows quicker detection of epidemic events in some regions of the world.

1 Introduction

The early detection of disease outbursts is critical for epidemic surveillance. One of the main sources of information is the press articles written all over the world, since diseases erupt anywhere. With the increasing amount of newspapers accessible on the Internet, tens of thousands of articles are available online daily. It has become one of the main lead to improve the early detection of epidemic events using computer driven information filtering and extraction.

Many projects use press articles for extracting epidemic events. ProMED [1] or GPHIN [2] rely on human intervention to extract epidemic events from press articles. Other systems are fully automated like BioCaster [3], EpiSpider [4], PULS [5] or DAnIEL [6]. Another approach is to propose an aggregation of events already collected by other systems, it is the choice of the researchers working on HealthMap [7].

One of the limitation encountered in classical natural language processing is the number of languages covered by any single system [8]. Table 1 shows the

different languages[1] processed by the previously cited systems and an estimation of the number of speakers for each language [9]. English (en) and Russian (ru) are handled by all the systems previously mentioned. Arabic (ar), Chinese (zh), French (fr), Portuguese (pt) and Spanish (es) are also well represented with 4 to 5 systems able to process them.The Japanese BioCaster system covers three Asian languages in addition: Korean (ko), Thai (th) and Vietnamese (vi). The DAnIEL system processes five European languages, including two, Polish (pl) and Greek (el), not available in other systems.

Table 1. Languages processed by existing epidemic surveillance systems and an estimation of their number of speakers (10^6)

	ar	cz	de	el	en	es	fi	fr	it	ko	nl	no	pl	pt	ru	sv	th	tr	vi	zh
#Speakers	255	10	166	13	1,000	500	5	200	62	78	21	5	46	240	277	8	60	75	86	1,151
GPHIN	✓				✓	✓		✓							✓					✓
HealthMap	✓				✓	✓		✓						✓	✓					✓
PULS					✓										✓					
Biocaster	✓				✓	✓		✓		✓				✓	✓		✓		✓	
DAnIEL				✓	✓			✓					✓		✓					✓
DAnIEL2	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓		✓		✓

Twenty languages are covered by at least one system, having a total of four billions of speakers. However, this number is over-estimated since people speaking two languages are counted twice. That means that the dropped-out languages represent more than 40% of the world's population. It is true that many events are eventually reported in English or another major language appearing in Table 1. However, two problems arise. First, it is extremely difficult to judge silence: some diseases can be ignored. Second, even in the case of delayed report, the problem is the time elapsed for response to be effective [10]. While medical reports use major languages for diffusion, press articles may report an epidemic event in a local language, but this valuable information is often by-passed.

Work has already been carried out on the impact of covering multiple languages on the informations extracted in different domains. Piskorski *et al.* [11] showed a significant improvement of the quality of the information extracted when more languages were processed. Lyon *et al.* [12] studied specifically this impact for epidemic surveillance by comparing BioCaster, EpiSPIDER and HealthMap. But, the number of languages involved (five) was quite small with respect to the number of languages for which press articles are published online (more than 20 languages on Google News for instance).

An important issue is to check to which extent the number of processed languages will give an added-value to the early detection of diseases and thus epidemic events. To this purpose, this study proposes a comparison between human-produced ProMED-mail, taken as a reference, and an extended implementation called DAnIEL2, based on the multilingual system DAnIEL [6,13]. ProMED approach is efficient, with many experts employed, but it is costly

[1] ISO 639-1 codes : `http://www.loc.gov/standards/iso639-2/php/code_list.php`

and slow. ProMED is an expert-based relevance gold standard. To the contrary, DAnIEL (Data Analysis for Information Extraction in any Language) claims to minimize the marginal cost for the analysis of new languages but handles only six languages, French, English, Russian, Greek and Polish (fr, en, ru, el, pl), plus Chinese (zh).It was used to process up to a total of 17 languages, allowing comparison of both geographic coverage and timeliness of event detection with ProMED.

The DAnIEL2 system processes 11 more languages than DAnIEL and 9 European languages not available yet in other systems. This study will focus on the impact of extended multilingual analysis for timeliness. Epidemic events detected through both ProMED and DAnIEL2 will be compared. The impact of multilingual automatic coverage will be measured on the timeliness of event detection and the coverage of different regions of the world.

ProMED and DAnIEL2 are presented in Section 2. The datasets for each approach are described in Section 3. The results obtained for both approaches are presented in Section 4. The conclusions and perspectives of this study are detailed in Section 5.

2 Multilingual surveillance with ProMED and DAnIEL2

This section presents the characteristics of the two compared information systems: Section 2.1 for ProMED and Section 2.2 for DAnIEL2.

2.1 The ProMED system

ProMED-mail publishes daily reports disseminating information on disease outbreaks worldwide [1,14]. ProMED moderators, with the help of ProMED subscribers, screen different sources of information to produce their reports. The main sources exploited are local media reports, official reports and information from local observers. The number of different languages used on the field is not known. ProMED published reports in English since the beginning of the project in 1994. Reports are now also available in French, Portuguese, Russian and Spanish. ProMED relies on the accuracy of its human experts analysis to produce highly reliable reports [15]. It is therefore used as a source in automatic information processing [16]. The question raised is whether the complexity of the reporting chain has a bad impact on the timeliness of public reporting.

2.2 The DAnIEL2 system

DAnIEL2 takes advantage of genre-based analysis, requiring little expert knowledge in medicine and making it easier to use for various languages [6]. A quick presentation of the approach used by DAnIEL is made here, more details can be found in a previous article [13]. DAnIEL uses the collective style of journalists in order to build one multilingual core analysis, relying on expert knowledge on news discourse rather than on specific languages syntax. Decisions are taken at

text-level. DAnIEL relies on a character-level analysis, so it can handle languages where graphical words do not exist (for instance Chinese).

DAnIEL2 compares the text and the disease names found in its database extracted from Wikipedia. Repetitions of substrings of disease names occurring at key positions are used to check if the document is relevant for epidemic surveillance and which disease is involved. The same algorithm is used for detecting the location. If no location is found *implicit location* heuristic is used: the location of the reported event is the same as the location of the source. DAnIEL2 extracts disease-location pairs in 17 languages, as given in Table 2. They include 7 European languages not yet reported to our knowledge, for automatic news filtering: Czech (cz), German (de), Finnish (fi), Norwegian (no) and Swedish (sv). The datasets used for this evaluation are described in the next section.

3 Datasets for ProMED and DAnIEL2

The ProMED reference dataset has been built with reports presented in Section 3.1. For DAnIEL2 a corpus of press articles was constituted using available data processed by DAnIEL and extra material presented in Section 3.2.

3.1 ProMED dataset

The reports produced by ProMED-mail from October 2011 to February 2012 have been automatically harvested on the ProMED website[2]. The data hereby obtained includes 2,558 structured reports in 5 main languages (English, Russian, Portuguese, Spanish and French). In this period, a few reports are also available in Thai and Vietnamese but for these particular languages, no reports were available after the 4th of November.

Each report from ProMED contains a triplet describing the event: the disease name, the location of the event and the date of the report. For each unique disease-location pair, the earliest report date in the period was kept to get *first reports*. The events appearing only in October were excluded from the dataset since the objective was to measure the delay between first reports of DAnIEL2 and ProMED, for the same disease-location pair. It was therefore necessary to extend the time-window to feed DAnIEL2 with press articles that give hints for the November 2011 events. In the same way, events reported by ProMED in February 2012 were kept, to check if they were connected to events reported by DAnIEL2 in January.

Details about the data collected are presented in Table 2. Reports in English represent more than 50% of the total number of reports published by ProMED. Most of these reports come from the analysis of a newswire in English. The importance of English in this corpus is due to the fact that is used as a *lingua franca* for many reports and news. Table 3 shows that English sources allow ProMED to cover a high number of different locations (153) and a great number

[2] `http://www.promedmail.org/`

of disease-location pairs. For English three locations (USA, Australia and United Kingdom) are involved in 40% of the reports.

Table 2. ProMED reports repartition by language and by month

	English	French	Portuguese	Russian	Spanish	Thai	Vietnamese
#Reports	**819**	148	129	127	220	25	78
#November 2011	285	3	26	49	68	25	78
#December 2011	291	33	15	28	78	0	0
#January 2012	193	62	48	37	37	0	0
#February 2012	54	50	40	33	38	0	0

Table 3. ProMED reports details: number of diseases, locations and disease-location pairs per language

	English	French	Portuguese	Russian	Spanish	Thai	Vietnamese
#Reports	819	148	129	127	220	25	78
#Diseases	**183**	33	34	47	58	10	31
#Locations	**151**	37	23	15	46	8	26
#Disease-location pairs	**366**	63	40	55	46	12	26

3.2 Corpus for DAnIEL2

The corpus used by DAnIEL2 in this study was constituted by downloadable data processed by DAnIEL, plus extra press articles belonging to the same time-window collected from Google News health category for Arabic (ar), Chinese (zh), Czech (cz), English (en), French (fr), German (de), Italian (it), Norwegian (no), Portuguese (pt), Russian (ru), Spanish (es), Swedish (sv) and Turkish (tr). For Finnish (fi), Greek (el) and Polish (pl), articles have been collected in health categories of national newspapers and health-related RSS feeds.

The DAnIEL2 corpus contains documents from the 1st of October 2011 to the 31th of January 2012. The repartition by language and by date are shown in Table 4. 40% of these documents are written in languages probably not covered by ProMED. Since DAnIEL2 does not process original html files, a phase of pre-processing was needed to allow the system to process the documents. For this purpose, an in house unpublished scrapping tool was used to clean non relevant content of original html pages. The scrapping quality differed according to the source/language of the documents. This could worsen results as compared with the original system.

DAnIEL2 needs a list of disease names and countries for each language to perform its analysis. These lists were obtained by translations provided by Wikipedia of a list of most common disease names in English. From the corpus mentioned above, DAnIEL2 extracted 1,571 epidemic events. They are detailed in Table 5 and Table 6. 32% of these events were extracted from documents in languages probably not covered by ProMED. Few events were found in Arabic, despite the high number of documents in this language reported in Table 4. Turkish was the only language in which the system extracted no event.

Table 4. Number of articles by language and by month for the DAnIEL2 dataset

	ar	cz	de	el	en	es	fi	fr	it	nl	no	pl	pt	ru	sv	tr	zh
Articles	3,093	208	2,509	1,380	4,742	4,389	132	2,132	703	876	311	801	1,362	1,896	196	239	1,122
10.2011	780	42	631	220	1,301	952	23	412	173	197	52	182	343	240	41	74	243
11.2011	819	99	809	289	1,181	1,020	37	506	100	172	61	199	205	312	72	79	174
12.2011	735	37	712	400	1,082	1,517	32	832	224	253	111	122	485	487	37	52	303
01.2012	759	30	357	471	1,178	900	40	382	206	254	87	298	329	857	46	34	402

Table 5. Number of epidemic events extracted by DAnIEL2 by language and by month

	ar	cz	de	el	en	es	fi	fr	it	nl	no	pl	pt	ru	sv	tr	zh
#Reports	30	15	63	83	285	230	7	142	54	24	11	140	92	296	26	0	73
October 2011	3	2	7	17	63	42	2	17	12	2	0	15	30	49	2	0	12
November 2011	5	7	13	25	75	62	0	50	27	4	4	37	22	84	10	0	25
December 2011	12	3	24	18	67	71	3	48	15	12	4	36	25	54	9	0	14
January 2012	10	3	19	23	80	55	2	27	8	6	3	52	15	109	5	0	22

Table 6 exhibits the different diseases and locations involved in the events extracted. Major languages permitted to extract events in many countries, e.g. the 285 events signaled in English cover 55 different locations. Less common languages like Finnish or Swedish seem to be more specific to their country.

Table 6. Details for epidemic events extracted by DAnIEL2: number of diseases, locations and disease-location pairs per language.

	ar	cz	de	el	en	es	fi	fr	it	nl	no	pl	pt	ru	sv	tr	zh
#Reports	30	15	63	83	285	230	7	142	54	24	11	140	92	296	26	0	73
#Diseases	7	6	12	13	33	29	6	32	22	9	6	19	23	21	7	0	16
#Locations	3	2	19	7	55	35	2	39	9	7	1	45	14	70	2	0	6
#Disease-location pairs	12	9	32	25	161	115	4	85	28	11	6	83	50	141	10	0	23

4 Evaluation

This evaluation aims to assess the benefit, if any, of the parsimonious scheme of DAnIEL for multilingual epidemic surveillance reproduced by DAnIEL2. The main hypothesis is that a local disease outburst is first reported in a local language. Consequently, there may be a delay between this very first report and the report in main languages processed by existing systems. The geographic repartition of the time elapsed between the publication of the event by ProMED and DAnIEL2 will also be studied.

From the datasets presented in Section 3, 167 events were in common between ProMED and DAnIEL2, which amounts to 15% (over 1,082). This figure is consistent with the study made by Lyon *et al.* [12]: the intersection between different epidemic surveillance systems is quite small. Table 7 shows a sample of events first reported by ProMED and by DAnIEL2.

Table 7. Examples of events detected by both systems with the differences in timeliness in number of days. A positive (resp. negative) value indicates that ProMED (resp. DAnIEL2) reported earlier. Language of detection and date of publication for ProMED and DAnIEL2 are indicated.

Pair		Timeliness	ProMED		DAnIEL2	
Disease	Location	(days)	Lang.	Date	Lang.	Date
Cholera	Zimbabwe	+43	en	2011-12-18	en	2012-01-30
Influenza	Canada	+27	en	2011-11-04	en	2011-12-01
Scabies	Spain	+18	en	2011-12-25	es	2012-01-12
Hepatitis	Russia	+14	en	2011-11-22	ru	2011-12-06
Botulism	Finland	-11	en	2011-11-01	fi	2011-10-21
Rabies	Russia	-12	ru	2011-12-21	fr	2011-12-09
Jap. Encephalitis	India	-22	en	2011-11-02	en	2011-10-11
Norovirus	Russia	-29	ru	2011-12-27	ru	2011-11-28

From the 167 events extracted by both approaches, 37% were first reported by DAnIEL2 (Table 8). DAnIEL2 gives better results than ProMED for regions where it processed more documents in local languages. Most examples are found in European countries, for instance Czech Republic, Finland or Greece. To the contrary, in America, ProMED is clearly better. Table 9 exhibits the comparison between the two systems. When DAnIEL2 shows a better timeliness it is frequently due to the fact its coverage in languages is complementary. However, the impact of this coverage was difficult to assess for some languages and some countries, such as Turkey and Norway, for lack of common data.

Table 8. Locations of events first reported by ProMED and DAnIEL2

	ProMED		DAnIEL2	
	Languages	#First reports	Languages	#First reports
France,Portugal,Spain,UK	en,es,fr,pt	31	en,es,fr,nl,pt	12 (28%)
Rest of Europe	en,fr	7	cz,de,el,fi,fr,it,sv	12 (63%)
Russia,Ukraine	en,ru	4	pl,ru	6 (60%)
North Africa	en,fr	5	ar,fr	3 (38%)
Rest of Africa	en,fr,pt	10	fr	3 (23%)
China,India	en	5	cn,en	3 (38%)
Rest of Asia	en	6	cn,ru	9 (60%)
North America	en,es	22	en,es	4 (15%)
Central,South America	en,es,pt	16	en,es,pt	9 (36%)
All locations	5	106	15	61 (37%)

Table 8 shows that the result of the comparison between the two systems is mainly affected by the locations. ProMED clearly outperforms DAnIEL2 in English-speaking and Spanish-speaking regions, specially North and Central America. ProMED also has a better timeliness in Portuguese-speaking and French-speaking regions in America and Africa. DAnIEL2 reports sooner on local events in African regions where news are published in Arabic.

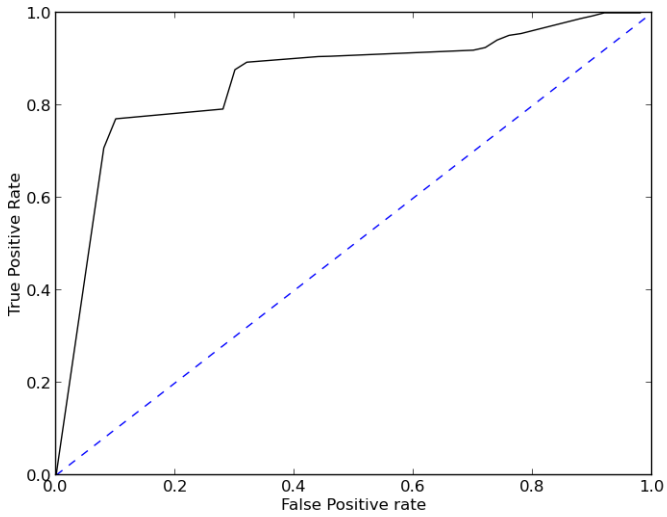
Table 9. Repartition by language of events first reported by ProMED and DAnIEL2, "-" means a non-covered language.

	ar	cn	cz	de	el	en	es	fi	fr	it	nl	no	pl	pt	ru	sv	tr
ProMED	-	-	-	-	-	54	27	-	5	-	-	-	-	15	6	-	-
DAnIEL2	1	3	2	3	4	8	8	2	8	3	3	0	2	4	9	1	0

ProMED is rarely slower than DAnIEL2 when reports are coming from articles in major languages (with the exception of Russian), confirming the idea that in major languages human analysis remains the reference [16]. Table 9 shows that the repartition of languages allowing DAnIEL2 to outperform ProMED is quite large. The hypothesis that a local language conveys information on its country of origin is valid, but not sufficient. It is also common that a neighbor country signals a disease when it spreads, thus becoming an epidemic event. This fact is correlated with the accuracy by zone presented in Table 8.

DAnIEL2 shows better results in Europe, in countries where it is the only system to analyze reports in the local languages, whereas it is outperformed by ProMED in other countries. The influence of Russian is difficult to assess since few events are reported by both systems. Their results are comparable in Russia and Ukraine. DAnIEL2, however, reported events occurring in Asia earlier than ProMED, thanks to documents in Russian.

Fig. 1. ROC curve for DAnIEL2 (plain black). Results for a manually annotated subset of 2,089 documents. The Area Under the Curve is 0.86.



We ran an evaluation on a manually annotated subset containing 2,089 documents in five languages (el, en, pl, ru and zh). Figure 1 presents the ROC curve of DAnIEL2 results. The area under the curve for this experiment is 0.86. One

can see that DAnIEL2 achieves a good equilibrium between True Positive (TP) rates and False Positive (FP) rates. For instance, for a 0.91 TP rate the system shows a 0.31 FP rate.

5 Conclusion

This paper proposed a comparative study of ProMED-mail, the reference human-based analysis for epidemic surveillance, and a multilingual automatic surveillance system, DAnIEL2, checked for 17 languages including little-studied ones. It was derived from an existing system called DAnIEL intended to process multiple languages at a limited cost. ProMED proposes highly reliable human-produced reports and seems to mostly use sources in the five major languages used to disseminate reports (English, French, Portuguese, Russian and Spanish). DAnIEL2, on the other hand, is a light automatic system using only parsimonious resources. It was indeed possible to extend the previous DAnIEL coverage by eleven languages in a very short time, each addition taking a couple hours once the crawling was done. This is a major breakthrough in disease monitoring.

From the events signaled by ProMED in a three-month time period, around 15% (167 over 1,082) were also extracted by DAnIEL2 and thus allowed comparison on a common set. The overlap between the two systems is quite small but the figures are comparable to those presented in previous studies [12]. The two characteristics studied here were the timeliness of the first description of events and their geographic repartition. Among the 167 epidemic events, roughly two out of three was first extracted by ProMED, leaving one third first detected automatically. DAnIEL2 gave worst results than ProMED for regions where English, French, Portuguese and Spanish are the main languages. The timeliness of the two approaches was comparable in Russian.

However, DAnIEL2 offers an important improvement for countries where it takes advantage of the local language news, mainly in Europe in the experiment related here. The human-based approach and the style-based automatic approach are complementary. When human analysts for one language are available, DAnIEL2 is outperformed. To the contrary, ProMED gives a great importance to English and Spanish. It is noteworthy that ProMED also relies on automatic surveillance systems, heavily if not exclusively based on English. This causes a bias in geographic coverage.

DAnIEL2 offers an interesting added-value for parts of the world where minor languages are used. It would clearly be worthwhile to test more languages for Africa and Asia, all the more so since the cost is low. The parsimonious approach behind this system seems to be well adapted for covering these regions. Therefore, the complementarity between opposite approaches seems to be important in terms of massive multilingual coverage. To complete this study, relevance tests to compare DAnIEL2 with reference data are needed to assess its sensitivity. This study shows that an automatic system does not replace manual systems, but could well assist experts to filter the web news and help detect epidemic events early.

References

1. Madoff, L., Freedman, D.: Detection of Infectious Diseases Using Unofficial Sources. In: Infectious Diseases: A Geographic Guide. Wiley-Blackwell (2011) 11–21
2. Mawudeku, A., Blench, M.: Global Public Health Intelligence Network (GPHIN). 7th Conference of the Association for Machine Translation in the Americas (AMTA) (2006) 7–11
3. Son, D., Quoc, H.N., Ai, K., Collier, N.: Global Health Monitor - a Web-based system for detecting and mapping infectious diseases. Proc. International Joint Conference on Natural Language Processing (IJCNLP) (2008) 951–956
4. Tolentino, H., Kamadjeu, R., Fontelo, P., Liu, F., Matters, M., Pollack, M.P., Madoff, L.: Scanning the Emerging Infectious Diseases Horizon - Visualizing ProMED Emails Using EpiSPIDER. Advances in disease surveillance **2** (2007) 169
5. Yangarber, R., von Etter, P., Steinberger, R.: Content collection and analysis in the domain of epidemiology. Proceedings of DrMED-2008: International Workshop on Describing Medical Web Resources (2008)
6. Lejeune, G., Doucet, A., Yangarber, R., Lucas, N.: Filtering news for epidemic surveillance: towards processing more languages with fewer resources. In: 4th Workshop on Cross Lingual Information Access. (2010) 3–10
7. Freifeld, C.C., Mandl, K.D., Reis, B.Y., Brownstein, J.S.: Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. Journal of the American Medical Informatics Association **15**(2) (2008) 150–157
8. Steinberger, R.: A survey of methods to ease the development of highly multilingual text mining applications. Language Resources and Evaluation (2011) 1–22
9. Katsiavriades, K., Qureshi, T.: The 30 most spoken languages of the world. `http://www.krysstal.com/spoken.html` (2007)
10. Mondor, L., Brownstein, J.S., Chan, E., Madoff, L.C., Pollack, M.P., Buckeridge, D.L., Brewer, T.: Timeliness of nongovernmental versus governmental global outbreak communications. Emerging Infectious Diseases **18**(7) (July 2012) 1184–1187
11. Piskorski, J., Belyaeva, J., Atkinson, M.: Exploring the usefulness of cross-lingual information fusion for refining real-time news event extraction: A preliminary study. In: Proceedings of Recent Advances in Natural Language Processing. (2011) 210–217
12. Lyon, A., Nunn, M., Grossel, G., Burgman, M.: Comparison of Web-Based Biosecurity Intelligence Systems: BioCaster, EpiSPIDER and HealthMap. Transboundary and Emerging Diseases **59**(3) (June 2011) 223–232
13. Lejeune, G., Brixtel, R., Doucet, A., Lucas, N.: DAnIEL: Language Independent Character-Based News Surveillance. In: JapTAL. (2012) 64–75
14. Cowen, P., Garland, T., Hugh-Jones, M.E., Shimshony, A., Handysides, S., Kaye, D., Madoff, L.C., Pollack, M.P., Woodall, J.: ProMED-mail as an electronic early warning system for emerging animal diseases: 1996 to 2004. JAVMA **229**(7) (2006) 1090–1099
15. Morse, S.S.: Public health surveillance and infectious disease detection. Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science **10**(1) (July 2012) 6–16
16. Collier, N.: Towards cross-lingual alerting for bursty epidemic events. Journal of Biomedical Semantics **2**(Supp5) (2011) 1–11