BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 8 2012, pages 1186–1188
doi:10.1093/bioinformatics/bts099

Databases and ontologies

Advance Access publication March 1, 2012

GENI-DB: a database of global events for epidemic intelligence

Nigel Collier[1,2,*] and Son Doan[1]

[1]National Institute of Informatics, ROIS and [2]PRESTO, Japan Science and Technology Corporation, Tokyo 101-8430, Japan

Associate Editor: Jonathan Wren

ABSTRACT

Summary: We present a novel public health database (GENI-DB) in which news events on the topic of over 176 infectious diseases and chemicals affecting human and animal health are compiled from surveillance of the global online news media in 10 languages. News event frequency data were gathered systematically through the BioCaster public health surveillance system from July 2009 to the present and is available to download by the research community for purposes of analyzing trends in the global burden of infectious diseases. Database search can be conducted by year, country, disease and language.

Availability: The GENI-DB is freely available via a web portal at http://born.nii.ac.jp/

Contact: collier@nii.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 5, 2011; revised and accepted on February 23, 2012

1 INTRODUCTION

Systems which gather information about disease outbreak events from informal digital sources such as news media are now seen as having high value to national and transnational public health agencies (Heymann and Rodier, 2001). Although agencies in wealthy countries have a sophisticated array of indicator sources such as over-the-counter sales or sentinel networks, not all countries possess the resources to implement or maintain such systems. With concerns about newly emerging diseases such as A(H5N1), there has been increasing attention on epidemic intelligence (EI) systems that can complement indicator networks by detecting events on a global scale so that they can be acted on close to source.

While there are several surveillance systems offering alerts and news browsing (Hartley *et al.,* 2010), to the best of our knowledge there are only a few databases where researchers, government health officials, physicians and public health practitioners can look for historical event data which are updated in real time. ProMED (Madoff and Woodall, 2005), a human network organized through the International Society for Disease Surveillance, is an excellent source with its wide-coverage open access database of human disease reports from 1995 onwards. Reports are gathered manually and reviewed by experts in a staged process before being sent out via email and stored in an online database. ProMED reports on both human and animal health events although its coverage of animal health is more limited in scope. The World Health Organization (WHO) offers a more specialized service through its weekly EuroFlu bulletins which are archived online as well as a news bulletin service through its Global Alert and Response site. Additionally, the World Animal Health Information Database (WAHID) provides access to reports of exceptional events submitted by member states of the OIE (World Organisation of Animal Health).

GENI-DB is a complementary service that provides additional support to those interested in understanding the context of ongoing disease outbreaks as well as analyzing the global burden of infectious diseases. We developed the GENI-DB database as a free, structured and searchable source on news event statistics reported in the global media. Information gathering has been done fully automatically without human intervention by the BioCaster EI system (Collier *et al.,* 2008) from thousands of sources in 10 languages. Our experiments have shown that aggregated event counts from news can provide valuable early warning alerts that in some cases are more timely than ProMED (Collier, 2010). An additional advantage is that since BioCaster is a single system with a common reporting standard, this allows users to obtain comparative estimates of disease outbreaks across disease conditions and geographic areas.

2 METHODS

The GENI-DB database and web server is implemented on a 24 × 2.66 GHz Xeon core server running on Ubuntu Linux version 9.04, Apache (version 2.2.11), PHP (version 5.2.9) and MySQL (version 14.14) and is viewable in all major web browsers and operating systems, e.g. Safari, IE, Firefox and Chrome on Linux/Windows/Apple OS. The database is freely available for users to view and download data 24/7. Updates to the database take place once every hour during normal operation but this can be shortened to 20 min as required during public health emergencies.

The BioCaster system comprises a modularized text mining pipeline running on a dedicated cluster linked to the backend of the web server. The modules consist of efficient natural language processing algorithms for web scraping, language detection, machine translation (Koehn *et al*., 2007), classifying documents into relevant or non-relevant (Bow toolkit: www.cs.cmu.edu/~mccallum/bow) as well as dedicated modules for identifying terms and their relationships (Simple Rule Language editor: http://code.google.com/p/srl-editor). These modules are implemented in various programming languages and glued together using Perl scripts. Various modules are integrated with a sophisticated knowledge model of the domain defining semantic categories for diseases, species, symptoms, agents etc. and the relationships between them. These relationships are assembled automatically into an event report comprising a slot filler template with a minimum fill of a country, province, disease, species and time element (Collier *et al.,* 2008). One event report is generated for each relevant news event. Reports cover 176 infectious diseases including under-specified types such as 'Unclassified disease'. At various points in the pipeline staged filtering heuristics are applied to ensure a minimum level of quality. For example, events with no identifiable province are entered into the database

*To whom correspondence should be addressed.

© The Author(s) 2012. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

but considered to be of lower quality and therefore not shown in the output of GENI-DB.

Currently BioCaster surveillances ~27 000 news items per day from Google News as well as various public and NPO sources such as the ProMed-mail, Hong Kong SAR Communicable Disease Watch list, the OIE alert lists, the European Media Monitor alerts and AlertNet.

One important change that has occurred in the system was in December 2011 when the freely available Google Translate service was deprecated. Up to this point, BioCaster used this service to translate articles to English in order to assess topical relevance. We have endeavored to work around this by implementing the freely available MOSES machine translation system and have currently trained translation engines for Arabic, Russian, French, Portuguese and Spanish to English. Chinese, Dutch, German, Italian, Korean and Vietnamese to English are expected to be ready by April 2012. However we have not been able to recover Thai to English due to lack of parallel texts needed to train the system. It is still too early to assess the quality impact on performance but we hope to report on this in future publications.

Domain modeling is encapsulated through the BioCaster ontology, a freely available public health applications ontology designed to integrate laymen's language of disease reporting across 12 languages (http://code.google.com/p/biocaster-ontology).

3 RESULTS

GENI-DB is a useful source for exploring media reporting patterns as well as following disease outbreaks. Figure 1 illustrates how aggregated multilingual reporting can be used to visualise media coverage and timeliness in different languages for the porcine foot-and-mouth epidemic in South Korea during 2010–2011. Ranking diseases by the number of reports (Table 1) and countries by the number of outbreak events detected per unit of population (Tables S2 and S3 in Supplementary Material) gives an indication of both the incidence of disease but also the characteristics of online media reports.

Ranking countries according to the language of the report also yields some interesting trends in media focus, supporting our view that multilingual reports are necessary to maximize sensitivity. For example, Haiti features in the top three reported countries between July 2009 and July 2011 for most languages except for Chinese where it appeared ranked at seven behind Japan, Taiwan, USA and France. In French both Angola and Canada were more widely reported than USA. A recent quantitative study by Lyon *et al.* (2011) provides further insights into the volume, geographic coverage, timeliness and sources of BioCaster's information and a comparison against two other systems: HealthMap and EpiSpider.

Within the GENI-DB database we have non-zero event counts for 170 states. Several states have unexpectedly low counts either because there were very few open source reports during the two year period (e.g. for sub-Saharan Africa or central Asia) or because of technical limitations in the system such as missing languages (e.g. Polish), out of vocabulary names for provinces (e.g. Egypt), failure to normalise diacritics to Roman and non-registration of small island states. These issues are now being addressed by adding automated detection for alternative Arabic Romanizations, extending our place names ontology to include all world states and provinces as well as a greater number of minor cities (populations under 100 000). Additionally, we are constantly looking at how we can improve place name disambiguation from evidence in the text which is one of the greatest technological challenges we face.

Several caveats need to be kept in mind when interpreting the data. Perhaps the most important is that the data have been sourced and analyzed automatically, i.e. no human moderation has taken place. In this respect, the events reported in the database are as is. De-duplication has not been attempted, except to exclude articles with the same URL, since the frequencies of reports may have something useful to say about the degree of concern felt about an event.

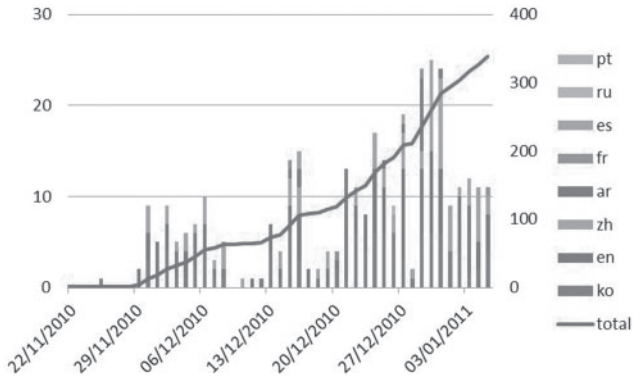

Fig. 1. Porcine foot-and-mouth outbreak in South Korea 2010–2011. Daily news event counts are shown for several languages as denoted by the ISO 639-1 codes. Event frequencies are given in stacked bar graphs and on the left-hand axis. The line graph and right-hand axis show cumulative event frequencies.

Table 1. Disease event frequency by species in GENI-DB between 15th July 2009 and 28th July 2011

Rank	Human disease	Reports	Animal disease	Reports
1	Unclassified influenza	20 982	Unclassified influenza	7519
2	Cholera	19 936	Foot-and-mouth	3827
3	Influenza A(H1N1)	17 759	Influenza A(H5N1)	2202
4	Dengue fever	14 064	Influenza A(H1N1)	1351
5	Measles	6378	West Nile fever	815
6	E-coli	2557	Anthrax	658
7	Anthrax	2123	Rabies	595
8	Influenza A(H5N1)	1946	Herpes	583
9	HFMD	1788	Brucellosis	568
10	Malaria	1716	Eastern equine encephalitis	512

4 CONCLUSIONS

The goal of GENI-DB is to offer a complementary service to extant databases helping provide insights and overcome information overload on experts. The database provides opportunities for comparisons against other sources as well as material for generating synthetic datasets. We hope that by making GENI-DB available, the data can aid in analysis of global trends, progress the state of the art in automated event alerting as well as helping those interested more generally in the patterns of media reporting on public health.

ACKNOWLEDGEMENT

We would like to thank all those who have contributed toward creating and maintaining the BioCaster system.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Funding: Japan Science and Technology Agency (JST) Sakigake fund.

Conflict of Interest: none declared.

REFERENCES

Collier,N. (2010) What's unusual in online disease outbreak news? *J. Biomed. Semantics*, **1**, 2.

Collier,N. *et al.* (2008) BioCaster: detecting public health rumors with a web-based text mining system. *Bioinformatics*, **24**, 2940–2941.

Hartley,D. *et al.* (2010) The landscape of international biosurveillance. *Emerg. Health Threats J.*, **3**, e3.

Heymann,D.L. and Rodier,G.R. (2001) Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases. *Lancet Infect. Dis.*, **1**, 345–353.

Koehn,P. *et al*. (2007) Moses: open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, demonstation session*, Prague, Czech Republic.

Lyon,A. *et al*. (2011) Comparison of web-based biosecurity intelligence systems: BioCaster, EpiSPIDER and HealthMap. *Transbound.Emerg. Dis.* doi: 10.1111/j.1865-1682.2011.01258.x

Madoff,L.C. and Woodall,J.P. (2005) The internet and the global monitoring of emerging diseases: lessons from the first 10 years of Promed-mail. Arch. Med. Res., **36,** 724–730.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016