Internet surveillance systems for early alerting of health threats

J P Linge[1], R Steinberger[1], T P Weber[1], R Yangarber[2], E van der Goot[1], D H Al Khudhairy[1], N I Stilianakis (Nikolaos.Stilianakis@jrc.it)[1]

1. Joint Research Centre (JRC), European Commission, Ispra (VA), Italy
2. Department of Computer Science, University of Helsinki, Helsinki, Finland

In order to gather a comprehensive picture of potential epidemic threats, public health authorities increasingly rely on systems that perform epidemic intelligence (EI). EI makes use of information that originates from official sources such as national public health surveillance systems as well as from informal sources such as electronic media and web-based information tools. All these sources are employed to enhance risk monitoring with the purpose of early alerting and initial risk assessment. In this context *Paquet et al.* [1] distinguish between *indicator-based* risk monitoring and *event-based* risk monitoring. As indicator-based monitoring relies on classical routine surveillance, many systems will use methods and data sources familiar to most epidemiologists and public health officials. The event-based component of EI is in contrast rather new; its methods, strengths and limitations are generally not widely known in the public health community. The purpose of this editorial is thus to provide an overview of the methods used in pro-active event-based monitoring and to put them into context with regard to the structured indicator-based monitoring such as that described in the article on the Lithuanian electronic surveillance system published in this issue of Eurosurveillance [2].

More and more national and international public health agencies employ systematic event detection systems using informal sources (news wires, media sources or websites) on the internet to monitor the potential threat of emerging and re-emerging infectious diseases. Such web-based event detection is the first step in EI systems designed to provide early warning signals to public health institutions. A number of different systems have been developed for this purpose. There is, however, still the need to emphasise some fundamental differences between the available systems and to identify the challenges that lie ahead. Existing event detection systems can be classified into three categories.

First, *news aggregators* collect articles from several sources, usually filtered by language or country. Users gain easy access to many sources through a common portal, but still need to examine each individual article.

Second, *automatic systems* such as the Medical Information System (MedISys) (http://medusa.jrc.it/) [3], Pattern-based Understanding and Learning System (PULS) (http://puls.cs.helsinki.fi/medical/) [3], HealthMap (http://www.healthmap.org/) [4], and BioCaster Global Health Monitor (http://biocaster.nii.ac.jp/) [5] go beyond the mere gathering task by adding a series of analysis steps. Automatic systems differ in their levels of analysis, in the range of information sources, their language coverage, the speed of delivering information and visualisation methods. HealthMap currently covers five languages, BioCaster seven languages, and MedISys more than 40 languages. While HealthMap mainly relies on Google News, World Health Organization (WHO) news feeds, ProMED-Mail (http://www.promedmail.org/) [6], and Eurosurveillance as sources, MedISys monitors ProMED-Mail, web sites of national public health authorities, specialist web sites (including Eurosurveillance), news from about twenty news wires, plus a balanced list of approximately 2,200 news sources from around the world, hand-selected with a view of ensuring a geographic balance.

Analysis steps may include: recognition of relevant terms (names of diseases, symptoms and organisations), recognition and disambiguation of geographical locations mentioned in the articles, grouping related articles into clusters, and extraction of full events from the news, providing the users with aggregated information about the disease, the number of cases, as well as time and place of an outbreak. Ideally, news items should be clustered across languages and national borders. Most systems focus on recognising communicable diseases and visualise the location of the extracted events on geographical maps. As a domain-specific application of the Europe Media Monitor (EMM) system, MedISys covers not only the whole range of chemical, biological, radiological and nuclear threats (CBRN), but also allows using a filter to only show outbreak-related information. MedISys additionally monitors trends and calculates alert levels per disease and per country, by comparing the number of recent news items with averages. PULS, which is integrated with MedISys, extracts event data from the English MedISys articles and produces searchable outbreak data in table format.

All automatic systems will clearly benefit from better machine-translation software so that a more diverse range of sources can be tapped. Ideally, a summary of each article should be shown in the original language together with its translation.

Third, *moderated systems* such as ProMED-Mail [6], GPHIN (Global Public Health Intelligence Network) [7] and ARGUS [8] rely on a group of analysts to scan available news sources. The analysts take into account information from individual web sites, aggregator sites, automatic systems, and other sources such as reports from medical practitioners and health authorities. In combination with its Rapid News Service (RNS) tool, MedISys also allows for manual moderation.

There are fundamental differences in these approaches. Non-moderated systems are able to search the web and display new articles without time delay in an unbiased manner. Moderated systems show fewer irrelevant news items (fewer false positives). However, moderator bias represents a risk (false negatives); users might have a different focus than the moderators.

For users who need to react to threats quickly and possess the man-power to entertain their own monitoring effort, automatic systems are appealing because of the detection speed. Other users might prefer to wait for human-moderated feeds.

Technical implementation of aggregators is straight-forward, but for both automatic and moderated systems, many challenges lie ahead. Redundancy is a major issue. Naturally, news agencies, online and printed news sources, national and international authorities or blogs may report the same event in different ways at various time points. This often leads to misclassification of events and overestimation of impact. Furthermore, feedback loops are created when automatic systems accept input from moderated systems (or vice versa). In any moderated approach, long-term funding or volunteer participation is necessary to maintain the analyst base.

Automatic approaches are the only option to sieve relevant information out of the abundant pool of multilingual media sources in real time. However, human moderation is needed eventually.

A further challenge for the future will be to improve the transition from risk monitoring to risk assessment. Recent approaches on extracting patterns of influenza-related search terms from queries stored by Google and Yahoo [9, 10] showed that patterns of searches matched with official influenza surveillance data, thus indicating that search-term analysis could be a useful complementary tool to surveillance. However, although search-term analysis and event-based monitoring can provide an important signal of a potential outbreak, the data gathered is usually not detailed or reliable enough to estimate relevant epidemiological parameters of incipient outbreaks and the methods are prone to false alarms.

Lithuania's electronic reporting system described in this issue of Eurosurveillance [2] is an example of an indicator-based component of EI which allows the collection of structured data at country level. Such national information is typically fed into the European Surveillance System (TESSy) [11] of the European Centre for Disease Prevention and Control (ECDC) which collects surveillance data on infectious diseases at the European Union (EU) level to support outbreak detection, risk assessment, outbreak investigation and control measures. This is complemented by the Early Warning and Response System (EWRS) which establishes permanent communication between public health authorities in the EU member states [12].

References

1. Paquet C, Coulombier D, Kaiser R, Ciotti M. Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. Euro Surveill. 2006;11(12):pii=665. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=665
2. Domeika M, Kligys, G, Ivanauskiene O, Mereckiene J, Bakasenas V, Morkunas B, Berescianskis D, Wahl T, Stenqvist K. Implementation of a national electronic reproting system in Lithuania. Euro Surveill. 2009;14
3. Steinberger R, Fuart F, van der Goot E, Best C, von Etter P, Yangarber R. Text mining from the web for medical intelligence. In: Fogelman-Soulié F. Perrotta D, Piskorski J, Steinberger R, editors. Mining Massive Data Sets for Security, IOS Press, Amsterdam, 2008. p. 295-310. Available from: http://langtech.jrc.it/Documents/2009_MMDSS_Medical-Intelligence.pdf
4. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. J Am Med Inform Assoc. 2008;15(2):150-7.
5. Collier N, Doan S, Kawazoe A, Goodwin RM, Conway M, Tateno Y, Ngo QH, Dien D, Kawtrakul A, Takeuchi K, Shigematsu M, Taniguchi K. BioCaster: detecting public health rumors with a Web-based text mining system. Bioinformatics. 2008, 24, 2940-2941.
6. Madoff LC, ProMED-mail: an early warning system for emerging diseases. Clin Infect Dis. 2004;39(2):227-32.
7. Mykhalovskiy E, Weir L. The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health. Can J Public Health. 2006;97(1):42-4.
8. Wilson JM, Polyak, MG, Blake JW, Collmann J. A heuristic indication and warning staging model for detection and assessment of biological events. J Am Med Inform Assoc. 2008;15(2):158–71.
9. Polgreen PM, Chen Y, Pennock DM, Nelson FD. Using internet searches for influenza surveillance, Clin Infect Dis. 2008;47(11):1443-8.
10. Ginsberg J, Mohebbi MH, Patel RD, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data, Nature. 2009;457(7232):1012-14.
11. Amato-Gauci A, Ammon A. ECDC to launch first report on communicable diseases epidemiology in the European Union. Euro Surveill. 2007;12(23):pii=3213. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=3213
12. Guglielmetti P, Coulombier D, Thinus G, Van Loock F, Schreck S. The Early Warning and Response System for communicable diseases in the EU: an overview from 1999 to 2005. Euro Surveill. 2006;11(12):pii=666. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=666

This article was published on 2 April 2009.

Citation style for this article: Linge JP, Steinberger R, Weber TP, Yangarber R, van der Goot E, Al Khudhairy DH, Stilianakis NI. Internet surveillance systems for early alerting of health threats. Euro Surveill. 2009;14(13):pii=19162. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19162