Review of Online Systems for Biosecurity Intelligence–Gathering and Analysis

ACERA Project 1003

Aidan Lyon

University of Maryland, College Park
Australian National University
University of Sydney

Final Report

December 2010

Acknowledgements

This report is a product of the Australian Centre of Excellence for Risk Analysis (ACERA). In preparing this report, the authors acknowledge the financial and other support provided by the Department of Agriculture, Fisheries and Forestry (DAFF), the University of Melbourne, Australian Mathematical Sciences Institute (AMSI) and Australian Research Centre for Urban Ecology (ARCUE).

The author appreciates comments and general feedback on earlier drafts of this report and work reported therein by Peter Black, Mark Burgman, Michael Cole, Brett Evans, Mark Gibbs, Neil Grant, Geoff Grossel, Mikael Hirsch, Tim Keeble, Claire Murray, Mike Nunn, Paul Pheloung, and Belinda Wright.

The author also acknowledges the helpful correspondence and assistance given by Nigel Collier (BioCaster); Herman Tolentino (EpiSPIDER), Michael Blench and Abla Mawudeku (GPHIN); John Brownstein and Clark Freifeld (HealthMap); and Josh Dein, Megan Hines, and Christine Marsh (WDIN).

Disclaimer

This report has been prepared by consultants for the Australian Centre of Excellence for Risk Analysis (ACERA) and the views expressed do not necessarily reflect those of ACERA. ACERA cannot guarantee the accuracy of the report, and does not accept liability for any loss or damage incurred as a result of relying on its accuracy.

Contents

List of Figures

List of Tables

1 Executive Summary

This report documents stages 1, 2, 4, and 5 of ACERA Project 1003 as well as initial progress on stage 7. The report assesses online systems for biosecurity intelligence–gathering and analysis against DAFF's intelligence needs (stage 2), which were evaluated in a workshop on 14/08/2009 (stage 1). The report also assesses existing online animal biosecurity intelligence systems (stage 4) and plant biosecurity intelligence options (stage 5). Work on stage 5 lead to the development of a prototype plant health intelligence system, thus making initial progress on stage 7. It also lead to the development of a more sophisticated prototype intelligence system for marine pests and diseases. The main findings of this report are:

1. The reviewed biosecurity intelligence systems do not satisfy DAFF's intelligence needs. There is virtually no coverage of plant pests and diseases, and the same is true for marine/aquatic pests and diseases. Although animal pests and diseases are better covered, the reviewed systems still fail to satisfy DAFF's biosecurity intelligence needs.

2. DAFF staff are crippled by current IT restrictions. The situation is so dire that in order to complete various stages of this research project, ACERA had to supply some members of DAFF with laptops with independent and unrestricted internet connections. These laptops have since proven to be indispensable to staff conducting modern biosecurity intelligence gathering and analysis.

3. There is a significant amount of important information pertaining to biosecurity intelligence in new online media such as Twitter, Facebook, YouTube, podcasts, and RSS feeds. DAFF staff are currently unable to access this information. Similarly, a wealth of information exists in scholarly journals, forums, discussion boards, and webinars. For various reasons, DAFF staff are unable to access this information and this is seriously affecting their ability to acquire and analyse biosecurity intelligence.

4. There is a huge opportunity for DAFF to be an international leader in modern biosecurity intelligence, especially for plant and marine/aquatic pests and diseases. Initial progress has already been made in developing a sophisticated marine and aquatic biosecurity intelligence system. This work was done during stage 5 of this project in collaboration with Dr Geoff Grossel from the Aquatic Division of Animal Biosecurity in the BSG. One early version of this system (developed by Dr Grossel) has already lead to a concrete decision that mitigated a serious biosecurity risk: oyster or ostreid herpesvirus (OsHV–1 or OsHV–1 μvar) entering Australia. The system gave early warning of the disease spreading through Europe, and the probability of it spreading to Australia through the importation of used oyster farming equipment. Based on this information, it was decided that all such equipment should be decontaminated before leaving quarantine. New Zealand was not in a position to make this decision, and the virus has now spread to New Zealand oyster farms.

Biosecurity information now exists in many forms on the internet. In the past, such information was mostly contained in news articles, which users would have to access directly or through a search engine (e.g., Google). However, RSS feeds have gained widespread use on the internet, and now articles from news media sites can be delivered to users as soon as they are published online. This means that biosecurity information can be brought to users as soon as it is published on the web.

This stream of information can be automatically collected, sorted, translated, and enriched. For example, an article in Spanish with the symptom terms 'apatia' (apathy), 'deshidratación' (dehydration), 'náusea' (nausea), and location term 'Peru' can be automatically detected in a newspaper's RSS feed and classified as likely to be about a cholera outbreak in Peru—even though the term 'cholera' is not in the original article. This enriched report can be published to a number of other RSS feeds, which can be tailored to the interests of health officials. One such RSS feed could be devoted to all articles on potential cholera outbreaks. Another could be for all articles on any public health issue in Peru. There is no limit to how RSS feeds can be filtered and recombined in this way.

Since the establishment of RSS feeds on the web, new types of feeds have emerged. Geo–RSS and KML feeds allow articles to be tagged with latitude and longitude co–ordinates. This means that information can be streamed to mapping systems (such as Google Maps and Google Earth) that allow information to be plotted on a map. Instead of a simple list of articles on potential cholera outbreaks around the world, a health official can view a map in which those articles are plotted. This allows the official to achieve so–called 'situational awareness" (i.e., to see potential clusters of reports, or other pattens in the data). This map can be updated automatically as new articles are published.

Automated systems will always produce errors. A system may ascribe a collection of symptoms to be indicative of cholera, when in fact those symptoms are more likely to be caused by some other disease. GPHIN and WDIN are two systems that use dedicated experts to review their reports in order to reduce such errors. In contrast, ProMED is completely expert based (it has no automation). At the other extreme, BioCaster is completely automated, not relying on any (direct) input from experts. An approach that is between these two extremes—partly taken by HealthMap—is to allow users to make contributions to the system with commentary, rankings of relevance, etc. This approach is sometimes called a 'Web 2.0' approach as it takes advantage of the knowledge that can be generated by large numbers of users interacting through a software system. (Wikipedia, for example, is a Web 2.0 application.) This sort of approach has the advantage of making use of human analysis for free.

A similar approach is taken by EpiSPIDER, which collects information from Twitter. By watching Twitter for biosecurity–related search terms, EpiSPIDER gathers articles that have been read and perhaps even categorised by users. Moreover, this can be done without the users knowing it. Someone travelling in Thailand could tweet a webpage on, say, a concert being cancelled, which also pertains to biosecurity—e.g., the concert is cancelled due to organisers fear of a dengue

outbreak. This may be the only source of information on this issue, and it can be detected by EpiSPIDER, even though no one consciously reported the webpage for this purpose.

Since such user input and feedback can reduce errors and increase the number of reliable sources, some biosecurity intelligence systems appear to be moving to use social media. There are important questions concerning the reliability of social media compared with standard news media. Such issues are studied in social epistemology (e.g., on the reliability of blogs versus news media, see Goldman [2008]). Probabilistic models from the literature on judgement aggregation and consensus formation have been also studied with respect to potential biosecurity applications (see ACERA Project 607). More research in formal methods of social epistemology (with a focus on applications to biosecurity intelligence) needs to be conducted as biosecurity intelligence systems move to social media.

The report confirms that the available software systems do not adequately meet DAFF's needs—particularly in the domains of plant, marine, and aquatic pests and diseases. None of the systems give substantial attention to plant and marine biosecurity—indeed, they give very little to no coverage. Although human, animal and zoonotic diseases are better covered, focus tends to be on diseases that attract significant coverage in the news media (e.g., H1N1). As part of stage 7, two prototype systems were developed for plant and marine biosecurity. Initial results from the prototypes suggest that further investigation and development of plant and marine biosecurity intelligence systems would be of significant advantage to DAFF. *It is highly recommended that these prototypes be further developed and supported.* The development of these prototypes should be guided by a study of the pros and cons of existing biosecurity intelligence systems as well as research in social epistemology. *It is crucial that the prototypes remain open–source and unhindered by DAFF's IT restrictions.* It would also be of considerable advantage to DAFF if there were systems that covered less well–reported animal diseases, and acquired intelligence from a greater range of conventional sources (e.g., scientific journals) as well as less conventional sources (e.g., social media). More detailed recommendations for how to further develop the prototypes are given in Section 9.1.

The marine prototype system analysed information from Twitter and it was noticed that topics that received a lot of attention on Twitter were typically already known or not of interest to DAFF. This suggests that Twitter could be used to filter out topics which DAFF staff do not need intelligence on, thus allowing them to focus on issues important to them. It was also noticed that Twitter was the first or sole source of information on some events which were potentially important to DAFF. Moreover, it was discovered that there are informal networks of people sharing biosecurity information within Twitter.[1]

Many of the systems reviewed in this report are biosecurity intelligence systems that are web–based and open–source. Enterprise search systems, which in contrast are closed–source, have not been reviewed in this report. However, it is likely that the implementation of an enterprise search system, such as MS FAST or ISYS, would be of great benefit to DAFF, since such a system would

[1]For example, http://twitter.com/biosurveillance is a Twitter feed devoted to operational biosurveillance.

allow information to be shared within DAFF in an intelligent way. It should be noted, though, that an enterprise search system does not perform the same functions that open–source web–based systems perform. Although enterprise search systems can search the internet for biosecurity information in more–or–less the same fashion as open–source web–based systems do, the *results* of such searches are not open–source and are accessible only to those who have access to the enterprise search system. Systems such as ProMED clearly demonstrate that there is great value in allowing the global community to view the results, since users of the system can contribute *back* into the system, by confirming/disconfirming initial reports, submitting new information missed by the system, etc.

These two types of systems are not in conflict, and in fact complement each other. An enterprise search system can be used to organise and share information within DAFF that may be of a sensitive nature and not to be shared with the rest of the world. A web–based open–source system can be used by DAFF staff to interact with experts in the global community to acquire intelligence that would otherwise not be able to be obtained. This intelligence can then be fed into the enterprise search system to add to the quality of that information. Implementation of both types of systems would probably best serve DAFF's biosecurity intelligence needs.

Access to the diverse and new sources of biosecurity intelligence is crucial to meet many of DAFF's biosecurity intelligence needs. Currently, DAFF staff are unable to access these sources due to content bans and IT restrictions. Access is needed to RSS and KML feeds; podcasts; webinars (online seminars); blogs; Twitter; online video and audio content; YouTube; Wikis (including Wikipedia); as well as journals, discussion boards, and online forums.

2 Introduction

2.1 The Problem

At any given moment, the probability of a new outbreak of an infectious disease or pest may be low, but over long enough timespans, and with many such diseases and pests, that low probability becomes an almost certainty. We don't know when or where, but we can be sure that such an outbreak will occur. A 2005 survey of 19 epidemiology and influenza experts reported an approximate 10% probability that H5N1 or a similar virus would become an efficient human–to–human transmitter within the next 3 years (Bruine de Bruin *et al.* [2006]). This probability increased to 50% for 10 years, 90% for 15 years, and 100% for 30 years. The current speed and volume of international trade and travel makes it possible for such a new outbreak to spread around the world very rapidly. The 2009 pandemic of influenza H1N1 is a case in point. Improved influenza surveillance was ranked most likely—out of six strategies—to reduce the severity of an influenza outbreak (Bruine de Bruin *et al.* [2006], p. 191).[2]

The global spread of an infectious disease is a problem for human health, ecosystems and agriculture. The spread of animal and plant diseases can occur through international trade and human travel, and also by natural processes, such as transport by wind and ocean currents. The spread of such diseases can have devastating consequences for ecosystems and agriculture, and these can have serious economic and security consequences (Strange and Scott [2005]). The spread of the cereal wheat rust is a good example (Hovmøller *et al.* [2008]). The spread of pests is equally problematic: for example, the spread of red fire ants, the Khapra beetle, or the peach–potato aphid (Margaritopoulos *et al.* [2009]). It is estimated that at least 33% of global food production is lost due to plant diseases, pests and weeds (James [1998]).

Clearly, the effects of a global spread of an infectious disease or pest can be devastating. Millions of lives can be lost, and so too can billions of dollars. Fortunately, various tools can mitigate this damage: vaccinations, quarantine controls, travel and trade controls, and eradication programs, to name a few. Unfortunately, these tools are often expensive, scarce, or not distributed/implemented in the right way (e.g., the limited supply of H1N1 vaccines during the 2009 flu season). It is therefore important to know how best to distribute and implement these tools. This often depends on knowing where the pest or disease is occurring and how it is spreading. This is known as biosecurity intelligence, biosecurity surveillance, or biosurveillance. This report reviews options for biosecurity intelligence, focusing especially on the potential of current open–source web–based systems to support DAFF's biosecurity intelligence.

[2] The other five strategies were: social distancing, barrier methods, ring antivirals, animal control, and mass vaccination of poultry workers.

2.2 Biosecurity Intelligence

Biosecurity intelligence can take many different forms. A traditional form has been for government agencies around the world to report outbreaks to international organisations such as the World Health Organisation (WHO) for human disease or the Organisation for Animal Health (OIE) for animal diseases, including zoonoses and diseases of aquatic animals. Biosecurity intelligence also uses information technology to detect trends and signals in distributed and unstructured open–source information. For example, by trawling thousands of websites, the (partly automated) intelligence system GPHIN detected SARS in China 3 months before the WHO announced it (Keller *et al.* [2009]). The (purely human based) intelligence system ProMED also found SARS early on, publishing an e–mail from a Chinese teacher asking if anyone had heard about a large number of flu–like cases that an acquaintance of the teacher had observed Section 5.5.1. This brought the event to the attention of ProMED's subscribers, who were then able to contribute and share information through the system. By examining search term data, Google Flu Trends can accurately detect flu activity two weeks ahead of CDC reports (Section 4.1). Such intelligence systems can enable us to get around barriers (possibly political barriers) to the flow of information.

Using information technology for biosecurity intelligence raises a number of challenges. First, it involves accessing information from a large number of sources, resulting in massive streams of information (from news feeds, journals, e–mails, etc.). This massive volume can overwhelm analysts. Second, this information is constantly changing, and needs to be analysed quickly. Initial reports of an event are often contained in a few articles that can easily be missed by simple keyword searches. Fourth, not all reports are in English. Systems need to be able to translate and analyse reports in as many languages as possible.

This report reviews a number of systems that attempt to solve these problems with a focus on how the systems might be used to meet DAFF's biosecurity intelligence needs. The systems reviewed differ in a number of ways. For example, they cover different types of pests and diseases, focus on different geographical regions, can understand different languages, vary in their reliance on automation, and use different information sources.

2.3 Systems

The systems that were initially identified as potentially useful for DAFF's biosecurity needs can be classified into three categories:

(i) Web–based systems.

(ii) Data analysis and mining systems.

(iii) Enterprise search systems.

Web–based systems

Web–based systems collect information over the internet, analyse that information, and disseminate it to users. Web–based systems differ in many respects, including how they collect, analyse and disseminate information. For example, some systems collect information from users, while others rely heavily on web-scraping news–media sites. Some use human–only data analysis, while others use purely automatic analysis procedures. Some report their results *via* e–mail and webposts, while others use more advanced communication methods such as RSS, JSON, KML and Twitter feeds, and blogs.

Data analysis and mining systems

Data analysis and mining systems are statistical software packages that can process large volumes of data and present trends and results in effective ways. They focus primarily on the *analysis* of data, not on the collection or dissemination of data. Respects in which data analysis and mining systems vary include the statistical algorithms they use, their graphical user interfaces, data visualisation techniques, and model–building processes. DAFF has little need for processing large data sets, so it was deemed that there is no need to review data analysis and mining systems. The systems initially identified as potentially useful for the DAFF's biosecurity needs were Rattle, SAS Enterprise Miner, PASW Modeler (previously known as SPSS Clementine), KNIME, and NeuralWorks Predict.

Enterprise search systems

Enterprise search systems are software systems that allow enterprises to search through all of their data types (e–mail, calendar events, word documents, spread–sheets, etc.) across their entire networks. Such a system could be useful for DAFF's biosecurity needs. However, to review an enterprise search system would require implementing that system over DAFF's network and assessing how effective it was in finding different data types in different sections of the network. This would require an IT specialist with the freedom to access the department's network and such a resources was not available. Thus, it was not possible to review enterprise search systems by evaluating their performance in this report. Enterprise search systems that may be of use to the department include ISYS Workgroup Web, Microsoft FAST, Autonomy, Endeca, and Vivisimo.

It was decided that a detailed review of the web–based systems should be the main priority of stage 2 of the project—i.e., this report. The web–based intelligence systems reviewed in this report are:

- Global Public Health Intelligence Network (GPHIN)
- Google Flu Trends
- ProMED
- HealthMap
- EpiSPIDER

- Wildlife Disease Information Node (WDIN)
- BioCaster Global Health Monitor
- National Agricultural Pest Information System (NAPIS)
- EUROPHYT
- Global Avian Influenza Network for Surveillance (GAINS)
- OIE
- North American Plant Protection Organisation (NAPPO)

In Section 3, the features of these systems are compiled in a table that compares their scope, data sources and languages, method of analysis, and method of reporting. In Section 4, each system is reviewed briefly with a general description, an overview of its data collection and analysis processes, and an overview of its visualisation tools. In Section 5, a retrospective study of the systems is reported. This includes an examination of how well some of the systems reported a range of past pest and disease outbreaks—Section 5.5. However, a retrospective study was not possible for many of the systems because they do not maintain searchable archives beyond (about) 30 days. For these systems, a statistical study over a seven day period was conducted and is reported in Section 5.2. A more thorough explanation of the retrospective study is provided in Section 5.1.

There are very few existing intelligence systems for plant pests and diseases. Those that do exist, cover only certain geographical regions and have other significant limitations. Over the course of stage 2 of the project, it was realised that some of the automated processes of the human and animal disease intelligence systems could be applied to the plant domain. Section 6 examines the possibility of developing a biosecurity intelligence system for plant pests and diseases. Section 9 provides the recommendations of the project.

2.4 DAFF's Biosecurity Intelligence Needs

Stage 1 of this project was a review of DAFF's biosecurity intelligence needs. Stage 2 of this project evaluates the aforementioned web–based biosecurity intelligence systems against these needs. The review was conducted through a project workshop on 14 August, 2009 at the University of Melbourne, and informal discussions with DAFF staff during Stage 2 of the project.

Attendees of the workshop were: Colin Grant BA/Plant Division, Roberta Rossely OCCPO/Plant Division, Mike Cole OCCPO/Plant Division, Karen Schneider BA/Livestock Div, Peter Beers OCVO/Livestock Div, Peter Black OCVO/Livestock Div, Belinda Wright OCVO/Livestock Div, Brant Smith BA/Animal Division, Mike Nunn BA/Animal Division, Jenni Gordon AQIS, Aidan Lyon, Andrew Speirs–Bridge, Terry Walshe, Andrew Robinson, and Mark Burgman. This section summarises some of DAFFS's intelligence needs that were identified during the workshop or through subsequent discussions.

Various groups within DAFF have very different intelligence needs in terms of time frames.

Some groups have real–time intelligence needs—i.e., intelligence concerning what is going on in the present moment, or within the past few hours. Other groups require more medium–term intelligence—i.e., intelligence concerning events evolving over days, weeks, or even months. Other groups require an even longer–term outlook—i.e., intelligence on trends and situations evolving over years or even decades for use in strategic planning. These needs are by no means exclusive. For example, some groups require both real–time and long–term intelligence.

The workshop and subsequent discussions determined that DAFF needs to improve its capability for biosecurity intelligence–gathering and analysis, including sharing of information for both immediate (tactical or operationa) and longer–term (strategic planning) use. Information–sharing needs to be done at several levels, both between business units within DAFF, and between jurisdictions, both nationally, and internationally.

DAFF also requires varying degrees of information quality. At one extreme, DAFF staff only need very low quality information to form hypotheses or make decisions and conduct further intelligence–gathering. At the other extreme, biosecurity intelligence needs to be of a quality that can be used to support decisions and defend them in international forums.

Groups and individuals within DAFF also require intelligence on different issues. Clearly, intelligence needs differ between the animal and plant divisions of DAFF. However, even within these divisions there is variation in the intelligence needs regarding issues. Some require constant intelligence on specific diseases, while others need intelligence on various groupings of diseases (by geographical regions, or by hosts, etc). Still others require intelligence on unknown or unidentified diseases. Furthermore, these requirements often change over time. For example, on one day a group may need intelligence on all diseases and pests for oranges, and on the next day they will require the same intelligence for tomatoes. Or, on the next day they may need intelligence on any emerging plant diseases from China. Or, they may require intelligence on any possibly new emerging diseases from anywhere.

The biosecurity intelligence needs of DAFF are so wide and varied that it is unlikely that any one approach will satisfy them all. It is likely that a range of approaches will be required (e.g., use of human networks, external surveillance systems, data mining, etc.). This report evaluates which of DAFF's needs are met by the existing web–based biosecurity intelligence systems, and how well these systems meet those needs. The report also recommends how the systems can be developed to meet DAFF's biosecurity intelligence needs in the future.

3 Comparison of the systems reviewed

System	Scope	Data Source (languages)	Data Analysis	Data Output
GPHIN	Human, zoonotic, plant, marine, food, water, bio–terrorism, natural disasters, product safety, drugs.	Factiva and Al Bawaba. (Ar, En, SC, ST, Fa, Fr, Po, Ru, Es)	Automated and human translation, categorisation, and geocoding.	Tailored e–mail alerts, filtered web posts, searchable archive.
Google Flu Trends	Influenza.	Google search data, historical flu activity.	Automated categorisation, timecoding, and geocoding.	CSV files, region specific activity graphs.
ProMED	Animal, human, marine, plant, zoonotic, toxins.	Reports from users, health departments, and media. (En, Po, Es, Ru)	Several levels of human analysis and editing.	e–mail alerts, web posts, searchable archive, RSS, Twitter, Facebook.
HealthMap	Animal, human, zoonotic.	Google, Moreover, WHO, ProMED, EuroSurveillance, WDIN, user inputs. (Ar, Ch, En, Fr, Po, Sp, Ru)	Automated translation, categorisation, timecoding, and geocoding, user comments and ratings.	Filtered maps, RSS feeds, Twitter feed, blog and iPhone and Android apps. (+ KML files bought by ACERA.)
EpiSPIDER	Animal, human, zoonotic, natural and manmade disasters, political, drugs.	Google, ProMED, GDACS, CIA Factbook, Moreover, DayLife, ReliefWebUpdates, Twitter. (En)	Automated translation, categorisation, timecoding, and geocoding.	Filtered maps, RSS, JSON and KML feeds, CSV files, filtered graphs, Twitter feed, blog.
BioCaster	Animal, human, zoonotic	Google, ProMED, EMMA, Meltwater. (Ar, Ch, En, Fr, Ja, Ko, Po, Ru, Es, Th, Vi)	Ontology searches and translations, natural language processing, geo and timecoding	Filtered maps, KML files, headlines, filtered graphs.
WDIN	Animal, zoonotic.	ProMED, Google, range of journals and newsmedia. (En)	Yahoo Pipes aggregation and filtering of RSS feeds + relevance screening by humans.	Filtered maps, e–mail list, digest, monthly bulletin, RSS feeds, KML feed, blog.
NAPIS	U.S. agricultural pests.	CAPS and PPQ surveys. (En)	Manual statistical analysis.	Maps, news archive, RSS feeds.
EUROPHYT[†]	EU agricultural pests and diseases.	EU member states' boarder control and customs + others (?). (EU languages)	Manual statistical analysis + other (?).	e–mail + other (?)
GAINS	Avian Influenza (H5N1).	Manual samples by international collaborators, bird census data. (NA.)	Manual statistical analysis.	Filtered interactive map, KML feeds.
NAPPO	Plant	US, Can. and Mex. official reports, journal articles (En.)	Manual screening, checking	e–mails, website.
OIE	Animal, human, marine, zoonotic	Official reports by members (En.)	Manual screening, checking	Filtered maps, e–mails.

4 Reviews of Web–Based Systems

4.1 Google Flu Trends

Description

Google Flu Trends is an automated intelligence system that monitors influenza activity based on certain terms that have been identified as good indicators of such activity. Google researchers have discovered a correlation between the number of influenza–related searches and how many people actually have influenza symptoms (Ginsberg et al. [2008]). Certain search terms tend to increase in frequency when an influenza season is occurring, and therefore can be used as reliable indicators for influenza activity.

Google Flu Trends has recently expanded to provide estimates for 20 countries: Australia, Austria, Belgium, Bulgaria, Canada, France, Germany, Hungary, Japan, Mexico, Netherlands, New Zealand, Norway, Poland, Russia, Spain, Sweden, Switzerland, Ukraine and the United States. Google Flu Trends is part of Google.org, which is an ongoing project that uses information technology to help forecast emerging threats before they become local, regional, or global crises. Google.org's initial focus is on emerging infection diseases, and this focus has produced Google Flu Trends. The service is free to users.

Data Collection, Analysis, Storage, and Retrieval

Two kinds of data are collected: (i) search query data, and (ii) historical influenza activity data. The former is relatively simple to collect. Every time a user makes a search query using Google, that search query's search terms and IP address are stored in Google's server logs. The latter is more difficult, as it requires cooperation with government departments, and those departments having the required data.

Google Flu Trends uses IP address information in their server logs to estimate the geographical origin of the search query. Key search terms are identified as good indicators of influenza activity. These are then fed into a model that has been trained on prior data sets to estimate the current influenza activity. More details of the method of data analysis can be found in Ginsberg et al. [2008]. Influenza activity estimates for the current week are updated daily as new search query data are collected. Users can view data on the Google Flu Trends website (e.g., Figure 1) or download them in raw form as a CSV file.

Google Flu Trends has proven to be up to two weeks ahead of traditional tracking systems (see Figure 5 for a comparison between Google Flu Trends and published CDC reports).

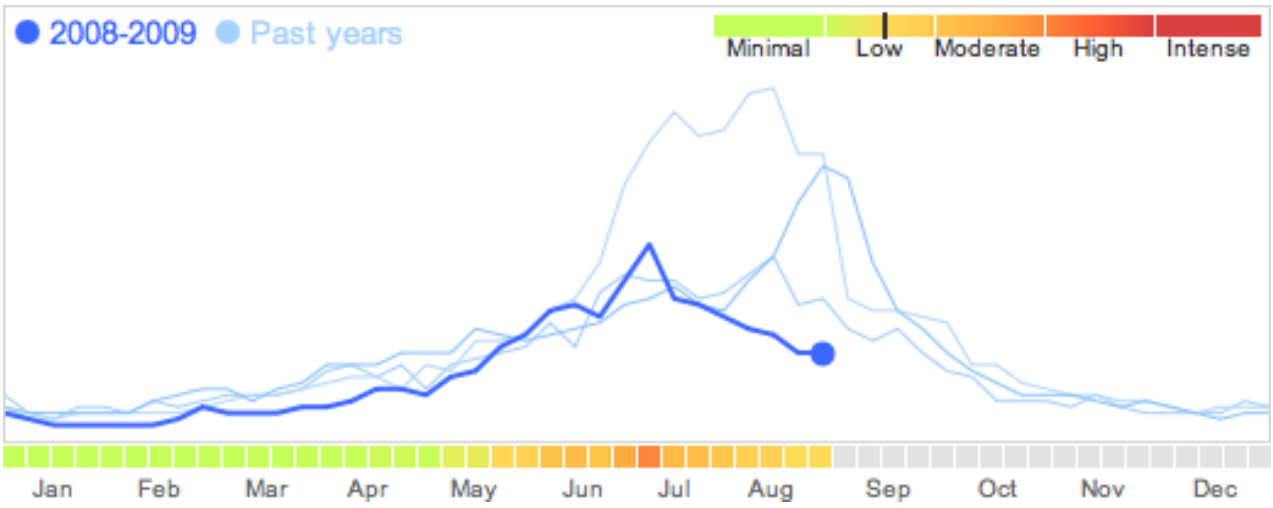


Figure 1: Visualisation of Flu Activity Estimates in Australia.

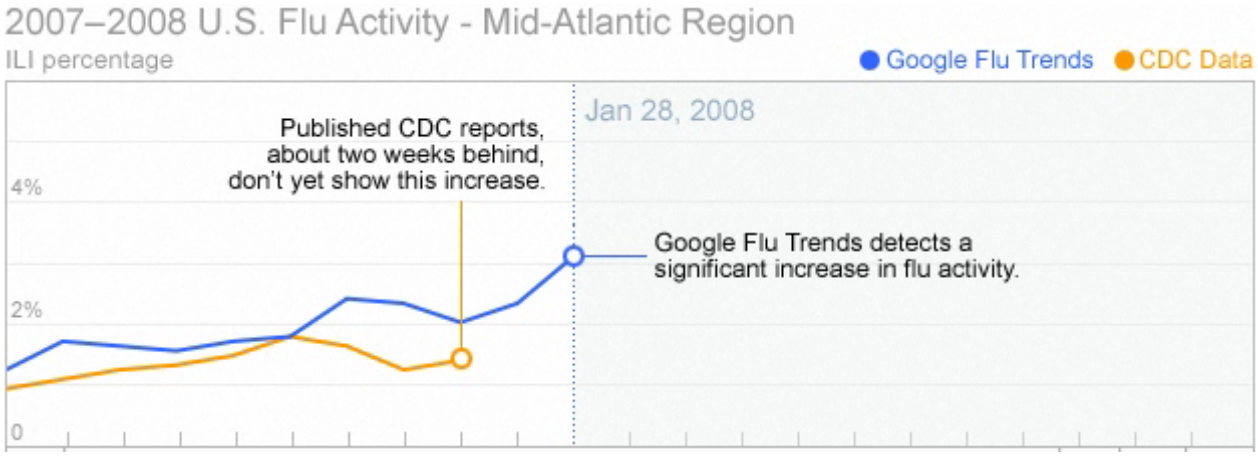


Figure 2: Snapshot of QuickTime video comparing Google Flu Trends with published CDC reports.

4.2 GPHIN

Description

GPHIN (the Global Public Health Intelligence Network) is a semi–automated, early warning system that reports information on human and zoonotic disease outbreaks. The system also covers other public health topics such as food and water contaminations, bioterrorism, exposure to chemical and radionuclear agents, natural disasters, and the safety of products, drugs and medical devices. GPHIN constantly monitors global media sources in several languages in real time. Users receive tailored e–mail alerts and can interact with filtered lists of articles using a webpage interface. GPHIN is a Microsoft/Java application compatible with Internet Explorer 6.0 (or better, except for Internet Explorer 8.0) and Netscape 6.2 (or better); it isn't compatible with any version of several popular browsers, including Safari, Firefox, and Chrome.

GPHIN was developed as a prototype by the Public Health Agency of Canada for WHO in 1997 and is now managed by the Agency's Centre for Emergency Preparedness and Response. Most GPHIN subscribers are from government organisations, but others come from non–governmental organisations interested in public health such as businesses, the North American Treaty Organisation (NATO), and WHO. GPHIN news reports comprise approximately 40% of WHO–verified disease outbreaks (Heymann *et al.* [2001]).

Data Collection, Analysis, Storage, and Retrieval

GPHIN collects information on disease outbreaks and other public health events by monitoring global news media aggregators Factiva and Al Bawaba. These aggregators gather news media from a large number of sources on the internet (including Reuters, Associated Press, New York Times, Sydney Morning Herald, Irish Times, etc.). Articles are filtered for relevance by an automated process and then analysed by GPHIN analysts. Notifications about public health events that may have serious public health consequences are sent immediately out to users as e–mail alerts.

News articles in English are posted in the system and translated into the other languages—Arabic, simplified and traditional Chinese, Farsi, French, Portuguese, Russian, and Spanish. News articles in any of the non–English languages are posted in the system and translated into English. This is both an automatic and manual process. Automated translations use dictionaries that are constantly refined by expert linguists and GPHIN analysts. GPHIN analysts (with topical expertise and linguistic skills) also manually translate what they deem to be the essence of the articles.

The automated part of the analysis begins every fifteen minutes, when GPHIN gathers articles from newsfeed aggregators (Al Bawaba and Factiva) that are determined to be relevant by established search syntaxes. The articles are then sorted into one or more of the following categories: animal diseases, human diseases, plant diseases, biologics, natural disasters, chemical disasters, radioactive incidents, and unsafe products. Each article is assigned a category–relevance score, which is a function of keywords that appear in the article that are associated with the categories

that the article has been assigned to. Each article is then automatically discarded, published, or presented to a human analyst, depending on its relevance scores (Blench [2008]). Articles that have especially high relevance scores are immediately e–mailed to GPHIN users (Mawudeku and Blench [2005]).

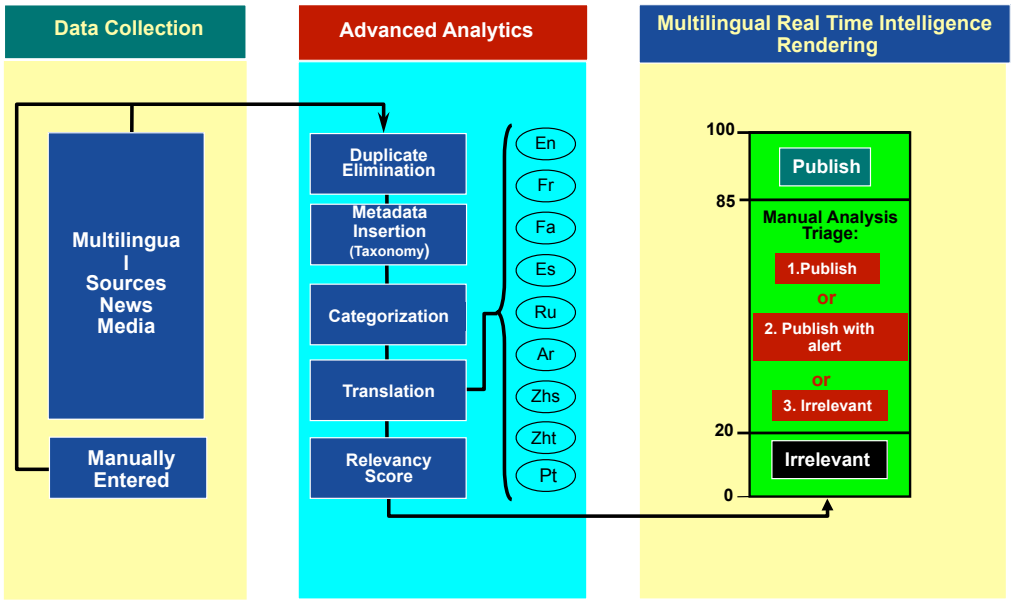


Figure 3: GPHIN Workflow (http://www.amtaweb.org/papers/Blench.pps).

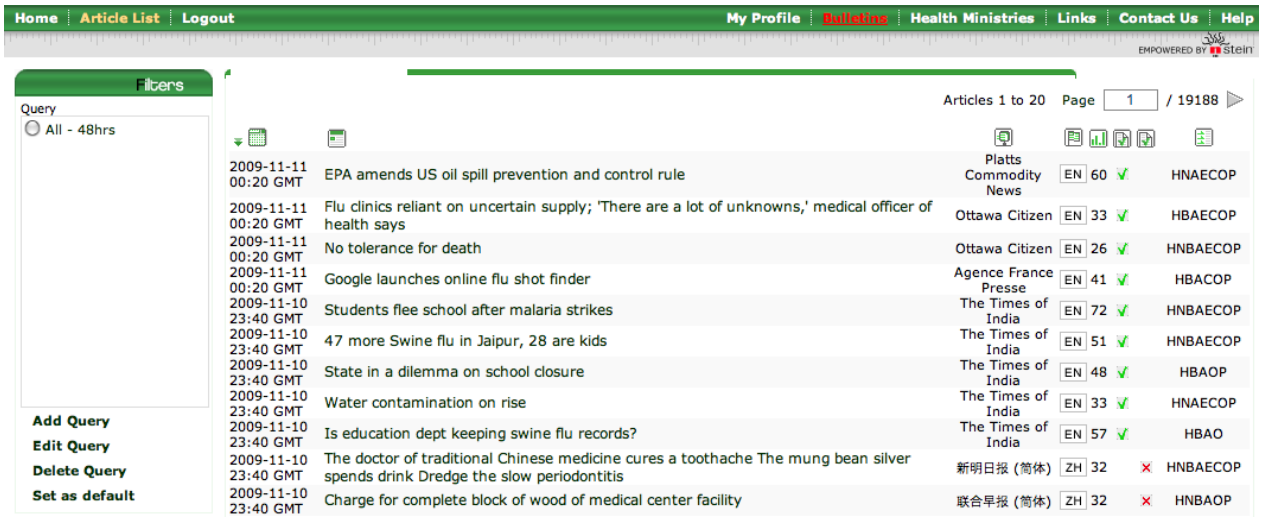


Figure 4: GPHIN Bulletins Screenshot.

There is also a manual component to the data analysis. If an article is not automatically discarded or published, it is presented to a GPHIN analyst who decides whether to discard or publish

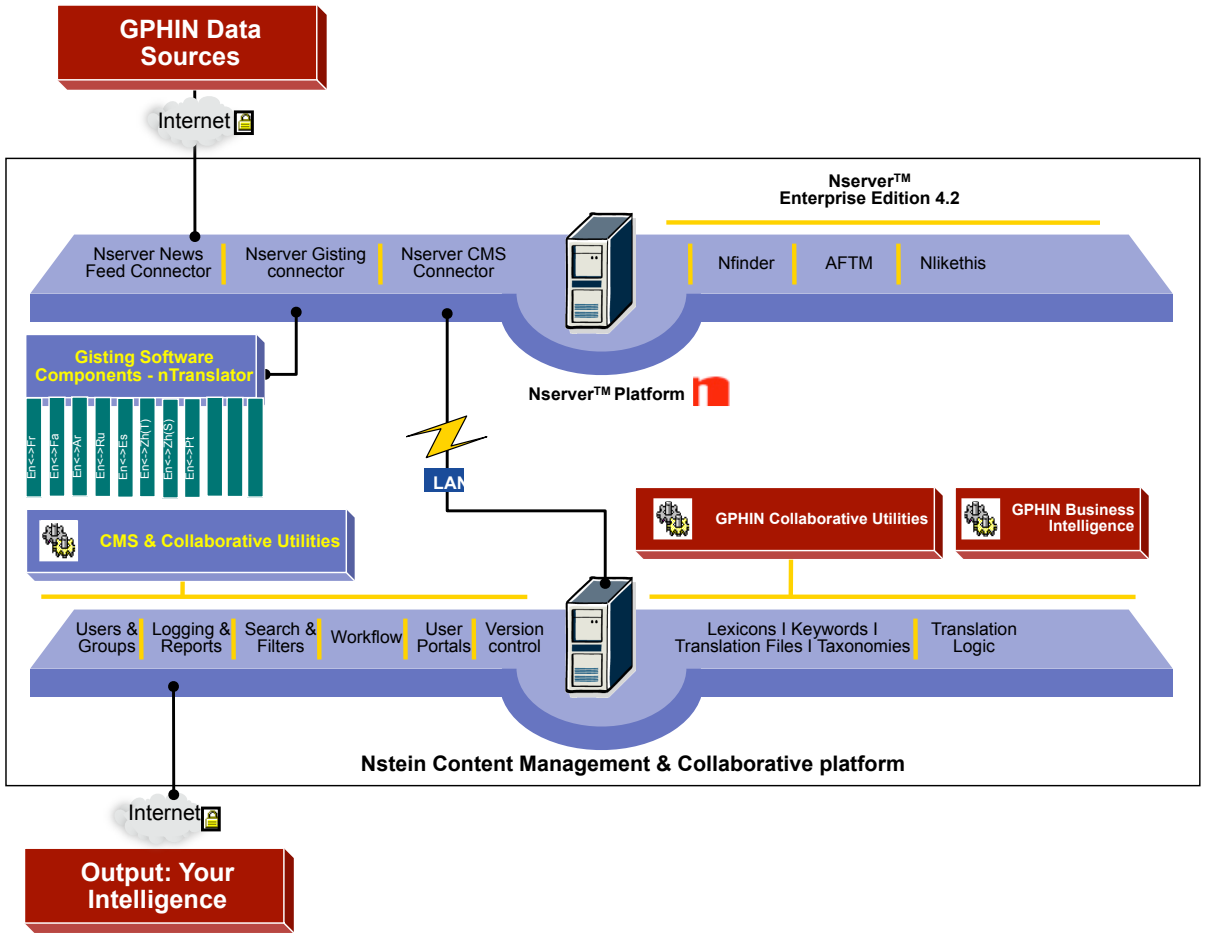


Figure 5: GPHIN Infrastructure (http://www.amtaweb.org/papers/Blench.pps).

the article. If the analyst deems the article to be of immediate concern, the article is immediately forwarded by e–mail to GPHIN users. Analysts also review automatically discarded articles to verify that relevant articles have not been discarded by the automated system. Figure 3 depicts the analysis process.

In addition to receiving tailored e–mails from the system, users can also review the latest list of published articles (screenshot in Figure 4), which can be filtered with a flexible search function. One limitation on the search function is that it only allows users to search for reports up to 5 years old. According to GPHIN, during the SARS outbreak in 2002, the GPHIN prototype was able to gather information about an unusual outbreak occurring in Guangdong Province, Mainland China, as early as 27 November, 2002 (Blench [2008]). However, because of the 5–year limitation, it is not possible to view this information.

4.3 ProMED

Description

Program for Monitoring Emerging Diseases (ProMED) is an early warning and disease–reporting system. ProMED (also known as ProMED–Mail) monitors emerging animal, plant and human diseases, and acute toxin exposures that are relevant to human health. The system gathers open–source information from subscribers, media sources, and manual searches conducted by ProMED staff. This information is then analysed by moderators and outside experts and then divided into seven reports, which are categorised into three levels of urgency. The reports are then published as e–mails sent to users and as posts on the ProMED website. All reports are also archived in a searchable database that dates back to 1994. The service is free to users.

Data Collection, Analysis, Storage, and Retrieval

Dozens of reports are submitted to ProMED daily. These reports come from a range of sources including government health departments, international organisations, subscribers' professional or personal observations, the media, and manual searches conducted by by ProMED staff.

Reports are initially examined by a 'top moderator' who decides whether to reject them or send them onto subject moderators. There are 12 subject moderators: 4 for veterinary and zoonotic diseases, 2 for viral diseases, and 1 each for bacterial diseases, parasitic diseases, plant diseases, epidemiology and surveillance, and medical entomology. Subject moderators check the accuracy of the reports, edit them for clarity and references, and frequently add brief commentary to highlight the importance of new information in the reports. Each report then goes back to the top moderator who audits the edited report and assigns the report one of three levels of urgency: green, yellow, and red—in order from least to most urgent. Green reports are sent to a copy editor for formatting and editing for grammar and consistency. The copy editor may also flag any questions concerning the report before it is sent back to the top moderator. Green reports are typically published within 24 hours. Yellow reports undergo expedited review and red reports circumvent sections of this review process and are published immediately (Cowen *et al.* [2006], pp. 1091–1092). Typically, 7 reports are published each day: 1 red, 1 yellow, and 5 green (Madoff [2004]).

Reports are published as e–mails sent out to users and as posts on the ProMED website. Users can choose to receive e–mails from specialised lists and search for them in the ProMED archive. This archive feature, housed on the ProMED website (developed and maintained by Oracle), allows users to search approximately 20000 reports using text, dates, and geographical locations (Madoff [2004], p. 230). Figure 6 depicts the collection, analysis and retrieval process. Users can also get ProMED reports through an RSS feed and they can follow ProMED on Twitter and Facebook.

ProMED itself has no visualisation system for its data. EpiSPIDER was initially designed to supplement ProMED with an interactive map that allows users to interact with ProMED data (Keller *et al.* [2009]). However, EpiSPIDER appears to cover only human and zoonotic diseases,

and so must ignore ProMED reports on plant diseases and toxin exposures.

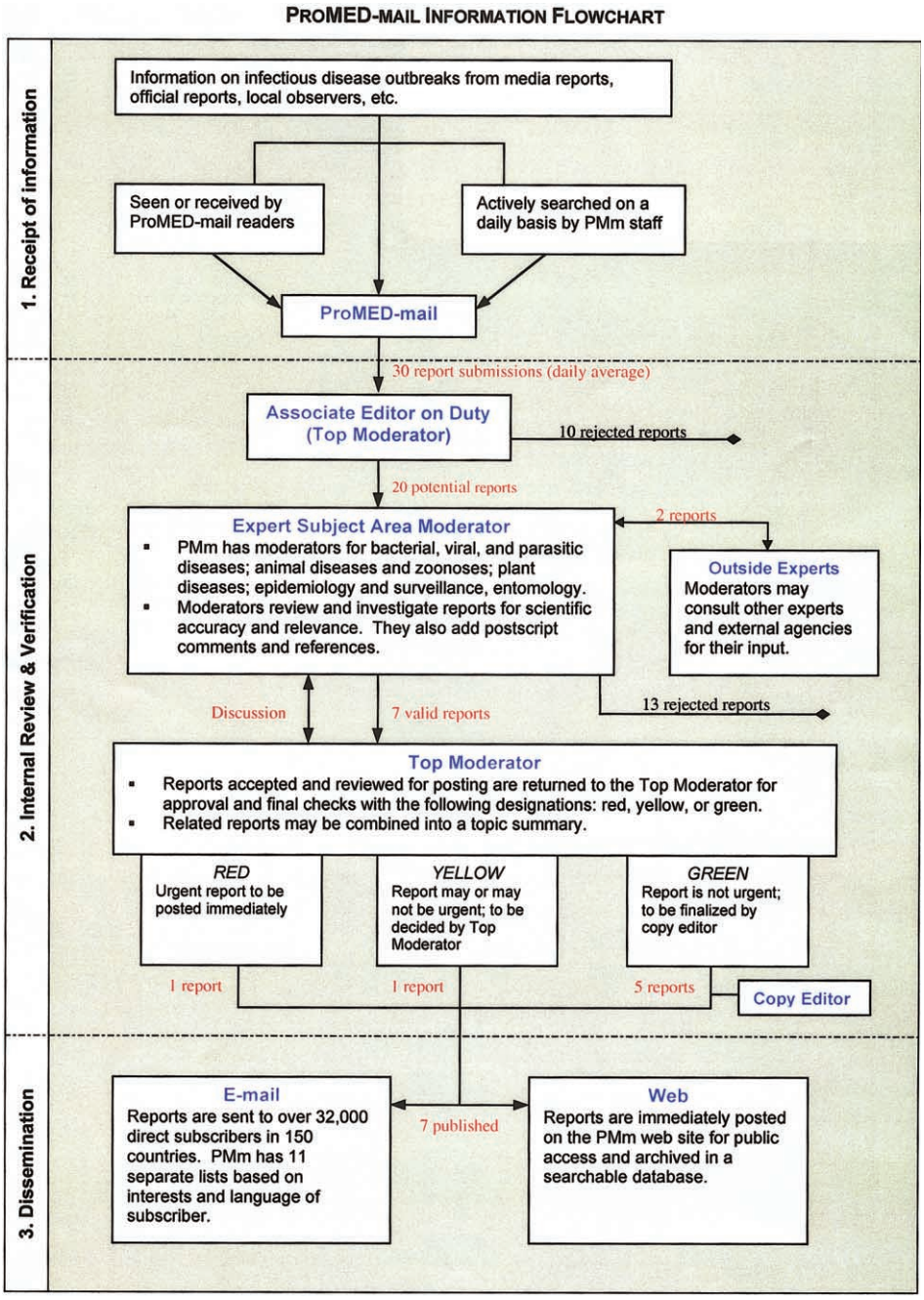


Figure 6: ProMED information flowchart (Madoff [2004]).

4.4 HealthMap

Description

HealthMap is an automated system that monitors global animal, human and zoonotic disease outbreaks. The service automatically collects data from a variety of sources and represents these data in the form of reports plotted on an interactive map using the Google Maps API. HealthMap has been in operation since September 2006, and has received funding from the Google.org project. The service is free to users.

Data Collection, Analysis, Storage, and Retrieval

As of February 2009, HealthMap automatically collects data from 17 sources—including Google News, Moreover, ProMED, WHO, EuroSurveillance, WDIN, HealthMap users, OIE, and GeoSentinel. Using these sources, HealthMap collects information on infectious diseases from more than 20 000 websites, every hour, 24 hours a day. The system collects an average of 300 reports per day. Sources of HealthMap's alerts are 92.8% news media, 6.5% ProMED reports, and 0.7% multinational agencies. (Keller *et al.* [2009]). HealthMap currently monitors outbreak news in Arabic, Chinese, English, French, Portuguese, Spanish, and Russian. The system uses Google Translate to translate original articles in other languages into English. It is expected that the service will expand coverage to outbreak news in Hindi in the near future.

The system uses Google Maps to represent the locations of events with a coloured marker. The Heat Index (a colour scale from yellow to red) represents a composite score of each event based on the recency of alerts, the number of disease outbreaks, and the number of sources providing information at a particular location. The algorithm used to determine the Heat Index of any given event applies an exponential weighting, yielding increased heat for more recent outbreak news.

Users are able to retrieve data from HealthMap by interacting with the map, its markers, and the marker filters. There are three types of marker filters: feeds, category, and diseases. An example of HealthMap's visualisation system can be seen in Figure 7. One notable restriction on the map interface is that it only allows users to view data up to one month old.

HealthMap also has a Twitter feed and a blog, where new entries are added each week, highlighting a disease outbreak or current headline–making alert. There is also an e–mail alert system that users can sign up to at info@healthmap.org. HealthMap also has applications for the iPhone App and any phone running the Google Android OS that uses the GPS of such phones to alert users in real–time of outbreaks near them. Users can also report outbreaks to the system using the iPhone and Android applications. Screenshots are included in Figure 8.

One significant difference between HealthMap and the other systems reviewed in this report (with the exception of ProMED) is that HealthMap allows users to contribute to the system. Users can report articles, rate them, add comments, and submit eye–witness reports using the web interface or from a smartphone.

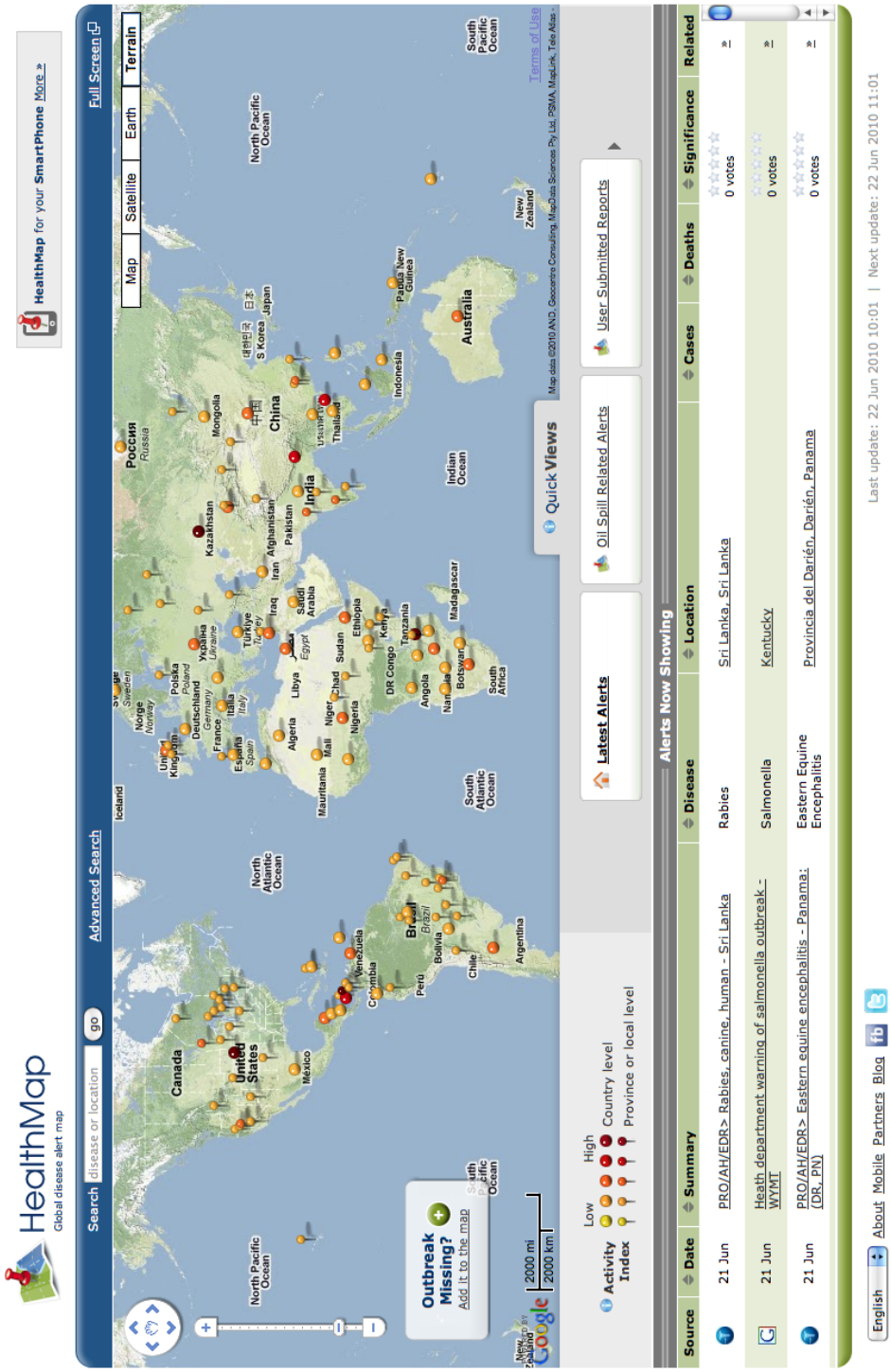
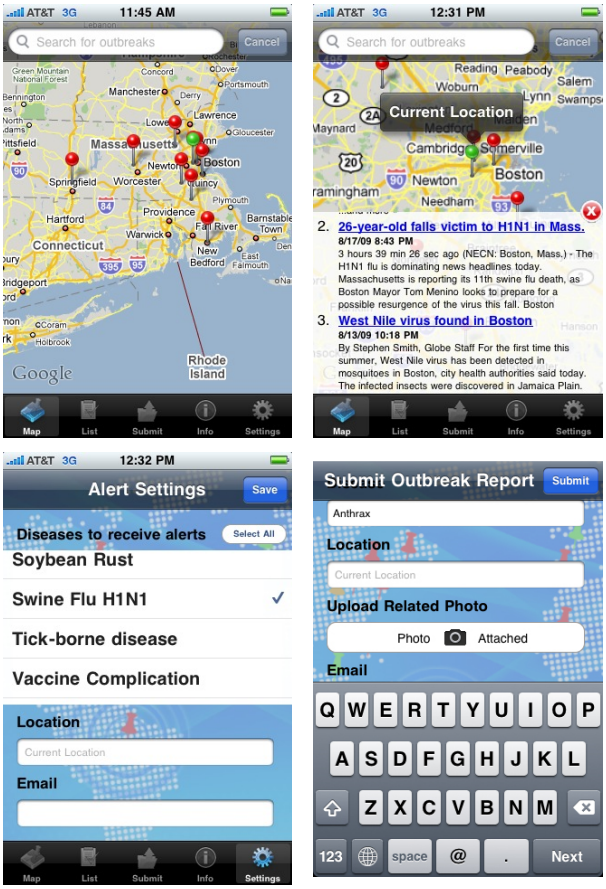


Figure 7: Snapshot of HealthMap's map interface taken at 10:06am, 22/06/10.

iPhone:



Android:

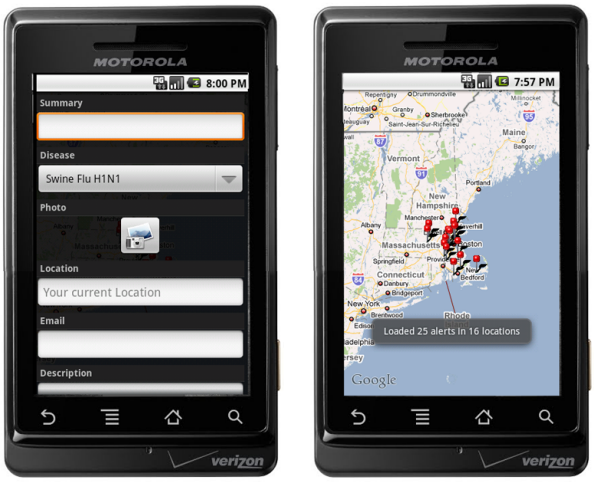


Figure 8: Snapshots of HealthMap's iPhone and Android Apps.

4.5 EpiSPIDER

Description

Semantic Processing and Integration of Distributed Electronic Resources for Epidemics and Disasters (EpiSPIDER) is an automated system that integrates online sources of information on emerging animal, human and zoonotic diseases (e.g., ProMED) with online sources of information on natural disasters (e.g., GDACS).[3] The project was designed in January 2006 to serve as a visualisation supplement to ProMED (Keller *et al.* [2009]), but now goes well beyond this initial goal. The service is free to users.

Data Collection, Analysis, Storage, and Retrieval

EpiSPIDER collects information from ProMED and GDACS. It also collects information from the RSS feeds of a variety of news sources—such as WHO, the European Surveillance Network, and Reuters. In addition to these data streams, EpiSPIDER also collects demographic and public health information (e.g., population sizes, per capita GDPs, public health expenditures, and physicians–to–population ratios) from the CIA Factbook and UNDP Human Development Reports.

EpiSPIDER then automatically processes this information in several ways. First, it extracts each report's date of publication, location information and topics. The system uses the location information to georeference the report using the georeferencing services of Geonames, Google Maps, and Yahoo Maps. A report's location information is also used to link the report to relevant demographic and public health information. EpiSPIDER feeds each report's topics into askMEDLINE to link the report to relevant scientific literature.

EpiSPIDER also converts the raw–text reports from ProMED and GDACS into semantic web format using natural language processing, some of which is outsourced to OpenCalais and the Unified Medical Language System. This allows EpiSPIDER to combine ProMED and GDAC information into one unified data stream. This means that users can be aware of disease outbreaks and natural disasters that are related to each other. This is an advantage over other forecasting systems as natural disasters can dramatically affect the spread and impact of infectious diseases. The system also collects articles from online news media through several sources, including Google News, Moreover, DayLife, and ReliefWebUpdates. EpiSPIDER also watches several Twitter feeds. Reports detected through Twitter account for over 50% of all the reports detected by EpiSPIDER.

The conversion of raw data into semantic web format allows EpiSPIDER to pass data onto other services. For example, the EpiSPIDER KML module was developed to enable the US Directorate for National Intelligence to distribute avian influenza event–based reports in Google Earth KML format to consumers worldwide (Keller *et al.* [2009]). Users can receive EpiSPIDER's output in a number of other formats, including RSS feeds, interactive maps, CSV files, filtered graphs, a Twitter feed, and blog. Users can plot activity of various diseases over a one–year period using

[3] GDACS is an organisation that provides near real–time alerts about natural disasters.

the filtered graph system, which is refreshed every 30 seconds. Figure 9 includes a depiction of the EpiSPIDER workflow and a screenshot of the graph system. Figure 10 includes a screenshot of the map exhibit.

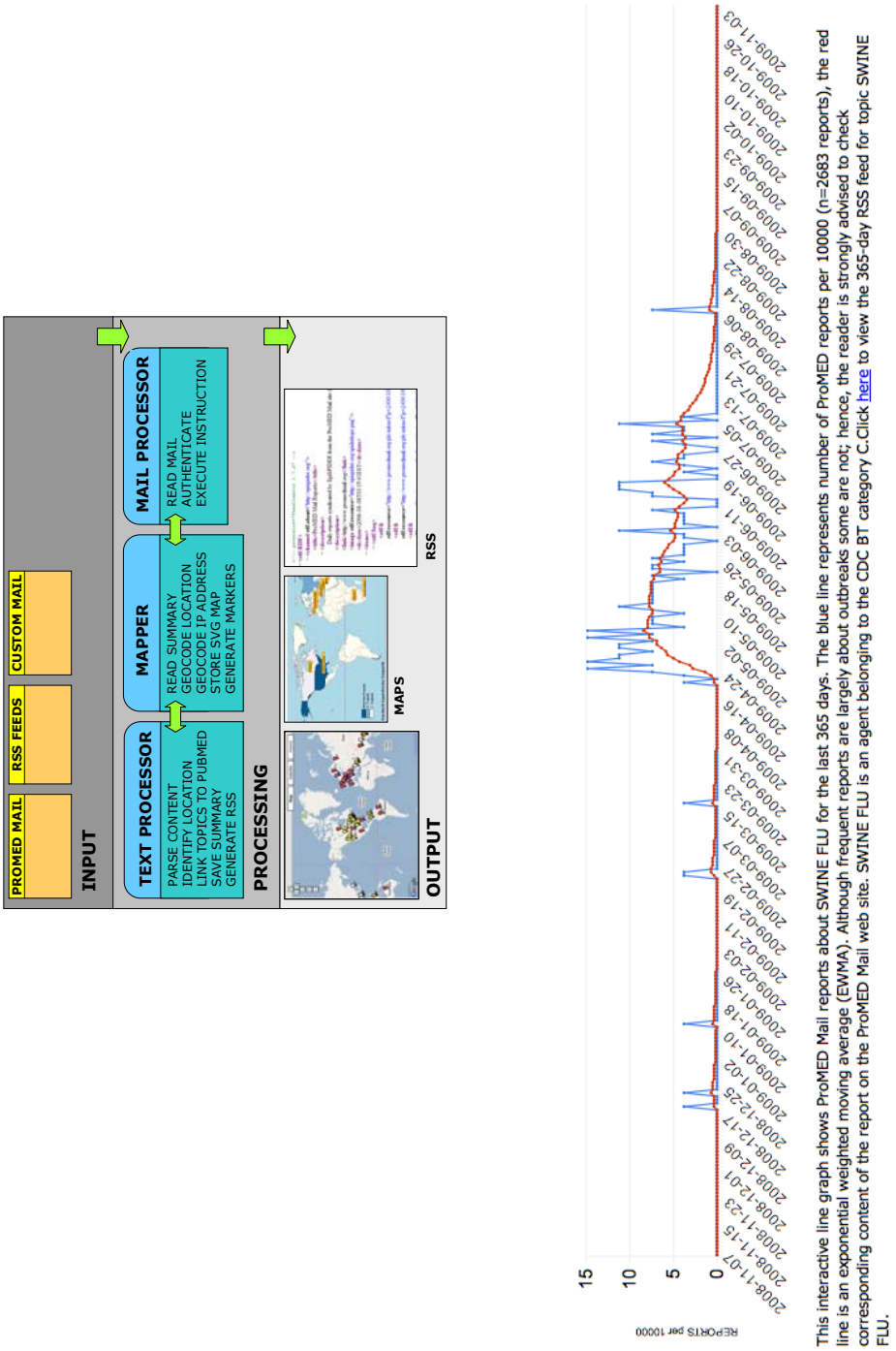


Figure 9: EpiSPIDER Workflow (Tolentino et al. [2007]) and EpiSPIDER Graph Screenshot.

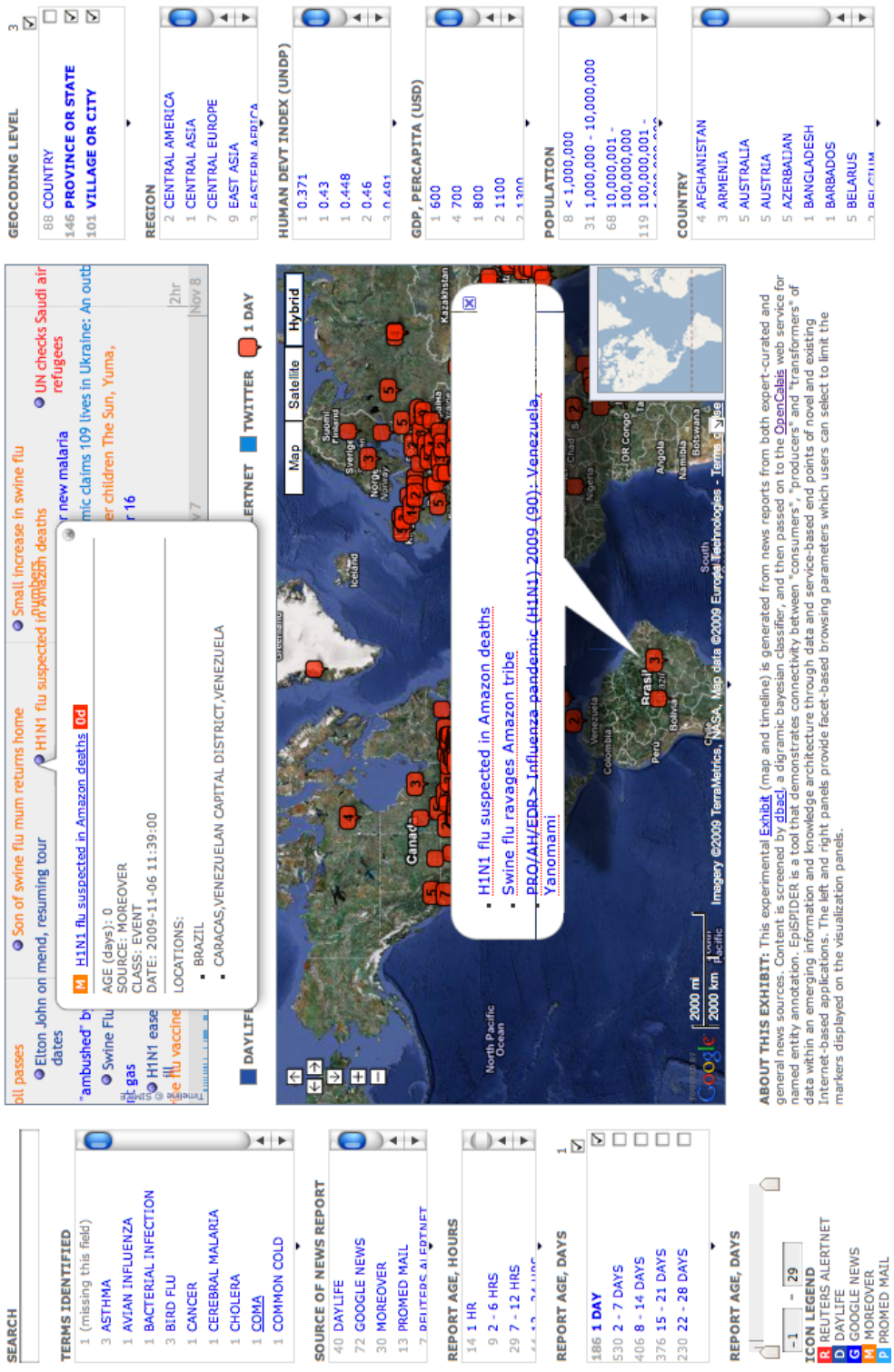


Figure 10: EpiSPIDER Map Exhibit Screenshot.

4.6 WDIN

Description

The Wildlife Disease Information Node (WDIN) is a monitoring and reporting system for information on wildlife diseases. The system provides a website portal for obtaining current information about wildlife health and wildlife–human–domestic animal disease interactions. Users include state and federal resource managers, animal disease specialists, veterinary diagnostic laboratories, physicians, public health workers, educators, and the general public. WDIN's major partners are the USGS National Wildlife Health Centre and the Nelson Institute for Environmental Studies at the University of Wisconsin–Madison.

Data Collection, Analysis, Storage, and Retrieval

WDIN automatically scans thousands of articles each day from a diverse range of online media sources. Articles are collected using a Yahoo Pipes feed (which was developed by WDIN). This feed has been published and can be viewed (both the output and the source) here. Inputs to the feed come in two types: (i) RSS feeds from over a broad range of general online news sources (e.g., New York Times, Science, Daily, Discovery, BBC, Associated Press); and (ii) wildlife disease–focused news sources, which come in the form of predefined searches of Google News and Google Videos (e.g. 'wildlife+surveillance'), and RSS feeds from online sources that specialise in wildlife news (e.g., ProMED, The Week In Wildlife at the Guardian).

The pipe feed then applies some automatic filtering to the articles that come from general online news sources. Articles whose description includes a biosecurity related key–term are allowed through the pipe. Articles from both streams are filtered for duplications and are sorted chronologically by publication date.

This automated process typically produces 200 to 300 articles. These are then scanned manually by WDIN staff. Usually 15 to 20 articles from this set are specific to wildlife health issues (as determined by the WDIN News Selection Policy) and are published in the Wildlife Disease News Digest. There are also typically 1 to 5 articles that discuss the detection or spread of diseases in a geographical area. These articles are manually indexed and geocoded by WDIN staff and then published to the WDIN News Digest GeoRSS and WDIN News Digest KML feeds. Users can also access these feeds through the Global Wildlife Disease News Map 2.0 (Figure 11 includes a screenshot of the mapping service).

No analysis is conducted beyond the automatic and manual relevance screenings. There is minimal verification of each article's accuracy. All information is in English, and there are no translation services. Aside from English abstracts of articles in other languages, no content is from sources in foreign languages. As such, the results of the service focus strongly on US events and sources.[4]

[4]There is an Australian counterpart to WDIN, the Australian Wildlife Health Network (AWHN), however it doesn't

4.7 BioCaster

Description

BioCaster is an automated early warning system aimed at providing relevant online news and research literature to public health workers, clinicians, and researchers interested in communicable diseases. The system, which began in 2006, covers human, animal and plant diseases. It is run at the Japanese National Institute of Informatics (NII) in Tokyo, by Nigel Collier with the help of Mike Conway (University of Pittsburgh) and Son Doanand (NII). The system is free to all users.

Data Collection, Analysis, Storage, and Retrieval

BioCaster gathers articles from a wide range of general news feeds and trusted sources such as Google News, ProMED, European Media Monitor Alerts, WHO, and Meltwater. BioCaster obtains this this information through more than 1700 RSS feeds each day (Collier *et al.* [2008]).

BioCaster's main distinguishing feature is its heavy use of text–mining and natural language processing technology. Part of the analysis process involves BioCaster's rich, multilingual ontology, which is open–source and downloadable from the home page. This rich structure allows the system to generate precise knowledge from unstructured articles written in many different natural languages. For example, the entry for 'avian influenza' in the ontology has 28 synonyms in English, French, Japanese, Korean, Chinese, Spanish, Thai, and Vietnamese, 8 causal agents, and 6 symptoms.

Users can interact with the results of the system in a number of ways. Upon first accessing the BioCaster site, the user is presented with a latest 'global round–up', which gives the headlines for the system's top stories and headlines for specific regions (Africa, Americas, Asia, Europe, Middle East, and Oceania).

The Global Health Monitor provides a geographical mapping of the news stories detected by the system over the last 30 days. The interface of BioCaster's map, which is based on the Google Maps API, is similar to the interfaces of the maps of other systems reviewed in this report. Users can choose date ranges of 30 days, the current day, the current week, and one, two or three weeks ago. Users can also filter the map for news genres, syndromes and diseases (Figure 13 includes a snapshot of the map interface). The most recent twelve months of news volume frequency related to infectious diseases tracked by BioCaster can also be viewed at the trends page. BioCaster also publishes its information in the form of a KML feed.

provide an analogous intelligence system based on open–source information.

Figure 11: Snapshot of Global Wildlife Disease News Map 2, taken 9 Jan, 4:00 pm (ET).

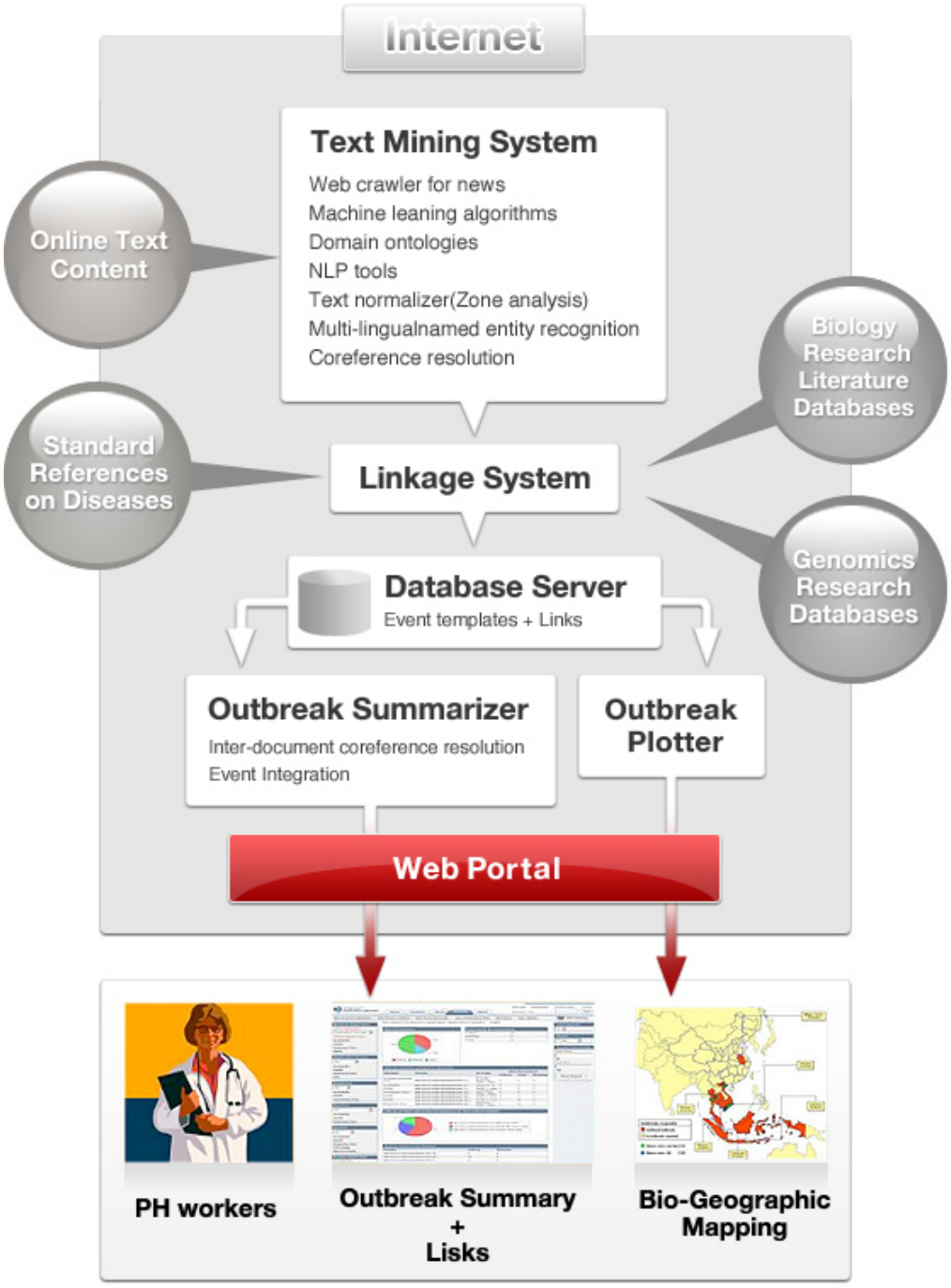


Figure 12: BioCaster Work Flow (http://born.nii.ac.jp/?page=about).

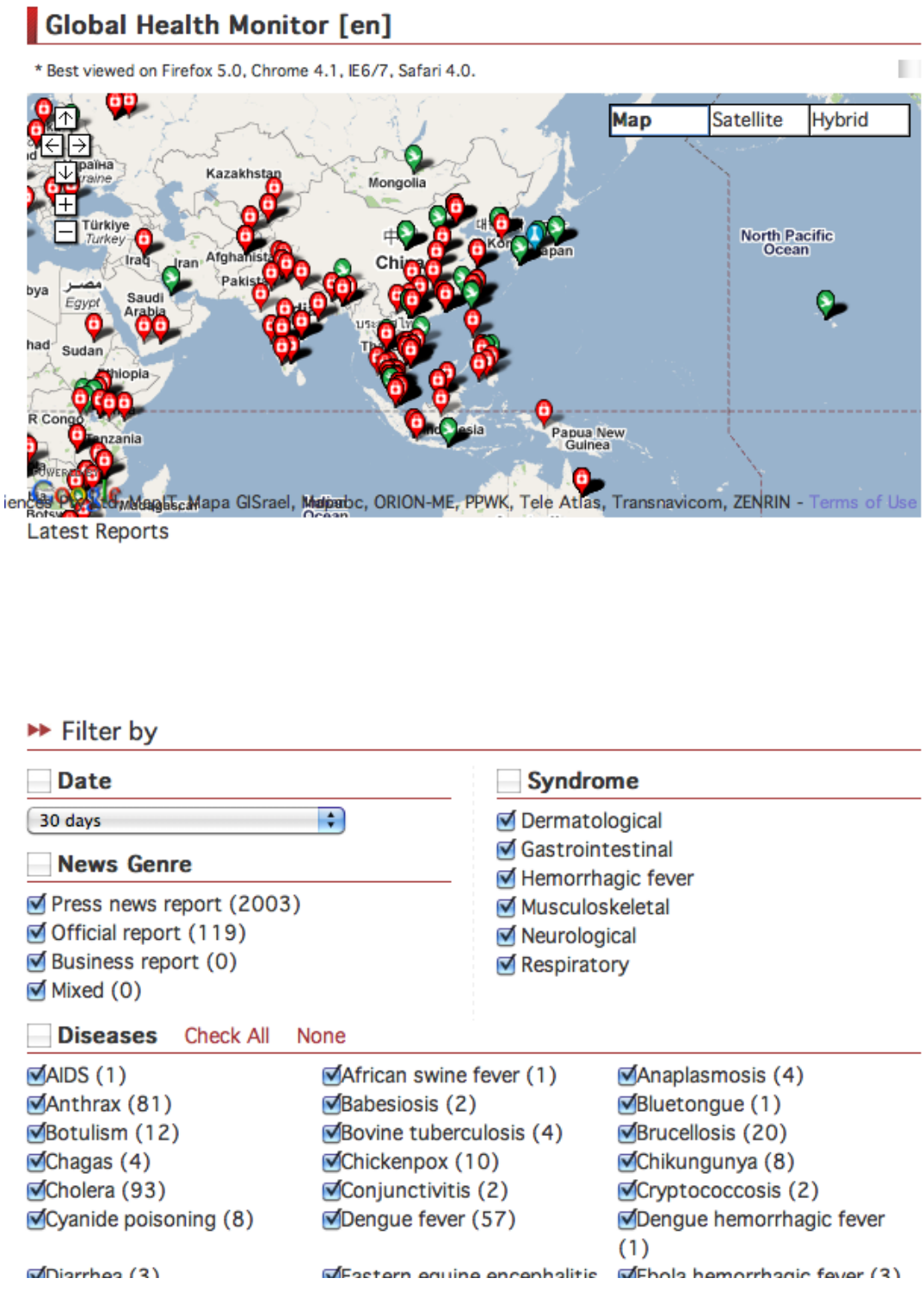


Figure 13: Snapshot of BioCaster Map, taken 14 May, 2010 6:28pm (ET).

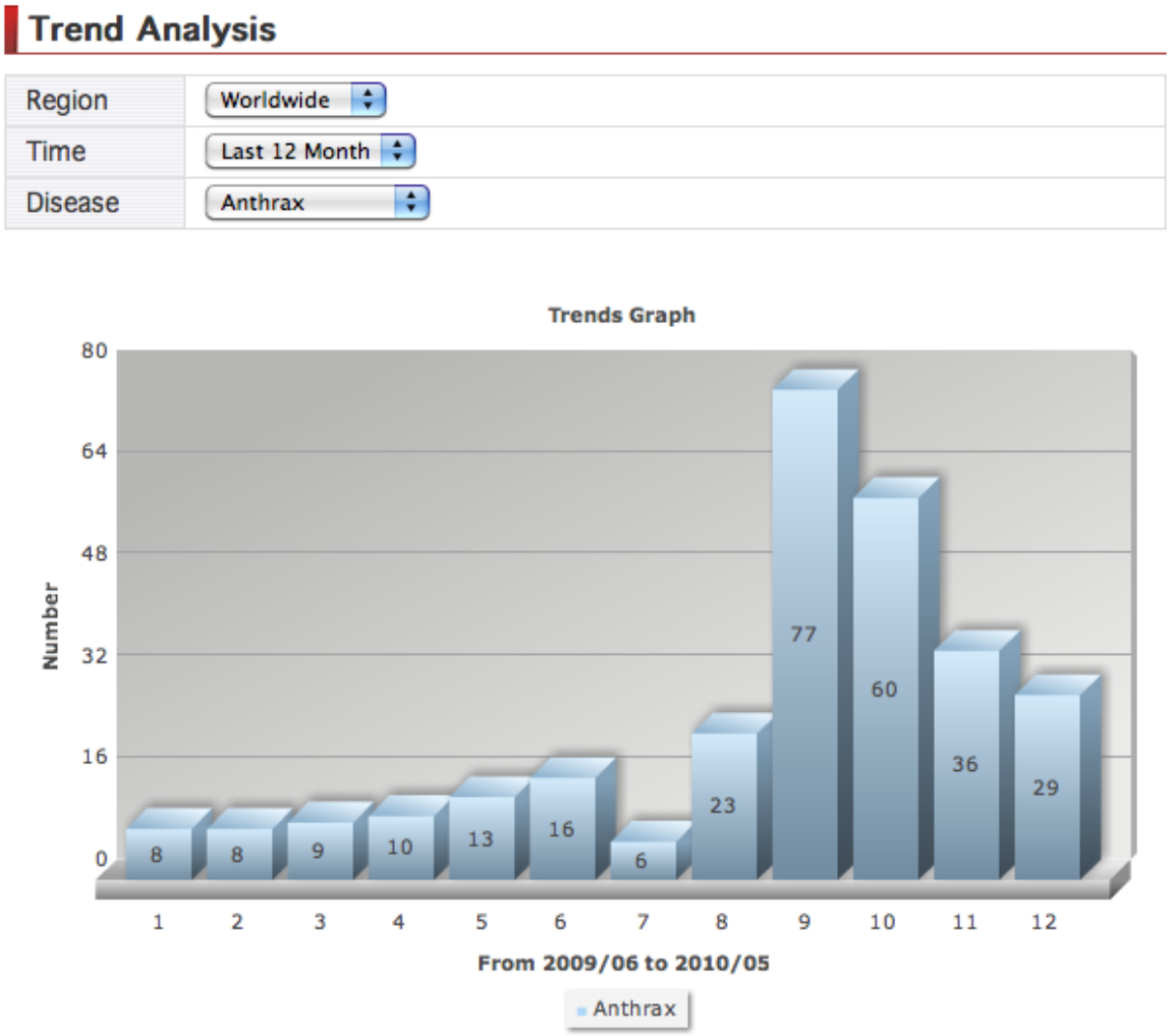


Figure 14: Trend for 'Anthrax' over the past 12 months.

4.8 NAPIS

Description

The National Agricultural Pest Information System (NAPIS) is a forecasting system that stores and manages pest survey data. NAPIS covers a wide variety of pests, but its scope is restricted geographically to the US. The service is free for all users.

Data Collection, Analysis, Storage, and Retrieval

Data are collected through Cooperative Agricultural Pest Surveys and other Plant Protection and Quarantine surveys. Given a particular pest, a county is given one of 7 statuses: blank if no survey was conducted, green if the pest was not found in the county, yellow if the pest is found, light brown if the pest is being eradicated, dark brown if the pest is eradicated, light purple if the pest is established by survey, and dark purple if the pest is established by consensus. A map for that pest is then generated with each county coloured according to its status. Maps can be filtered to present survey information from the past three years, all time, and for each year. Figure 15 includes an example map for the 2008 survey data for Australasian Soybean Rust.

Users can view pest maps by clicking on the Pest List, and then the map link for the pest of interest. In addition to viewing data through maps, users can get pest news or subscribe to an XML newsfeed.

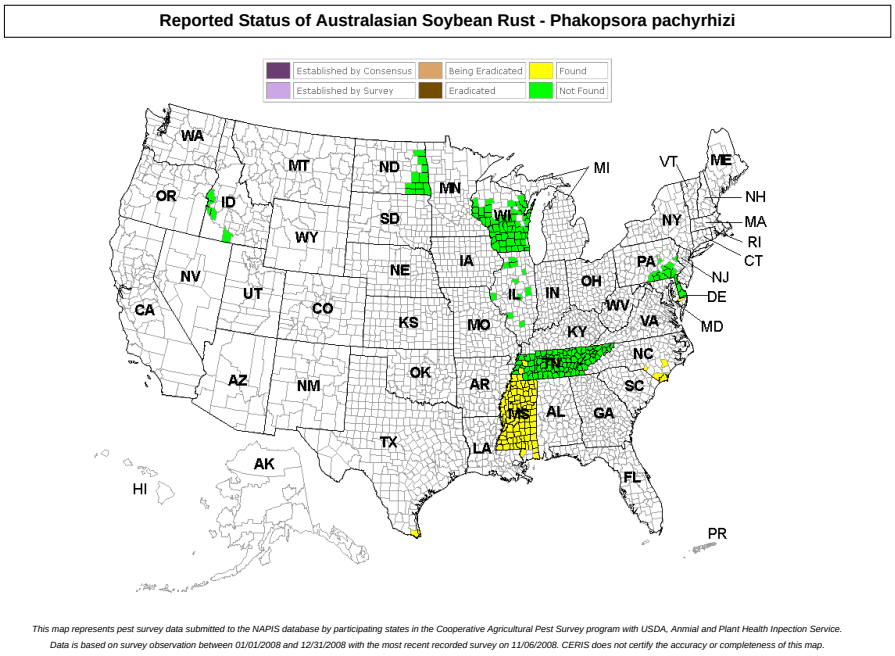


Figure 15: NAPIS 2008 map for Australasian Soybean Rust.

4.9 EUROPHYT

Description

EUROPHYT is a network of plant health information systems restricted geographically to Europe. It is a web–based network that is designed to support the protection of human, animal and plant life and health by the European Commission and Member States.

EUROPHYT has two components—EUROPHYT–PHY and EUROPHYT–CIRCA. The former is a database for biosecurity events. The latter is a database for technical and biological information, plant health legislation, and guide books for plant health inspectors.

Data Collection, Analysis, Storage, and Retrieval

EUROPHYT–PHY is a database that manages notifications of interceptions of plants or plant products that do not comply with EU legislation. Users can enter, edit and consult notifications using the interactive interface or the message exchange facility. Notifications are e–mailed to all Member States as soon as they enter the system. Due to current lack of access, the GUI and analysis process cannot be reviewed in this report.

4.10 GAINS

Description

The Global Avian Influenza Network for Surveillance (GAINS is a global network for the surveillance of avian influenza in wild birds. It is an early warning system for the spread of highly pathogenic avian influenza, which threatens poultry and human health, and biodiversity.

Data Collection, Analysis, Storage, and Retrieval

Data are collected through biological samples of wild birds (which are caught and released). These samples are collected by GAINS collaborators, who include US–based and international institutions, foreign governments, NGOs, universities and businesses. Users can interact with data using the GAINS WISDOM Map Explorer (based on the Bing Maps API). Users can also subscribe to various KML feeds. The Map Explorer has highly flexible filters. Users can view data for any date range. A screenshot of the Map Explorer is included in Figure 16.

Figure 16: GAINS WISDOM Map Explorer Screenshot.

4.11 NAPPO

Description

The National American Plant Protection Organisation (NAPPO) is a forum for public and private sectors in Canada, the US, and Mexico to collaborate in the protection of agricultural, forestry, and other plant resources against regulated plant pests, while facilitating trade. It sets phytosanitary standards that are recognised by the North American Free Trade Agreement. The organisation has been in operation since 1976.

Data Collection, Analysis, Storage, and Retrieval

NAPPO provides a Phytosanitary Alert–System (PAS), which is freely available to everyone. This alert system provides two types of alerts, official pest reports, and unofficial alerts. The unofficial alerts are news articles from various public sources, which are typically journal articles. Unofficial alerts are published roughly once a month. The official pest reports are relatively more frequent—about 5 reports are published each month. These are confirmed reports and meet the organisation's standard for pest reporting. Both sorts of alerts can be received by e–mail subscriptions. Apart from that, users need to otherwise visit NAPPO's site to receive the alerts. There are no mapping services or RSS or KML feeds available. This point is discussed further in Section 6.

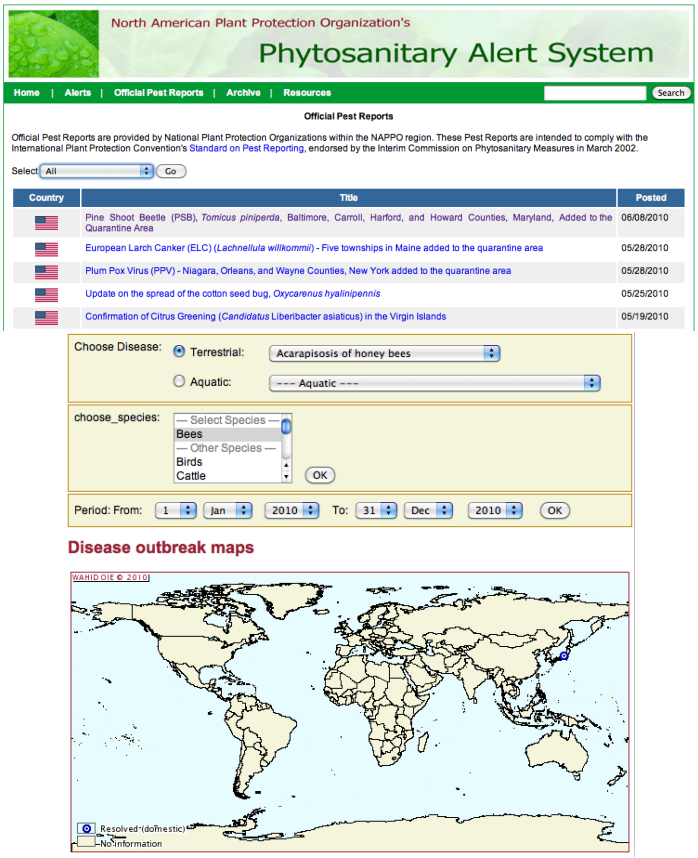


Figure 17: NAPPO Phytosanitary Alert–System Screenshot — Official Pest Reports (top); WAHID mapping interface (below).

4.12 OIE

Description

The World Organisation for Animal Health (OIE) is an intergovernmental reference organisation for improving animal health worldwide. It was created in 1924, and originally named the Office International des Epizooties. OIE is significant in that it covers pests and diseases of aquatic animals, along with zoonotic diseases, and diseases of other animals.

Data Collection, Analysis, Storage, and Retrieval

Data are collected from official reports submitted by Member Countries to the World Animal Health Information Database (WAHID) on outbreaks of specified pests and diseases. The system allows users to view disease outbreaks and distribution maps using WAHID's mapping interface. This is done by using a series of dropdown menus to select disease, and date ranges—a screenshot is included in Figure 17.

5 Retrospective Study

5.1 Introduction

Part of Stage 2 of this project evaluated the software systems against their utility for gathering, filtering, analysing, storing, retrieving and sharing biosecurity intelligence. There has been some previous evaluation of the systems reviewed in this report. A 2009 article compares the systems GPHIN, HealthMap and EpiSPIDER, and reports on some aspects of these systems (Keller *et al.* [2009]). Some of these results are from Heymann *et al.* [2001]. In that article, the following facts are reported:

- From July 1998 to August 2001, GPHIN was the first to detect 56% of 578 reports, which were subsequently verified by the WHO. These reports were for outbreaks in 132 countries (Heymann *et al.* [2001]).
- In 1998, GPHIN was the first to provide preliminary information about a new strain of influenza in northern China.
- WHO declared the 2003 SARS outbreak in March 2003, while GPHIN detected an unusual respiratory illness outbreak in Guangdong Province in November 2002.
- Through the duration of the 2003 SARS outbreak, GPHIN's information was approximately 2 to 3 days ahead of official WHO reports.
- HealthMap processes an average of 133.5 disease alerts each day. About 50% are categorised as breaking news.
- Over a period of 30 days, the system displayed more than 800 breaking news alerts for any given day.
- From October 2006 to 20 November 20, 2007, HealthMap processed more than 35749 alerts across 171 disease categories and 202 countries or semi–autonomous or overseas territories.
- Sources of HealthMap's alerts are 92.8% news media, 6.5% ProMED reports, and 0.7% multi-national agencies.

Heymann *et al.* [2001] reported that:

- From July 1998 to August 2001, ProMED was the first to detect (approximately) 8% of 578 reports, which were subsequently verified by the WHO. Percentages for other organisations are in Figure **??**.

These results show the clear utility and potential of open–source biosecurity intelligence systems. However, they raise many questions, and don't include several of the systems explored in this project.

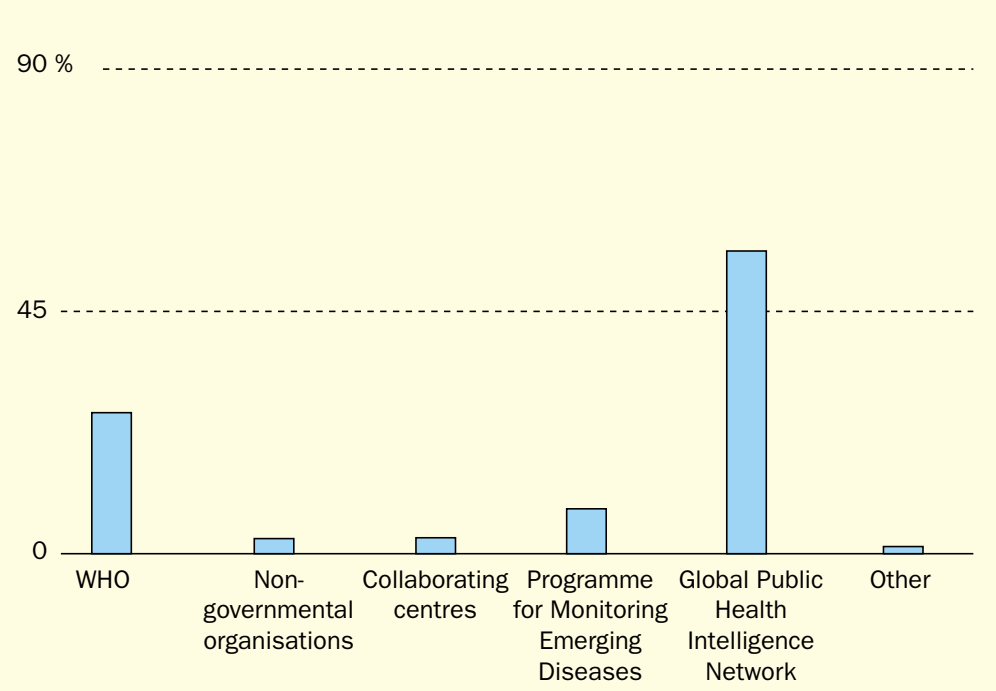


Figure 18: Percentages of reports first detected by organisations and later verified by WHO. Image taken from Heymann *et al.* [2001] (p. 349).

To evaluate the software systems in this project, it was initially proposed that a retrospective study of the systems should examine what each of the systems detected and reported about select set of past biosecurity events. It was intended to use this information to assess the degree of correspondence between what is currently known about those events and what the systems reported. This would then provide some measure of the biosecurity intelligence utility of each of the systems.

It was thought that the select set of past biosecurity events should concern pests and diseases that are representative of those that are of interest to DAFF, while ensuring the inclusion of pests and diseases for which there isn't a plethora of information available to the biosecurity intelligence systems reviewed in this project. The following pests and diseases were identified as potentially good candidates for the retrospective study:

- *Phytopthora ramorum* (plant disease)
- Guava rust (plant disease)
- Cassava mosaic virus (plant disease)
- Citrus canker (plant disease)
- SARS (zoonotic disease)
- Nipah virus (zoonotic disease)
- *Taenia solium* (zoonotic disease)

- Chikungunya (human disease)
- Japanese encephalitis (zoonotic disease)
- Black–striped mussels (marine pest)
- *Hydroides sanctaecrucis* (marine pest)
- *Caulerpa taxifolia* (marine pest)
- Yellowhead virus (aquatic animal disease)

Only five of the systems reviewed in this report covers plant pests or diseases—EUROPHYT, GPHIN, NAPIS, NAPPO and ProMED. Only ProMED and GPHIN provide international coverage of plant pests and diseases, but this coverage is minimal as these systems mostly focus on animal, human and zoonotic diseases.[5] Access to EUROPHYT is restricted to members only. Of the plant diseases mentioned above, NAPIS covers only citrus canker and guava rust. NAPPO has covered *Phytopthora ramorum* and citrus canker—one report for each (19 January 2006 and 19 January 2005, respectively). For these reasons, no retrospective study was conducted for plant pests and diseases.

A similar difficulty exists for including marine/aquatic pests or diseases in the retrospective study: only OIE gives some coverage and ProMED has some minor coverage. This leaves only the animal and zoonotic diseases as potential candidates for retrospective study.

One difficulty with performing a retrospective study on these systems is that some of them have significant limitations regarding the age of the data that users can access. HealthMap only lets users interact with data that are at most 1 month old, although ACERA was able to obtain slightly more flexible access to HealthMap's Data through KML files (see below). EpiSPIDER graphs go back only one year, and the EpiSPIDER Map Exhibit only lets users interact with data that are at most 29 days old. The WDIN mapping service allows only two date filters: data 45 days old; and data 7 days old, although ACERA was able to obtain slightly more flexible access through a KML feed.) GPHIN only lets users search for reports in their archive that are up to 5 years old. BioCaster allows a number of date filters, but allow access to data up to 30 days old. These limitations make it difficult to assess the performance of these systems on detecting and reporting biosecurity events in the past. However, it is possible to compare ProMED and GPHIN over the previous 5 years with respect to the selected animal and zoonotic diseases.

It should also be noted that all of the intelligence systems are constantly being updated and are improving very rapidly. A study of how well the systems performed on events 5 years ago will likely not reflect how well those systems perform now or will perform in the future.

Another significant difference among the systems is relevant to conducting a retrospective study. Some of the systems—HealthMap, EpiSPIDER, WDIN, and BioCaster—do not *add* content; rather, they collect, organise and analyse (mostly automatically) information that is online or contributed by users. This typically results in a simple link that takes the user to the original report (e.g., a news media article). In contrast, ProMED and GPHIN can, and often do, add content to what they collect (manually, using various biosecurity experts), producing their own reports. In-

[5] BioCaster claims to cover plant diseases, but no evidence for this claim has been found.

deed, HealthMap, EpiSPIDER, and BioCaster often link to reports generated by ProMED. (GPHIN is currently a closed source system, and so its reports can't be shared publicly.) This suggests that it may be more useful to compare ProMED with GPHIN while separately studying HealthMap, EpiSPIDER, WDIN, and BioCaster.[6]

Another point of difference is that HealthMap, EpiSPIDER, WDIN, and BioCaster all have interactive mapping features, while ProMED and GPHIN do not. (GPHIN maps its reports, but not side–by–side, nor in a way in which users can apply filters.) The former systems publish their data in the form of KML files, so it is possible to compare them side–by–side in one interface, such as Google Earth. This means that it is possible, for example, to see what the four systems reported for Mexico City in a given month. Unfortunately, EpiSPIDER only publishes KML files for reports it gets from ProMED and WAHID (OIE), missing the vast majority of reports that it detects. However, EpiSPIDER now publishes CSV files for the reports that it detects over the most recent 7–day period from all its sources—and it does the same for its ProMED and WAHID reports, though for 14–day periods. By converting these CSV files into KLM files (KMLCSV Converter), it is possible to compare all of the systems with interactive mapping features in Google Earth. This means that it is possible to compare the systems over a large sample of geographical locations as a way to measure how often one system picks up a report that the others miss.

Therefore, a dual strategy was used for the retrospective study. ProMED and GPHIN can be compared over the last 5 years by studying their archived reports on past events involving the select zoonotic diseases. The other systems—HealthMap, EpiSPIDER, WDIN, and BioCaster—can be compared over a 7–day period (the limiting factor being EpiSPIDER), using two methods. The first method is to simply search each system's corresponding CSV file(s) for the terms 'SARS', 'Nipah', 'Taemia', 'Chikungunya', and 'Japanese Encephalitis'—along with any synonyms. The second method involves randomly sampling small longitude and latitude regions and examining what reports for those regions, if any, are detected by the systems.

This second method allows the systems to be compared in other ways. For each report detected in the sample, the original language of the report was recorded, thereby giving a measure of how well the systems detect reports in different languages. Reports in the sample were also classified into five types: human, animal, general, food, irrelevant/mistaken. This gives a measure for the different focuses of the systems, and a measure of their accuracy. Also, by reading the original article, it is possible to judge whether the report was accurately plotted (e.g., a report on influenza A H1N1 in New York, plotted in Sydney would not be accurate).

A difficulty with retrospectively studying ProMED and GPHIN is that there is an ambiguity between data *signals* and *interpretations*. If a system is collecting and disseminating information on an emerging disease outbreak, it doesn't follow that the system *knows* that it is doing this. For example, ProMED published initial reports of 'atypical pneumonia' cases in Guangzhou (20030210.0357) and Hanoi (20030311.0595) before it was realised that these were all cases of what was later known

[6]HealthMap does now allow users to add comments to articles that it reports. Users can also rate articles using a 5 star ranking.

as SARS. ProMED now recognises these initial reports as reports of SARS cases, even though at the time, it didn't. This is not a point of criticism of the system, but it creates a problem for conducting a retrospective study. If a system now reports that, say, on 1 January 2001 there was a spike in Nipah virus activity, does this mean that the system, on 1 January 2001, detected the spike. It may simply have detected an increase in equine fatalities that were only later recognised as being caused by the Nipah virus.

Initial results from the comparison of ProMED and GPHIN turned out to be largely negative. So a new list of pests and diseases and particular outbreak events of them, which were recommended by DAFF staff, were added to the study. These pest and disease events are:

- New strains of UG99 in South Africa, May 2010 (plant disease).
- *Drosophila Suzukii* in the US, 2009 (plant pest).
- Guava/Myrtle rust in New South Wales, Australia, 2010 (plant disease).
- African swine fever in the Caucus region, 2007 (animal disease).
- Bluetongue virus (BTV8) in Europe, 2006 (animal disease).
- Black–striped mussel, *Mytilopsis sallei*, Darwin, Australia, 1999 (marine pest).
- The Caribbean tubeworm, *Hydroides sanctaecrucis*, Cairns, 2001 (marine pest).
- Infectious myonecrosis, Brazil and Indonesia, mid June 2010 (marine disease).
- *Caulerpa taxifolia*, New South Wales, Australia, 2001 and South Australia 2002 (marine pest).

The outbreak events associated with *Mytilopsis sallei*, *Hydroides sanctaecrucis* and *Caulerpa taxifolia* occurred outside of the 5–year window, so they were rejected. Results concerning the remaining pest and disease events are discussed in Section 5.5.

Google Flu Trends does not cover any of the diseases that are candidates for the study, so it was not included in the study. Since EUROPHYT, NAPIS and NAPPO have geographically restricted coverage of plant pests and diseases, they are also excluded. All of WDIN's reports are now included in HealthMap's reports, and since the study covers HealthMap, WDIN was not included as an independent system. Only OIE covers pests and diseases of aquatic animals, however its reports are based on official confirmed outbreak reports and not on (potentially earlier) media reports, so it too was excluded from the study. As mentioned earlier, over the course of the project, a prototype system to detect reports on plant pests and diseases was developed using Yahoo Pipes. How this system was developed and tested is discussed in Section 6.

5.2 Statistical Testing of BioCaster, EpiSPIDER, and HealthMap

5.2.1 Goal

Three systems—BioCaster, EpiSPIDER and HealthMap—perform a number of different tasks. They collect information from different sources. They get that information using search terms in various languages. They filter that information for irrelevant information. They extract locations and focus on information relevant to particular locations. And so on. Knowing how well each system performs these tasks helps understand how they might be used to meet DAFF's biosecurity intelligence needs.

This section attempts to measure the following attributes of the systems:

- With what frequency do the systems detect information for different geographical regions? (Asia, Middle East, Europe, etc.)
- With what frequency do the systems detect information in languages other than English?
- How often do the systems detect the same information?
- How often does one system detect useful information that no other system detects?
- How well does each system extract locations from the information it gathers?
- How well does the system use those locations to plot reports?
- What kinds of information are the systems detecting? How often do they detect information on human, animal, zoonotic, and plant pests and diseases, or on marine pests?

Some of these questions can be answered by simply examining all of the reports of the systems. However, others require more in–depth analysis. For example, to find out whether a system has detected an article on an emerging human disease in Beijing, it is necessary to read that article and judge whether or not it is actually an article on an emerging human disease in Beijing. Sometimes the system will include a brief summary or classification of the article it has detected. However, those summaries or categorisations cannot be relied on because they are produced by an automated processes and are occasionally or more frequently incorrect.

On average, it takes about 12 minutes to study each report. Over the week that the systems were studied, there were about 6200 reports. It would therefore take about 51 days to study each report for that week alone. It is thus necessary to study a sample of those reports, and use that sample to make general conclusions about the systems. However, it is possible to answer some questions by looking at all of the reports. Comparing these answers with answers to the same questions determined by the sample will reflect how representative the sample is.

5.2.2 Method

KML and CSV files for the reports generated by each system over a week (6–13 May, 2010) were created. The following summarises how this was done for each system.

EpiSPIDER

EpiSPIDER publishes a CSV file for all of its reports over the most recent 7 days. The CSV file for the period 6–13 May was downloaded and a KML was generated from it using a KML–CSV converter (KMLCSV Converter).

BioCaster

BioCaster published a KML feed for all its reports (in 12 different languages) over the most recent 30 days. The KML file for the period from 23 April to 23 May was downloaded and a CSV file was generated from it. This CSV file was loaded into a spreadsheet program (Excel) and reports for dates outside 6–13 May were deleted. A new KML file was then generated from this edited CSV file.

HealthMap

HealthMap generates KML files that can have the following filters applied:

- Source types: any combination of official reports, news reports, community reports, and scientific reports.
- Languages: any combination of English, Spanish, French, Portuguese, Chinese, Russian, and Arabic.
- Start date: any date that is a maximum of 2 months ago.
- End date: any date after the start date (up to the present).

A KML file for the period 6–13 May, for reports in all languages and from all sources, was downloaded, and a CSV file was generated from it.

Sample Points

To help randomly sample reports plotted by each system, a CSV file with randomly generated latitude/longitude pairs was created (in Excel). To ensure various geographical regions were adequately sampled, the sample was stratified over 12 zones as shown in Figure 19. The zones roughly correspond to natural geographical regions: Oceania: Zone 1, Asia: Zone 2, Russia: Zone 3, Middle East: Zone 4, Africa: Zones 5 and 6, Europe: Zone 7, South America: Zone 8, Central America: Zone 9, US: Zones 10 and 12, Canada: Zone 11. Sample points were generated uniformly over latitudes, -180° to 180°, and longitudes, -90° to 90°. These were then re–scaled to avoid any clustering towards the North and South poles. The CSV file of sample points was also converted into a KML file and imported into Google Earth, alongside the KML files for the systems (Figure 20).

By analysing the CSV files for each system, various facts about the systems were obtained:

- How often various news sources were used by each system.
- How often each system reported for each zone.
- The total number of reports for each system.
- How many unique locations had reports for each system.
- In the case of HealthMap, how often articles in languages other than English were detected.

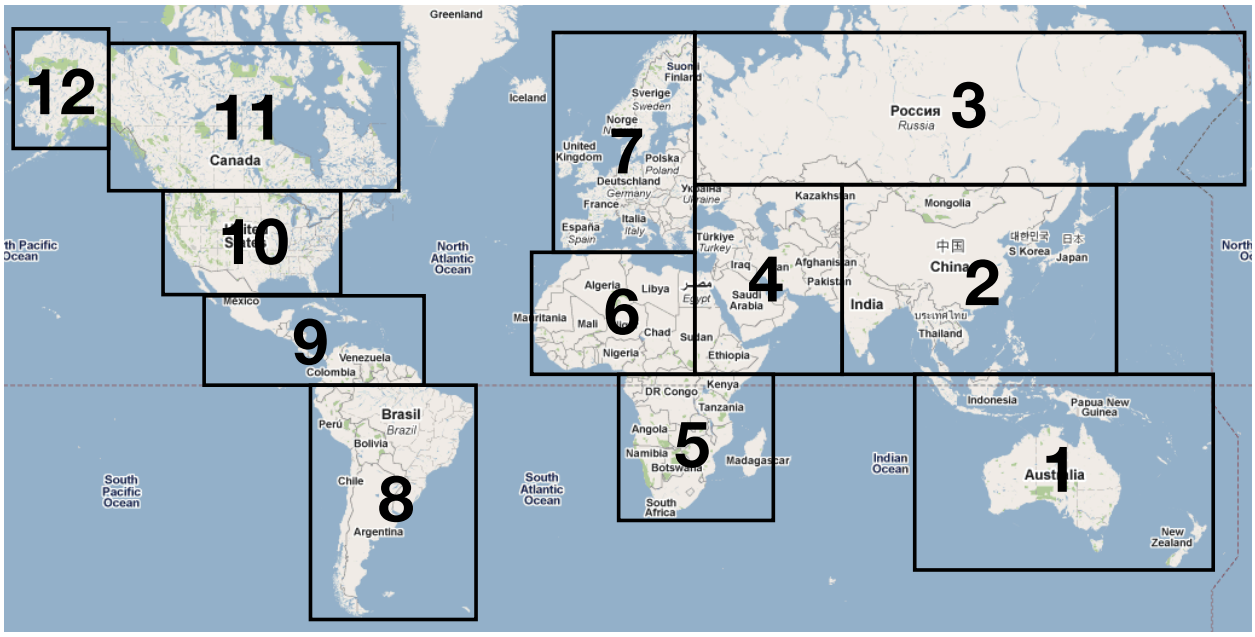


Figure 19: 12 Zones.

To answer questions from Section 5.2.1, it was necessary to investigate manually the articles reported by the systems. However, due to the large number of reports, it was necessary to study a sample. In addition to shedding light on the systems, these facts (reported in Section 5.2.3) were used to gauge how representative the sample (to be described below) is. For each report selected, the following was recorded:

- The latitude and longitude (from these, the zone can be determined).
- The system that generated the report.
- The date the system generated the report.
- The article link status in the report.
- The source the system used to obtain the article.
- The type of the event in the article.
- The date of the article.
- The original language of the article.
- How well the system translated the article into English.
- Which system was the first to report the event described in the article, with the same or better geographical precision and accuracy (only if the article was accurately plotted by the system).
- Which system was the first to report a signal for the event described in the article, with the same or better geographical precision and accuracy (again, only if the article was accurately plotted by the system).

- How well the location of the report matched a relevant location in the article.
- How many relevant locations were described in the article.
- How many relevant locations were plotted by the system, and to what accuracy and precision.
- How many irrelevant locations were plotted by the system.

Some of these attributes require further explanation.

First, sometimes a report's links are broken (e.g., the page is taken down or moved to another location). Or sometimes the only link is to some literature, which may or may not be relevant. If the link is not broken and is not merely a link to literature, then the source is recorded and the event type is assessed.

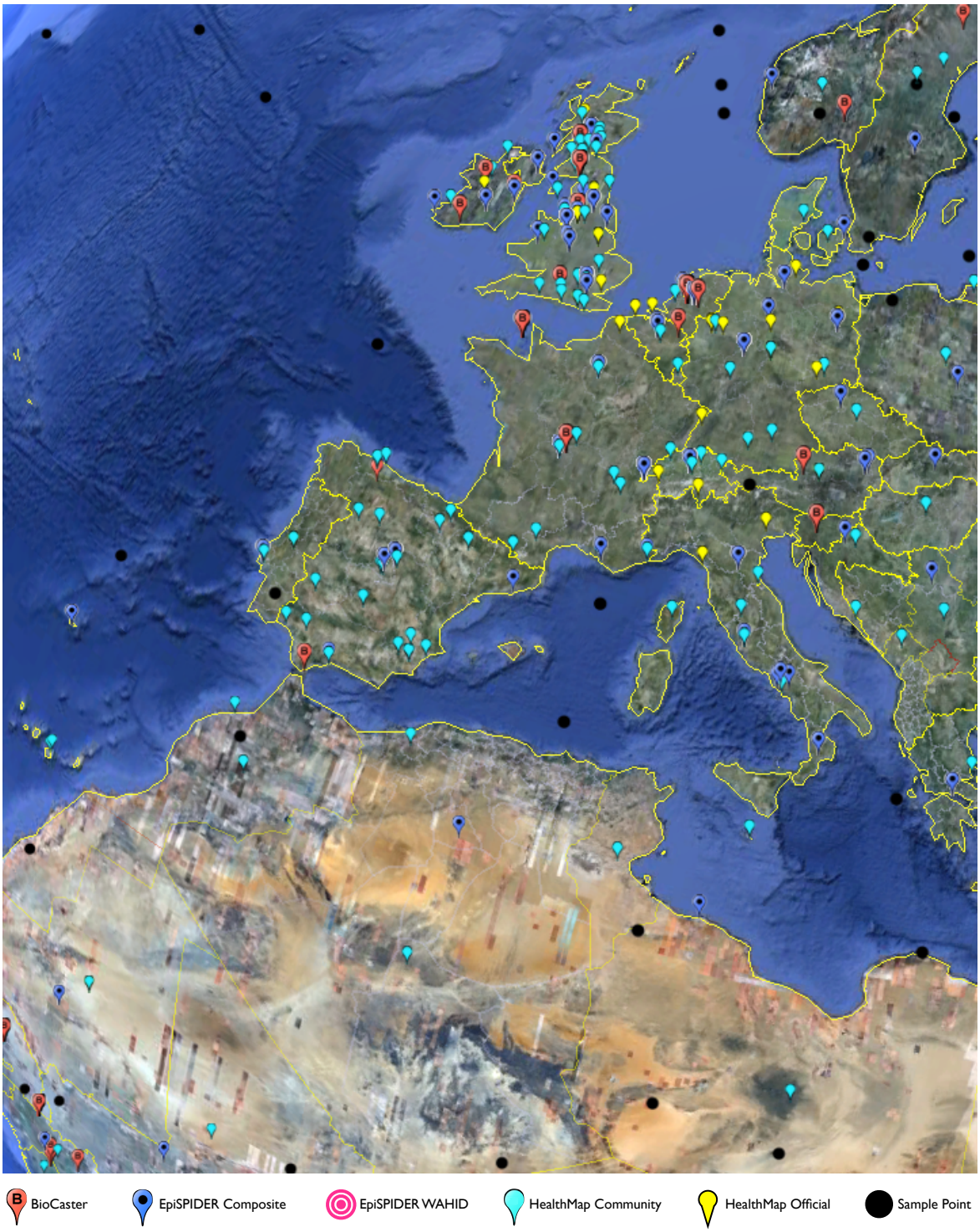


Figure 20: KMLs mapped in Google Earth.

Events were classified into the following types:

A	Particular outbreak of an animal disease.
Z	Particular outbreak of a zoonotic disease.
H	Particular outbreak of a human disease.
P	Particular outbreak of a plant pest or disease.
M	Particular outbreak of a marine or aquatic disease or pest.
F	Particular food security event.
Po	Political event.
ND	Particular natural disaster.
MD	Particular man–made disaster.
HN	Particular humanitarian event.
HH	Particular event concerning human health (but not an outbreak of a disease).
AH	Particular event concerning animal health (but not an outbreak of a disease).
PH	Particular event concerning plant health (but not an outbreak of a disease or pest).
V	News concerning production or distribution of vaccines.
I	Irrelevant news ((i.e., news that did not pertain to biosecurity in any way).
N	Not able to be determined for whatever reason.
O	Other.
G∗	Event types are prefaced with a 'G' if the article is general news. (Example: GZ would be the event type of an article on the WHO reminding people attending the 2010 World Cup to get vaccinated for the Rift Valley Fever virus.)

Classifying events into these event types necessarily requires interpretation of the articles describing the events, which introduces a degree of subjectivity. In some cases, there could be some dispute about how to classify an event described in an article. This in turn could result in different overall statistics. However, with a large enough sample size, and with interpretations only done by experts, such disputes would be relatively infrequent and would not have a significant effect on the overall statistics. Only if the event of the article is an A, Z, H, P, M, or F event are the other attributes recorded.

To determine which system was the first to report an article about a particular event, it is necessary to have some conception of an event. The general notion of an event is vague and difficult to analyse. For this particular purpose, focus can be restricted to events classified as A, Z, H, P, M, or F events. All events of these types have a *first case* associated with them–first case of *E.coli* poisoning, first case of H1N1, etc. If two articles mention the same first case, then they are (at least partially) about the same event. Similarly for subsequent cases. In some situations, two articles may disagree over what the case is. For example, Article 1 reported a case of dengue fever in Comoros in early July and Article 2 reported a case of chikungunya in Comoros in early July and also mentions that other reports initially mistook the disease to be dengue fever. In such cases, Article 1 was treated as a possible *signal* for the event, and Article 2 was recorded as the first to

be about the event. In some situations, an article may report a mysterious disease or undiagnosed illness, and a subsequent article may report what the disease or illness is. Again, we say that the first article was about a possible signal for the event, while the second was about the event.

The order of the articles does not matter for this aspect: if the event is an outbreak of chikungunya, the first article reported a case of chikungunya, and the second reported a case of dengue fever (or a mysterious fever), the second was treated as a possible signal for the event. In cases where these guidelines are not enough to determine whether a report is about an event detected by a system, a '?' was noted next to the system that detected the ambiguous article. This is not only to be careful. It could be the case that one of the systems regularly detects ambiguous articles before any other system detects any articles (about a given event).

There are many aspects to the accuracy of location. An article can mention more than one relevant location. So, for each article (on an A, Z, H, P, M, or F event), all of the relevant locations are recorded. Sometimes locations are mentioned, but are not relevant to the outbreak of an event—e.g., 'a sample from Delhi was examined in Washington D.C., U.S.A' (Delhi is relevant, Washington isn't). Such locations are ignored (purely automated systems often don't ignore them).

Locations come in varying degrees of precision—e.g., Europe; France; Burgundy, France; Beaune, Burgandy, France; etc. Only locations that are at least as precise as a country are recorded. This means that locations such as 'Asia', 'Northern Europe', 'Sub–Saharan Africa', etc. are ignored. Locations were assigned one of three categories of accuracy: 'C' if only the country was specified, 'CS' if the state or province was specified, 'CST' if the city, town, or village was specified. For a given article, the number of locations that fall into each category was recorded. The number of relevant locations that the system plots at each level of accuracy was also recorded.

Both sets of numbers were recorded as a sequence of ratios. If an article had 2 CST locations, 1 CS location and 3 C locations (and the system accurately plotted the article at both CST locations and the CS location, but only 2 of the C locations), this was represented by the sequence: 2/3–1/1–2/2. If an article plotted a relevant location accurately, but at the wrong level of precision, then this was recorded in square brackets. For the hypothetical article just mentioned, if the system accurately plotted one of the CST locations at the CS level, and the other CST location accurately at the C level, this was represented as: 2[0,1]/3–1[1]/1–0/2. If the CS location was also plotted accurately at the C level, this was represented as: 2[1,1]/3–0[1]/1–0/2.

The total number of irrelevant locations plotted by the system was counted and recorded; the levels of precision were ignored. Finally, the accuracy of the sampled marker was recorded. If the relevant location in the article had precision CST and the system accurately plotted the marker to this precision, this was recorded as CST/CST. If the system plotted the article at the CS level of precision, this was recorded as CST/CS. If the system plotted the article accurately at the CS level and inaccurately at the CST level, this was recorded as CST/CSX. Other values include: CST/CXX, CST/XXX, CS/CS, CS/CX, CS/XX, CS/XX, CS/C, C/C, C/X, C/XX, and C/XXX. If there was no relevant location in the article, then possible values are: N/X, N/XX, N/XXX, depending on the level of precision of which the marker was plotted at. This is summarised in the following table:

	Country	**State/Province**	**Town/City/Village**	**No Relevant Location**
1 missing	–	CS/C	CST/CS	–
2 missing	–	–	CST/C	–
1 wrong	C/X	CS/X	CST/X	N/X
2 wrong	C/XX	CS/XX	CST/XX	N/XX
3 wrong	C/XXX	CS/XXX	CST/XXX	N/XXX

Finally, accuracy of language was assessed. Again, this was recorded only if the event of the article was of the A, Z, H, P, M, or F types. If the system performed no translation on an article at all, this was recorded as 'NT' (not translated). If the title of the article was translated in a way that was understandable and accurate, this was recorded as 'UT' (understandable title). If it was translated in a way that was understandable but not accurate, this was recorded as 'MT' (misinformative title). If it was not even understandable, this was recorded as 'NUT' (not understandable title). If only the event of the article was translated in a way that was understandable and accurate, this was recorded as 'UE' (understandable event). If it was not accurate, it was recorded as 'ME' (misinformative event). If it was not understandable, it was recorded as 'NUE' (not understandable event). 'NA' was reserved for all other cases.

5.2.3 Total Population Results

The first set of results describes the total number of reports for each system and their geographical origin—see Figure 21.

Of the three systems compared, EpiSPIDER had by far the most reports. With more than seven times the number of reports from BioCaster and more than 19 times the number of reports from HealthMap, it clearly had the largest volume of information. This was robust across the geographical zones. The zone it performs most poorly on — in terms of total percentages for zones — is Zone 6 ('South America'), with 58% reports by EpiSPIDER, 11% by HealthMap and 21% by BioCaster. The zone it performs best on — again in terms of total percentages for zones — is Zone 12 ('Alaska'), with 100% of reports by EpiSPIDER. In terms of absolute numbers, though, Zone 10 ('USA') clearly stands out: 8048 reports by EpiSPIDER, compared with 132 and 279 by HealthMap and BioCaster, respectively; moreover, 66% of all of EpiSPIDER's reports are for Zone 10. The worst EpiSPIDER does in terms of absolute numbers is Zone 12, with 58 reports—however, the other systems produced *no* reports for this zone.

In some zones, BioCaster dominates HealthMap (in terms of total number of reports) and in others this is reversed. In Zone 2 ('Asia'), BioCaster performs overwhelmingly well, with almost 7 times the reports by HealthMap. Also, as a percentage of BioCaster's total number of reports, Zone 2 stands out, with 38% of all of BioCaster's reports from Zone 2. This is substantially larger than the other systems (14% for HealthMap and 9% for EpiSPIDER), probably due to the fact that BioCaster is able to analyse reports in a number of Asian languages—Chinese, Vietnamese,

Thai, and Japanese—whereas the only Asian language that HealthMap can analyse is Chinese, and EpiSPIDER records only articles in English.

HealthMap produces more reports than BioCaster in Zones 3, 4, and 11 ('Russia', 'Middle East', and 'Canada'). It is not clear that there is any particular feature of the two systems that causes this, and the difference may simply be due to sampling error (e.g., in Zone 11, there were only 16 reports from BioCaster and 18 from HealthMap). Zones 3 and 4 are adjacent and both contain regions with populations that speak Russian (particularly Zone 3!). Both systems can scan for articles in Russian, but perhaps HealthMap does this better than BioCaster and, so there may be a systematic difference between the two systems for these zones. Further testing would be required to evaluate this.

Once reports are broken down by source (see Figure 23), another trend is clear: EpiSPIDER gets a large majority of its reports from Twitter—about 68% (8275 reports). Neither of the other systems obtains reports from Twitter, so it is interesting to compare EpiSPIDER minus its Twitter reports—which will be denoted as 'EpiSPIDER –T'—with the other two systems. Comparative results for the systems when Twitter reports are removed are shown in Figure 22.

Once the Twitter reports are removed, EpiSPIDER still produces the vast majority of reports: 64% are from EpiSPIDER –T, 26% are from BioCaster, and 10% from HealthMap. A significant proportion of the Twitter reports detected by EpiSPIDER are for the Zone 10. The proportion of EpiSPIDER's reports for Zone 10 ('USA') decreases by 18% once the Twitter reports are factored out; this percentage for every other zone either increases or remains about the same.

All three systems fared poorly for Zone 6, which roughly corresponds to the Western regions of North and Equatorial Africa. About 1% of BioCaster reports are for this zone, and about 0.5%, 0.5%, 1.4% for HealthMap, EpiSPIDER and EpiSPIDER –T, respectively. Zones 11 and 12 ('Canada' and 'Alaska') also received little attention from all three systems.

Looking at the reports by source (Figure 23), it is clear that by far the largest source is Twitter, comprising 57% of all reports. It should be stressed that this is the *number* of reports by each system. A system can report an article from a source multiple times, some of these in different locations, and sometimes even in a single location. Many reports from Twitter are duplications—either in the one location or for multiple locations. This mostly happens because several people will 'tweet' the same news article, and EpiSPIDER displays all of those 'tweets' as separate reports. A brief assessment of the utility of the articles that EpiSPIDER finds via Twitter is included in the next section.

After Twitter, the next largest source is Google News, with about 15% of all reports. BioCaster relies most heavily on Google News, with 67% of its reports coming from that source—compared with 43% for HealthMap, 7% for EpiSPIDER and 21% for EpiSPIDER –T. The next two largest sources are two news aggregators, Moreover and DayLife, with 9% and 8% of all reports, respectively. EpiSPIDER –T relies most heavily on Moreover, with 31% of its reports coming from that source—compared with 10% for EpiSPIDER, 15% for HealthMap, and 0% for BioCaster, which does not use Moreover as a source. DayLife is used exclusively by EpiSPIDER with about 10%

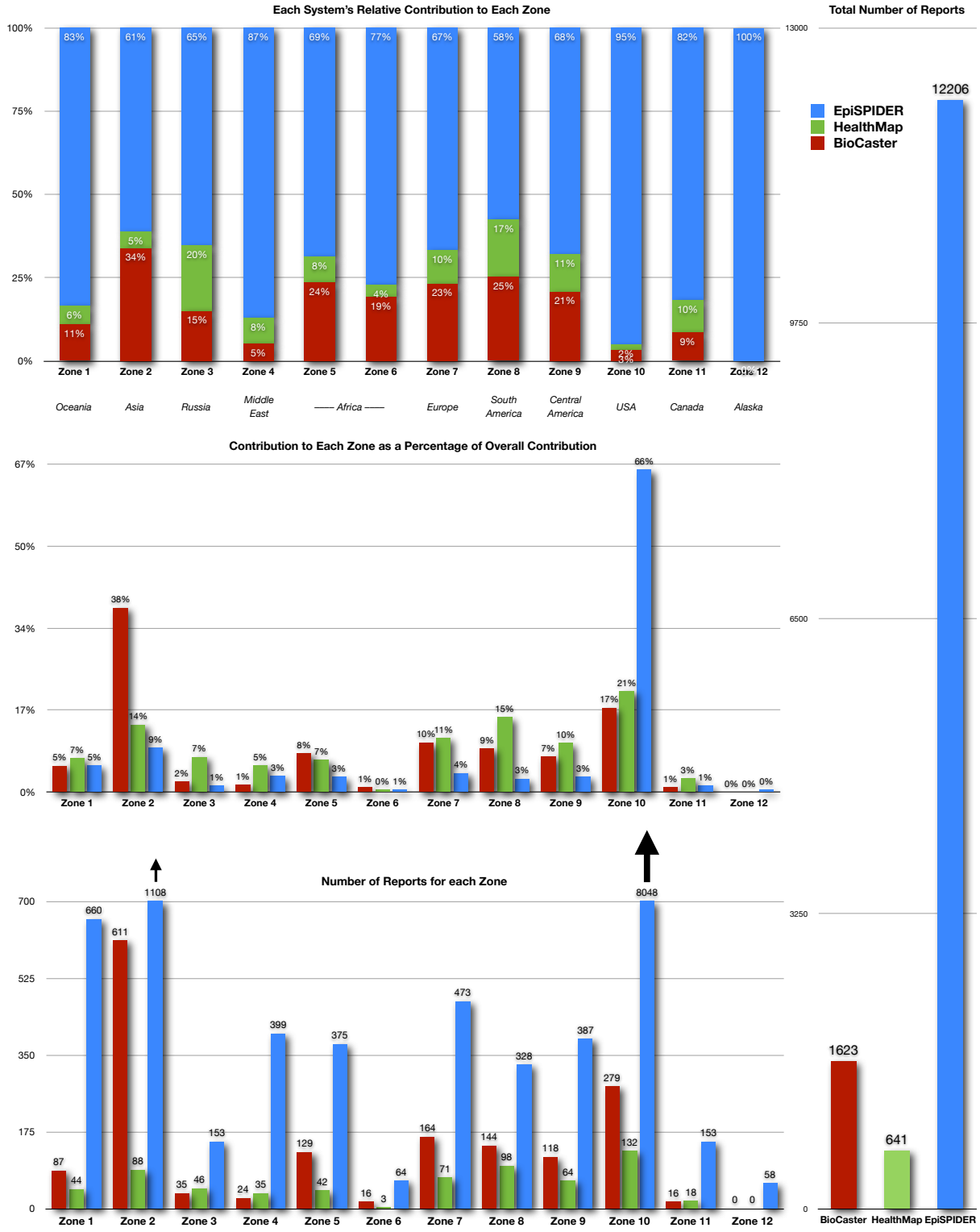


Figure 21: Geographical Zones and Total Numbers of Reports.

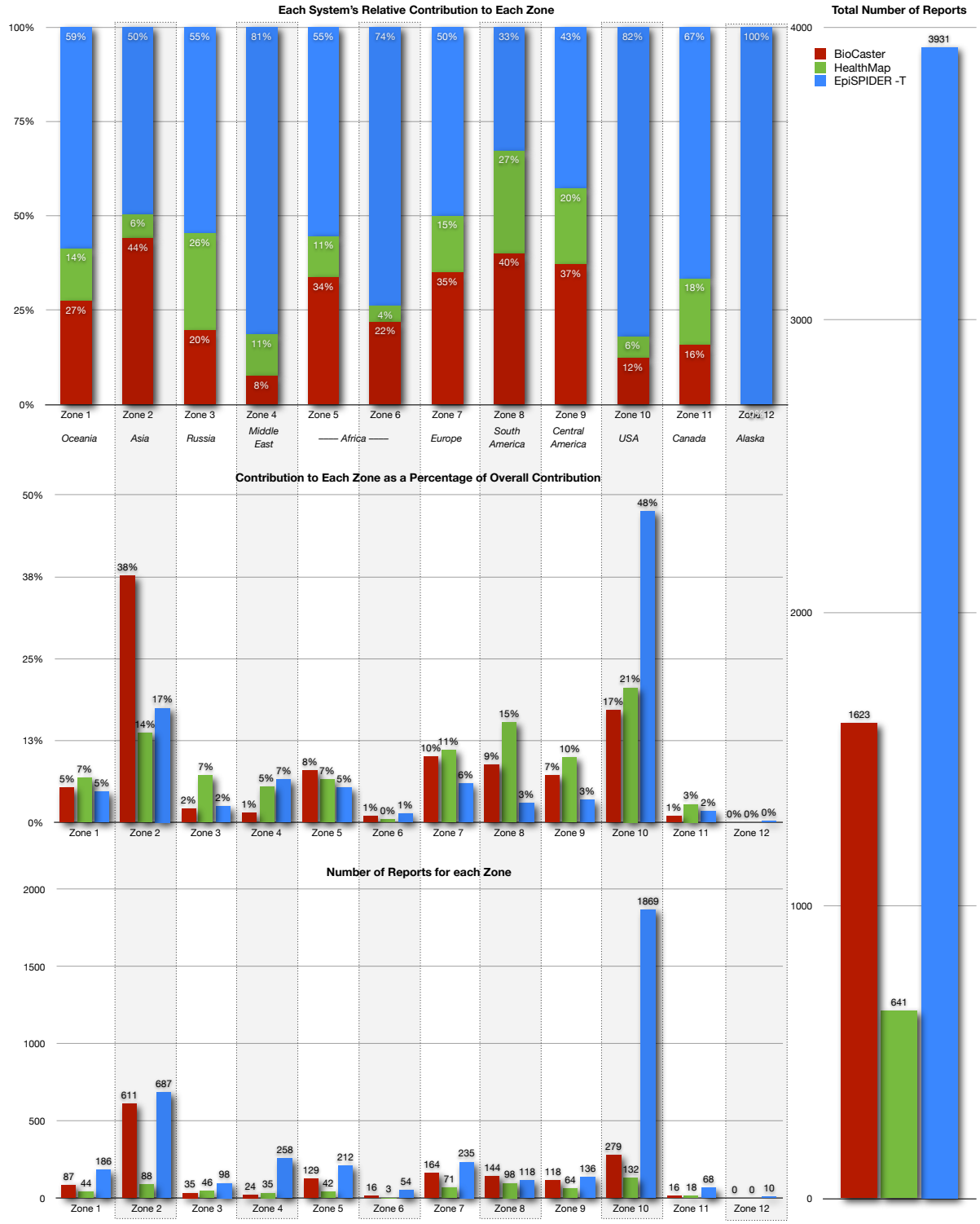


Figure 22: Geographical Zones and Total Numbers of Reports — DeTweeted.

of the system's reports—20% when Twitter reports are factored out. The next largest source is ProMED, which accounts for 4% of all reports. HealthMap relies most heavily on ProMED, with 23% of its reports coming from that source—compared with 2% for BioCaster, 3% for EpiSPIDER and 9% for EpiSPIDER –T. A variety of other sources account for the remaining 6% of reports.

There are some sources that are used by one only one system and contribute to that system's reports in a significant way. DayLife by EpiSPIDER has already been mentioned. EpiSPIDER also uses ReliefWeb Updates, which accounts for about 9% of EpiSPIDER –T's reports (345 reports in total). Similarly, BioCaster uses Meltwater, which accounts for about 25% of its reports. HealthMap uses contributions by its HealthMap Community, which account for about 8% of HealthMap's reports.

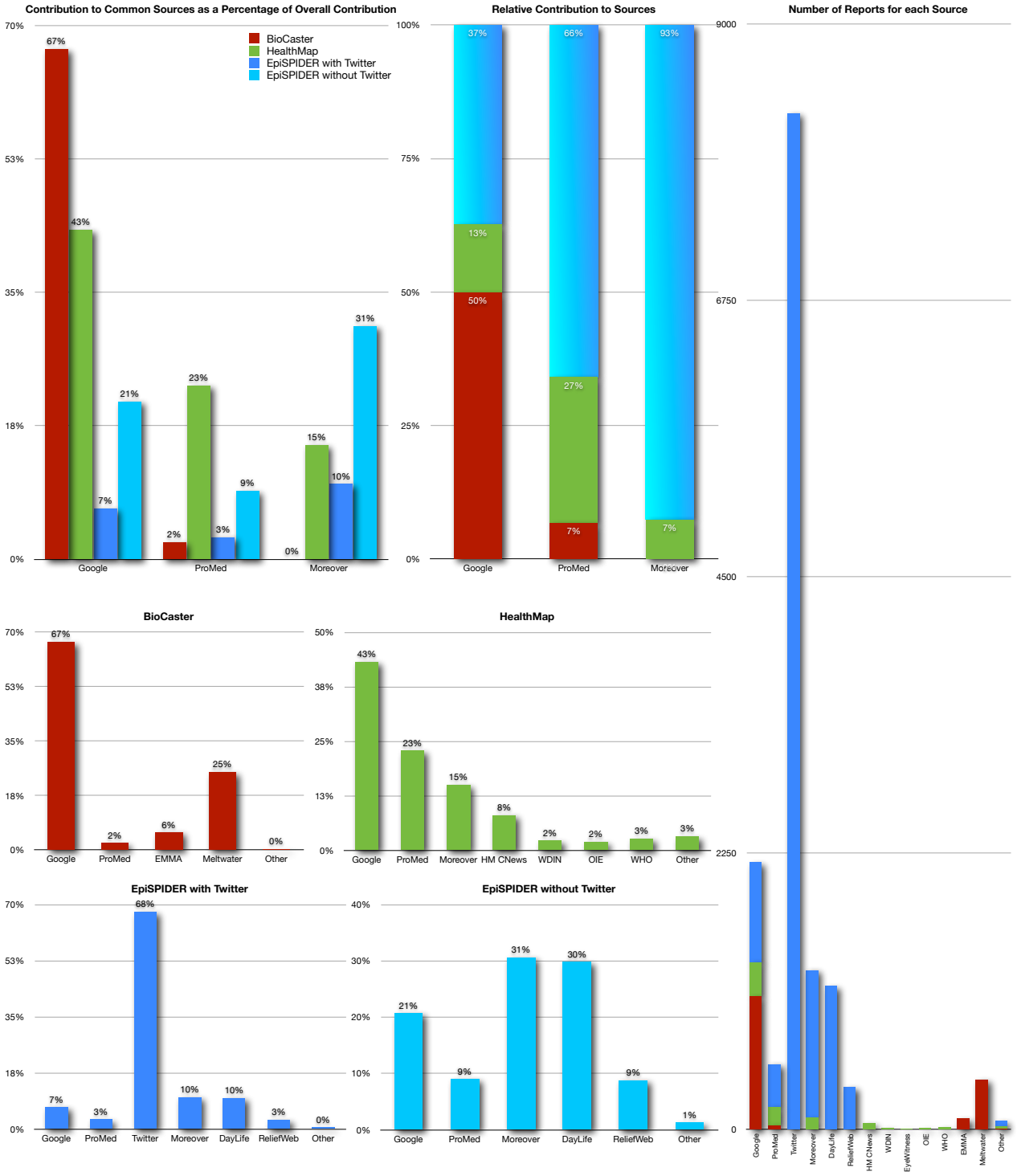


Figure 23: Geographical Zones and Total Numbers of Reports.

By examining the CSV file for HealthMap's reports, it was possible to identify the languages of the articles that HealthMap found. EpiSPIDER detects only articles written in English, and original language information was not contained in the BioCaster CSV file. These results are shown in Figure 24.

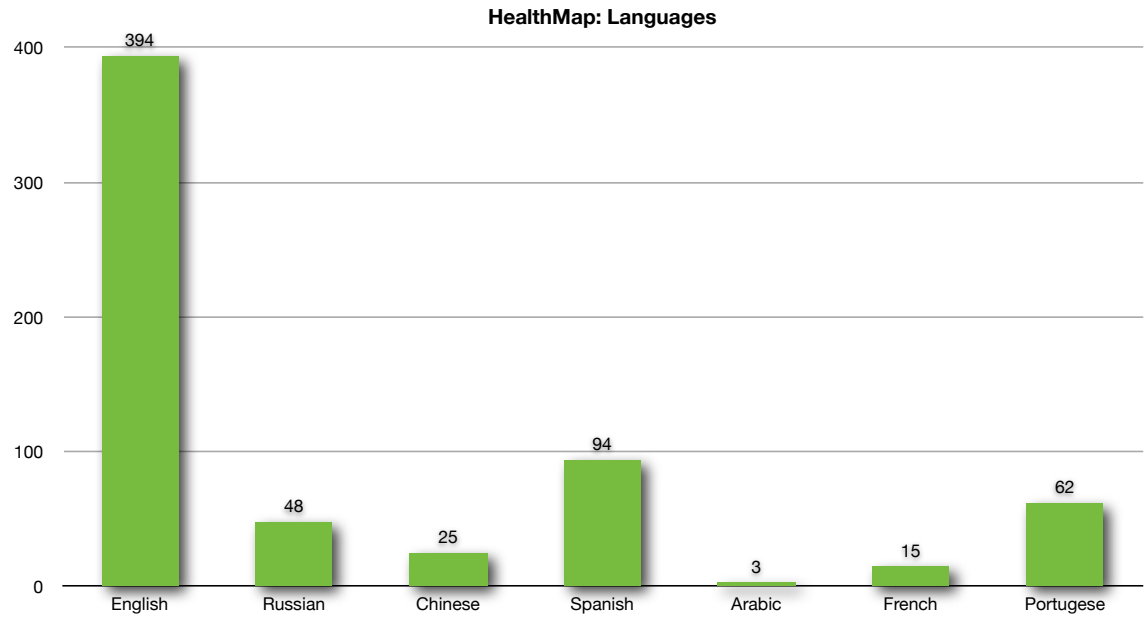


Figure 24: Languages and Total Numbers of Reports.

As expected, articles written in English made up the large majority of reports. Spanish and Portuguese were the next most common languages in the systems that were compared.

5.2.4 Sample Results

In total, 101 reports were sampled. Of these, 85 did not have any technical problems such as broken links. Of these 85 reports, 36 were articles on a specific event pertaining to an animal, human, or zoonotic disease, or food security issue. Out of these 36 reports, 29 were plotted with a reasonable level of accuracy and precision. That is, each report had to at least get a location of its article within one level of precision and accuracy. For example, if the location of the article was CST, then the report had to at least get CS or CSX. The contributions of each system to these numbers are shown in Figure 25.

There was a substantial difference between the relative contributions to the total sample and the relative contributions to the collection of all reports. These are compared in Figure 26. One reason for this is that BioCaster and EpiSPIDER exhibit significantly more clustering than HealthMap (as mentioned earlier). Generally, HealthMap reports are spread more uniformly over the surface of the Earth, and so they are more often selected by a uniform random sample.

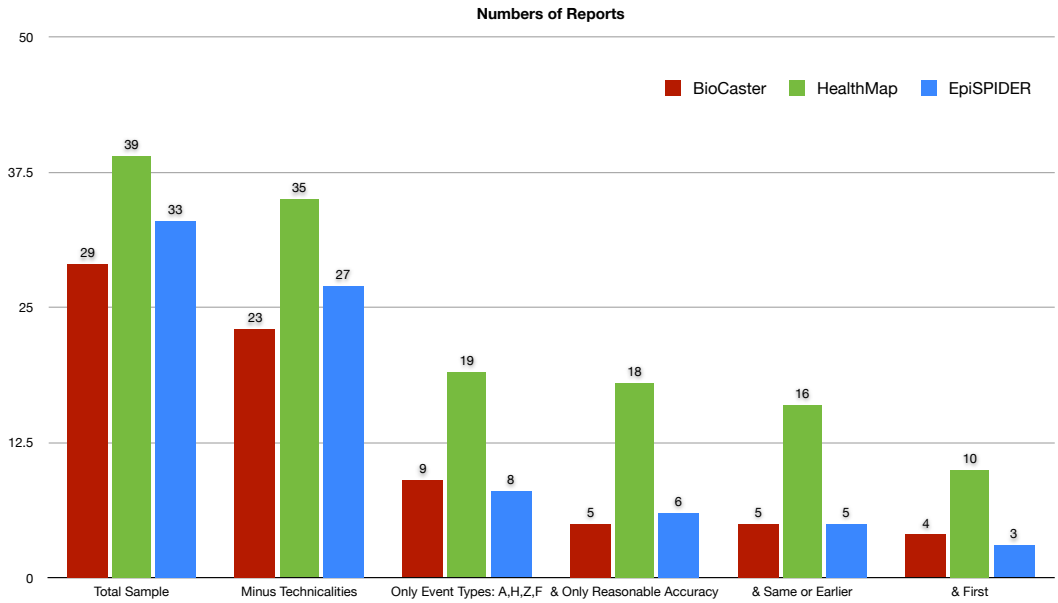


Figure 25: System Statistics.

Once technical difficulties, event types other than A, H, Z and F, and inaccurately plotted reports were factored out, 17% of BioCaster's reports remained, 46% of HealthMap's, and 18% of EpiSPIDER's. HealthMap had far more 'quality' reports than the other two systems—even when their reports were combined (18 vs. 5 and 6). 10% of all reports in the sample were HealthMap 'quality' reports that were the *first* to report on an event. This was the case in 4% and 3% of the time for BioCaster and EpiSPIDER, respectively.

109 reports from EpiSPIDER sourced through Twitter for the period 20 to 28 June were also sampled. The event type for each report was determined by reading each article. Most of the reports were for natural and/or man–made disasters or general news on political events, vaccinations, etc.—this is because EpiSPIDER looks for those sorts of reports. Factoring out those left 37 reports—8% of which were on animal diseases, 24% on food/water security, 22% on human diseases, and 30% on zoonotic diseases. There were no irrelevant reports (i.e., reports that did not pertain to biosecurity in any way).

The sample distribution of reports across different languages for HealthMap was close to the distribution of the total population—see the comparison in Figure 27—with one exception. In the sample, there was a significantly higher percentage for reports in Arabic than in the total population. This is probably because the reports in the Middle East happened to get oversampled. However, it is still possible to draw conclusions about BioCaster's ability to detect reports in other languages. Perhaps surprisingly, BioCaster did not detect any articles written in Arabic, even though the Middle East was oversampled. Although BioCaster does detect articles in Arabic, it seems it is not as effective at this as is HealthMap. BioCaster also detects relatively more articles

Figure 26: Clustering rates.

in Russian and Chinese. No reports in BioCaster's focus languages—Japanese, Korean, Thai, and Vietnamese—were collected in the sample. This might be a result of the small sample size.

However, reports by BioCaster in the Asia–Pacific region were collected in the sample.

5.2.5 Conclusions

Due to the amount of time it takes to analyse each report and the large number of reports from all the systems, only a small sample of reports was collected. A further study would be required to make more definitive conclusions regarding the systems that were compared. However, these initial sample results plus the total population results suggest the following:

- Social media networks, such as Twitter, are new and potentially important sources of biosecurity intelligence. Currently only EpiSPIDER takes advantage of this new type of source.
- BioCaster has a significant focus on the Asia–Pacific region. This can be seen in the total population statistics, and explained by the fact that its ontology deliberately includes terms in Asian languages besides Chinese.
- HealthMap has a lot less noise and plots reports more accurately than do BioCaster and EpiSPIDER.
- A significant proportion of the reports are on potentially useful information that is not directly related to outbreak events (e.g., general health news, politics, and disasters).
- The systems compared all make heavy use of Google.
- The systems compared use a variety of other search engines and news aggregators.
- A significant percentage of HealthMap's reports come from its community of users. None of the other systems compared takes advantage of inputs from users in this way.
- No reports from BioCaster and in BioCaster's focus languages—Japanese, Korean, Thai, and Vietnamese—were collected in the sample. This suggests that the system is not yet at a stage where it is successfully accessing news written in these languages. However, the total population results show that BioCaster does focus many of its reports on the Asia–Pacific region.

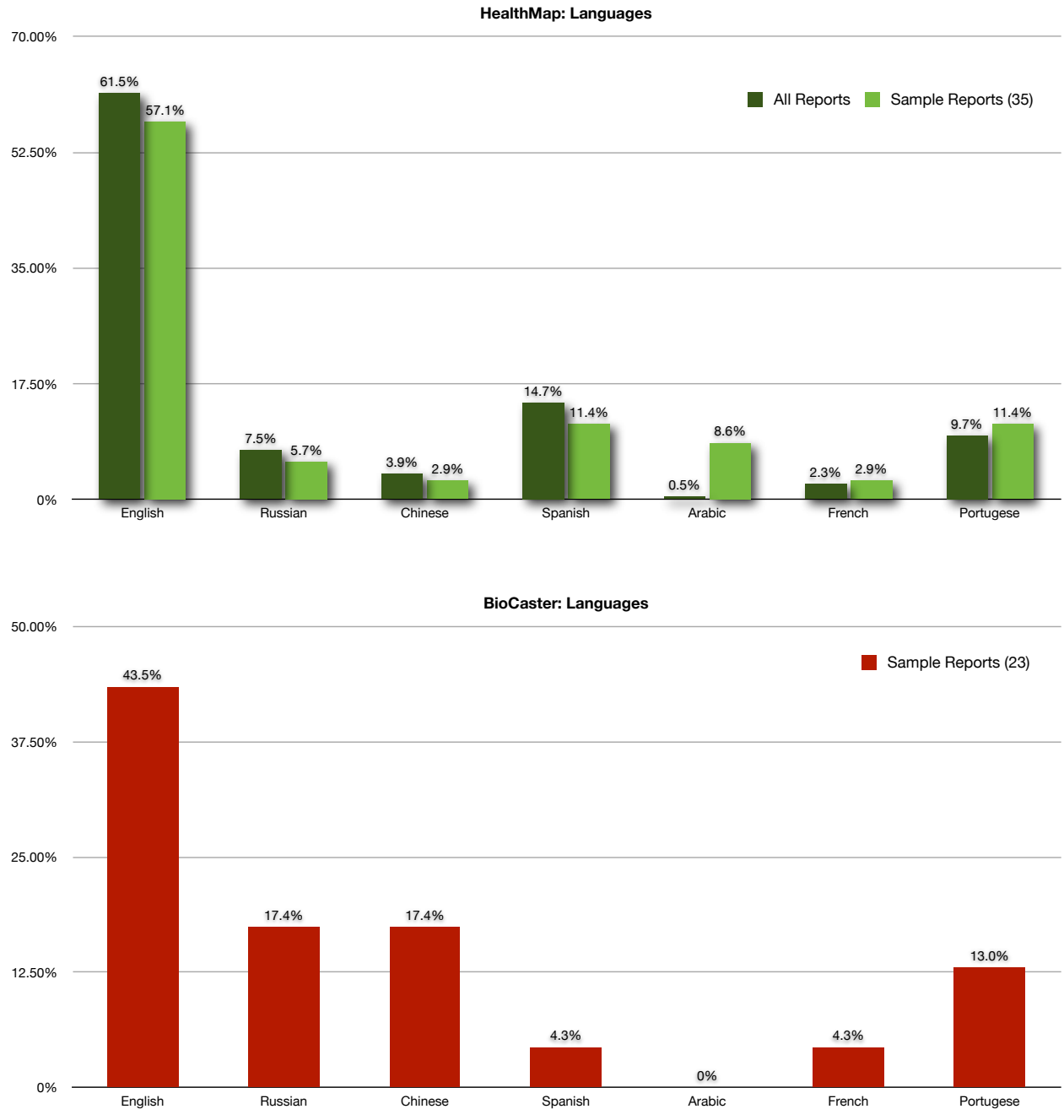


Figure 27: HealthMap and BioCaster Languages.

5.3 Second Comparison of BioCaster, EpiSPIDER and HealthMap

A second comparison of BioCaster, EpiSPIDER and HealthMap, which didn't involve manual sampling and analysis, was also conducted. To do compare the systems, it was necessary to count

and classify the articles that each system finds. Each article found by a given system is linked to in a report published by that system in a KML or CSV file. Each report has a link to an article, a publish date, the source by which the article was found, latitude and longitude coordinates, and various other pieces of information, depending on the system. To determine the number of articles a system has found over some period of time, it isn't enough to simply count the total number of reports published by that system over that time period. This is because often a single article will be linked to by a number of different reports, which may have different latitude and longitudes, publish dates, etc.

So a better way to determine the number of articles that a system has found is to count the number of unique links in all of that systems' reports. However, in some cases, two different links can redirect to the same article. A common example of this is when two people use two different URL shorteners to link to an article. So it is necessary to follow all of the redirections of the links to arrive at their final links and then count how many unique final links there are. However, this still wouldn't be the number of articles that a system has found, for a single article can be associated with multiple URLS. For example, `http://www.promedmail.org/pls/otn/f?p=2400:1001:4293294425104239::NO::F2400_P1001_BACK_PAGE,F2400_P1001_PUB_MAIL_ID:1050,84399` and `http://promedmail.org/pls/otn/f?p=2400:1001:462275334141704::NO::F2400_P1001_BACK_PAGE,F2400_P1001_PUB_MAIL_ID:1055,84399` both link to the same ProMED article, even though they are different links. So it's also necessary to compare the pages of each link to see if they are the same. The pages, however, often vary slightly and in irrelevant ways (as with the previous two links). So to compare the pages of two links to see if they are links to the same article, the pages' contents must be scraped clean of any irrelevant surrounding material such as headers, sidebars, advertisements, etc. In this study, this was done using Alchemy's Text Extraction / Web Page Cleaning API.

Even once the scraped contents of pages are compared, they can still vary in irrelevant ways. One common way for this is to occur is when two news media sites share an article, but apply different editing and/or formatting standards. Another way is if an article appears on two sites, but different timestamps are applied (e.g., due to different timezones). So it's also necessary to compare the scraped contents of pages in a way that allow for them to vary slightly. In this study, this was done using the Python SequenceMatcher Class. If two scraped contents had a high similarity ratio—defined as $2M/T$>0.9 where M is the number of matches and T the total number of elements—then they were judged to be the same article.

So to estimate the number of articles a system has found, the number of scraped contents whose pairwise similarity ratios were below 0.9 was determined. Doing this involved following each original link through any redirections and scraping it's content. Both of these processes were not always possible. Sometimes the original links were broken when they were followed, and sometimes webpages were too messy for their contents to be scraped. Let b be the fraction of links that were broken, m the fraction of pages that were too messy to scrape the contents of, and S the number contents that were determined to be unique (by similarity). The expected number of

unique articles, U, is then: $U = S/(1-b)(1-m)$.

The following sections use variations of the above method to compare the systems over the period August 2–30, 2010 in the following respects: their numbers of unique articles, their overlaps and comparative timeliness, their usage of different sources, and their focus on different countries and languages.

5.3.1 Unique Articles

Table 2 and Figure 28 show the counts at various stages of the estimation of the number of articles for each system. While BioCaster had a larger overall reduction (from number of original links to expected unique articles) than HealthMap (49% and 18%, respectively), it still reported more unique articles than HealthMap—about 28% more. EpiSPIDER had by far the largest overall reduction with 89% of its links removed as repetitions. Despite this large reduction, EpiSPIDER still had more unique articles than BioCaster and HealthMap—about 54% more than BioCaster and 96% more than HealthMap.

Table 1: Numbers of Articles

	BioCaster	**EpiSPIDER**	**HealthMap**
Total Original	6,860	58,046	3,856
Unique Original	4,682	8,235	3,302
Unique Original & Live	4,382	7,250	3,174
Unique Real	4,352	7,148	3,137
Unique Real & Scrapable	4,192	6,784	3,073
Unique Content (exact)	3,843	5,818	2,997
Unique Content (similarity)	3,620	5,367	2,960
Expected Unique	4,015	6,193	3,144
Percentage Reduction	49%	89%	18%

5.3.2 Overlaps and First Reports

To determine the overlaps between systems, the unique contents (by similarity) of each system were compared. First, each unique content (by similarity) for a given system was combined with all the contents that were similar to it, along with all the final links that linked to it, along with the dates of the corresponding original links. Call each combination a *story*. (There were 3,620 stories for BioCaster, 5,367 for EpiSPIDER and 2,960 for HealthMap.) When two systems had two stories with at least one matching final link, the two stories were defined to be a match. When two systems had two stories which had contents judged to be the same (by similarity), the two stories were defined to be a match. When two stories matched, the earliest dates of the stories were compared and the difference between them was recorded.

Table 2 contains the numbers of stories that matched by links and by similarity. About 10% of BioCaster's stories matched a HealthMap story, and about 13% *vice versa*. On average, BioCaster's

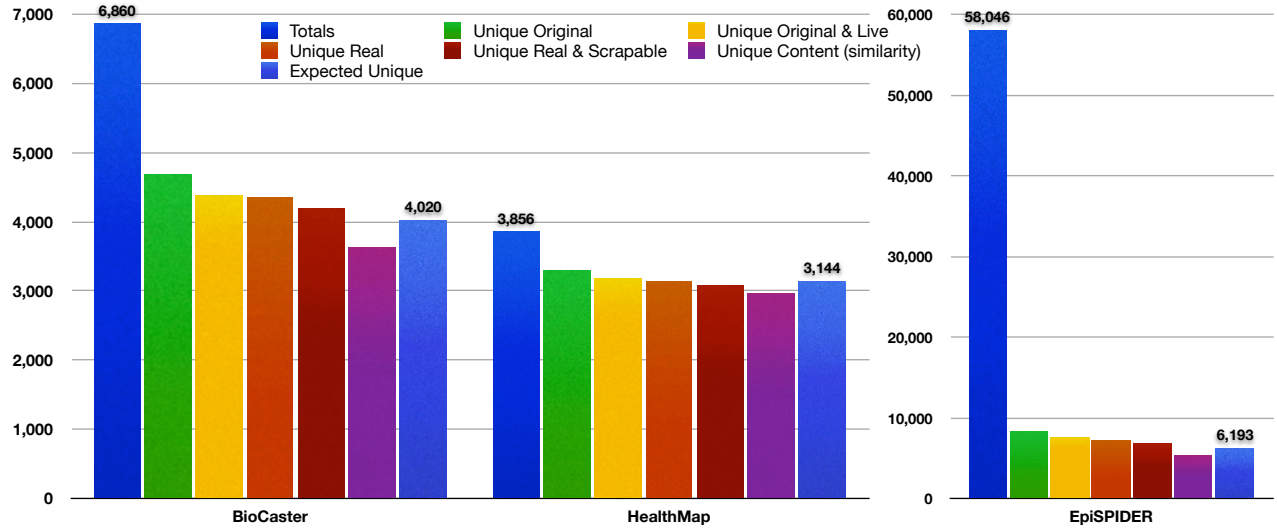


Figure 28: Numbers of Articles.

Table 2: Overlaps

	BioCaster	**EpiSPIDER**	**HealthMap**
BioCaster	–	294 + 166	151 + 228
EpiSPIDER	272 + 156	–	267 + 264
HealthMap	145 + 229	277 + 271	–

(E.g., EpiSPIDER had 294 stories that matched a BioCaster story by link, and 166 by similarity.)

publish date for a story that both it and HealthMap found was after HealthMap's publish date by 0.4 of a day. However, the two systems operate in different timezones (BioCaster in Japan, HealthMap in the US), so this difference is not significant and probably mostly an artefact of the difference in timezones.

About 10% of EpiSPIDER's stories matched a HealthMap story, and about 18% *vice versa*. On average, EpiSPIDER's publish date for a story that both it and HealthMap found was before HealthMap's publish date by 0.2 days. Both systems are based in the US, so this might mean that EpiSPIDER is slightly faster at detecting and publishing articles than HealthMap.

About 9% of EpiSPIDER's stories matched a BioCaster story, and about 12% *vice versa*. On average, EpiSPIDER's publish date for a story that both it and BioCaster found was before BioCaster's publish date by 0.8 days. Again, the systems are in different timezones so it is difficult to determine if this is an artefact of that difference. However, the figure confirms, to some degree, that EpiSPIDER finds and publishes articles slightly faster than HealthMap, since the difference between EpiSPIDER and BioCaster is roughly twice that of the difference between HealthMap and BioCaster.

5.3.3 Languages

EpiSPIDER finds articles solely written in English, so it was only necessary to compare BioCaster and HealthMap with respect to the systems' focus on different languages. To determine the language of each article reported by the systems, Alchemy's Language Detection API was used. For HealthMap, it was also possible to determine the language of each article directly. The system has a translation schema that it uses in the links of reports to automatically translate the articles it has detected. For example, if "trto=en&trf=zh" appears in one of HealthMap's links, then when a user clicks on the link, HealthMap gets Google Translate to translate the article that the link points to from Chinese to English. So it was assumed that whenever "trto=en&trf=zh" appeared in a link, the original article was in Chinese—and similarly for other languages. Using this translation schema, it was then possible to determine another distribution of HealthMap's reports over languages—Fig. 30. This second distribution was strongly correlated with the distribution determined by using Alchemy's API (Fig. 29). This strong correlation lends support to the accuracy of the distribution for BioCaster which was only able to be determined using the API. A notable exception in the correlation between the two distributions for HealthMap's languages was Portuguese. Alchemy found 115 fewer pages in Portuguese than those determined using the translation schema in HealthMap's links, and it was unable to obtain the language of 93 of HealthMap's links. It was also unable to obtain the language of 113 of BioCaster's links. It's therefore likely that the actual number of BioCaster's links in Portuguese is significantly higher (by about 100). Incidentally, the Alchemy API also detected a handful of pages in languages not reported by HealthMap—German, Indonesian, Japanese, Ukrainian, and Vietnamese—and some for BioCaster—Dutch and German.

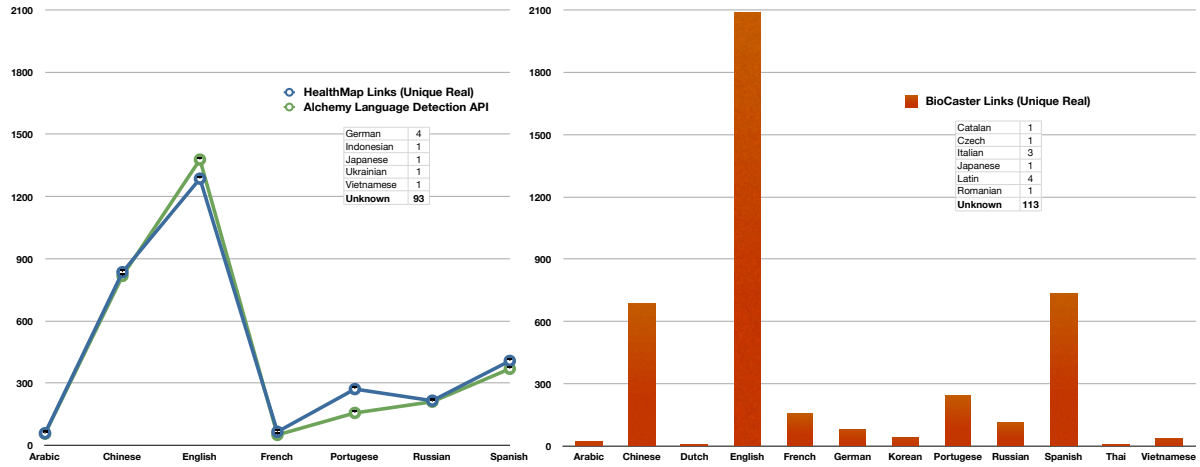


Figure 29: Left: HealthMap languages determined by HealthMap and Alchemy. Right: BioCaster languages determined by Alchemy.

English and Chinese are the two most common languages for HealthMap reports, with 46%

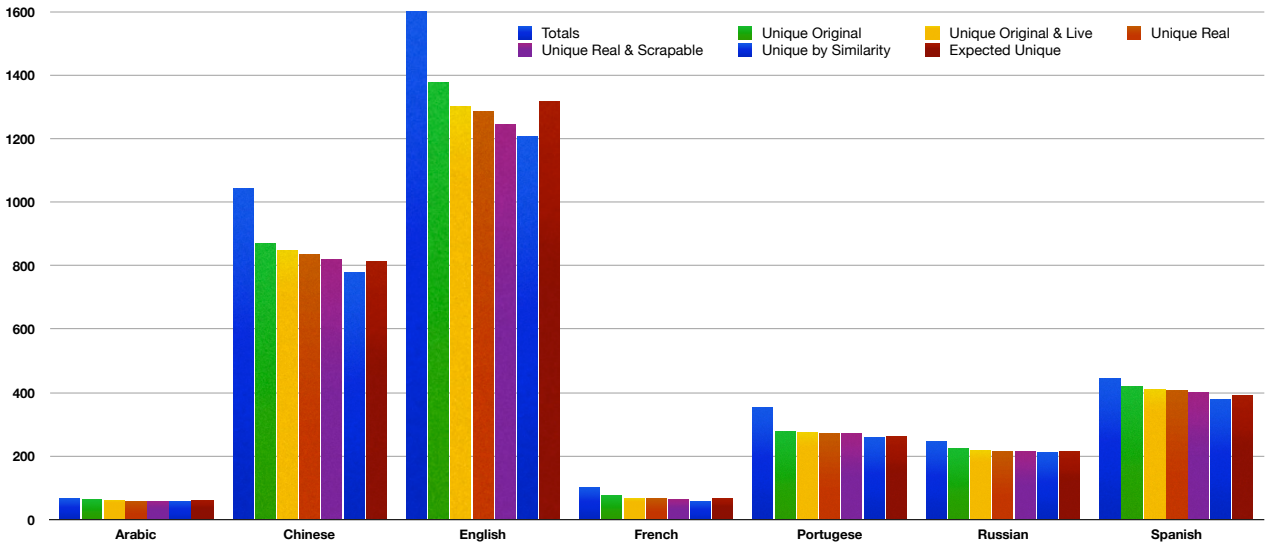


Figure 30: HealthMap Languages (via URL Translation Schema).

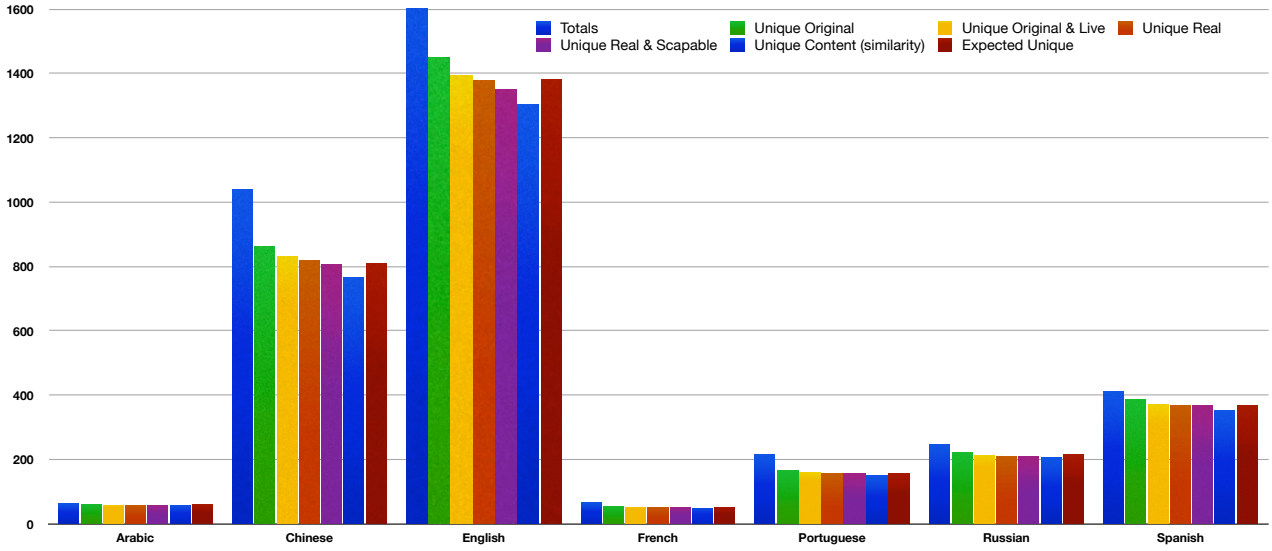


Figure 31: HealthMap Languages (via Language Detection API).

and 26% of the expected unique articles, respectively. BioCaster has a similar focus on Chinese and English (45% and 18%, respectively) except that it finds slightly more articles in Spanish than HealthMap (both in absolute terms (696 vs. 367) and relative terms (17% vs. 12%). Given that BioCaster has a priority for finding articles in Asia–Pacific languages—Chinese, Japanese, Korean, Thai, and Vietnamese—it finds surprisingly few articles in these languages, with the exception of Chinese, although it finds more articles in these languages than HealthMap (again with the exception of Chinese).

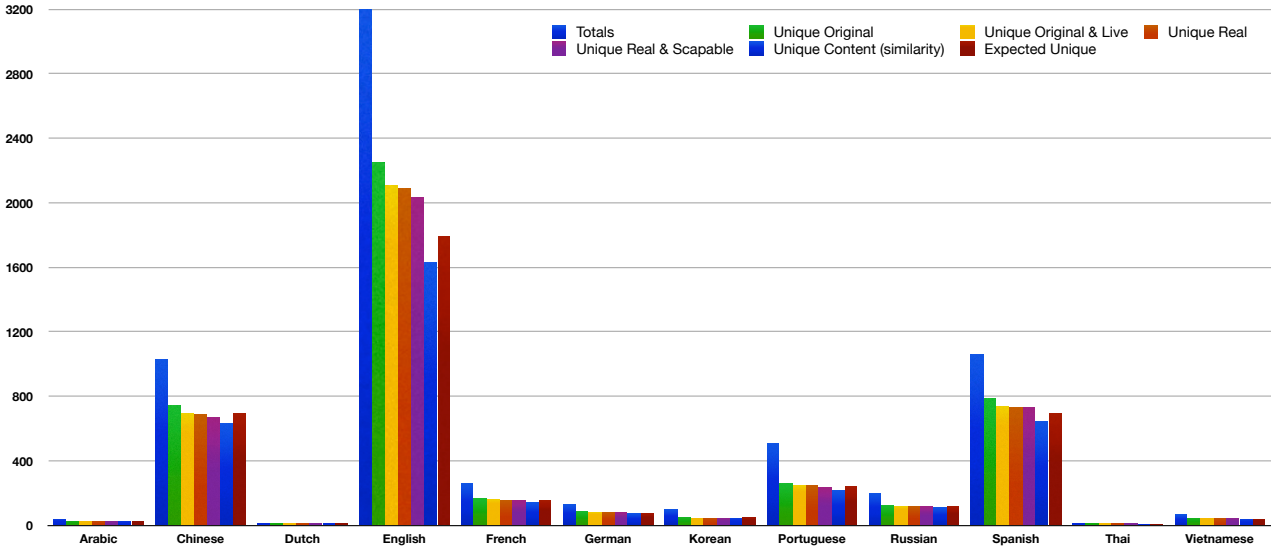


Figure 32: BioCaster Languages (via Language Detection API).

5.3.4 Geographic Distributions

The country in which each report was determined by reverse geocoding each report's latitude and longitude with Geonames' Reverse Geocoding API. In some cases, the coordinates of a report corresponded to a location that was not on land, so a 20km buffer radius was used. All other reports were treated as errors and ignored (e.g., some reports had 0,0 coordinates). Since a given article can refer to multiple locations, repetitions of links across countries weren't removed. All of the systems report an article in multiple locations if they detect that the article makes reference to multiple location. In what follows, the expected numbers of unique articles (by similarity), within country categories, are reported.

In total, there were 215 countries that had an article reported by at least one of the systems. Fig. 33 contains six choropleth maps representing different aspects of the distribution of articles over countries. In general, EpiSPIDER reported the most articles for each country, although there were some notable exceptions—e.g., China, France, Brazil, Argentina, and Spain. In some cases, EpiSPIDER reported a significant number of articles where BioCaster and HealthMap reported none or very few—e.g., Afghanistan, Haiti, and Switzerland.

There were several interesting differences between BioCaster and HealthMap. HealthMap had more articles in two Asia–Pacific countries: China and Laos. It's not clear what the difference was in Laos, but the difference in China was probably due to the systems' different sources (see next section). HealthMap also had nearly 150 articles for Pakistan, while BioCaster had none. However, in the original list of BioCaster's reports, there were 62 instances of "Pakistan" in the title of the report. This suggests that the system is finding articles about events in (or relating to) Pakistan, but failing to plot them in Pakistan. HealthMap also reported 20–50 articles each for Egypt, Cameroon,

Martinique, Nepal, South Africa, and Ukraine, whereas BioCaster had at most 3.

In the other direction, BioCaster had 203 articles for France in contrast to HealthMap's 13, more than double than HealthMap for Mexico and the UK, and more than 50% more for India. BioCaster also had 3 to 6 times more articles for Taiwan, Angola, Bangladesh, and the Dominican Republic. In Uruguay, BioCaster had 87 articles, in contrast to HealthMap's 1. And generally, it did better in the Asia–Pacific region: Indonesia (63 to 38), Hong Kong (52 to 13), Japan (46 to 20), Malaysia (57 to 22), Philippines (48 to 38), Singapore (36 to 7), South Korea (22 to 4), Thailand (55 to 40), Vietnam (53 to 29)

BioCaster also found more articles in much of South America than both EpiSPIDER and HealthMap. This may be due to BioCaster's better ability find articles in Spanish (it found more articles than HealthMap in Mexico and Spain too). However, some of the South American countries it did better in are predominantly Portuguese speaking countries—e.g., Brazil—and some that it did worse in are predominantly Spanish speaking—e.g., Peru. The difference also doesn't seem to be due to the system's different sources since both systems used mostly Google and ProMED as sources for South America. The difference may simply be due to a difference in topics of articles that the systems search for.

5.3.5 Source Distributions

All of the systems acknowledge where they get each of their articles from. Using these acknowledgements, it was possible to determine how many articles each system acquired from each source. Figures 34, 35, and 36 show each system's distributions over its sources (with only repetitions within sources were removed).

HealthMap's three most used sources (in terms of expected unique articles) were Google (48%), ProMED (18%), and Moreover (16%). BioCaster's three most used sources were Google (67%), MeltWater (18%), and EMMA (10%). HealthMap also uses Baidu (6%) and SOSO (3%)—both are Chinese search engines—while BioCaster doesn't. This is probably why HealthMap found slightly more articles in China than BioCaster. EpiSPIDER's three most used sources were Twitter (43%), Google (18%), and Moreover (16%). EpiSPIDER is the only system for which Google is not the primary source. EpiSPIDER had the highest overall reduction from original links to unique original links, and this was reflected in each of its sources. Twitter had the largest percentage reduction (91%), the next largest being Moreover (88%).

EpiSPIDER is the only system to have a social media platform—*viz.*, Twitter—as a source. Most of the reports from Twitter were plotted in the US (30%), Pakistan (11%), India (10%), Mexico (4%), UK (3%), and China (3%). However, there were a number of countries for which a large percentage, if not all, of all reports were from Twitter (this is including BioCaster, HealthMap, and the rest of EpiSPIDER's reports). Interestingly, these countries were not the US, China, India, Pakistan, the UK, and Canada. Figure 37 contains a choropleth that shows Twitter's contribution to all of the reports for each country. The countries for which Twitter generated more (unique original) re-

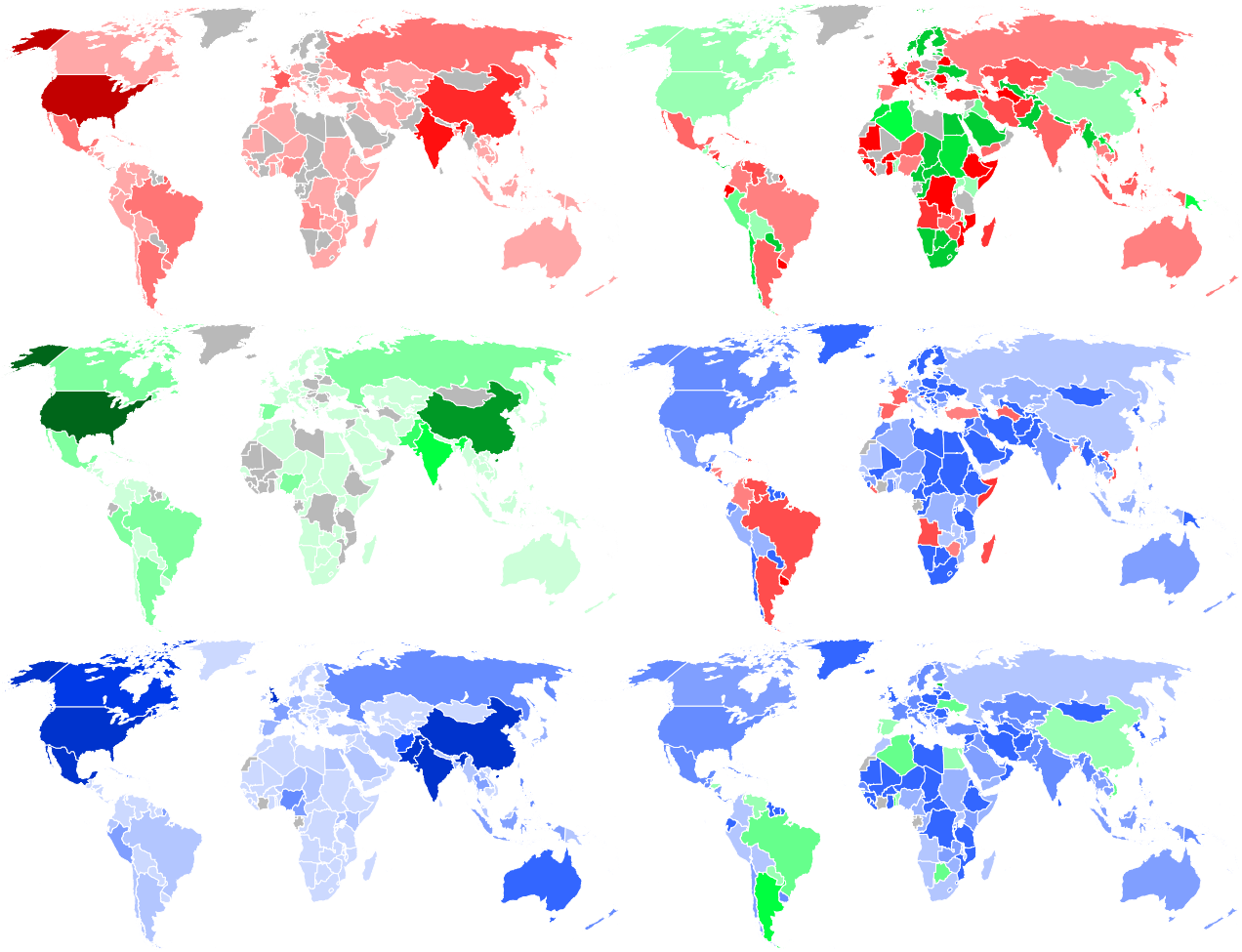


Figure 33: Top to bottom left: Each system's distribution of unique original reports over countries (EpiSPIDER's is on a log scale). Top to bottom right: Pairwise comparisons of the systems. The darker the shade, the more reports. Shades come in 10% brackets. BioCaster is red, HealthMap green, and EpiSPIDER blue.

ports than all other reports combined were: Faroe Islands, Micronesia, New Caledonia, Seychelles, Suriname, Vanuatu, Northern Mariana Islands, Guam, Latvia, American Samoa, Bermuda, Haiti, Iceland, Trinidad and Tobago, Papua New Guinea, Tonga, Mauritius, Iran, East Timor, Isle of Man, Denmark, Montserrat, Anguilla, Tajikistan, Iraq, Syria, North Korea, Sri Lanka, and the Central African Republic. The US came in at 46%, Pakistan 45%, India 29%, Mexico 32% UK 35% and China 16%.

5.4 Discussion

The three systems reviewed here have various strengths and weaknesses, but all are improving over time in a number of ways—e.g., by including new sources, refining analysis techniques, adding new ways of visualising data, etc. One interesting refinement has been the move to make

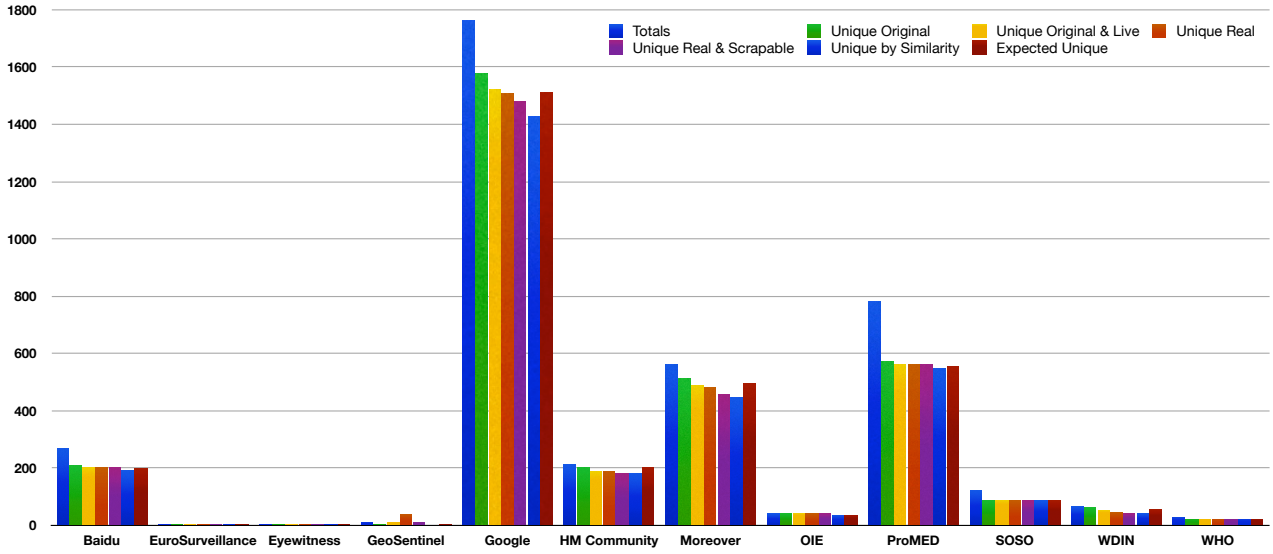


Figure 34: HealthMap's Sources.

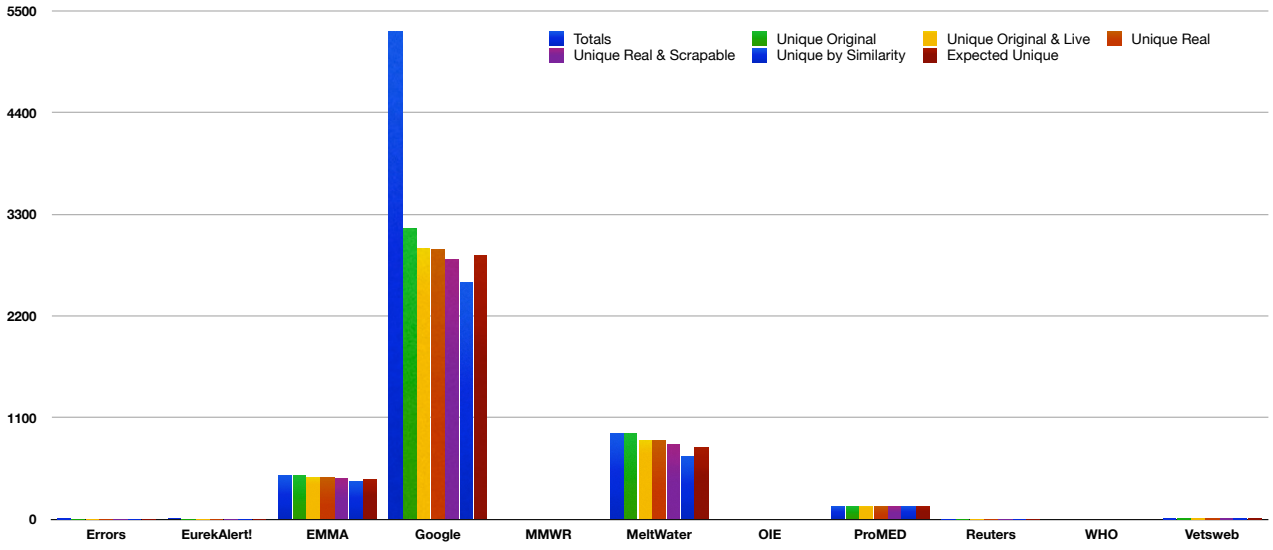


Figure 35: BioCaster's Sources.

use of social media. HealthMap allows users to add commentary and rank articles on a five–star system of significance, and EpiSPIDER scans Twitter for biosecurity information.

There is clearly a large volume of information in Twitter that pertains to biosecurity. However, the vast majority of this information is in the form of multiple tweets about single articles—at least, this is true for the tweets detected by EpiSPIDER and studied here. Nevertheless, for a number of countries, Twitter was the sole source of information or it was the source of most of the information (even with repetitions within countries removed).

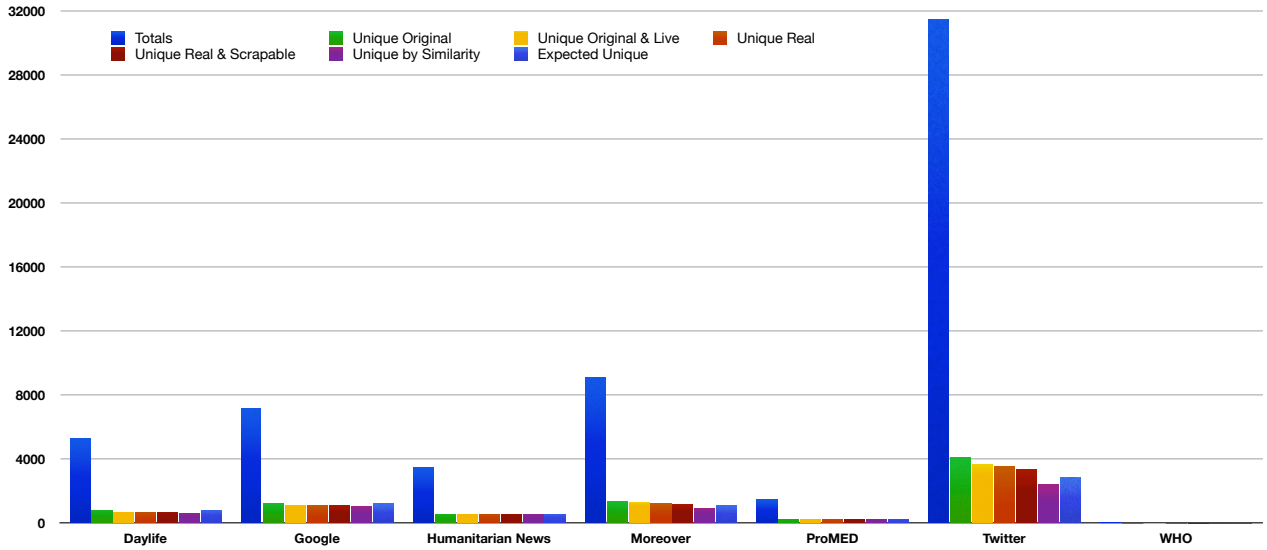


Figure 36: EpiSPIDER's Sources.

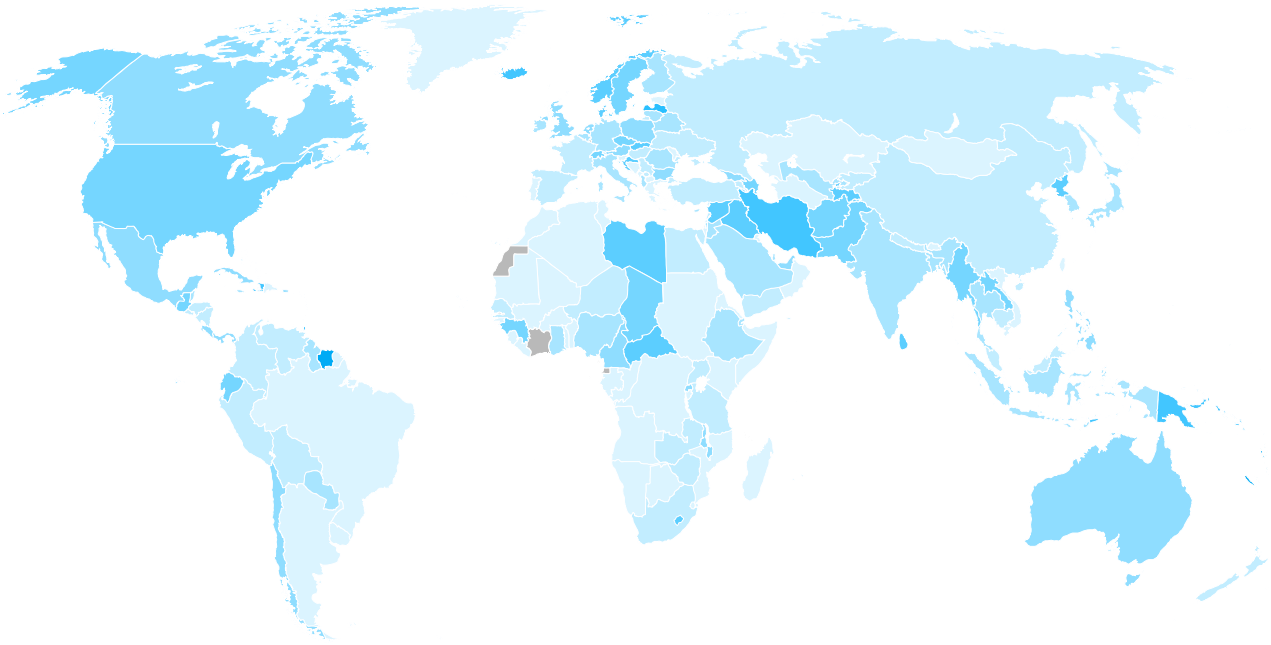


Figure 37: Twitter's contribution to the reports of each country as a percentage of all reports by BioCaster, HealthMap, and EpiSPIDER (without its Twitter reports). Each shade represents a 10% band with the lightest representing 0–10% and the darkest representing 90–100%.

Using Twitter as a *source* of information is just one way to use it; it may be possible to also use Twitter for *analysis* purposes. Instead of simply using 'tweets' as a way of detecting articles, it might be worthwhile, for example, to use the *number* of tweets of an article to measure its

significance. Each tweet can be thought of as a vote of significance of an article—after all, someone has taken the time to tweet it. As just mentioned, HealthMap allows users to rate articles on a five–star ranking system for significance. This allows for the possibility of getting the benefit of the "wisdom of the crowds": if lots of people rank an article as significant, then it's quite likely to be significant. However, it seems that currently few votes are actually cast each day. An alternative approach would be to look at how many times an article has been 'tweeted'. This approach wouldn't rely on users of a system to *actively* contribute to it, rather it relies only on members of the general public to do what they are already doing—so it is more of a *passive* intelligence–gathering approach.

Another way to use the information in Twitter would be to look at how *few* 'tweets' there are that link to an article. If a system finds an article that has received very little attention on Twitter, then this might be reason for the system to bring the article to the attention of its users. It could be the information that everyone misses that is really important. Also, public health officials will typically already know about an issue that thousands of the general public are 'tweeting'.

Yet another way to use the information in Twitter would be to track the 'tweet' record of users who 'tweet' about articles that pertain to biosecurity. There are a number of ways in which this could be beneficial. For example, a given user may 'tweet' only about, say, food security issues, and this could be determined automatically by analysing all of the articles that they've 'tweeted'. If that's the case, then if there happens to be a new article that is difficult to classify automatically, a system could check if that particular Twitter user has 'tweeted' it, since that could be a good sign that the article is about food security.

5.5 Retrospective Comparison of GPHIN and ProMED

On recommendation by DAFF staff, GPHIN and ProMED were compared with respect to what they detected and reported on the following events:

- SARS, worldwide, 2003.
- Chikungunya, India, 2005/2006.
- Nipah virus, India and Bangladesh, 2007.
- *Taenia solium* —
- Japanese encephalitis —

No particular events associated with *Taenia solium* and Japanese encephalitis were recommended to be studied by DAFF staff, so GPHIN and ProMED were compared simply in terms of the number of reports they had on each disease. A second list of pest and disease events were also recommended by DAFF staff:

- New strains of UG99 in South Africa, May 2010 (plant disease).

- *Drosophila Suzukii* in the US, 2009 (plant pest).
- Guava/Myrtle rust in New South Wales, Australia, May 2010 (plant disease).
- African swine fever in the Caucus region, 2007 (animal disease).
- Bluetongue virus (BTV8) in Europe, 2006 (animal disease).
- Infectious myonecrosis, Brazil and Indonesia, mid June 2010 (marine disease).

The following compares GPHIN and ProMED's ability to detect and report information on these events.

5.5.1 SARS Worldwide Outbreak (2003):

ProMED
First report with 'SARS' in title or message was:

- 15 March 2003: PRO/ALL> Severe acute respiratory syndrome – Worldwide:alert

The first report relating to SARS was:

- 10 February 2003: PRO/EDR> Pneumonia - China (Guangdong): RFI

> [1] 'Date: 10 Feb 2003
> From: Stephen O. Cunnion, MD, PhD, MPH <cunnion@erols.com>
>
> This morning I received this e–mail and then searched your archives and found nothing that pertained to it. Does anyone know anything about this problem?
>
> 'Have you heard of an epidemic in Guangzhou? An acquaintance of mine from a teacher's chat room lives there and reports that the hospitals there have been closed and people are dying.'
>
> – Stephen O. Cunnion, MD, PhD, MPH
> International Consultants in Health, Inc
> Member ASTM&H, ISTM
> <cunnion@erols.com>'

GPHIN:
It is not possible to search GPHIN's archives back to 2003. The earliest report found in GHPIN's archive is:

- 02 July 2005: SARS, bird flu to be included as occupational diseases in HK.

However, GPHIN claims the 2003 SARS outbreak as a showcase for how useful the system can be. According to Keller *et al.* [2009], GPHIN detected the SARS outbreak 3 months before the WHO officially reported it, and before any other system.

5.5.2 Chikungunya, India, 2005/2006

ProMED

The first ProMED report on Chikungunya with India mentioned in the title is dated 20/02/06 (20060220.0551). The report was:

> 'Date: Mon 20 Feb 2006
> From: ProMED-mail <promed@promedmail.org>
> Source: Newsline, 17 Feb 2006 [edited]
> <http://cities.expressindia.com/fullstory.php?newsid=170348>
>
> NIV team detects Andhra Pradesh's mysterious fever, preliminary probe shows mosquito-transmitted chikungunya virus behind it
>
> ---
>
> The preliminary investigations into the mysterious fever that wreaked havoc in at least 5 districts of Andhra Pradesh in January this year [2006] have pointed towards a mosquito-transmitted virus as the cause behind it. This arthropod-borne virus – chikungunya virus – was found to be behind the mysterious fever by a team of scientists from the National Institute of Virology (NIV) here.
>
> The team that visited the districts and collected the samples has yet to finalize its report, but investigations have all pointed towards chikungunya virus. They will soon be sending a detailed report to the Union Ministry of Health.
>
> Chikungunya virus is transmitted by the _Aedes aegypti_ mosquito, the same species of mosquito that [transmits] dengue. High fever, chills, severe headache followed by acute joint and muscle pain are the symptoms of the infection. Fever persists for 3 days and pain for 7 days or more, depending upon the resistance of the patient.
>
> Chikungunya virus is highly infectious and disabling. The name comes from Swahili and means "that which bends up," a reference to the positions that victims take to relieve the joint pain. Chikungunya is responsible for extensive _Aedes aegypti_-transmitted urban disease in Africa and is also the cause of epidemics in Asia. The crippling arthralgia and frequent arthritis that accompany the fever and other systemic symptoms are clinically distinguishing.'

The only report that could count as being on a 'mysterious' fever is a report on suspected cases of leptospirosis, dated 23 January 2006 (20060123.0226). The report was:

> [1] 'Date: Wed 18 Jan 2006
> From: A-Lan Banks <A-Lan.Banks@thomson.com>
> Source: NewKerala.com [edited]
> <http://www.newkerala.com/news.php?action=fullnews&id=89819>
>
> Leptospirosis infection in Karnataka villages
>
> ---
>
> The mystery disease that has affected 258 people, including 156 women, at Siddammanahalli village in the Bellary district has been diagnosed as Lapto Psoriasis [as printed; the disease is, of course, leptospirosis, and the proper spelling will be used below. - Mod.LL]
>
> On Wed 28 Jan 2006, the National Institute for Communicable Diseases Deputy Director and Microbiologist, Dr. Sohan Lal, said that an unclean and unhygienic environment, a lack of a proper drainage system, and the drinking of contaminated water led to the outbreak of the disease. Animals and rodents were the main carriers of the bacteria that cause the disease.
>
> He said the disease was also prevalent at Yadgiri in Gulbarga district, some villages in Bidar district, Madanapalle in Andhra Pradesh, and some villages in Maharashtra.'
>
> ...

> Date: Mon 23 Jan 2006
> From: ProMED-mail <promed@promedmail.org>
> Source: WebIndia123.com [edited]
> <http://news.webindia123.com/news/showdetails.asp?id=227975&cat=India>
>
> 53 new cases of leptospirosis reported in Shahapur
> ———————————————
>
> As many as 53 new cases of suspected leptospirosis have been reported in 3 villages of Shahapur taluk in Gulbarga District. Official sources said here today [23 Jan 2006] that 38 patients from Kumakanur, 12 from Arjungi and 3 from Bilar are undergoing treatment for symptoms of the infection.
>
> Over 170 patients were treated for leptospirosis in the neighboring Badiyal and Wonegera villages in Yadgiur taluka in December 2005. Although the incidence of the disease has been brought under control in Badiyal and Wonegera, it has now spread to new areas. Sources said there are no primary health centers in Kumakanur, Arjunagi, and patients have to go to the one at Wadegera, which is 13 km away, or visit the Yadgir government hospital for treatment.

The moderator notes that 'The methods for diagnosis of leptospirosis are not stated nor is the amount of morbidity and/or mortality associated with this sometimes fatal infection linked to water contaminated with rodent (or other animal) urine.'

GPHIN
No reports.

5.5.3 Nipah Virus, India and Bangladesh, 2007

GPHIN — Nadia, India (Apr, 2007):
First possible signal, and two interpretations:

- 08 May 2007 — Alert: 'Rare Nipah virus outbreak kills 5 in eastern India'.
- 04 May 2007 — Article: 'Nipah virus threat'.
- 28 April 2007 — Article: 'Fear of new dengue'.

04 May article is the first to mention "Nipah".

GPHIN — Kushtia, Bangladesh (Mar–Apr, 2007):
First possible signals, characterised as encephalitis:

- 12 Aril 2007 — Article: 'Encephalitis kills 6 in western Bangladesh'.
- 11 April 2007 — Alert: 'Encephalitis kills 6 in Kushtia'.

Neither of these mention "Nipah".

GPHIN — Thakurgaon, Bangladesh (Jan–Feb, 2007):
First possible signals:

- 16 February 2007: 'Bangladesh mystery deaths not caused by bird flu: official'.

- 14 February 2007: 'Bird flu experts join probe of mystery Bangladesh deaths'.
- 12 February 2007: 'Couple dies of mysterious disease'.

Again, no mention of 'Nipah".

ProMED — Nadia, India (Apr, 2007):

First possible signal:

- 30 April 2007: PRO/EDR> Undiagnosed deaths - Bangladesh, India (02)

No mention of 'Nipah'. Mentioned as a 'possible dengue outbreak'.

ProMED — Kushtia, Bangladesh (Mar–Apr, 2007):

First possible signal and correct interpretation:

- 12/4/2007: PRO/AH/EDR> Undiagnosed deaths, encephalitis - Bangladesh (Kushtia): RFI

The moderator on this report notes that this could be henipavirus (genus of nipah virus):

> 'Upon reading the above newswire, this moderator is reminded of the types of descriptions of the newswires that accompanied the henipavirus outbreaks in Bangladesh and India in 2001, 2004 and 2005.'

ProMED — Thakurgaon, Bangladesh (Jan–Feb, 2007):

First possible signal was:

- 17/02/2007: PRO/AH/EDR> Undiagnosed illness - Bangladesh: RFI

Moderator comment:

> 'The description of the illness is too slight to hazard an opinion as to its nature at this juncture.'

Summary

In each outbreak — Thakurgaon, Kushtia, and Nadia — GPHIN was the first to report a (possible) signal for Nipah virus. However, ProMED was the first to realise that the disease was potentially Nipah virus when a moderator reflected on the similarity of the reports with previous reports in the area for henipavirus. This occurred on 12 April 2007, during the Kushtia outbreak.

5.5.4 *Taenia solium*

No particular outbreak of *Taenia solium* was recommended by DAFF staff for study. However, it appears GPHIN generally publishes more reports on the disease than ProMED:

GPHIN: 41 reports for 'Taenia Solium' — December 2006 to June 2010.

ProMED: 14 reports for 'Taenia Solium' — August 1997 to August 2009.

5.5.5 Japanese Encephalitis

No particular outbreak of Japanese encephalitis was recommended by DAFF staff for study. However, it appears GPHIN publishes substantially more reports on the disease than ProMED:

GPHIN: 1,378 reports for 'Japanese encephalitis' — February 2005 to June 2010

ProMED: 100 reports for 'Japanese encephalitis' — August 2006 to June 2010

5.5.6 UG99, South Africa, May 2010

GPHIN and ProMED both had initial reports on 26 May 2010 (though GPHIN had 3 other (but similar) reports whereas ProMED had just 1).

5.5.7 *Drosophila Suzukii, US 2009.*

Both systems had no results for 'drosophila suzukii' or 'spotted wing'.

5.5.8 Guava/Myrtle rust, New South Wales, Australia, May 2010

GPHIN

First report on 5 June 2010.

ProMED

No reports.

5.5.9 African swine fever, Caucas, 2007

GPHIN

First report on 6 June 2007.

ProMED

First report on 7 June 2008.

5.5.10 Bluetongue virus (BTV8), Europe, 2006

GPHIN

No reports

ProMED

First report mentioning BTV8 was 21 August 2006. First report of possible BTV in the Europe was 17 August. The report was a suspicion of BTV in the Netherlands. Confirmation was published 18 August 2006.

5.5.11 Infectious myonecrosis, Brazil and Indonesia, mid June 2010

GPHIN

No reports.

ProMED

No reports.

5.5.12 Conclusions

For SARS (2003) and Nipah virus (2007), GPHIN was the first to detect a possible signal of an outbreak, but in the latter case, ProMED was the first to supply an interpretation to the signal. There were no reports in GPHIN's archive for 'Chikungunya' dating back to 2006/2005 during the time of the outbreak in India. (It's not known whether this is due to limitations on the archive or because the system wasn't 'watching' for Chikungunya.) ProMED, in contrast, had several reports covering the event. GPHIN seems to have better coverage of *Taenia solium* than ProMED and substantially better coverage for Japanese encephalitis. Both systems reported on the new strains of UG99 in South Africa (May, 2010) on the same day. Neither system had any reports for *Drosophila suzukii*. GPHIN had one report on the Guava/Myrtle rust outbreak in Australia, though it was almost 1 month after the outbreak was first reported by the media. The systems performed roughly equally well with respect to African swine fever in the Caucas region (2007. ProMED did much better than GPHIN on Bluetongue virus (BTV8) in Europe (2006), with substantial coverage and GPHIN having no reports. Neither system had any reports for infectious myonecrosis in Brazil and Indonesia (2010)/

5.6 Conclusions

To varying degrees, the existing systems can be used to meet some of DAFF's biosecurity intelligence needs.

ProMED has at least some coverage of the pests and diseases (chosen for this study) that are relevant to the plant, animal, livestock, and marine areas of DAFF (although focus on plant and marine is quite limited). The system provides short–term, medium–term, and some occasional long–term intelligence on these topics. The system is easy to use, and information in the system can be accessed and filtered in a number of ways. If approved by the moderators, users can contribute information to the system.

GPHIN was the first system of its type, and is credited by WHO as being the first to identify the 2003 SARS outbreak. However, the system's web interface is seriously outdated. Users using some of today's most popular web browsers—Internet Explorer 8, Firefox, Safari, Chrome—cannot even access the system. To use the system, one must be using Internet Explorer 6–7, which means one also must be using Windows XP or a version of Windows 7 or Vista that can run in XP mode. The system interface is not streamlined, and there is no filtered mapping service—unlike HealthMap or

BioCaster—even though all reports are geocoded. The system covers animal, human, and zoonotic diseases, and has very limited coverage of plant and marine issues. Access to reports can be only obtained via the web interface, although minimal alerts can be obtained through an e–mail subscription service. GPHIN can be used to serve some of DAFF's biosecurity intelligence needs. However, it appears that many of the services that GPHIN can provide can also be provided by the other systems reviewed in this report.

Of the three purely automated systems with mapping features, HealthMap seems to have the highest information integrity. It stands out as the system with fewest location extraction errors and irrelevant reports. The system also takes advantage of user inputs by taking submissions of reports and eye–witness–accounts through its website, and also through its iPhone and Android applications. This gives the system a source of biosecurity intelligence that the other automated systems currently do not share, although these 'eye–witness' reports tend to be of human diseases or natural disasters rather than of animal or plant biosecurity events of interest to DAFF staff. It also gives the system an ability to improve continually the quality of its information—by users ranking and adding comments to reports. These features give the system a high degree of integrity and the capacity for the international community to share biosecurity intelligence easily (and anonymously).

However, HealthMap does not cover reports written in languages from the Asia–Pacific region, with the exception of Chinese. Identifying reports of diseases and pests in the Asia–Pacific region is one of DAFF's biosecurity intelligence needs. BioCaster specialises in identifying reports for the Asia–Pacific region—its ontology was built with this focus (Collier *et al.* [2006])—and the system currently performs intelligent scans for articles in Chinese, Japanese, Korean, Thai, and Vietnamese.

BioCaster serves another of DAFF's intelligence needs. Since BioCaster performs ontology–based searches, it has a particular ability to search for disease signs and causal agents, without searching for particular diseases. Combined with its focus on the Asia–Pacific region, this means BioCaster has a particular ability to detect reports on unknown or unidentified diseases in the Asia–Pacific region. This feature of the system could be useful for those who wish to detect outbreaks of unknown diseases in this region.

EpiSPIDER takes advantage of a new potential source of biosecurity intelligence: Twitter. The study in Section 5.2 indicates that this is a reliable source of biosecurity intelligence. It should not be surprising that this is the case. Each report picked up by the system through Twitter originates with some individual who has read the report and identified it as relevant to biosecurity in some way, and effectively categorised the report with a 'hashtag'.[7] Even though each individual may only 'tweet' a few reports, the overall effect is thousands of reports on biosecurity–related topics each day being pre–analysed by humans of varying expertise. This is a potentially valuable source of high quality information, which can be further analysed by software and human experts.

[7]Hashtags are used by the Twitter community to group 'tweets' together under topics. For example, if a user writes a message about dengue fever, they might include '#dengue' in the message to denote that it is abut dengue.

WDIN reports are focused largely on the US. Reports by the system are almost always relevant and accurately plotted—due in large part to the fact that all reports are scanned by a human analyst. WDIN reports are now incorporated into HealthMap, so there is no need to watch WDIN in addition to HealthMap.

With the exception of EUROPHYT, NAPIS, and NAPPO, none of the systems pay much attention to plant pests and diseases. It is not possible to obtain access to EUROPHYT at this time, and NAPIS and NAPPO both focus on North America. The next section, shows that it is possible to develop an open–source biosecurity intelligence system for plant pests and diseases with international coverage. A intelligence system for weeds is currently being developed by Christopher Auricht (see http://www.auricht.com/awrc/).

There is also very little attention paid to marine pests and diseases. With the exception of OIE and ProMED, none of the systems reviewed in this report covers marine pests. ProMED has some coverage of pests and diseases of aquatic animals, and OIE has slightly more coverage, however its user interface is not very streamlined. To obtain intelligence on an outbreak from OIE, one must use a series of drop–down menus, or scan e–mails sent through the e–mail subscription service. There is no interactive mapping service. EpiSPIDER helps fill this gap by including OIE reports in one of its KML feeds, however no OIE reports on pests or diseases of aquatic animals detected by EpiSPIDER were observed during the time of Stage 2 of the project.

Although more attention is paid to animal, human, and zoonotic diseases, most of the systems reviewed in this report provide little coverage of the more obscure or even unknown diseases of these types—with the exception of BioCaster. All of the systems, with the exception of EpiSPIDER and HealthMap, scan standard news media reports. It is expected that a significant amount of biosecurity intelligence exists in other online forms and is currently not being detected by the systems reviewed in this report—e.g., forums, discussion boards, blogs, etc.

Various combinations of the systems reviewed in this report can serve a number of DAFF's needs. For example, those in the Animal Division of DAFF may satisfy most of their biosecurity intelligence needs by systematically using ProMED, HealthMap, BioCaster, and OIE. None of the systems reviewed meet all of DAFF's biosecurity intelligence needs. However, this is to be expected, given how wide and varied those needs are.

6 Intelligence–Gathering on Plant Pests and Diseases using Yahoo Pipes

6.1 Introduction

As mentioned earlier (Section 5.6), there is very little coverage of plant pests and diseases by the existing systems. This is particularly true in an international setting. The systems reviewed in this report that cover plant pests and diseases to any extent are EUROPHYT, NAPIS, NAPPO, and ProMED. The first three of these systems have restricted geographical foci. EUROPHYT is restricted to Europe, NAPIS is restricted to the US, and NAPPO is restricted to North and Central America. In addition, access to EUROPHYT is restricted to officials of member states. ProMED does have some international coverage of plant pests and diseases, although this is not its main focus. These systems are not automated; they rely on manual collection, analysis and reporting of information. There is also currently no mapping system for plant pests and diseases that could serve as a plant health analogue to HealthMap, BioCaster, WDIN and EpiSPIDER.

One reason why so little attention is paid to plant pests and diseases is probably because they don't usually generate the same public concern and attention as human and animal diseases. There may simply be very little information on plant pests and diseases for an online system to detect, or the information may be of a form such that it is very difficult for a system to detect. However, although plant pests and diseases may attract relatively less public attention, biosecurity information exists online that could be of use for intelligence–gathering and analysis in DAFF. A variety of sources of information on plant disease and health could be automatically captured, aggregated and analysed—and even plotted on a map—including articles recorded by news aggregators, predefined searches on Google news, RSS feeds to contents of journals and magazines, RSS feeds to a variety of blogs, ProMED reports, and content that is 'locked' in webpages (such as the NAPPO Official Pest Alerts). NAPIS already generates some RSS news feeds for the pests that it covers (`http://pest.ceris.purdue.edu/pestlist.php`).

An automated plant pest and disease intelligence system that is akin to the systems in the human and animal disease domains is likely to be of great utility to DAFF. The absence of such a system and the presence of information that can be detected by such a system suggests that it is worthwhile exploring the possibility of developing one. As an initial attempt, mimicking one of the systems in the human and animal disease domain could be the easiest way to proceed.

Out of the automated systems that cover animal and/or human diseases reviewed in this report, WDIN is based on the simplest technology. As mentioned in that system's review (Section 4.6), a crucial part of the WDIN's data collection and analysis involves using Yahoo Pipes, a web–based tool that collects, organises, and analyses online content, and that can be used to publish that content in a variety of ways. WDIN uses Yahoo Pipes to collect RSS feeds on wildlife diseases from a large number of sources, merge those feeds, filter them for relevance, and produce one final feed that a human expert then examines. This same approach could be used for plant pests and

diseases to build a prototype biosecurity intelligence system for plant health. Using that prototype it would then be possible to see if a system for plant pests and diseases could have the same sort of success that systems such as HealthMap, BioCaster, WDIN, and EpiSPIDER have achieved.

There are two main—and complementary—ways that Yahoo Pipes can be used to build a rudimentary plant biosecurity intelligence system for plant health. The first is simply by collecting RSS feeds from a number of sources such as journals, news aggregators, blogs and predefined searches. These feeds can be merged, organised, and further filtered in a variety of ways. This would be a straightforward replication of the WDIN data collection and analysis process. However, Yahoo Pipes also has the ability to extract location information from the contents of feed items, thus enabling the possibility of automatically plotting feed items on an interactive map.

Yahoo Pipes also offers 'web scraping' tools that can automatically extract information from webpages. It also has tools that allow that information to be structured and organised in various ways. For example, the Regex Module allows items in RSS feeds and pipes to be manipulated in numerous ways, and then to publish that information into RSS feeds, which can then have location information added and be merged with feeds from other sources. This makes it feasible to build a map displaying reports on plant pests and diseases that includes content that is 'locked' in webpages, such as the NAPPO Official Pest Alerts.

The following two sections demonstrate these approaches to create a rudimentary open–source biosecurity intelligence system for plant health. The next section shows how the Official Pest Alerts published on the NAPPO website can be 'unlocked', have their locations extracted, and be imported into a mapping system such as Yahoo Maps or Google Earth. The following section then shows how a pipe can be built to aggregate a large number of sources on plant pests and diseases and filter them to exclude irrelevant reports. Some preliminary statistical tests are performed on the results of using the pipe, to demonstrate that even a very simple pipe can be effective.

6.2 NAPPO Pest Alerts Unlocked and Mapped

Figure 6.2 is a screenshot of the Official Pest Reports published by NAPPO's Phytosanitary Alert System. To obtain the information in NAPPO reports, the user needs to go to the NAPPO website and click on the links. To know if a new report has been published, the user needs to check the page (or subscribe to the e–mail list). By just looking at the list, it is hard to achieve so–called 'situational awareness' whereby brief inspection results in the observer understanding the importance of the information.

However, using Yahoo Pipes, one can build a pipe that automatically collects this information, and extracts the date and location information in the report to create a geotagged RSS feed that can be used to plot the Official Pest Reports on a map. Two screenshots of such a feed, plotted using the Yahoo Maps API are included in Figure 28. The result can also be imported as a KML feed into Google Earth; the screenshot is included in Figure 40. The pipe can be accessed using the following url: `pipes.yahoo.com/asl4acera/nappo_getaround`.

Country	Title	Posted
	Anastrepha ludens (Loew) (Mexican fruit fly), Removal of Quarantined Area in Cameron County, Texas – United States	05/11/2010
	Detection of gladiolus rust, *Uromyces transversalis* (Thum.), in Manatee County, Florida - United States	05/05/2010
	Detection of Huanglongbing (*Candidatus* Liberibacter asiaticus) in the Municipality of Tecoman, Colima, Mexico	04/30/2010
	Bactrocera dorsalis (Oriental fruit fly) - Removal of Quarantined Area in Los Angeles County, California	04/28/2010
	Tilletia indica (Mitra) Mundkur (Karnal Bunt) – Removal and Addition of Quarantine Areas in Arizona, California, and Texas, United States	04/19/2010
	Detection of Huanglongbing (*Candidatus* Liberibacter asiaticus) in the Municipality of Calakmul, Campeche, Mexico	04/12/2010
	Detection of orange rust of sugar cane (*Puccinia kuenii*) in Mexico	04/12/2010

Figure 38: A sample of NAPPO's Official Pest Reports.

Such a feed can be merged with other feeds, for example the RSS feeds from NAPIS, or RSS news feeds from Google, or the RSS feeds of plant pathology journals. These can also have location information added to them. The result can be a rich source of up–to–date, geocoded reports on plant pests and diseases.

6.3 News Feeds and ProMED Reports

This section describes how a number of different feeds can be aggregated to generate a single stream of information on plant pests and diseases.

The following feeds are first collected:

- ProMED — interglacial.com/rss/promed-mail.rss
- New Scientist, Environment — feed://feeds.newscientist.com/environment
- New Scientist, Science News — feed://feeds.newscientist.com/science-news
- The Guardian — feed://feeds.guardian.co.uk/theguardian/environment/rss
- Biology News — feed://feeds.biologynews.net/biologynews/headlines
- Digg, Environment — feed://feeds.digg.com/digg/topic/environment/popular.rss
- Discover News — feed://feeds.feedburner.com/DiscoveryNews-Top-Stories
- Google News: 'plant+disease' — feed://news.google.com/news?um=1&cf=all&ned=us&hl=en&q=plant%2Bdisease&cf=all&output=rss
- Google News: 'plant+pest'— feed://news.google.com/news?um=1&cf=all&ned=us&hl=en&q=plant%2Bpest&cf=all&output=rss
- Google News: 'plant+pests' — feed://news.google.com/news?um=1&cf=all&ned=us&hl=en&q=plant%2Bpests&cf=all&output=rss
- Google News: 'canker' — feed://news.google.com/news?um=1&cf=all&ned=us&hl=en&q=canker&cf=all&output=rss
- Google News: 'gypsy moth' — http://news.google.com/news?um=1&cf=all&ned=us&hl=en&q=gypsy+moth&cf=all&output=rss
- Google News: 'UG99' — http://news.google.com/news?um=1&cf=all&ned=us&hl=en&q=UG99&cf=all&output=rss
- Google News: 'fire ant' — http://news.google.com/news?um=1&cf=all&ned=us&hl=en&q=fire+ants&cf=all&output=rss

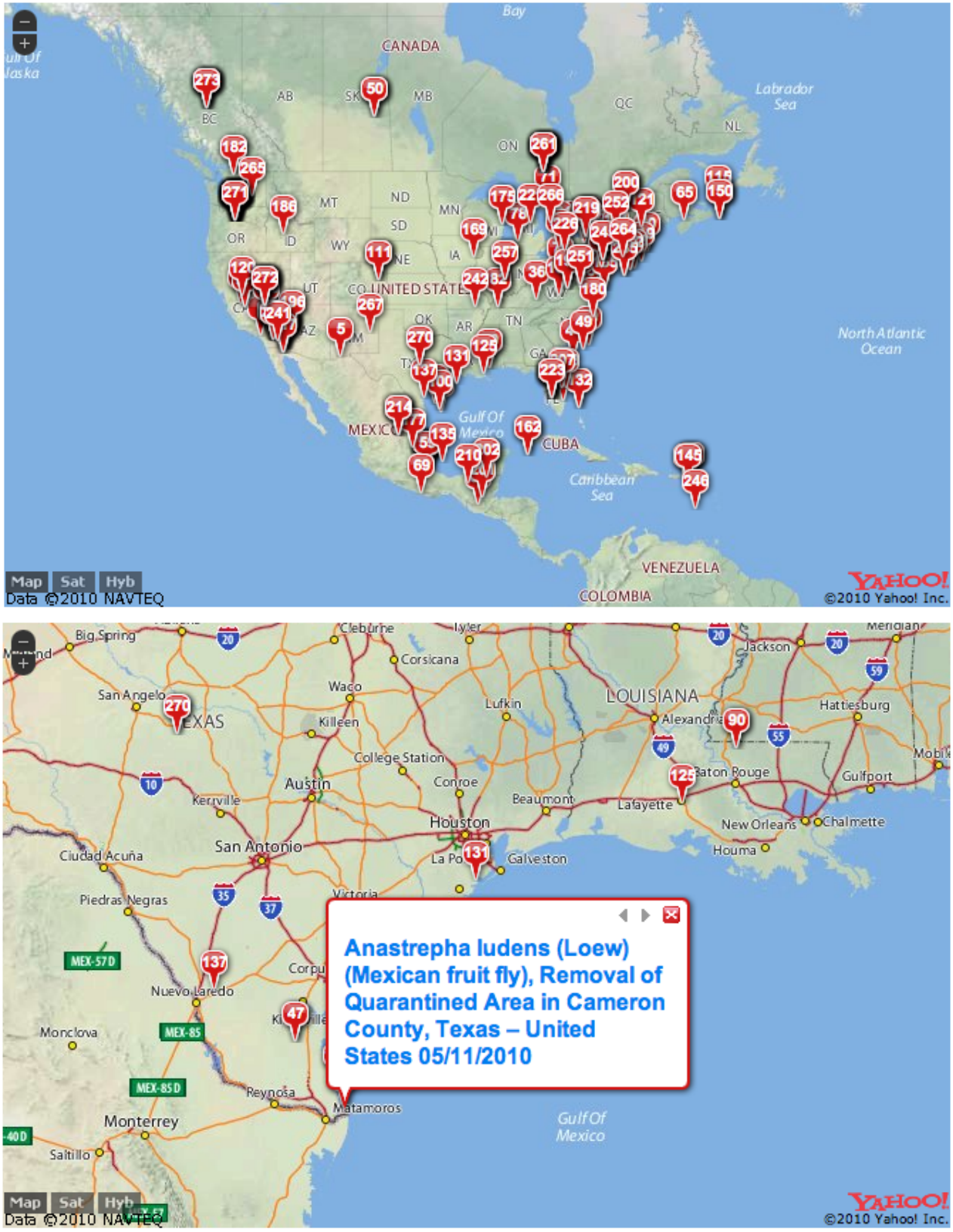


Figure 39: NAPPO's Official Pest Reports Map.

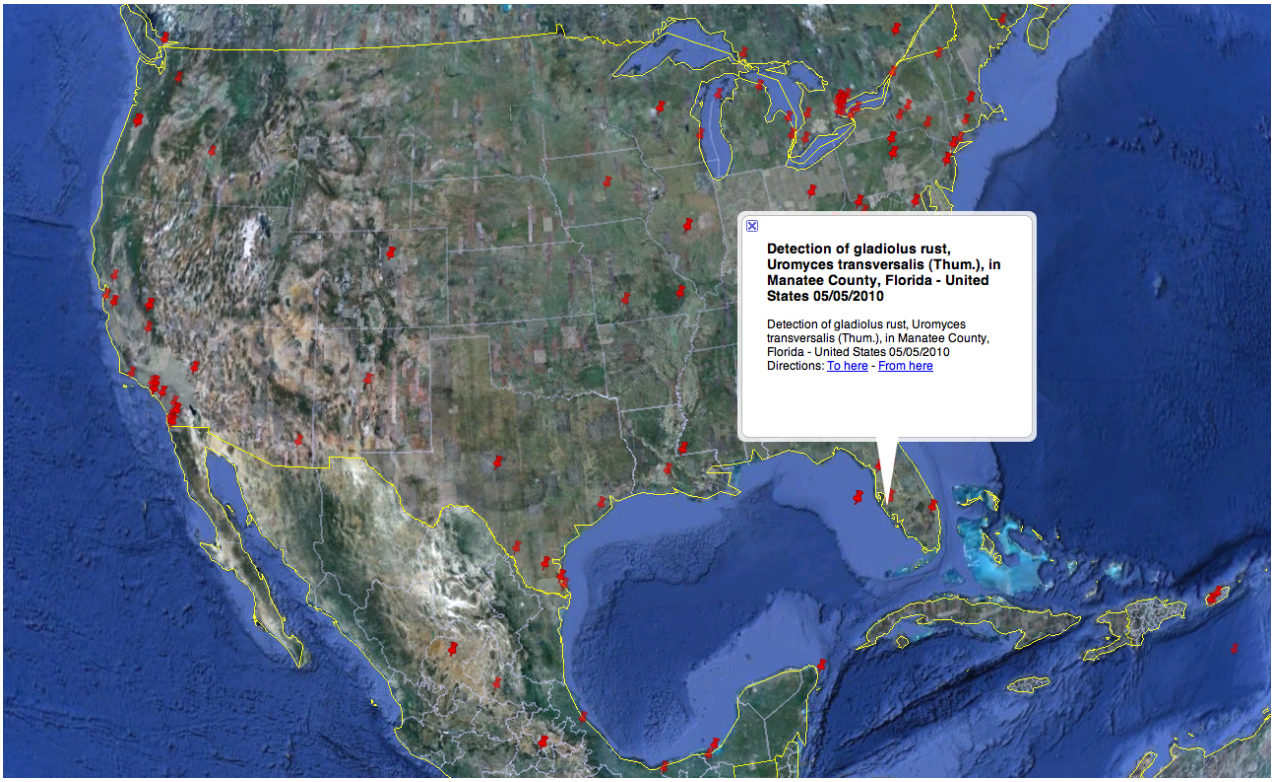


Figure 40: NAPPO's Official Pest Reports Map in Google Earth.

- Google News: 'citrus greening' — feed://news.google.com/news?um=1&cf=all&ned=us&hl=en&q=citrus+greening&cf=all&output=rss
- Google News: 'leaf spot' — http://news.google.com/news?um=1&cf=all&ned=us&hl=en&q=leaf+spot&cf=all&output=rss

This is essentially what WDIN does for wildlife diseases—albeit on a much larger scale.

These feeds are then merged and duplicate items are removed. A crude filter is then applied to remove irrelevant items. Items with any of the following terms are blocked: swine flu, bird flu, H1N1, fever, anthrax, avian. This constitutes the source side of the pipe.

Another pipe then applies a second layer of filters to the feed. Items that contain any of the following terms are allowed through—all others are blocked: guava rust, gypsy moth, canker, cassava mosaic, ramorum, lyme, citrus greening, psyllid, ramora, fire ant, UG99, wilt, leaf spot. Items are then sorted by date and finally sent to the pipe output. Figure 41 includes an example screenshot of the results produced by the pipe.

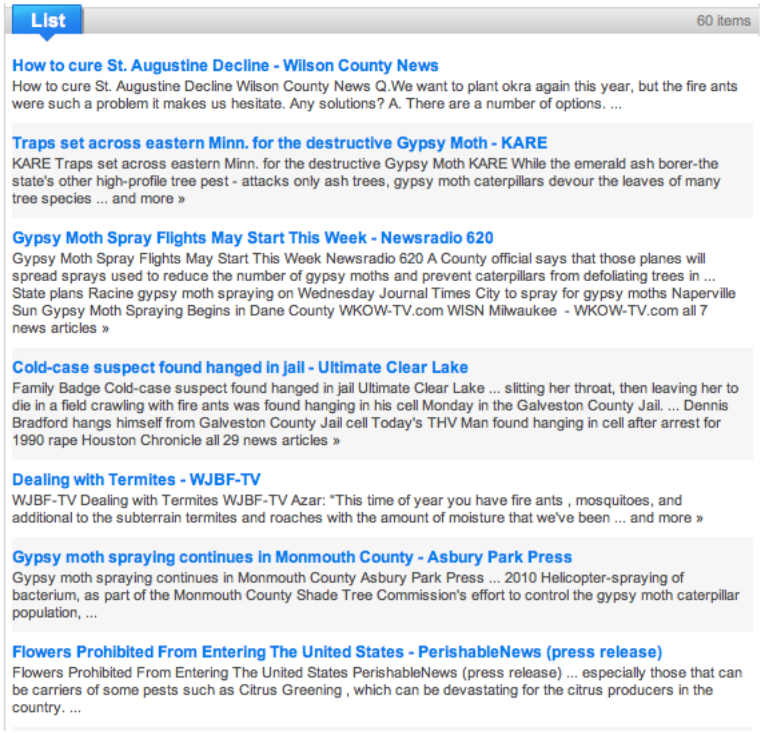


Figure 41: Sample of Results Produced by Plant Disease Pipe.

Despite the small number of sources, limited range of search terms, and crudeness of the filters, the pipe produces a significant number of reports per day, with surprising accuracy (i.e., surprisingly few irrelevant reports were produced). From a sample over ten days of results, the pipe produced an average of 11 reports per day, of which 69% were relevant to plant pests and diseases in some way. Of these relevant reports, 55% were about some specific event, and 45% were general news about plant pests and diseases (Figure 42).

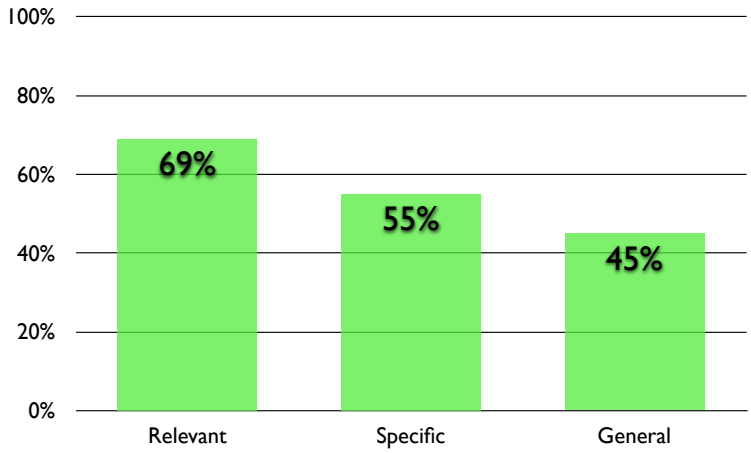


Figure 42: Percentage of reports that were relevant, and percentages of these that were about a specific event or general news.

Results from the pipe focused largely on events in the US. Of the relevant reports, 67.1% were focused on the US, 10.5% on the UK, 6.5% on Canada, 5.2% on Australia, and 9.2% on other, including general global news (Figure 43). These figures are somewhat misleading because a large number of reports for the UK and other regions clustered around single events that were newsworthy enough to make it into the news media in the US. Many of the UK reports were clustered on the UK introducing an insect to combat Japanese knotweed, and many of the 'other' results were clustered on Korea beginning to export tangerines again, after a 15–year hiatus.

These results suggest that with an improved list of search terms, a wider variety of sources for plant pest and disease news, sources inlcuding other languages, and a more sophisticated pipe, a substantially better prototype could be built. If substantial improvements could be made by moving in these directions, this would suggest that developing a system from the ground up (i.e., not relying on Yahoo Pipes), might be worthwhile.

6.4 Conclusions and Future Research and Development

By studying the strengths and weaknesses of a number of existing open–source biosecurity intelligence systems that focus on animal, human, and zoonotic diseases, it is possible to build a more sophisticated intelligence system for plant pests and diseases. The example outlined in the preceding section demonstrates that building such a system is feasible by copying the basic technology that WDIN uses and building a short list of search terms. The quality of the output can be dramatically improved in a number of ways.

Improving the list of search terms is an important first step. Expanding the list of terms will result in more reports being brought into the system. Refining the list of terms will result in less

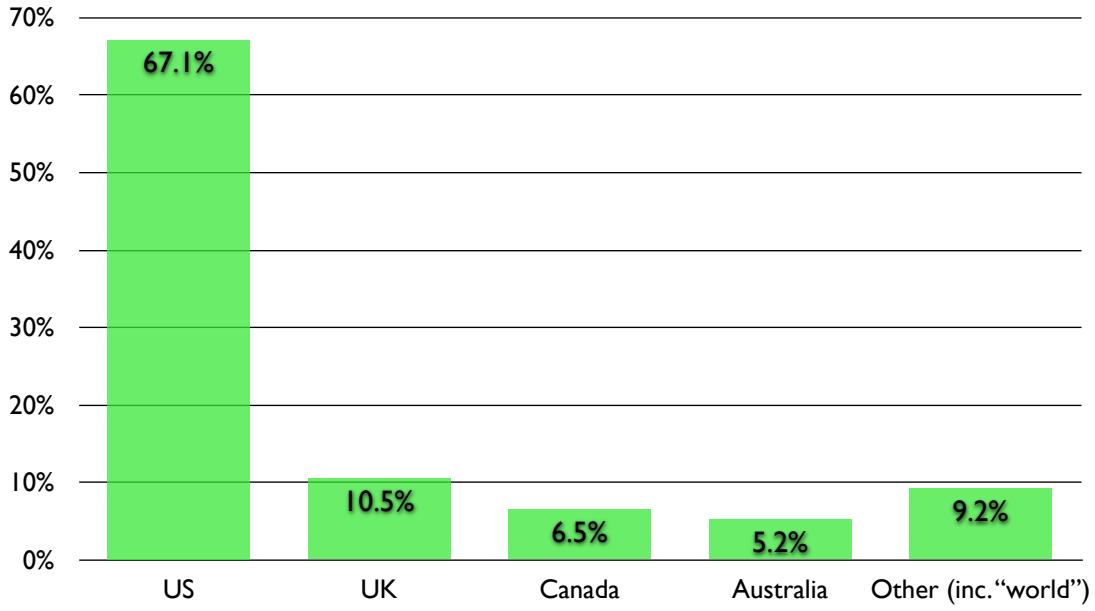


Figure 43: Geographical coverage of the pipe.

noise (e.g., the Google News feed on 'canker' systematically results in irrelevant reports, since 'canker' is used in English in a number of different ways). Adding structure to the list of terms, in the form of an ontology, will simultaneously reduce noise and increase the number of reports. BioCaster has demonstrated the utility of using such an ontology. Research in this direction has already begun as part of Stage 7 of this project. However, the list of search terms and their structure has generated too much information for Yahoo Pipes to run the sophisticated pipe robustly. Future development may require writing the software from the ground up, instead of relying on Yahoo Pipes' interface and servers (e.g., using Python, which has a rich set of webscraping tools).

Expanding the list of sources is another obvious way to improve the system. This could be done by including more search engines and news aggregators, such as Meltwater, Moreover, European Media Monitor, and DayLife. It would also be advantageous to include content from plant pathology journals, especially those that publish articles in a timely manner (e.g., New Disease Reports and Plant Pathology Journal). In some cases, this is a simple matter of collecting the RSS feeds for those journals. In other cases, RSS feeds are not available and some sort of web–scraping tools would need to be used to unlock that content—in the way that Yahoo Pipes was used to unlock the NAPPO official alerts (Section 6.2).

As mentioned earlier (Section 4.4), HealthMap stands out as a system that accepts input direct from users. A plant plant biosecurity intelligence could use the same sort of Web 2.0 approach. This would have a number of advantages. First, users could report relevant news articles that the system might miss. Second, users could reject irrelevant articles that the system might have

accidentally included. Third, users could also add semantic content to reports by classifying the articles. Fourth, users could report their own eye–witness accounts. Fifth, users could contribute to the quality of the information in the system by verifying the reports. For example, an amateur gardener could photograph a plant showing disease and upload it to the system with GPS coordinates. Another user who is an expert in plant pathology could then suggest a diagnosis and link the report to relevant online articles (if there are any). Other users could then signal whether they agree with the suggested diagnosis and comment on differential diagnoses. However, for such a social media approach to work, it is crucial that such a system is open and free to everyone.

It is also important to move beyond a system based on simple searches for key terms. By applying various techniques from computational linguistics and natural language processing, the quality of the information in the system could be dramatically improved. Techniques such as looking for clusters of synonyms could be used to rank articles in terms of probability of relevance. Allowing the system to 'learn' by updating a 'Bayesian prior' would allow it to improve in quality over time, and respond to changing circumstances. These methods have been well studied in computer science, linguistics, and statistics and are already being used in varying degrees by BioCaster, EpiSPIDER, and HealthMap. Such probabilistic methods can be combined with the social media approach described in the preceding paragraph by using results from formal social epistemology. Probabilistic models from the philosophical literature on judgement aggregation and consensus formation could be used effectively in such a domain (see ACERA Project 607 Report, *Evaluation and Development of Formal Consensus Methods*).

By taking advantage of user inputs and probabilistic approaches, it should be possible to approximate the same effect as having human curators of the system, with lower cost. A significant advantage to curated systems such as ProMED or GPHIN is that they have human experts contributing valuable analysis to reports. However, this comes at a price: the price of employing those experts, the price of being limited to what those experts can accomplish, and in some cases, a delay in publishing outputs. An analogy between biosecurity intelligence systems and encyclopaedias is useful here. ProMED and GPHIN are to Encyclopaedia Britannica as HealthMap and EpiSPIDER are to Wikipedia. The latter capitalise on the power of thousands of distributed users adding and correcting information, at the price of some error. The former may contain less error, but at the cost of timeliness, paying staff, and the expense of being restricted to the limited processing abilities of that staff. Interestingly, in investigating how the prototype from Section 6.3 could be improved, it was discovered that some DAFF staff are successfully using Wikipedia to obtain some biosecurity intelligence—especially the versions of Wikipedia for other languages, as they often contain content that is not available anywhere else.

Finally, it should be emphasised that although the primary focus here has been restricted to plant pests and diseases, future research and development should also focus on systems that cover marine pests and pests and diseases of aquatic animals—as well as other animal, human, and zoonotic pests and diseases. With the exception of ProMED and OIE, marine pests and diseases receive no coverage, and while animal/human/zoonotic diseases are better covered, the systems

tend to focus on the already well–known diseases. It would be beneficial to DAFF if intelligence systems along the lines just described were developed for these diseases too.

7 Prototype Marine Biosecurity Intelligence System

There is also very little coverage of marine pests and diseases. This section documents a prototype marine biosecurity intelligence system that was developed during the course of the project. This system was based on a program written from the ground up using Python and it implemented a number of features that the biosecurity intelligence systems reviewed in this report use.

The system used two types of sources: search sources and trusted sources. The trusted sources consisted of a range of RSS feeds and websites devoted to marine pest and disease news. The search sources were Google News, Google Blogs, Moreover News, Moreover Blogs, and Twitter. These sources were used to conduct searches based on a preselected list of search terms, which were refined over the course of the project in collaboration with Geoff Grossel from DAFF.

Each day the system acquired a list of news articles from the trusted and search sources. Articles that were more than 1 day old were removed, since they were assumed to have been recorded in the previous day's scan. Then each article was tagged with the search term that resulted in it, if it came from a search source. Repetitions of articles were then removed, with the tags of duplicate articles being combined. This resulted in some articles receiving multiple tags, which allowed the user to quickly work out what the article was about by scanning those tags. Using the search results to tag articles in this way provided a degree of content analysis, but without the need of doing any actual content analysis.

The links for each of the articles were sent to Alchemy's API to extract the main text of the articles from their webpages and to detect any locations mentioned in the article. (EpiSPIDER also uses Alchemy to conduct its extract locations from the articles it finds.) Each article and its tags were combined with the extracted text and locations to form a report. Reports were then filtered for relevance by removing any reports that contained terms in a preselected list of 'irrelevant terms', which was refined over the course of the project in collaboration with Geoff Grossel. Daily lists of reports were emailed to some DAFF staff members. They were also published on a map using Google's Map API and Google Docs. See Figure 44 for a snapshot of the map.

From 23 August, Geoff Grossel compiled a weekly logsheet that documented the successes and failures of the system. By comparing how well the system fared against other systems in the review and also against his manual scans, it was possible to determine the strengths and weaknesses of the system. By the end of the course of the project, the system was finding nearly everything that the manual scans were finding, and vastly outperforming the other systems.

There are a number of ways in which the system could be dramatically improved:

1. Develop and use a multilingual ontology, like that of Biocaster's.

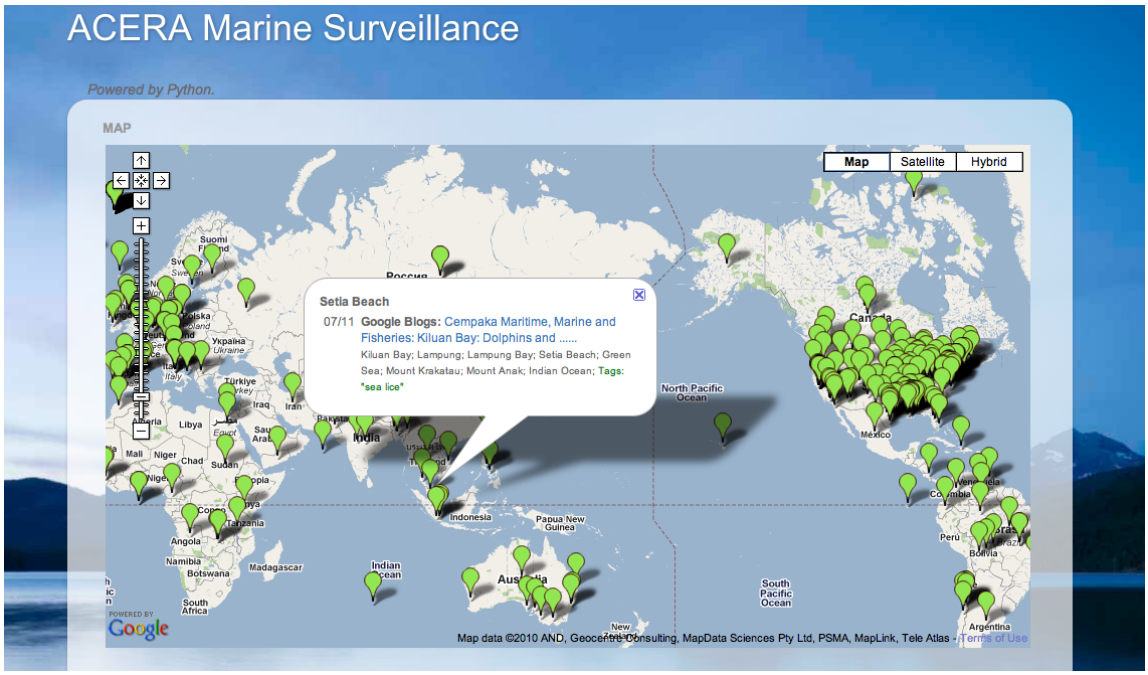


Figure 44: Snapshot of the prototype marine biosecurity intelligence program.

2. Incorporate some natural language processing to determine the content of articles and their relation to each other.
3. Expand search sources and search terms.
4. Expand trusted sources by discovering more RSS feeds and using webscraping tools.
5. Watch and analyse market prices for sudden changes. Changes in market prices can be good indicators of outbreaks of marine diseases.
6. Use Twitter to filter out articles that contain already known information. If an article is receiving a lot of attention on Twitter, then DAFF staff probably already know about it.
7. Develop a web interface for the system in order to allow users add their own commentary, tags, build their own maps and set up customised alerts.
8. Develop a retrieval function, so that users can search for old reports.
9. Develop visualisation and trend analysis tools, such as dynamic choropleths and wordclouds.

Since the space for a marine biosecurity intelligence system is currently unoccupied, the development of one could be of significant benefit to DAFF.

8 Biosecurity Intelligence in Other Countries

This section briefly documents some of the biosecurity intelligence–gathering and analysis practices of some other countries. The term 'biosecurity' tends to be used to refer to issues involved with defence against bioterrorism. 'Biosurveillance' is more commonly used to refer to issues involved with detecting/monitoring/forecasting pest and disease outbreaks.

8.1 Canada

As mentioned earlier, GPHIN was developed by the Public Health Agency of Canada, who have a number of other intelligence programs.[8] The Canadian Food Inspection Agency (CFIA) also have the Canadian Animal Health Surveillance Network (CAHSN). CASHN is establishing a network of federal, provincial and university animal health diagnostic laboratories to improve the capacity to quickly detect emerging animal disease threats.[9] Apart from GPHIN, no open–source biosecurity intelligence systems used by Canada were found.

8.2 New Zealand

AsureQuality is a company owned by the New Zealand government that provides food safety and biosecurity services to the food and primary production sectors.[10] Part of their biosecurity information service includes the use of AgriBase.[11] Their website describes AgriBase as follows:

> "AgriBase is a voluntary system and AsureQuality is continually in touch with rural properties in order to collect and update information. In the event of a rural, regional or national emergency AgriBase can be quickly populated with any additional data necessary to control and manage the situation. Because AgriBase is fully linked with geospatial systems, real-time analysis and problem solving can be easily and expertly handled.
>
> Geospatial technology allows us to render information from AgriBase (such as property location, size, management, operations, land use, stock numbers, homestead location, farm gate points and address points) into virtual features for display and geospatial analysis."

No open–source biosecurity intelligence systems used by New Zealand were found.

8.3 UK

The Department for Environment Food and and Rural Affairs (DEFRA) conducts international disease monitoring.[12]. They get their information from the Veterinary Administrations of their trading partners, including EU Member States, as well as the European Commission; the OIE; the reports of reference laboratories; High Commissions and Embassies; referenced publications. Their

[8] http://www.phac-aspc.gc.ca/surveillance-eng.php
[9] http://www.inspection.gc.ca/english/anima/surv/cahsnrcsze.shtml
[10] http://www.asurequality.com/about-us/company-overview.cfm
[11] http://www.asurequality.com/online-customer-services/search-agribase-database.cfm
[12] http://www.defra.gov.uk/foodfarm/farmanimal/diseases/monitoring/index.htm

website also mentions that they get information from a range of other sources, including Internet searches. Apart from that, it doesn't appear there is any formal biosecurity intelligence–gathering and analysis program.

8.4 US

The National Bio–surveillance Integration System (NBIS) was established by the Department of Homeland Security (DHS) in 2004. The purpose of the system is

> "to provide early detection and situational awareness of biological events of potential national consequence by acquiring, integrating, analyzing, and disseminating existing human, animal, plant, and environmental biosurveillance system data into a common operating picture that represents a comprehensive depiction of the global biosurveillance security environment."[13]

By 2008, the system was renamed as 'Biosurveillance Common Operating Network (BCON)'.[14] It now appears to be called 'Biosurveillance Common Operating Picture (BCOP). Grady *et al.* [2008] report that NBIS (BCON/BCOP) is:

> "an application designed to facilitate the identification of events of national significance through the harvesting of open–source information and to facilitate the collaboration on and dissemination of the details of these events with a broader community." Grady *et al.* [2008], p. 23.

Grady *et al. et al.* [2008] describes the system in further detail. It appears to have some similarities with the systems reviewed in this report. For instance, it's built upon an ontology for infectious diseases (Grady *et al.* [2008], p. 26) and uses open–source information. The system doesn't appear to be accessible by members of the public.

A report in 2007 by the DHS concluded that the program had, to that date, been mismanaged.[15] More information about the system can be found at: http://www.emergencymgmt.com/health/Biosurveillance-Common-Operating-Picture.html. It was apparently brought online in 2009 as the H1N1 pandemic began.

9 Conclusions

This report has reviewed a wide range of open–source online biosecurity intelligence systems: Google Flu Trends, GPHIN, ProMED, HealthMap, EpiSPIDER, WDIN, BioCaster, NAPIS, EUROPHYT, GAINS, OIE, and NAPPO. None of these systems completely meet DAFF's biosecurity intelligence needs, falling short in a number of respects. They fall short in terms of the pests and diseases that they cover. There is little coverage of plant, marine, and aquatic pests and diseases. And while there is better coverage of animal, human, and zoonotic diseases, the systems tend to

[13] http://www.dhs.gov/xlibrary/assets/mgmt/e300-prep-nbis2008.pdf

[14] http://www.archives.gov/records-mgmt/rcs/schedules/departments/department-of-homeland-security/rg-0563/n1-563-08-018_sf115.pdf

[15] http://www.dhs.gov/xoig/assets/mgmtrpts/OIG_07-61_Jul07.pdf

focus on diseases and pests that are well–reported in the media (e.g., H1N1). The systems also fall short in their geographical coverage. The systems tend to focus on the US and other Western countries, and some exclusively focus on the US and/or other Western countries (e.g., NAPIS). Many of the systems also fall short in terms of there ability to meet DAFFS intelligence needs in regards to long–term time frames. This is because many of the systems do not allow users to access archive information beyond a moving wall that's typically not longer than 90 days. Although the systems do not completely meet DAFF's intelligence needs, various combinations of them do so to some degree. Indeed, DAFF staff already use some of the systems on a regular basis, e.g., ProMED, OIE, and (to a lesser extent) GPHIN.

Although there are currently no existing open–source online biosecurity intelligence systems that meet DAFF's intelligence needs, it was shown that such systems can most likely be developed. Prototype systems for plant and marine biosecurity were developed with promising initial results. Some other countries have developed their own systems—e.g., the Public Health Agency of Canada first developed GPHIN as a prototype and the US Department of Homeland Security has recently developed BCOP. It's therefore plausible that DAFF could develop its own systems and that such systems would be of great benefit to DAFF. The prototype systems for plant and marine biosecurity could be further improved by expanding search terms and sources, by creating a web interface to allow users to interact with the system and add or modify its information, add retrieval functions so that users can search for old reports in order to find trends, develop visualisation and trend analysis tools, and in the case of the marine system, add the ability to watch and analyse market prices.

Many of the systems reviewed in this report are biosecurity intelligence systems that are web–based and open–source. Enterprise search systems, which in contrast are closed–source, have not been reviewed in this report. However, it is likely that the implementation of an enterprise search system, such as MS FAST or ISYS, would be of great benefit to DAFF, since such a system would allow information to be shared within DAFF in an intelligent way. It should be noted, though, that an enterprise search system does not perform the same functions that open–source web–based systems perform. Although enterprise search systems can search the internet for biosecurity information in more–or–less the same fashion as open–source web–based systems do, the *results* of such searches are not open–source and are accessible only to those who have access to the enterprise search system. Systems such as ProMED clearly demonstrate that there is great value in allowing the global community to view the results, since users of the system can contribute *back* into the system, by confirming/disconfirming initial reports, submitting new information missed by the system, etc.

These two types of systems are not in conflict, and in fact complement each other. An enterprise search system can be used to organise and share information within DAFF that may be of a sensitive nature and not to be shared with the rest of the world. A web–based open–source system can be used by DAFF staff to interact with experts in the global community to acquire intelligence that would otherwise not be able to be obtained. This intelligence can then be fed into

the enterprise search system to add to the quality of that information. Implementation of both types of systems would probably best serve DAFF's biosecurity intelligence needs.

9.1 Recommendations for Future Research and Development

DAFF has the opportunity to make a strategic entry into the field of online biosecurity intelligence. Although the systems that cover human, animal, and zoonotic biosecurity intelligence do not currently satisfy DAFF's intelligence needs, that space is currently occupied by several systems, which are continually improving. On the plant and marine/aquatic side, however, there is little to no coverage at all. There is, therefore, real potential for the development of plant and marine/aquatic intelligence systems. By further developing the prototypes developed in this project (in collaboration with DAFF staff), DAFF could become a world leader in plant and marine/aquatic biosecurity intelligence. Such systems would be of significant benefit to DAFF, and there is potential for international collaboration in building them (e.g., through the QUADs alliance).

There are three main components to such a system: the automated analysis, the web interface, and the intelligence community. All three components need to be developed in order for a system to be successful, and they should be developed with an understanding of how they are implemented in other systems (even though those systems may not focus on plant or marine/aquatic biosecurity). Much of the work in studying the advantages and disadvantages of other systems has been done in this project, but as the systems are constantly changing, they should be continually reviewed. The current development of the plant and marine/aquatic prototype systems (in collaboration with DAFF staff) are being guided by the review of the systems in this report. The following are some recommendations for the next stages of this process.

The most important next stage is to build a web interface. A web interface will allow users to easily view the output of a system (e.g., news articles detected by the system) and organise it in different ways—e.g., maps and trend graphs. It will also allow them to add articles that the automated scan misses, add commentary to articles, add documents that summarise patterns that they have detected, remove junk articles, edit locations, ask questions, etc. In designing the web interface, it is of the utmost importance to make it fast, simple, and streamlined. One problem common to many of the systems studied in this report is that their interfaces are so clunky that DAFF staff don't or can't use them. GPHIN's website is incompatible with all modern web browsers (it accepts only Internet Explorer 6.x–7.x and Netscape 6.x), EpiSPIDER crashes on many of DAFF's computers and takes a long time to load even on modern computers with fast internet connections, HealthMap can take a long time to load its map, and BioCaster's interface is unintuitive. One way to make sure the web interface is fast and streamlined is to not be obsessed with the mapping feature and refrain from putting a map on the front page. Maps can take a long time to load, and, while they are an important visualisation tool, they are not what every user is after. It would be better to have a simple list of reports on the front page, which can be filtered by categories, and give users the option of viewing the map. A simple list of reports loads much

faster than a map.

In terms of data visualisation tools, maps that contain the system's reports should be included, and they should be able to be exported as KML and Geo–RSS feeds so that users can view that content in any way they see fit. There are other mapping tools that should also be included which other systems currently do not include. Map overlays of data such as recent rainfall, temperatures, changes in market prices, etc., can be informative, especially when combined with the articles that the system detects. Users should also have unlimited access to the system's archive. ProMED has this feature, but all of the other systems do not—many have roughly 30 day windows. Unlimited access to the archive would allow the system to serve more of DAFF's intelligence needs. One clear message from the workshop (University of Melbourne, August 2009) was that different sectors of DAFF require intelligence on very different timescales. Some need to know what is happening in real time, while others need to be able to detect patterns that span over periods of years, even decades. Unlimited access to the archive would help serve these different needs.

In terms of the automated analysis (the so–called "backend" of the system), there are a number of obvious areas for improvement: expand sources (e.g., to scholarly journals), improve list of search terms, build an ontology, and include languages other than English. There are also some less obvious ways to improve the automated analysis. One is to use Twitter for analytical purposes. News stories that receive a large amount of attention on Twitter (e.g., the BP oil spill and nearby fish kill) are typically of little interest to DAFF staff. If a story is receiving a lot of attention on Twitter, the system can recognise this, and filter out all articles it finds on that particular story—perhaps putting them aside in case DAFF staff are interested in it. Another way to improve the system is include some natural language processing of the articles that the system detects in order to match the articles by similarity. The advantage of this is that users can then be told when two different articles are about the same or similar stories, of if one article is a follow–up to another.

In terms of the intelligence community, it's very important that it's possible for anyone to contribute to the system, and anonymously if they wish. This introduces the possibility of introducing error and disinformation, but that can be mitigated by appropriately structuring the community. One way to do this is by having a three–tiered community. At the top tier are users that have been verified by DAFF staff (or perhaps by other members of the top tier) as experts in the area. The second tier could be other users who have been unverified, but have user profiles that include information about their expertise, where they work, etc. The third tier, could be the general public, who can contribute to the system anonymously and without having to login to the system. In addition to this tiered system, users could rate each other on their expertise and reliability. These two features would give users a sense of the reliability of the comments or postings of any particular user. The task of designing the structure of the community should be guided by results from epistemic institutional design, an area of social epistemology. Building a well–structured community of experts is very important. Automated data collection and analysis will always have its limits and be prone to error. Having a community of users constantly adding and refining information will ensure that the system is able to assist in building genuine biosecurity intelligence.

References

Blench, M. (2008). Global Public Health Intelligence Network (GPHIN). *MT Summit XI*, 45–49.

C., J. (1998). Global Food Security. In *International Congress for Plant Pathology, 7th Edingburgh, UK. No. 4.1GF*.

Collier, N., S. Doan, A. Kawazoe, R. Goodwin, M. Conway, Y. Tateno, Q. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, et al. (2008). BioCaster: Detecting Public Health Rumors with a Web-Based Text Mining System. *Bioinformatics 24*(24), 2940.

Collier, N., A. Kawazoe, L. Jin, M. Shigematsu, D. Dien, R. Barrero, K. Takeuchi, and A. Kawtrakul (2006). A Multilingual Ontology for Infectious Disease Surveillance: Rationale, Design and Challenges. *Language resources and evaluation 40*(3), 405–413.

Cowen, P., T. Garland, M. Hugh-Jones, A. Shimshony, S. Handysides, D. Kaye, L. Madoff, M. Pollack, and J. Woodall (2006). Evaluation of ProMED-Mail as an Electronic Early Warning System for Emerging Animal Diseases: 1996 to 2004. *Journal of the American Veterinary Medical Association 229*(7), 1090–1099.

De Bruin, W., B. Fischhoff, L. Brilliant, and D. Caruso (2006). Expert Judgments of Pandemic Influenza Risks. *Global Public Health 1*(2), 179–194.

Ginsberg, J., M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant (2008). Detecting Influenza Epidemics Using Search Engine Query Data. *Nature 457*(7232), 1012–1014.

Goldman, A. (2008). The Social Epistemology of Blogging. In *Information Technology and Moral Philosophy*, pp. 111–122. Cambridge University Press.

Grady, N., L. Vizenor, J. Marin, and L. Peitersen (2008). Bio-surveillance Event Models, Open Source Intelligence, and the Semantic Web. *Biosurveillance and Biosecurity*, 22–31.

Heymann, D., G. Rodier, et al. (2001). Hot Spots in a Wired World: WHO Surveillance of Emerging and Re–Emerging Infectious Diseases. *The Lancet Infectious Diseases 1*(5), 345–353.

Hovmoeller, M., A. Yahyaoui, E. Milus, and A. Justesen (2008). Rapid Global Spread of Two Aggressive Strains of a Wheat Rust Fungus. *Molecular Ecology 17*(17), 3818–3826.

Keller, M., M. Blench, H. Tolentino, C. Freifeld, K. Mandl, A. Mawudeku, G. Eysenbach, and J. Brownstein (2009). Use of Unstructured Event-Based Reports for Global Infectious Disease Surveillance. *Emerging Infectious Diseases 15*(5), 689.

Madoff, L. C. (2004). ProMED–Mail: An Early Warning System for Emerging Diseases. *Clinical Infectious Diseases 39*(2), 227–232.

Margaritopoulos, J., L. Kasprowicz, G. Malloch, and B. Fenton (2009). Tracking the Global Dispersal of a Cosmopolitan Insect Pest, the Peach Potato Aphid. *BMC Ecology 9*(1), 13. doi:10.1186/1472-6785-9-13.

Mawudeku, A. and M. Blench (2005). Global Public Health Intelligence Network (GPHIN). In *7th Conference of the Association for Machine Translation in the Americas*, pp. 8–12.

Strange, R. and P. Scott (2005). Plant Disease: A Threat to Global Food Security. *Annual Review of Phytopathology*.

Tolentino, H., R. Kamadjeu, P. Fontelo, F. Liu, M. Matters, M. Pollack, and L. Madoff (2007). Scanning the Emerging Infectious Diseases Horizon—Visualizing ProMED Emails Using EpiSPIDER. *Adv Dis Surveill 2*, 169.