Using Internet Intelligence to Manage Biosecurity Risks: A Case Study for Aquatic Animal Health

Aidan Lyon,[1,2] Geoff Grossel,[3] Mark Burgman,[2] and Mike Nunn[3]

1: Department of Philosophy, University of Maryland.

2: Australian Centre of Excellence for Risk Analysis, University of Melbourne.

3: Biosecurity Division, the Australian Federal Department of Agriculture, Fisheries and Forestry.

Contact: alyon.umd@gmail.com

Forthcoming in *Diversity and Distributions*

Abstract

AquaticHealth.net is an online intelligence system dedicated to tracking and forecasting aquatic animal disease outbreaks. It scans the internet on a regular basis for opensource content (such as news articles, journal articles, and tweets from Twitter) related to aquatic animal health. It also allows anyone to manually submit content. The system automatically detects location information in the content so that it can be plotted on a Google Map and tags it with useful key terms. All users can browse content and generate reports and maps by search terms, filtered date ranges, tags, locations, etc., and receive RSS feeds and email updates. Authenticated users can tag, edit, and classify all reports; and they can add and refine the search terms that the system uses to find content. AquaticHealth.net also includes a wiki-blog devoted to emerging issues in aquatic animal health. Authenticated users can edit and update blog entries on diseases and other topics related to aquatic animal health. The result is an online intelligence system for aquatic animal health that is social at almost every level, and provides real-time and relevant information to decision-makers in a practical, efficient and timely manner.

1 Introduction

The prevalences and damages caused by invasive pests and emerging diseases have escalated internationally in concert with burgeoning trade and travel (Hulme *et al.*, 2008). Animal and plant pests and diseases can devastate ecosystems, agriculture and aquaculture, with serious consequences for the environment, the economy and food security (Strange & Scott 2005). Invasive pests and diseases remain one of the world's most important threats to biodiversity loss globally (McGeoch *et al.*, 2010) and James (1998) estimated that at least 33% of global food production is lost to plant diseases, pests and weeds. For example, the rust species Puccinia psidii, a plant pest that originated in South America and subsequently spread to the southern

United States and Hawaii, recently established in Australia where it threatens a substantial proportion of the continent's dominant tree and shrub species, as well forestry plantations, orchards and other economic activities (Carnegie and Lidbetter 2012).

Such emerging biosecurity risks create special imperatives to identify hazards early and to intervene effectively. A great deal of relevant information is available on the internet and in social media that may assist people to take early preventative action to protect the environment and productive activities. The task confronting efforts to gather and interpret intelligence that gives early warning is to sift important information from the vast volume of electronic communications and to render it in a form that local analysts can use efficiently. During the past two decades, new internet technologies have emerged that substantially improve our ability to detect reports of new outbreaks of pests and diseases quickly. For example, by tracking people's searches and IP addresses, Google Flu Trends can report influenza activity 1–2 weeks ahead of traditional CDC reports (Ginsberg *et al.*, 2008), and by automatically scanning thousands of websites for signals of disease outbreaks, the Global Public Health Intelligence Network (GPHIN) detected SARS in China 3 months before the World Health Organization officially announced it (Keller *et al.*, 2009). (More examples of such technologies are discussed in section 2 of this paper.)

Many *online biosecurity intelligence systems* have been developed to track reports of invasive pests and emerging diseases. They include: BioCaster, the BedBug Registry, EpiSPIDER, the Early Detection and Distribution Mapping System (EDDMapS), the USDA's Emerging Animal Disease Notices, Google Flu Trends, GPHIN, HealthMap, the North American Plant Protection Organisation (NAPPO), ProMED, the OIE's World Animal Health Information Database (WAHID), and the Wildlife Data Integration Network (WDIN). These systems have many differences (see, for example, Lyon *et al.*, 2011; Keller *et al.*, 2009), but they focus primarily on human, animal and zoonotic diseases, with some coverage of plant pests and diseases by NAPPO and ProMED, and invasive species by EDDMapS.

Marine environments and protected areas are particularly susceptible to invasion from ballast water and the movement of vessels (Molnar *et al.*, 2008). The aquaculture industry also suffers from devastating disease outbreaks. For example, infectious salmon anemia virus (ISAV) emerged as a disease issue in Chile devastating the salmon industry between 2007 and 2009 (Godoy *et al.*, 2008). Until recently, there were no online biosecurity systems devoted to the pests and diseases of aquatic animals (particularly finfish, molluscs, and crustaceans). Important information pertaining to aquatic animal health on the Internet can assist regulators and governments to anticipate emerging threats and make better decisions about allocating limited aquatic biosecurity resources. OIE's WAHID specifically aims to cover some aquatic animal diseases and CEFAS' International Database on Aquatic Animal Diseases (IDAAD) aims to provide a more complete epidemiological picture by adding data from the scientific literature to OIE's information. However, both systems do not take advantage of the real-time and open-source information on the internet; they collate only official information and information from the scientific literature that is relevant to an internationally agreed list of disease agents. There is thus a need for an online an open-source biosecurity intelligence system for aquatic animal health, and this need has become more urgent with the rapid expansion of aquaculture globally (Halpern *et al.*, 2007; Oidtmann *et al.*, 2011).

This paper reviews the availability and utility of internet-based, open-source systems for gathering biosecurity information for environmental protection and explores what may be needed to better target these tools towards a broader range of environmental threats. It describes the development, deployment and performance of a case study, AquaticHealth.net, a system that gathers and analyses information on aquatic animal health and aquatic ecosystems, and discusses the advantages and disadvantages of alternative approaches to internet intelligence-gathering for biosecurity generally. The development of AquaticHealth.net has been guided by previous in-depth studies of the strengths and weaknesses of existing systems (Lyon *et al.*, 2011; Lyon 2010), and this paper describes how AquaticHealth.net tries to integrate their best features. We thus start with an overview of some of the existing systems.

2 Methods Review

2.1 ProMED

ProMED, established in 1994, was the first online biosecurity system. It is, roughly speaking, a moderated email list with the goal of finding and disseminating information about human health. Dozens of reports—in the form of news articles, first-hand accounts, official reports, etc.—are submitted to ProMED daily. This information comes from a range of sources, including government health departments, international organisations, subscribers' professional or personal observations, the media, and manual online searches conducted by ProMED staff. Reports are initially examined by a 'top moderator' who decides whether to reject them or send them on to 'subject moderators'. Subject moderators check the accuracy of the reports, edit them for clarity and references, and frequently add a brief commentary to highlight the importance of new information. Each report then goes back to the top moderator who audits the edited report and assigns a level of urgency—green, yellow, or red (the most urgent). Green reports are sent to a copy editor for formatting and editing for grammar and consistency, and 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). On a typical day, 7 reports are published: 1 red, 1 yellow, and 5 green (Madoff 2004). These are e-mailed to users and posted in the ProMED archive.

2.2 GPHIN

GPHIN (Mawudeku & Blench 2005) 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.[1] It is a semi–automated, early warning system that reports information on human and zoonotic disease outbreaks, and other public health topics such as food and water contamination, bioterrorism, exposure to chemical and radionuclear agents, natural disasters, and the safety of products, drugs and medical devices. GPHIN collects information on disease outbreaks and other public health events by monitoring global news media aggregators (including Reuters, Associated Press, New York Times, Sydney Morning Herald, Irish Times, etc.) that gather news media from a large number of sources on the internet. GPHIN uses an automated process to filter reports for relevance before passing them to GPHIN analysts. Notifications about public health events that may have serious public health consequences are sent to users immediately 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. 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 15 minutes, when GPHIN gathers articles from newsfeed aggregators (e.g., from Al Bawaba and Factiva) that are determined to be relevant by established search syntaxes.

The articles are then sorted into one or more categories: animal diseases, human diseases,

[1] At the time of writing, GPHIN is undergoing an extensive update and is currently offline.

plant diseases, biologics, natural disasters, chemical disasters, radioactive incidents, and unsafe products. Each article is assigned a categoryrelevance score, which is a function of keywords that appear in the article that are associated with the category to which the the article has been assigned. Each article is then automatically discarded, published, or presented to a human analyst, depending on its relevance score. Articles that have especially high relevance scores are immediately e-mailed to GPHIN users.

If an article is not either automatically discarded or published, it is presented to a GPHIN analyst who decides whether to discard or publish it. If the analyst deems the article to be of immediate concern, the article is forwarded by e-mail to GPHIN users. Analysts also review automatically discarded articles to verify that relevant information has not been discarded by the automated system. As mentioned in section 1, during the SARS outbreak in 200203, the GPHIN prototype was able to detect and gather information about an unusual outbreak occurring in Guangdong Province of China as early as 27 November 2002 (Mawudeku & Blench 2005) well before official reports became available.

2.3 BioCaster

BioCaster (Collier *et al.*, 2006, 2008) has been in operation since 2006 and is similar to GPHIN in structure and focus, but is completely automated and has a particular focus on languages of South-East Asia. BioCaster also differs from GPHIN in using a rich, open-source multilingual ontology to help categorise the content that the system finds. This means that it uses an explicit system to decide how the content should be grouped, related or subdivided. Such an approach helps the system to categorise articles by their semantic content (i.e., the encapsulated meaning of an article) rather than merely by their syntactic content (i.e., their sentence structure). 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 clinical signs or symptoms . Using such an ontology allows the system to make reasonably sophisticated inferences concerning the contents of the articles. For example, if an article mentions people in a region suffering from the 6 clinical signs or symptoms of avian influenza listed in the ontology, then the system could warn users of a potential outbreak of avian influenza in the region, even if the article itself makes no explicit mention of avian influenza per se. BioCaster collects information from EurekAlert!, European Media Monitor Alerts (EMMA), Google, the CDC's Morbidity and Mortality Weekly Report (MMWR), MeltWater, OIE, ProMED, Reuters, WHO and Vetsweb. It scans for articles in Arabic, Chinese, English, French, Japanese, Korean, Portuguese, Russian, Spanish, Thai and Vietnamese. The system publishes its content on a Google Map, which users can filter in various ways.

2.4 HealthMap

HealthMap (Brownstein *et al.*, 2009, 2010; Keller *et al.*, 2009; Wilson & Brownstein 2009) was founded in 2006. The system collects information from Baidu, Community News Reports, EuroSurveillance, Google, HealthMap, OIE, ProMED, SOSO, User Eyewitness Reports, WDIN and WHO. It scans for articles in Arabic, Chinese, English, French, Portuguese, Russian and Spanish. It also has a mapping system that allows users to view reports and apply a number of filters. Users can also comment on articles and rank them for significance. HealthMap is similar to BioCaster in that it scans for articles in different languages, analyses their content, and publishes links to them on an interactive map. However, in contrast to BioCaster, HealthMap allows users to report articles that they find to the system and these can then be published on the map. Users can report articles by using HealthMap's website, by calling a hotline, or by using HealthMap's telephone application, Outbreaks Near Me (which runs on iPhones and Android phones). Users can also submit eyewitness accounts using these methods, with the telephone application allowing them to attach photographs and GPS coordinates. In addition, HealthMap allows users to comment on reports and rank them for 'significance'. Thus users can both contribute information and make judgements concerning the quality of that information that other users can see and assess themselves.

2.5 EddMapS

EDDMapS was launched in 2005 by the Center for Invasive Species and Ecosystem Health at the University of Georgia. It is an alert system that focuses on invasive plant and animal species. Alerts can be set up to e-mailed to users when a new or existing invasive plant or animal is reported in a particular state or country. An alert can also be set to be e-mailed when a specific species has been reported or when it is reported in a new county. Alerts are sent each morning and include updates in the previous 24 hours. Users can submit reports together with a user profile (which includes a name, organisation or company, and an e-mail address). In cooperation with the United States Forest Service Forest Health Protection, EddMapS is expanding the call for invasive plant distribution data in the United States. This will allow EddMapS to display the range of known invasive plants as well as add new species to the distribution maps as they are reported.

2.6 Summary

One of the most significant ways in which the systems differ is their reliance on automated software (rather than humans) for content-gathering and analysis. For example, BioCaster is a completely automated system. It collects information through RSS feeds and then uses natural language processing to filter and classify the information it gathers. In contrast, ProMED is much more human-based. Its moderators conduct manual searches and accept reports from e-mail subscribers. The results of these searches and reports are then reviewed by the moderators (and sometimes external experts) to assess whether the information is accurate, relevant and timely. As just one example among many, consider the following comment on a newswire:

> "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. (see references below). While in the earlier outbreaks, Japanese encephalitis (JE) was often speculated as the etiology of the outbreaks, the clinical picture described (with henipavirus infection) was more severe, more rapid in evolution and was associated with a higher case fatality rate than usually seen with JE. In the initial outbreak in 2001, cases occurred in all age groups; in subsequent outbreaks there was a preponderance of cases in the childhood population. In the January 2005 outbreak, illness was associated with drinking juice made from local palm fruit. Speculations were that the juice was made from fruit contaminated by fruit bat droppings or from fruit the bats had half-eaten. In the above outbreak, there is mention of 6 deaths out of 50 cases (a case fatality rate of 12 percent - much lower than the observed 40 percent case fatality rates associated with henipavirus).The above newswire mentions a fairly wide spread age distribution (ages 12 through 60 years) . That fact, combined with an observation that this is not the usual JE season, suggests that this outbreak is most likely not due to JE. Given the geography, and prior history of outbreaks in Bangladesh, henipavirus seems a possible candidate as an etiology of this outbreak. Another virus that has been implicated in outbreaks in neighboring India is Chandipura virus. A caution to an interpretation of the differential diagnosis here is the mention of high fever and then an apparent accompanying delirium... a clinical picture that could also be consistent with cerebral malaria. Rather than continue to speculate, ProMED-mail would like to request further information from knowledgeable individuals in the region. [...]" ProMED-mail 2007.

This review process can be incredibly valuable, since human moderators can assess the quality of information and add to it in a way that automated algorithms cannot, as demonstrated by the above comment. Between these two extremes of pure automation and pure human content-gathering and analysis, systems such as GPHIN collect information automatically and use natural language processing while also using a dedicated team of curators to assess and verify reports.

The existing systems have many useful and innovative tools, and they continue to evolve in terms of their generality and the sources of information they access (see Table 1 for a summary of the systems reviewed in this paper). However, each system is constrained by the types of sources it accesses and the way in which reports are processed, assessed and published. A biosecurity intelligence system for aquatic animal health needs to focus on—or have the potential to be directed towards—relevant sources, to use appropriate key words and search terms and concepts, and to appeal to appropriately trained and experienced moderators and assessors. None of the systems examined had the appropriate focus, or was sufficiently flexible to serve these needs. We thus designed AquaticHealth.net to integrate the systems components that have the most utility for aquatic animal health, adding new tools where the demands of aquatic health required them or when opportunities emerged during the development phase of the work to innovate beyond the capabilities of existing systems.

3 Methods

This section documents how AquaticHealth.net gathers (§3.1), processes (§3.2), disseminates (§3.3), and retrieves (§3.4) information. See Fig. 1 for a diagram of AquaticHealth.net's information flow.

3.1 Information Gathering

Every hour, AquaticHealth.net uses Google and Twitter to search the internet for new information on aquatic animal health. Using these search tools, the system searches for relevant information on webpages, news media sites, and personal blogs. The search terms are generated by the system's users. Search terms are categorised in several ways to help classify the search results. Registered users can edit the search terms and add or edit tags that are relevant to the search terms. For example, the search term 'crayfish plague' is tagged with '*Aphanomyces astaci*', the scientific name of the agent that causes this disease. Any article found by using the search term 'crayfish plague' is automatically tagged with 'crayfish plague' and '*Aphanomyces astaci*'. Over time, the tags may develop to be sensitive to a more complex set of ideas related to the initial content. The evolution of the list of search terms represents one of the ways the system may

System	Scope	Data Sources	Data Analysis	Data Output
ProMED	Animal, human, some aquatic, some plant, and zoonotic diseases, and toxins.	Reports from users, health departments, and media.	Several levels of human analysis and editing.	e–mail alerts, web posts, searchable archive, RSS, Twitter, Facebook.
GPHIN	Human, zoonotic, plant, marine, food, water, bio–terrorism, natural disasters, product safety, drugs.	Factiva and Al Bawaba.	Automated and human translation, categorisation, and geocoding.	Tailored e–mail alerts, filtered web posts, searchable archive.
HealthMap	Animal, human, zoonotic.	Baidu, Google, Moreover, WHO, ProMED, Euro-Surveillance, WDIN, user inputs, and others.	Automated translation, categorisation, timecoding, and geocoding, user comments and ratings.	Filtered maps, RSS feeds, Twitter feed, blog and iPhone and Android apps.
BioCaster	Animal, human, zoonotic	Google, ProMED, EMMA, Meltwater, and RSS feeds.	Ontology searches and translations, natural language processing, geo and timecoding	Filtered maps, KML feeds, headlines, filtered graphs.
EddMapS	U.S. Invasive pests.	User GPS submissions	Manual statistical analysis.	Maps, email alerts, and a searchable archive.
OIE	Animal, human, marine, zoonotic	Official reports by members (En.)	Manual screening, checking	Filtered maps, e–mails.

Table 1: Summary of the scope, sources, and analysis and output methods of the systems reviewed.

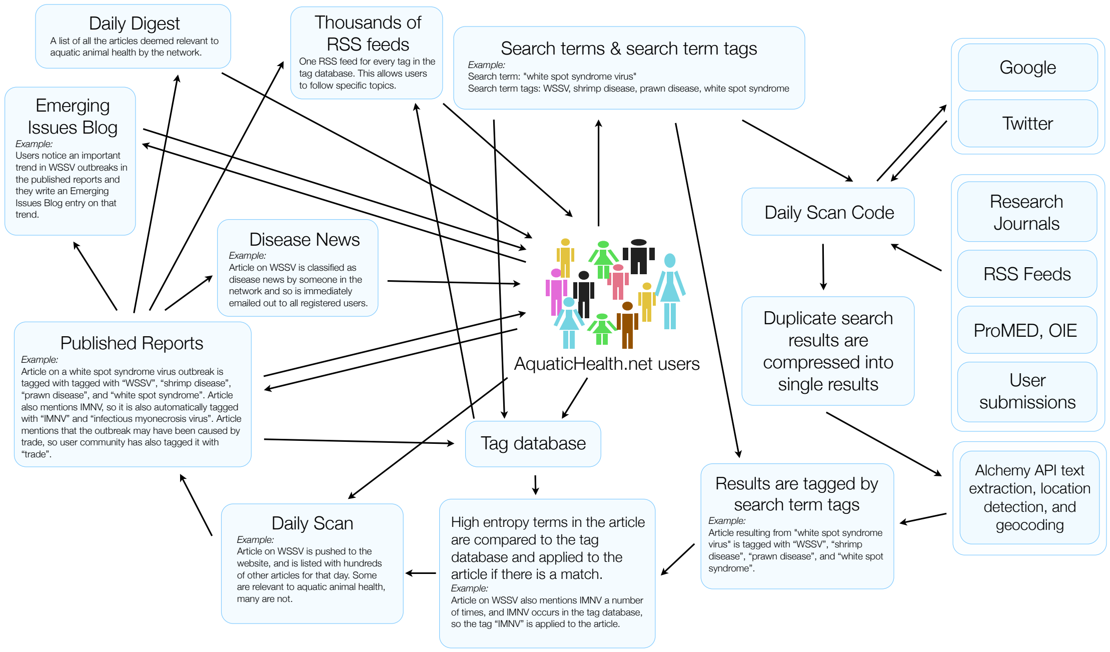

Figure 1: AquaticHealth.net's information flow.

learn and adapt to changing contexts and information needs.

By scanning Twitter, the system can find information that would be missed by using a regular search engine. This is because when a Twitter user tweets a link to, say, a new article, the user will often describe the basic contents of the news article, and in doing so, they may use a term not used in the article. If that term is one of AquaticHealth.net's search terms, the system will detect the article through Twitter, but not necessarily through a regular search engine. In addition to these searches, the system also follows a number of RSS feeds from websites that are devoted to news relating to aquaculture, fisheries and the fishing industry. These sites include: fishinfo.com, fishupdate.com, growfish.com, intrafish.com, and thefishsite.com. The system also collects relevant information from ProMED and OIE. AquaticHealth.net gathers information from a number of scholarly research journals including: *Aquaculture*, *Aquaculture Research*, *Diseases of Aquatic Organisms*, *Fish and Shellfish Immunology*, *Journal of Aquatic Health*, *Journal of Fish Diseases*, *Journal of the World Aquaculture Society*. The system also accepts reports manually from users (including users who are not registered). These reports can be news articles that the system missed in its scan of the internet, or first-hand accounts.

3.2 Processing

AquaticHealth.net uses both automated and manual processing methods. Initially, the system compares the results from RSS feeds, Google and Twitter and compresses duplicates. This comparison involves three stages. First, raw URLs are checked for duplicates. Second, remaining URLs are followed to their final URLs, which are then checked for duplicates. Third, the contents of the pages of the URLs are compared pairwise using a similarity measure (the Python 2.7 SequenceMatcher Class). Any two pages that are sufficiently similar to each other are considered duplicates. It is important to remove duplicates to mitigate potential information overload. Our anecdotal assessment of the users' experiences have found the system's output is much easier to digest than the raw search results from Google and Twitter.

Intelligent compression involves retaining information in the duplicates that could be useful. Any RSS feed result or Google search result that is also a Twitter result is tagged by the username of the Twitter user who sent it. Any two results that are considered a match by the system are tagged by the search terms that generated each result and also by the tags that the system automatically associates with the search terms (e.g., the search term 'WSSV' is tagged with 'Shrimp/Prawn'). Registered users decide which tags the system automatically associates with each search term. These collection and compression processes are also applied to existing results in the database. For example, if a new Twitter search result already exists in the database as a Google search result, then the result in the database is tagged by the Twitter username and then the new Twitter result is discarded. Once all duplicate results have been compressed, remaining results are deemed to be reports, which the system retains.

A report consists of a number of fields—date, url, content, location information, and a range of tags. Some of the automated analysis is conducted during the compression process (e.g., when a report is tagged by Twitter usernames, search terms, and search term tags). This range of tags enables the system to get a sense of the content of each article without employing any natural language processing. The location information of a report is determined by using the Alchemy Location Extraction application programming interface (API), which EpiSPIDER uses. Location terms that are extracted are then geocoded using the Google Maps API, so that reports can be presented on a Google Map on the system's website.

After compression and automated analysis, all remaining reports are uploaded to a MySQL database, which can be viewed using the system's website (built using the open-source Drupal [drupal.org] content management system). All reports automatically receive a 'published' or ‘unpublished’ status. Reports containing certain key search terms (usually scientific names for diseases) automatically receive a ‘published’ status; all other reports automatically receive an ‘unpublished status’. This strategy relies on the assumption that authors using correct scientific terminology are very likely to be disseminating authentic and relevant communications. Users may still revise their status, if this turns out to be false. The ’Daily Scan’ is a list of the articles and other information the system has found through the auto mated search and that have not been verified by the system or a user.

The design of AquaticHealth.net aims to combine the cheapness and efficiency of automation with the accuracy and potential for value adding by humans. To achieve this, we developed the system so that it crowd-sources content and analysis. We strove to make the system as open as possible, by providing the potential for users to rank, comment on and add value to reports, to modify search capacities, tag and classify content, to add new information that the automated scan misses, and to add location information to reports. As with GPHIN and ProMED, there is also a manual element to the information processing, but it is not necessarily conducted by an administrator or a dedicated team of experts. Rather, the analysis is left to all of AquaticHealth.net’s users. This means that part of the analysis of the reports is crowd sourced. Users can change any aspect to any report, including translating it from another language, adding or refining its locations, and classifying it as ‘Disease News’ (when an article is classified as ‘Disease News’ all of the users are automatically emailed that article). Users can be anonymous or registered. All users can view the system’s reports, add reports, and add commentary. Registered users have additional privileges, including the ability edit report content and to classify and tag reports.

There is a constant feedback between the human element and the automated element. This occurs in a variety of ways. Users can refine search terms based on the results that automatically appear in the Daily Scan, and tags that users add to search terms are used to automatically tag reports. New tags that users add to the system’s reports are then used by the system to tag other reports automatically. Also, over time the system learns which Twitter users generally produce reliable information and which do not. How often a Twitter user sends an article that makes it into the published reports or ‘Disease News’ sections provides an estimate of the Twitter user’s reliability. This allows the system automatically to grade the expected quality of the information it gets from Twitter, making it easier for the user to analyse this information. If the source is a reliable Twitter user the report is pushed to the top of the ‘Daily Scan’; if user is unreliable the report is pushed to the bottom.

This approach allows the system to keep all of the search results that it obtains—even those by unreliable Twitter users—while making it easy for users to see the reports that are probably relevant. The results from even unreliable Twitter users are retained because they may report something relevant and may be the only source of some information. Once a report from the 'Daily Scan' is published by a user, it is removed from the 'Daily Scan' and placed in the 'Published Reports' section. This means that subsequent users can focus on the remaining reports of the 'Daily Scan'.

3.3 Communications

A key element of AquaticHealth.net as an intelligence network is the timely communication to users of current, and relevant information. Users have the option of receiving a variety of automated communications. These include disease news alerts, a daily digest of published reports, comment e-mails on any report, chat rooms, and RSS feeds. The alerts are sent out to users' e-mail addresses as soon as the sysem detects them up in an easy-to-read format containing a link to each report and its original source. There are a number of RSS feeds, one for each tag in the system, and dedicated feeds for particular topics. These include the latest disease news, aquaculture news, industry news from trusted sources, latest comments, published reports, and the 'Daily Scan'. Emerging trends in aquatic animal diseases over each six-month period are collected and entered into an emerging diseases wiki and other report formats.

3.4 Information Retrieval

Users can retrieve the system's information through: the 'Daily Scan' page, the 'Published Reports' page, the research page, the map, and a general search page. The 'Daily Scan' is in an 'accordion' format, which allows users to click on the titles of articles to obtain a preview without having to leave the page. Published reports are the articles that have been verified either by the system or by a user. Registered users can choose to publish reports from the 'Daily Scan' if they deem them to be relevant to aquatic animal health, and can also edit reports and add locations, classifications and tags. Published reports are listed on the front page of the site (in the same 'accordion' format as the 'Daily Scan' list). The research page has a same format as the published reports and 'Daily Scan' pages, but it focuses on information from the research journals that the system follows.

As with many of the existing systems, AquaticHealth.net uses a Google Map to plot its reports. A basic map is presented on the front landing page, which summarises the most important and recent information that the site has found. In addition to this, there is a separate Map page, which allows users to filter and search for reports in various ways. Registered users can also enter a general search workspace where they can use the filters to generate any type of report. For example, a user can search, sort and create a chronological report for all information on outbreaks of abalone diseases and articles relating to *Xenohaliotis californiensis* captured by the system during the previous year. These filters can be developed by users to make the system sensitive to their individual information needs.

A feature that has recently been added to a site is a wiki-based blog devoted to emerging issues, called the Emerging Issues Blog. The primary purpose of the blog is to capture trends in the information that the system gathers and identify any emerging issues, such as new potential disease threats, substantial changes to the aquaculture industry, etc. Each blog entry contains a summary of the issue identified and a list of related reports that are in the system's database. Any registered user can edit any blog entry, and all edits are saved so that the entire history of each blog entry can be viewed by the system's users. The Emerging Issues Blog is a feature that evolved from a regular wiki system that was previously implemented. Although the wiki had some successes and contained a lot of valuable information, most of the pages remained fairly inactive, with the exception of the page that was devoted to emerging diseases and issues. It was concluded that this was the primary interest of the users, and so the wiki-based Emerging Issues Blog was developed.

4 Results

Between 100 and 150 news articles, web pages, blog postings, etc. are gathered daily by the current search terms and listed in the 'Daily Scan'. Of these, about 5 to 10 articles are published and tagged by users as useful, and of the published reports the system will tag about 10 articles each week as 'Disease News'. Over time, this collection of disease-related information builds a bank of information that allows analysts to identify and interpret emerging trends. It can sometimes be difficult to determine what counts as relevant information (e.g., it is not always clear whether a fish kill is due to a disease or, say, an environmental cause), so the system keeps all of the information it gathers. It takes only one user to identify an event that has been incorrectly classified to make an intelligence breakthrough.

In its short history, AquaticHealth.net has been used to capture emerging disease information, analyse disease trends, map diseases, organise data, perform basic forecasting or 'predictive modelling', contribute to strategic planning, provide biosecurity alerts, build biosecurity risk profiles, and support decision-making relating to imports and exports. Some of the forecasting applications have proven to be very useful for improving biosecurity planning. Emerging trends in each six-month period are entered into the 'Emerging Issues' wiki-based blog. A report and map are produced from the blog and provided to the Australian Government committees responsible for aquatic animal health, and has a particular focus on emerging diseases and issues outside Australia and their implications for Australia. For example, a report generated from AquaticHealth.net emerging issues is provided to the Australian Government Sub-committee for Aquatic Animal Health for their consideration. The report allows the committees to keep abreast of potential emerging disease threats to inform future biosecurity planning and to recommend potential preventive and preparedness actions.

AquaticHealth.net also supports decision-making for biosecurity issues relating to imports and exports. For example, interrogation of AquaticHealth.net provided evidence of the unregulated movement of used aquaculture equipment, which is a widely recognised direct entry and exposure pathway for pests and diseases of concern for aquatic animals and the aquatic environment. The first report from the website (http://aquatichealth.net/report/4102) to highlight the problem was originally communicated by ProMED (6 August 2010) about an oyster farmer from the UK who had an outbreak of oyster herpes virus after deploying equipment previously used in France to refurbish oyster beds (ProMED-mail 2010). Although the emergence of aquatic animal diseases globally is multifactorial (associated with movement of stock, climate change, etc.), the role of used imported aquaculture equipment in disease spread had not previously been prominent. Internet sites specialising in the online trade of aquaculture equipment, including used equipment, provide producers such as salmon farmers, abalone farmers or oyster farmers with an easy avenue to liquidate farm assets after their stock has been wiped out by disease. As a result of the risk posed by the movement of used aquaculture equipment from disease-affected areas overseas, within weeks of the threat becoming recognised, the Australian Government introduced preventive measures to ensure that all used equipment exported to Australia is decontaminated on arrival (DAFF 2010).

One unexpected benefit from AquaticHealth.net is the provision of information important to users who are primarily concerned with environmental issues. The system was designed to track information relevant to aquatic animal health and aquaculture, and some of the search terms that users created were terms such as 'fish kill', 'fish die-off' and 'millions of dead fish'. Although fish kills can be caused by disease outbreaks, they are frequently caused by algal blooms, which can in turn be caused by environmental factors, such as pollution. Without specifically intending to do so, AquaticHealth.net developed an extensive database of news about algal blooms, and users concerned with environmental issues found that information to be of value. After discovering this, the search terms were expanded so that the system now explicitly scans for new information on harmful algal blooms ('HABS') and is now building an even more extensive database on this topic.

5 Discussion

There is a trade-off between automated and human-based analysis. Automated analysis is cheap and efficient, but can be error-prone and typically does not add value to content. Errors arise especially when ambiguous terms appear in the text being analysed. For example, an article containing the term 'van' that is referring to a vehicle can be determined as being relevant to the location Van, in Turkey. Another typical error can occur when someone from a particular location is cited in connection with a disease-related event occurring at a different location. For example, if a prawn disease researcher from the University of Texas is interviewed about an outbreak of WSSV in the Philippines, the article may erroneously be determined to be about an outbreak in Texas. Human experts are not as error-prone because they are better conditioned to deal with linguistic and contextual complexities than are automated filters. They can typically add value to content, but can also be expensive (in terms of both money and time). This creates an opportunity for future research, to develop vector-conditioned algorithms that can detect the semantic content of text more efficiently under a broader range of cultural and social settings.

Crowd-sourcing is an open call to an undefined group, usually people appropriate for a specified task, to contribute to an analysis or to solve a problem (Brabham 2008). Crowd-sourcing takes advantage of human analysis while avoiding the expense of more conventional group deliberations. Of the existing internet-based biosecurity systems, HealthMap makes the most effective use of crowd-sourcing to provide human analysis. HealthMap allows users to add commentary and rank articles for relevance. This is an efficient and effective approach to generate human-mediated analysis and has been emulated in AquaticHealth.net. A potential pitfall with this approach, however, is that misinformation can also be generated by members of the crowd, either intentionally or unintentionally. However, since anyone can comment on any report and any registered user can edit any report, any such errors can be quickly corrected. To date, the most common source of errors are in the news articles that the system collects. When reports with such errors are published, users have been quick to add corrections (e.g., see http://aquatichealth.net/node/50309#comment-1736). All such corrections are emailed out to all registered users, so that everyone who has registered with AquaticHealth.net is aware that there has been an error and that it has been corrected.

Most of the existing systems plot information on an interactive map that can be filtered in various ways (e.g., by disease or host categories). Such maps allow users to see geographical patterns over time easily and help to mitigate information overload (Brownstein *et al.*, 2008). However, although interactive maps have great utility, maps in some systems can take a long time to load. Some also then continue to run slowly and occasionally crash, even when run in modern browsers on modern computers. Such impediments are likely to be resolved as computing technology improves but are also likely to be superseded by other technical and speed issues as developers take advantage of greater computing power. Thus any intelligence system needs to be maintained continuously to adapt with the changing conditions and assumptions of internet technology.

Initial statistical analysis of AquaticHealth.net's extensive database has indicated that it is possible to detect significant events by examining the relative frequencies of tag occurrences. For example, an emerging issue recorded by the system's users was the Cohen commission (see http://aquatichealth.net/issues/the-cohen-commission), following the controversy of whether Infectious salmon anemia virus (ISAV) had been found in wild salmon in British Columbia, Canada. During this period, there was a spike in the frequencies in several of the system's tags relating to ISAV (see Fig. 2). This suggests that it is possible to use such spikes in activity to automatically notify users of potential new emerging issues before the users are aware of them. However, this is an open area of research, and much work needs to be done (see e.g., Collier *et al.*, 2008; Thrush *et al.*, 2011).

This brief review of biosecurity intelligence systems is not comprehensive. However, it illustrates the breadth of applications for such internet-sourced systems. The discussion has focused on systems that track disease outbreaks but there are similar systems with different goals. For example, Ushaidi is a crisis mapping tool that allows users to come together to quickly generate crowd sourced maps in crisis situations. We will continue to monitor these sites and

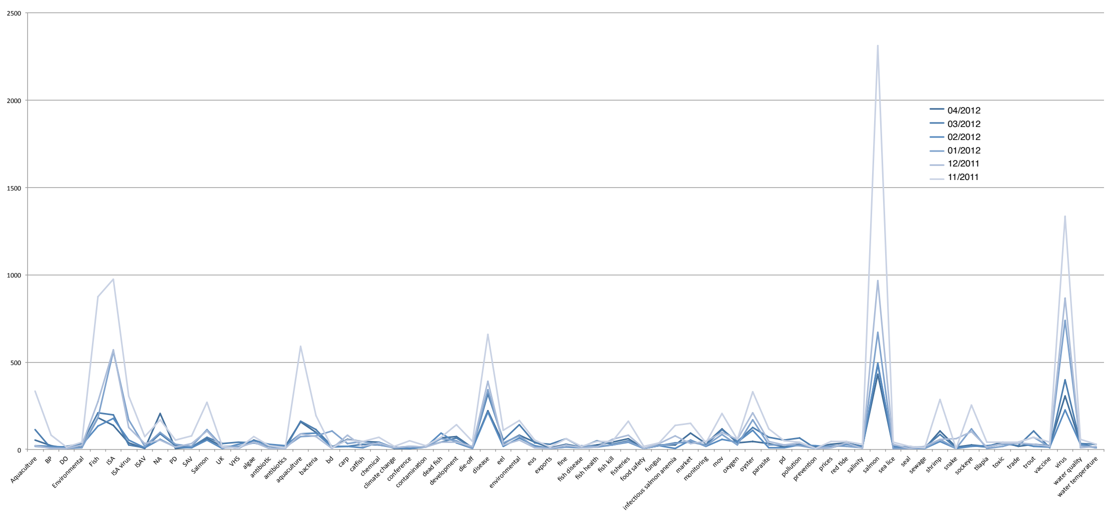

Figure 2: The frequencies of tag occurrences in AquaticHealth.net's database during the period of the Cohen commission (see http://aquatichealth.net/issues/the-cohen-commission).

implement ideas that will serve biosecurity interests. This initiative forms an important part of timely strategies to detect and intervene, to mitigate emerging environmental threats.

6 Conclusion

The power of raw data feeding into discussion and analysis in real time is immense. AquaticHealth.net offers the means to process raw disease information intelligently in real time. AquaticHealth.net's aim is to create a community of practice and professional exchange of informtion on pests and diseases of aquatic animals through a transparent system that relies on public participation, collaboration and trust. Openness will strengthen information on aquatic animal health and promote efficiency and effectiveness in aquatic animal biosecurity. It will also build an effective intelligence community and create critical mass for implementing action. The global aquatic animal health intelligence community is growing stronger and adding value to this internet-based intelligence system every day.[2]

[2] If you are interested in becoming a registered user, please contact the corresponding author by email, or go to http://aquatichealth.net/user/register.

7 Acknowledgements

This work was financially supported by the Australian Centre of Excellence for Risk Analysis and the Australian Government Department of Agriculture, Fisheries and Forestry. Thanks to Max Bialek, Dr David Johnson, Alistair Herfort, Kirsten McKimmie, and Dr Alan Mooney for their help in developing the website and its contents.

References

Brabham, D. (2008). Crowdsourcing as a model for problem solving an introduction and cases. *Convergence: The International Journal of Research into New Media Technologies 14*(1), 75–90.

Brownstein, J., C. Freifeld, B. Reis, and K. Mandl (2008). Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS medicine 5*(7), e151.

Brownstein, J. S., C. C. Freifeld, E. H. Chan, M. Keller, A. L. Sonricker, S. R. Mekaru, and D. L. Buckeridge (2010). Information technology and global surveillance of cases of 2009 h1n1 influenza. *New England Journal of Medicine 362*(18), 1731–1735.

Brownstein, J. S., C. C. Freifeld, and L. C. Madoff (2009). Digital disease detection —harnessing the web for public health surveillance. *New England Journal of Medicine 360*(21), 2153–2157.

Carnegie, A. and J. Lidbetter (2012). Rapidly expanding host range for puccinia psidii sensu lato in australia. *Australasian Plant Pathology*, 1–17.

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.

(DAFF) Australian Department of Agriculture, Fisheries and Forestry (2010). Change to AQIS policy on the import of used aquaculture equipment. *Import Clearance Notices to Industry 2010 69*. http://www.daff.gov.au/aqis/import/general-info/ian/10/69-2010.

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.

Godoy, M., A. Aedo, M. Kibenge, D. Groman, C. Yason, H. Grothusen, A. Lisperguer, M. Calbucura, F. Avendaño, M. Imilán, et al. (2008). First detection, isolation and molecular characterization of infectious salmon anaemia virus associated with clinical disease in farmed atlantic salmon (salmo salar) in chile. *BMC veterinary research 4*(1), 28.

Halpern, B., K. Selkoe, F. Micheli, and C. Kappel (2007). Evaluating and ranking the vulnerability of global marine ecosystems to anthropogenic threats. *Conservation Biology 21*(5), 1301–1315.

Hulme, P., S. Bacher, M. Kenis, S. Klotz, I. Kühn, D. Minchin, W. Nentwig, S. Olenin, V. Panov, and J. Pergl (2008). Grasping at the routes of biological invasions: a framework for integrating pathways into policy. *Journal of Applied Ecology 45*(2), 403–414.

James, C. (1998). Global Food Security. In *7th International Congress of Plant Pathology, Edinburgh, UK, August*, Volume 4.1.

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.

Lyon, A. (2010). ACERA II Project 1003 - Biosecurity Intelligence–Gathering and Analysis. Technical report, Australian Centre of Excellence for Risk Analysis, the University of Melbourne.

Lyon, A., M. Nunn, G. Grossel, and M. Burgman (2011). Comparison of web-based biosecurity intelligence systems: Biocaster, epispider and healthmap. *Transboundary and Emerging Diseases 59*(3), 223–232.

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

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.

McGeoch, M., S. Butchart, D. Spear, E. Marais, E. Kleynhans, A. Symes, J. Chanson, and M. Hoffmann (2010). Global indicators of biological invasion: species numbers, biodiversity impact and policy responses. *Diversity and Distributions 16*(1), 95–108.

Molnar, J., R. Gamboa, C. Revenga, and M. Spalding (2008). Assessing the global threat of invasive species to marine biodiversity. *Frontiers in Ecology and the Environment 6*(9), 485–492.

Oidtmann, B., M. Thrush, K. Denham, and E. Peeler (2011). International and national biosecurity strategies in aquatic animal health. *Aquaculture 310*(1-2), 22–33.

ProMED-mail (2007). Undiagnosed deaths, encephalitis - bangladesh (kushtia): Request for information. *ProMED-mail* (http://www.promedmail.org/direct.php?id=20070412.1221).

ProMED-mail (2010). Oyster herpesvirus - England, Ireland: emerging, OIE (20100817.2858). *ProMED-mail* (http://www.promedmail.org/direct.php?id=20100817.2858).

Strange, R. and P. Scott (2005). Plant Disease: A Threat to Global Food Security. *Annual Review of Phytopathology 43*, 83–116.

Thrush, M., P. Dunn, and E. Peeler (2011). Monitoring emerging diseases of fish and shellfish using electronic sources. *Transboundary and Emerging Diseases*.

Wilson, K. and J. Brownstein (2009). Early detection of disease outbreaks using the internet. *Canadian Medical Association Journal 180*(8), 829–831.