cebra
Centre of Excellence for
Biosecurity Risk Analysis

CEBRA 1503B: Intelligence tools for regulated goods traded via e-commerce

CEBRA 1503B Deliverable 1

Matthew Chisholm, CEBRA, University of Melbourne

March 6, 2017

Acknowledgements

This report is a product of the Centre of Excellence for Biosecurity Risk Analysis (CEBRA). In preparing this report, the author acknowledges the financial and other support provided by the Australian Department of Agriculture & Water Resources (DAWR), the New Zealand Ministry for Primary Industries (NZ MPI) and the University of Melbourne.

The author thanks the anonymous reviewers of this report for their valuable feedback, and the following people who provided advice, contact details and much of the information contained in this report:

- Hernando Acosta, NZ MPI
- Wendy Beltz, USDA–APHIS
- Carolyn Bleach, NZ MPI
- Mark Burgman, CEBRA
- Barney Caton, USDA–APHIS
- Abby Clearihan, DAWR
- Cheryl Dollard, Canada Food Inspection Agency
- Geoff Grossel, DAWR
- Julio Hernandez-Castro, University of Kent
- Greg Hood, DAWR
- Chris Howard, DAWR
- Franziska Humair, ETH Zürich
- Luc Humair, ETH Zürich
- Fabian Kuhn, University of Freiburg
- Steve Llewellyn, Invasive Species Council
- David Mackay, DAWR
- Cory Marker, USDA–APHIS
- Craig Martin, U.S. Fish & Wildlife Service
- Tania McCrae-Steele, International Fund for Animal Welfare
- Shaun Moss, Astro Multimedia
- Françoise Petter, EPPO
- David Roberts, University of Kent
- Andrew Robinson, CEBRA
- Yin Shan, Aust. Dept. of Human Services
- Barney Stephenson, NZ MPI.

Contents

Executive Summary

As approved by the Biosecurity Research Steering Committee, the aims of this project are to

1. identify software that can meet NZ MPI's requirements for monitoring international trade in biosecurity-regulated goods, including the ability to detect breaches of the biosecurity system, and
2. investigate the experiences of other jurisdictions in managing biosecurity risk from e-commerce.

Of the biosecurity agencies successfully contacted only USDA–APHIS has an existing program of automated e-commerce monitoring. The monitoring service they use is provided on contract by a third party. The Great Lakes Commission has been concurrently using and refining its software to monitor e-commerce of invasive species into the Laurentian Great Lakes region. The Australian Department of Agriculture & Water Resources has a manual monitoring program, which is augmented by intelligence received from a non-profit organisation, the Invasive Species Council.

The systems that most closely meet the software requirements advised by NZ MPI, or whose available descriptions suggest they may, are

1. IBIS Replacement — an as yet unnamed system that can perform fully customisable searches of the internet, potentially including e-commerce monitoring
2. GLDIATR — the system developed by the Great Lakes Commission to monitor e-commerce in its member jurisdictions
3. iTrade — a system in the early stages of development by the University of Kent for detecting probable sales of illegal wildlife items; and
4. ETH Zürich's unnamed system developed for researching e-commerce trade in invasive plants.

In selecting a system for use, or in developing an entirely new system, the following factors — beyond normal software procurement practice — should be considered:

1. the technical and legal issues with scraping data from websites
2. the inability to obtain details of persons who actually purchase goods online
3. the need to invest in developing and maintaining effective query parameters; and
4. if automated text processing techniques are used, the need to invest in the development and maintenance of effective algorithms and any associated ontologies.

Finally, there are several opportunities for collaboration on the further development of automated e-commerce surveillance capability, notably with organisations and researchers active in biosecurity and wildlife trade regulation and advocacy.

1

Introduction

As approved by the Biosecurity Research Steering Committee, the aims of this project are to

1. identify software that can meet NZ MPI's requirements for monitoring international trade in biosecurity-regulated goods, including the ability to detect breaches of the biosecurity system; and
2. investigate the experiences of other jurisdictions in managing biosecurity risk from e-commerce.

The approved aims of the project mean that some areas of potential interest were deemed out of scope. These include

- Legal issues associated with monitoring online trade, such as authorisation to use scraping technology on trade sites, proxy membership of social media forums, etc.
- Analysis of the scale and nature of online trade in regulated goods (some brief background information is provided for context where necessary)
- Identifying specific e-commerce sites that warrant monitoring; and
- Testing the degree to which biosecurity outcomes may be improved by deploying suitable e-commerce monitoring software — an expectation stated in the project proposal.

This report contains a summary of the research findings, which should inform future decisions by biosecurity agencies on how to deal with risks posed by e-commerce.

E-commerce is a broad term encompassing the buying and selling of goods and services online, i.e. electronically on the internet. This project is specifically interested in sales, offers of sales, and purchases of goods that are subject to regulatory biosecurity import restrictions in the destination country. Such restrictions may include prohibition, mandatory treatment on or before arrival, and statutory evidence of manufacturing standards.

The biosecurity risks associated with e-commerce have been described in many papers [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. These studies suggest that the online trade in goods of biosecurity concern is not yet effectively regulated, and recommend that action be taken to improve regulatory awareness and compliance. The call to implement systematic monitoring of e-commerce is common among these papers.

In 2014 the International Plant Protection Convention (IPPC) made several recommendations to member organisations in response to the growing biosecurity risk arising from e-commerce [15]. These include developing and establishing mechanisms for identifying e-commerce traders and the products of concern that they sell, and to explore options for promoting regulatory compliance.

Many live plants and animals, and their products, fall under both biosecurity and illegal wildlife trade (IWT) regulations. The enforcement and public relations activities of IWT regulators can therefore play a role in reducing the biosecurity risk for their respective jurisdictions. However, the overlap between IWT-regulated and biosecurity-regulated goods is imperfect: not all items of biosecurity concern are of IWT concern (e.g. soil, weeds and tramp ants), and not all items of IWT concern are necessarily of biosecurity concern (e.g. finished ivory and tortoise shell products).

E-commerce is recognised as a major wildlife conservation challenge, but its ability to facilitate biological invasions seems often to be overlooked [5]. IWT can have serious biosecurity implications:

> *Illegally imported animals and plants can also carry diseases or become pests with significant negative impacts on human health and agriculture. The highly dangerous H5N1 avian flu virus appeared for the first time in Europe when it was discovered in a smuggled pair of eagles seized in Belgium in 2004.* [16]

Several publications describe the scale and characteristics of IWT via e-commerce [17, 16, 18, 19, 20]. There is a strong international effort to control IWT, centred around the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES, https://www.cites.org/). In addition to the regulatory bodies of CITES-member countries, there are many non-government organisations and academic researchers actively involved in CITES-related issues. Some of these use, or have used, software to detect and monitor IWT via e-commerce. For example, the International Fund for Animal Welfare (IFAW) has conducted regular assessments of IWT via e-commerce since 2004 [16].

Given the scale of the e-commerce monitoring task, deploying software to automatically detect sales of biosecurity-regulated goods is an important part of addressing these concerns. One recent study found that automated surveillance of selected e-commerce sites is relatively easy, but that greater complexity is required to capture a broader range of platforms [13]. The authors of that study suggest that efforts could be coordinated among jurisdictions with similar goals. The present study indicates that there is considerable overlap between biosecurity and IWT regulators in terms of their e-commerce monitoring goals. This may present valuable opportunities for collaboration.

2

Methods

High-level software requirements were sought from the project leaders in the New Zealand Ministry for Primary Industries (NZ MPI), and from the teams within the Australian Department of Agriculture and Water Resources (DAWR) known to perform manual surveillance of some e-commerce sites.

A systematic review of publications was performed via the University of Melbourne Library and Google Scholar using combinations of the search terms [*internet trade*, *online trade*, *e-commerce*] AND [*biosecurity*, *quarantine*, *wildlife*]. Of the 780 papers found via the searches, including duplicates, 25 proved to be unique and relevant. An additional 37 relevant references were identified from citations, by searching for the names of systems as they were identified, and via social and professional contacts. Broad-term searches for software that automatically finds sales of any types of goods were unsuccessful.

To identify what comparable organisations are doing, direct contact was made with representatives wherever possible. Initial contacts were brokered by the NZ MPI project leaders, and were augmented via social and professional networks. Where literature indicated the use of automated data collection methods, the authors were contacted for further details. The entities contacted (some unsuccessfully) are listed in Table 2.1.

Customs agencies were not contacted directly, initially due to the likely restrictions on releasing information about their intelligence systems and processes. The literature subsequently indicated that the 'hard' contraband targeted by Customs agencies – narcotics, cash and weaponry – are not generally traded on the public websites (or 'surface web') that are the focus of this study, but rather on the dark web or other non-public forums. This contrasts with IWT and biosecurity-regulated items, which the literature indicate are openly traded on the surface web (see Section 1 and [21, 22]). Information subsequently obtained from IWT organisations and literature indicated that no known Customs agencies (which usually have a large IWT enforcement role) have any automated capability for detecting online IWT sales. However, one system was identified that collates Customs-related news items.

To evaluate software a set of evaluation criteria are required. A recent study lists evaluation criteria useful for assessing biosurveillance systems [23], which were developed from publications by the U.S. Center for Disease Control [24, 25]. In preparation for the present study those criteria were adapted to the subject matter at hand, and are listed below.

Organisation/researcher	Jurisdiction(s)
Dept. of Agriculture & Water Resources	Australia
Dept. of Human Services	Australia
Invasive Species Council	Australia
Dept. of Agriculture & Food	Western Australia
European Plant Protection Organisation	Europe
Canada Food Inspection Agency	Canada
International Fund for Animal Welfare	International
Global Eye	International
TRAFFIC	International
DEFRA	UK
U.S. Fish & Wildlife Service	USA
USDA–APHIS	USA
Great Lakes Commission	USA & Canada
NSF CIPM	USA
K. Kaminski	
F. Kuhn & L. Humair	
S. Moss	
D. Roberts & J. Hernandez-Castro	

Table 2.1: Agencies, groups and individuals consulted.

1. Acceptability: the willingness of persons and organisations to participate in the system
2. Data quality: the completeness and validity of the data recorded in the system
3. Flexibility: how well the system can adapt to changing information needs or operating conditions with little additional time, personnel, or allocated funds
4. Portability: how well the system could be duplicated in another setting
5. Representativeness: the system is representative if it accurately describes the occurrence of targeted events over time and their distribution in the population by place and person
6. Resources needed
7. Sensitivity: the proportion of actual targeted events that are successfully detected by the system
8. Simplicity: in terms of both structure and ease of operation
9. Specificity: the proportion of reported events that actually relate to the target criteria
10. Stability: the reliability (i.e., the ability to collect, manage, and provide data properly without failure) and availability (the ability to be operational when it is needed) of the system
11. Timeliness: the speed between steps in the system.

In practice, the information available on most systems was insufficient to allow full evaluation against these criteria, so only the list of software requirements presented in Section 3 was considered.

3

Software requirements

This section is based on information provided by the NZ MPI project leaders. DAWR contacts did not provide specific requirements, but based on what is known about their monitoring activities, the requirements listed below would be sufficient for this horizon-scanning project.

Note that these requirements are not a substitute for those that would be developed by a business analyst or similar IT professional. It is recommended that the services of such a professional be engaged before embarking on any software procurement or development project.

Overall description The system will be used by NZ MPI and/or DAWR to automatically identify sales or sale offers of certain goods to buyers in their respective jurisdictions. The target goods are those that are of biosecurity concern to each jurisdiction. The system will automatically collect information on the sale offers, and — as far as technical and regulatory restrictions allow — collect contact details of the buyers and sellers. The information collected will be used within the agency to assist with targeting inspection resources, public relations activities and other risk mitigation measures.

Product functions The core function required of the software is the automatic identification of items of biosecurity concern that are traded on e-commerce sites for sale to New Zealand or Australia.

The software **must**

1. Trawl selected e-commerce sites, such as online auction sites, nursery websites, enthusiast forums etc.
2. Identify items posted publicly for sale that match one or more keywords in a list of targets.
3. Retrieve any contact details associated with each sale item, but only those details that are publicly available and that will not breach privacy regulations etc.
4. Filter and prioritise the potentially large number of hits according to relevance criteria such as organism, quantity, origin, destination etc.
5. Allow the target sites and target keywords to be changed easily as needed.

In addition, the software **should**

6. Identify any URLs of interest, not just known e-commerce sites.
7. Continually refine the specificity and sensitivity of the search parameters based on machine learning or other algorithms.
8. Detect risks that aren't explicitly declared, such as hitch-hiker pests and contaminants.

User characteristics The software would be used by a small number of staff within government departments. Users will probably be familiar with biosecurity regulations and the nature of the import pathways associated with e-commerce, namely postal mail and air freight. Users may also have specialist knowledge of exotic organisms.

4

Activity of relevant organisations

4.1 NZ MPI

The New Zealand Ministry for Primary Industries (NZ MPI, https://www.mpi.govt.nz/) is the regulatory agency responsible for biosecurity functions in New Zealand. NZ MPI initiated this project, which was first focused on seed incursions, but was broadened to include all biosecurity risk material imported via e-commerce. They have looked into software and services available from third party providers for monitoring e-commerce, as well as the potential to perform some mapping of keywords in social media using geographic information system (GIS) software.

It is expected that the intelligence gathered from e-commerce monitoring will be used to

a. increase awareness among online traders of biosecurity requirements by methods such as banner advertising and check-boxes acknowledging understanding of requirements before purchase
b. enhance MPI's ability to detect potential breaches of the Biosecurity Act in online trades.

4.2 DAWR

The Australian Department of Agriculture and Water Resources (DAWR, http://www.agriculture.gov.au/) is the regulatory agency responsible for biosecurity functions at the national level in Australia.

Contacts in DAWR were not aware of any use of web-crawling technology to monitor e-commerce trade in regulated goods. However, two parts of the organisation are known to manually monitor e-commerce sites; their activities are described further below. DAWR has also been involved in the development of an intelligence-gathering system as a successor to the IBIS system. These systems are described in Section 5.

Compliance Division The Engagement, Assurance and Governance Branch within DAWR's Compliance Division conducts manual searches on Google to identify suppliers who ship to Australia. The aim is to gather email addresses of the suppliers, rather

than identifying whether they are selling goods of concern, in order to provide them with information about Australia's import requirements.

The Google searches are for key words, and they focus on the first ten pages of results returned. Their experience is that after the first ten pages the relevance of the matches diminishes greatly. Finding the email addresses on these sites is probably the most time consuming aspect, with some sites only having on-line forms, while others substitute 'at' for '@' in their email addresses to mitigate against automated address scraping. They also use border interceptions data to identify commodities and suppliers for both targeting on eBay and providing compliance information.

With eBay they look at specific commodity types, e.g. certain plants and seed types, and then request eBay to block the site or the commodity type to Australian buyers.

This team also investigates e-commerce sales identified by the Invasive Species Council, whose activities are described in Section 4.9.

Service Delivery Division A recent audit report recommended that DAWR should allow relevant operational staff to access appropriate internet sites such as online shopping sites, to research, develop profiles and assess unidentifiable mail items at the border. [26]

Relevant staff were granted access to selected internet sites in September 2014. This access is reportedly used to

a. research previously unseen commodities to determine their import requirements, and
b. determine categories of mail that are more likely to contain regulated goods, and therefore the level of screening required.

The internet sites were selected in consultation with the DAWR IT Security team, as were the access arrangements for staff. The searches are conducted manually as needed, as opposed to ongoing systematic monitoring. Searches can be both proactive and reactive depending on the situation. DAWR has indicated that these arrangements have resulted in detections of regulated material [26].

4.3 UK DEFRA

The UK Department for Environment, Food and Rural Affairs (DEFRA, https://www.gov.uk/government/organisations/department-for-environment-food-rural-affairs) is the regulatory agency responsible for biosecurity functions in the UK.

The only information obtained about DEFRA's activities relating to e-commerce monitoring is from a 2009 publication, quoted below. Requests for further information were unsuccessful.

> *Increased routine monitoring of Internet sites has been instigated* [in the UK] *in an attempt to identify potential problems at an early stage when effective preventive action is still possible. Although some initial surveys of potentially infringing websites have been conducted at CSL (S. Bishop, pers. comm., 2008), effective monitoring of the Internet would be extremely labour intensive. For example, a search of UK plant sellers on eBay in 2008 produced more than 9000 returns.* [3]

4.4 Germany

The Julius Kühn Institute (JKI, http://www.jki.bund.de/) 'is the Federal Research Centre for Cultivated Plants in Germany and an autonomous superior federal authority directly subordinated to the Federal Ministry of Food and Agriculture'.[1]

The JKI conducted a brief investigation into plants sold into Germany via e-commerce [10]. The study used manual online searches, and found that 89% of the consignments did not comply with phytosanitary import requirements. To the knowledge of the lead author, there has been no further activity in Germany regarding the biosecurity risks posed by e-commerce trade.

4.5 USDA–APHIS

USDA–APHIS (https://www.aphis.usda.gov/aphis/home/) 'is a multi-faceted Agency with a broad mission area that includes protecting and promoting U.S. agricultural health, regulating genetically engineered organisms, administering the Animal Welfare Act and carrying out wildlife damage management activities. These efforts support the overall mission of USDA, which is to protect and promote food, agriculture, natural resources and related issues.'[2]

The following information has been compiled from unpublished literature [27] and personal communication with USDA–APHIS representatives.

USDA–APHIS has a Smuggling Interdiction and Trade Compliance (SITC) section, which seeks to detect and prevent the unlawful entry and distribution of biosecurity-regulated goods. This section includes an Internet team, which handles matters arising from internet trade and works with internet marketplaces to promote compliance.

This team has contracted a third party service provider to perform automated searches of certain e-commerce sites for sales of a number of regulated commodities. SITC originally tried to develop an in-house capability (see Section 5.12), but decided that an external provider could do it faster, cheaper and with better performance. Searches are performed for a small number of commodities, which can be changed at SITC's discretion. The service has increased the number of prohibited items detected, and led to the closure of a number of entry pathways.

The SITC team within USDA–APHIS provides a list of search terms to the service provider. Each commodity search can cover multiple key words. For example, the 'Bamboo' search would include common and alternate names, non-English equivalents, and the names of all relevant genera. The service provider performs the web-trawling function and provides SITC with the results on a regular basis after filtering out unwanted hits. Results can include all sites that the search terms can be found on, but SITC currently ignores foreign-based sites because there is nothing they can do to stop the sales. SITC provide feedback about any hits that aren't relevant, and the service provider refines its search terms accordingly. The service provider makes ongoing enhancements to its web-trawling and triage algorithms to achieve the desired levels of sensitivity and specificity.

[1] http://www.jki.bund.de/en/startseite/ueber-uns.html, retrieved 13 Feb 2016.
[2] https://www.aphis.usda.gov/aphis/banner/aboutaphis, retrieved 13 Feb 2016.

SITC have used two different companies for these services and have been satisfied with both. They have found that the outsourced service has the following benefits:

a. SITC don't have to worry about the details of the technology
b. SITC can readily change search terms as they like
c. The provider caters to SITC's needs, including customising search terms and sites of interest
d. The service is reliable
e. Results are provided in any format requested
f. SITC can contact the provider at any time to discuss, make changes, etc.

The weaknesses identified relate not to the outsourced service itself, but rather to other limitations, namely that the budget restricts the number of search terms, and there is no power to influence the behaviour of foreign-based sites.

When potentially prohibited goods are detected, further action is taken either to reduce the likelihood of the goods being imported, or to determine whether landed goods are compliant. Information about buyers is obtained from the seller, which usually requires a subpoena due to privacy laws in most jurisdictions. This enables follow-up with the buyers. SITC also works with listing companies (such as eBay, Amazon, etc.) in a bid to remove non-compliant listings and to bring the seller(s) into compliance.

4.6 US Fish & Wildlife Service

The U.S. Fish & Wildlife Service (USFWS, http://www.fws.gov/) enforces federal wildlife laws, including those that prohibit the import and interstate transport of listed species.

The Aquatic Invasive Species Branch of the USFWS previously investigated the potential of e-commerce surveillance technology to help understand which species may be in trade that are not otherwise being detected. They currently do not have any projects developing in-house tools, but have worked with the Great Lakes Commission, which is developing a tool of this nature (described further in Section 4.8).

4.7 Canada Food Inspection Agency

The Canada Food Inspection Agency (CFIA, http://www.inspection.gc.ca/) is the regulatory agency responsible for biosecurity functions in Canada.

A spokesperson for the CFIA was not aware of any software being used by the agency to monitor e-commerce trade of regulated goods.

4.8 Great Lakes Commission

The Great Lakes Commission (GLC, http://glc.org) is a compact agency whose stated purpose is 'to promote the orderly, integrated, and comprehensive development, use, and

conservation of the water resources of the Great Lakes Basin.'[3] Membership comprises the U.S. states of Michigan, Minnesota, Pennsylvania, Illinois, Indiana, New York, Ohio and Wisconsin, and the Canadian provinces of Ontario and Québec.

The GLC has been developing and using a system to monitor the e-commerce trade of species that are of concern to the states and provinces in the Laurentian Great Lakes region. This system, GLDIATR, is described further in Section 5.11.

4.9 Invasive Species Council

The Invasive Species Council (ISC, http://invasives.org.au/) is an Australian not-for-profit charitable organisation. It was formed in 2002 'to campaign for stronger laws, policies and programs to keep Australian biodiversity safe from weeds, feral animals and other invaders.'[4]

The ISC has a program of manually searching certain e-commerce sites for sales of nationally prohibited plants. The query strings are based on the list of prohibited species (drawn from DAWR's BICON database, https://bicon.agriculture.gov.au/BiconWeb4.0), and refinements based on experience. Searches are also performed using wildcard characters and fuzzy matching. The sensitivity of the searches is affected by sellers' spelling variations, their use of plants' common names, and the complete absence of plant names. Some search engines impose a maximum character count on the query string, which can limit the opportunities for refining the query strings.

Each sale item returned by each search is manually validated. Items that appear to be genuine offers of prohibited species are collated and forwarded to DAWR for follow-up. All successful search strings are saved for re-use, along with the corresponding URIs for each marketplace.

The ISC states that DAWR

> *has done a great job reducing the sale of prohibited seeds via eBay and other online sellers by engaging with the administrators of these sites and with international suppliers to promote compliance with Australian quarantine laws. In recent searches on eBay we have found far fewer examples of prohibited seeds advertised for sale to Australia.*[5]

4.10 NSF Center for Integrated Pest Management

The National Science Foundation (NSF) Center for Integrated Pest Management (CIPM, http://www.cipm.info/) was established in 1991 to serve a lead role in technology development, program implementation, training, and public awareness for integrated pest management at the state, regional, and national (USA) level. It is based at the North Carolina State University.

[3] http://glc.org/about/, retrieved 1 Dec 2015.
[4] http://invasives.org.au/who-we-are/, retrieved 8 Dec 2015.
[5] https://invasives.org.au/blog/international-company-selling-weed-seeds-into-australia/, retrieved 8 Dec 2015.

Researchers at CIPM have developed and hosted software over the years for use by several agencies in the USA, including USDA–APHIS. The relevant systems, PestLens and ISIMS are briefly described in Sections 5.8 and 5.12.

Requests for further information on the e-commerce surveillance activities of this organisation were unsuccessful.

4.11 International Fund for Animal Welfare

Founded in 1969, the International Fund for Animal Welfare (IFAW, http://www.ifaw.org) aims to rescue and protect animals around the world. Research into IWT via e-commerce forms part of their work.

In 2014 IFAW conducted research on the scale of IWT in New Zealand via popular e-commerce sites [16]. They manually performed broad preliminary searches to determine commonly-traded items and commonly-used keywords, and then used these keywords as search terms for more intensive manual searches.

IFAW would like to have an automated tool that can be used to identify instances of wildlife trade across a variety of platforms, including online marketplaces and social media forums. They have been working with a number of consultants and partners on this problem for several years. These include researchers from the University of Kent who are currently developing the custom software described in Section 5.7.

4.12 TRAFFIC

TRAFFIC (http://www.traffic.org/) is an international NGO which monitors wildlife trade. Some of its publications use survey data to describe the online trade of certain organisms and products, such as the Indian star tortoise *Geochelone elegans* in Malaysia [28], and ivory in China [20].

TRAFFIC's experts began routinely surveying online markets in China for illegal wildlife products in 2010. They initially used manual methods, searching for product names and code words on monitored websites, and recording findings in a database. Data were subsequently screened to remove duplicates and to verify potential IWT advertisements.

An automated program was initiated at the beginning of 2012. This is more comprehensive than the manual approach, and is described in Section 5.6. TRAFFIC is also developing new survey methods and research techniques to overcome the challenges of monitoring the wildlife trade on social media platforms [20].

4.13 Academic research

Several papers were identified that document the results of investigations into the online trade of regulated animal and plant products. All these studies focused on particular

species/commodities, destinations or marketplaces, and the majority of studies relied on manual methods of searching for sale items. These papers are summarised in Table 4.1.

Locale	Commodity/Species	Search method	Reference
Australia	Ornamental marine spp.	Manual	[7]
Brazil	Amphibians & reptiles	Unknown	[11]
Brazil	Ornamental fish	Manual	[6]
China	Ivory	Unknown	[17]
China	Ivory	Unknown	[19]
China	Ivory	Auto	[20]
Germany	Terrestrial plants	Manual	[10]
Great Lakes	Aquatic organisms	Manual	[2]
Italy	Aquatic organisms	Manual	[14]
New Zealand	IWT	Manual	[16]
South Africa	Aquatic plants	Manual	[8]
UK	Vertebrates	Manual	[4]
USA	Endangered plants	Auto & Manual	[29]
USA	*Caulerpa* spp.	Manual	[1]
Unknown	Cactaceae	Manual	[30]
Int'l	Orchidaceae	Manual	[31]

Table 4.1: Attributes of papers researching online trade in regulated animal and plant products.

5

Existing solutions

The intelligence-gathering software identified through this project fall within five subject domains:

1. public health surveillance (PHS)
2. illegal wildlife trade (IWT) monitoring
3. animal and plant health (APH) surveillance
4. biosecurity-regulated trade (BRT) monitoring, and
5. Customs news monitoring (CNM).

The systems in this chapter are described in the order above, and then in alphabetical order.

5.1 BioCaster

The following information has been collated from published literature [32, 33, 34, 35, 36, 37, 38].

Introduction BioCaster is a fully automated experimental system for near real-time 24/7 global health intelligence based at the National Institute of Informatics in Tokyo. Its development was started in 2006. At the beginning of the project the focus was on Asia-Pacific languages due to the perceived risk of newly emerging and re-emerging health threats in the region.

Data acquisition BioCaster automatically collects and processes approximately 30 000 reports per day from over 1700 RSS feeds from

a. EurekAlert! (http://www.eurekalert.org/)
b. European Media Monitor Alerts (http://emm.newsbrief.eu/)
c. Google (http://google.com)
d. Morbidity and Mortality Weekly Report (http://www.cdc.gov/mmwr)
e. MeltWater (http://www.meltwater.com/)
f. OIE (World Organisation for Animal Health, http://www.oie.int/)
g. ProMED (http://www.promedmail.org/)
h. Reuters (http://www.reuters.com/)

i. World Health Organization (WHO, http://www.who.int/)
j. Vetsweb (URL unknown).

The languages of the reports in order of frequency are English, Chinese, German, Russian, Korean, French, Vietnamese, Portuguese, Thai, Italian and Arabic.

Automatic processing The system uses text mining methods, including named entity recognition, to scan the reports for 182 human diseases, 143 zoonotic diseases, 46 animal diseases and 21 plant diseases, as well as 40 chemicals and 9 radionuclides. Approximately 200 reports will be considered relevant after full analysis has taken place.

Central to the system is a rich multilingual ontology, which contains terms such as diseases, agents, symptoms, syndromes and species, as well as geographical locations with their latitudes and longitudes. The ontology allows the text mining algorithms to identify the key concepts, locations and relationships in each report, and to fill in gaps not mentioned explicitly.

Non-English articles are translated into English to determine topical relevance. This was done via Google Translate until December 2011, when the free service was deprecated. Since then the freely available Moses machine translation system (http://www.statmt.org/moses/) has been used, which necessitated the training of custom translation engines for the various languages using parallel texts.

Human processing Nil described for the system *per se*, but the development and ongoing maintenance of the ontology requires human input. The core team involved in BioCaster's development is usually three or four members with expertise in computational linguistics and software engineering. Other academic partners in Japan, Thailand and Vietnam provide expertise in software engineering, public health, genetics and computational linguistics across several languages.

Dissemination The website allows users to view event maps and trend graphs, perform advanced searches and filtering, and subscribe to RSS and Twitter feeds. Visualisation, using Google Maps, allows users to gain a geographically contextualised view of an outbreak anywhere in the world which can be filtered by pathogen, syndrome or text type. Users can drill down to source evidence by clicking on map points which display associated headlines for the event along with links to scientific databases, such as PubMed, HighWire and Google Scholar. News events are also collated into the GENI-DB database, which allows for searching and visualisation of event statistics.

Usage status BioCaster appears to be a defunct project. The most recent tweet (https://twitter.com/biocaster) is dated 12 July 2013, and the system's website (http://born.nii.ac.jp/) is no longer available. The lead author's personal webpage (https://sites.google.com/site/nhcollier/projects/biocaster) indicates that the project was active from 2006 to 2012.

5.2 EpiSPIDER

The following information has been collated from published literature [39, 40, 37].

Introduction The EpiSPIDER project ('Semantic Processing and Integration of Distributed Electronic Resources for Epidemiology') was initiated in January 2006. The system retrieves news items relating to infectious disease threats to public health, and uses natural language processing to transform free-text content into structured data. It is built with open-source software components.

Data acquisition The primary data source for EpiSPIDER is ProMED-mail (http://www.promedmail.org/). Information is also collected from

a. Daylife (http://daylifenews.com/)
b. EuroSurveillance (http://www.eurosurveillance.org/)
c. Humanitarian News (http://humanitariannews.org/)
d. Moreover (http://www.moreover.com/)
e. World Health Organization (WHO, http://www.who.int/)
f. Google (http://google.com)
g. Twitter (http://twitter.com)

The system scans for articles in English only.

Automatic processing Data pre-processing and natural language processing, including concept annotation, is performed via OpenCalais (http://www.opencalais.com) and the Unified Medical Language System (UMLS, https://www.nlm.nih.gov/research/umls/). Location names are parsed and geocoded using Yahoo Maps (http://maps.yahoo.com), Google Maps (http://maps.google.com) and Geonames (http://www.geonames.org). Reports with location information are linked to relevant auxiliary data from the CIA World Factbook (https://www.cia.gov/library/publications/the-world-factbook/) and the United Nations Human Development Reports (http://hdr.undp.org/en). Links to relevant literature are added to each report topic by means of the askMEDLINE query tool (http://askmedline.nlm.nih.gov/ask/ask.php). The SIMILE Exhibit API (http://simile.mit.edu) is used to produce interactive scatter plots, maps and timelines.

Human processing Nil described.

Dissemination Filtered and unfiltered RSS and geoRSS (geocoded RSS) feeds containing summaries of daily ProMED reports are published through the EpiSPIDER web site (http://www.epispider.org). The website provides country-level maps for all countries covered by ProMED reporting, and state-level maps for some countries.

Usage status EpiSPIDER appears to be a defunct project. The last tweet from the Twitter account (https://twitter.com/epispider) was on 6 December 2011, and the website (http://www.epispider.org) is no longer available. In 2009 it received 50–90 visits per hour from users in North America, Europe, Australia, and Asia, including government departments, universities and health research organisations.

5.3 GPHIN

The following information has been collated from published literature [41, 40, 34, 38, 42, 43] and from pages on the GPHIN website (https://www.gphin3.net/).

Introduction The Global Public Health Intelligence Network (GPHIN) was first developed by the government of Canada and the World Health Organization in 1997. The system gathers information about possible disease outbreaks worldwide from various news media sources, and alerts international bodies of such events. After the outbreak of Severe Acute Respiratory Syndrome (SARS), a new, robust, multilingual GPHIN system was developed and was launched in November 2004 at the United Nations. The system is maintained by the Public Health Agency of Canada. The annual operational cost of the system and its administration is C$3 million.

Data acquisition The GPHIN software application continuously retrieves articles from news-feed aggregators that together represent 20 000 news media sources, namely Al Bawaba (http://www.albawaba.com) and Factiva (http://www.factiva.com). Articles are retrieved in nine languages, namely Arabic, English, Farsi, French, Portuguese, Russian, simplified Chinese, Spanish and traditional Chinese. The system tracks events such as disease outbreaks, infectious diseases, contaminated food and water, bioterrorism and exposure to chemicals, natural disasters, and issues related to the safety of products, drugs and medical devices and radioactive agents. Approximately 3000 news reports are processed per day.

Automatic processing English articles are machine-translated into Arabic, Chinese, Farsi, French, Russian, Portuguese, and Spanish. Non-English articles are machine-translated into English. The quality of the translations is generally sufficient to provide the essence of the articles, but an algorithm automatically rates the comprehensibility of the translations.

Articles retrieved by the search queries are categorised as animal, human, or plant diseases; biologics; natural disasters; chemical incidents; radiologic incidents; and unsafe products. Duplicates are identified and removed. Relevancy algorithms assign a score to each article. Articles with high relevancy are immediately released to users, and low score articles are dismissed, but retained in the database. Articles with mid-range relevancy are passed to a human analyst for assessment.

Human processing A multidisciplinary team of analysts provides linguistic, technical, and analytical expertise that are critical to the success of the system. These analysts

a. review and edit machine-translated reports with low comprehensibility scores
b. assess articles with low relevancy to verify that none constitute an alert
c. manually augment the GPHIN database with articles obtained from open access web sites
d. identify trends or possible relationships between events, and
e. develop and maintain the search queries, keyword lists, custom dictionaries and algorithms.

Dissemination Users access the system via a multilingual website that requires user credentials. Users can view reports, generate queries, and build and store personalised search criteria. Notifications about events that might have serious public health consequences are immediately sent by email to users in the form of an alert.

Usage status The website is in active use, with users including government departments and non-governmental organisations involved in public health. Any organisation can request access to the system at no cost, but potential users must demonstrate that they belong to an institution able to apply the information for public health purposes.

5.4 HealthMap

The following information has been collated from published literature [44, 45, 46, 40, 34, 37, 38] and the HealthMap website (http://www.healthmap.org/).

Introduction HealthMap automatically retrieves, parses, filters and reports information about outbreaks of infectious diseases. The system has operated since 2006.

Data acquisition The system retrieves data every half hour from multiple electronic sources, namely

A. health organisations
    a. EuroSurveillance (http://www.eurosurveillance.org/)
    b. Food and Agriculture Organization (FAO, http://www.fao.org/)
    c. GeoSentinel (http://geosentinel.org/)
    d. ProMED (http://www.promedmail.org/)
    e. World Health Organization (WHO, http://www.who.int/), and
    f. World Organisation for Animal Health (OIE, http://www.oie.int/);
B. news feeds
    a. Baidu News (http://www.baidu.com/)
    b. Google News (http://news.google.com/)
    c. Moreover (http://moreover.com/)
    d. SOSO Info (http://www.soso.com/), and
    e. Wildlife Data Integration Network (http://www.wdin.org/newsmap); and
C. user eyewitness reports.

The search criteria include scientific and colloquial disease names, symptoms, keywords, and phrases. The system collects information in Arabic, Chinese, English, French, Portuguese, Russian and Spanish.

Automatic processing Text mining algorithms determine the diseases and locations described by each document, identify and remove duplicate documents, and group related documents together. Once the documents are classified by location and disease, they are stored in a database that is designed to perform a variety of queries and display different views of the data.

To assist with filtering, additional category tags are applied to each document. The categories indicate whether the report relates to a newly discovered outbreak (defined as

an 'alert'), the possibility of an outbreak, or an outbreak that is already known; whether the report is background/contextual information about a disease and its management; or whether it does not relate to any disease or health condition. Alerts are automatically geocoded.

Translation methods are not described in the literature, but the HealthMap website includes a credit for Google Translate.

Human processing After automatic processing, human curators check all reports and correct any misclassifications. Human analysts also develop and maintain the dictionaries and ontologies required by the text mining algorithms. Users can submit URLs or articles directly for consideration. This may be done via the website, by calling a hotline, or by using a smart-phone app.

Dissemination Alerts are reported on the freely-accessible website (http://www.healthmap.org/) via an interactive map, categorised by geography, time, and disease agent. Non-alert reports are accessible via 'related information' links. The website also includes search functionality and simple time-series charts for selected diseases. Tweets are also published on Twitter (https://twitter.com/healthmap).

Usage status The system is in active use. In 2008 the website had approximately 20 000 unique visitors per month, and users included libraries, academia, government departments and multinational agencies. The website has been redeveloped since the last known literature was published.

5.5 Wildlife Trade on HealthMap

The following information has been collated from published literature [47].

The HealthMap system (see 5.4) was adapted to provide a similar service relating to illegally traded wildlife and wildlife products. It is published as a subsidiary site to HealthMap (http://www.healthmap.org/wildlifetrade).

Reports relating to the illegal wildlife trade are retrieved via RSS feeds from

A. Official sources
    a. TRAFFIC (http://www.traffic.org/)
    b. WildAid (http://wildaid.org/)
    c. the Coalition Against Wildlife Trafficking (http://www.cawtglobal.org)
    d. World Wildlife Fund (http://www.worldwildlife.org/), and
    e. the International Fund for Animal Welfare (IFAW, http://www.ifaw.org/); and
B. Freely and publicly available unofficial sources
    a. websites
    b. discussion forums
    c. mailing lists
    d. news media outlets, and
    e. blogs.

During the proof-of-concept period (2010–2011) reports were gathered in English only, with additional monitoring planned for Japanese, Chinese, Malay and Indonesian. Each report collected was reviewed manually by an analyst. The analyst also added geographical details and species information as needed, removed duplicate reports, and checked that reports were correctly tagged. No documentation has been found for the system's current processes.

Users access reports via a website with a similar look and feel to the main HealthMap site. The wildlife map differs by including a geographical layer featuring international airports, due to their importance as entry points for illegal wildlife products.

5.6 Train Browser

Information in this section is taken from published literature [20] and personal communication with TRAFFIC China staff.

The online monitoring system used by TRAFFIC China searches selected e-commerce websites for sales of wildlife items, and retrieves details of matches. This automated program has been in place since January 2012.

The software used to perform this monitoring is a Chinese system called Train Browser, available at http://www.locoyposter.com/ (in Chinese only). The website includes a brief system description, a translation of which is given below.[1]

> *Train Browser is an automatic visual scripting tool. We can set up a script to automatically log in, discern user identification codes, retrieve data, submit data, click on pages, download files, manipulate databases, send e-mails and other operations. It can also use logical operators to test conditions, loop and branch. All features are customisable, powerful and unique script features can be written to assist your work, and separate executable programs can be generated to perform sales.*

At end of September 2014, TRAFFIC was monitoring 25 websites every month for eight wildlife products using 64 search terms including code words. Matches are recorded along with web links, title, description and seller details. Duplicate matches are rationalised, having been determined from product descriptions and seller details. Matches are supplied to the relevant marketplace owners and enforcement agencies.

5.7 iTrade

Information in this section is drawn from webpages (http://csac.anthropology.ac.uk/Research/iTrade and http://gotw.nerc.ac.uk/list_full.asp?pcode=NE%2FL00075X%2F1) and personal communication with the researchers.

Researchers at the University of Kent's Centre for Social Anthropology and Computing (CSAC) have developed software, 'iTrade', that automatically identifies likely IWT items posted for sale on a major e-commerce website. This builds on an algorithm, reported in published literature [48], which identifies likely elephant ivory items in online sales listings

[1] Accessed 2 Feb 2016.

with 90% accuracy. The system obtains data via the marketplace's API, and there are plans to expand the system scope to include other online marketplaces.

The CSAC is now working to further develop this software and make it available to IWT regulators and other agencies via a new business entity, which is envisaged to be non-profit. They have established partners in the UK National Wildlife Crime Unit, the UK Border Force and the International Fund for Animal Welfare. The expansion plans also consider the potential for its application to other areas of regulated trade, such as BRT.

5.8 PestLens

The following information has been collated from the PestLens website (https://pestlens.info/), published literature [49], and personal communication with USDA–APHIS representatives.

Introduction PestLens is a web-based early-warning system used by the Plant Protection and Quarantine (PPQ) program of USDA–APHIS. It is used to retrieve, filter, evaluate and disseminate plant pest information, and to keep track of the agency's actions in response to that information.

PestLens was developed by NSF CIPM. It is a successor to two earlier systems, namely Exotic Pest Information Collection and Analysis (EPICA) and the Offshore Pest Information System (OPIS). These legacy systems are described in [50, 51, 52, 53].

Data acquisition Information is collected regularly by a small team of specialist analysts, who monitor selected online sources including scientific journals, plant-health websites, specialist e-mail lists and NPPO reports. They also use automated custom queries in Google Alerts (https://www.google.com/alerts) to search other sources such as newspapers, trade journals and blogs. Results of these queries are received via email. Users can also submit information for consideration, and can make comments on published articles.

Automatic processing Nil described.

Human processing Analysts assess the relevance, value, specificity and reliability of all collected information. Where dissemination is warranted, the analysts summarise the information into a succinct article, along with regulatory context and pertinent background information about the pest. Citations, metadata and links to related PestLens reports are also included.

Dissemination Articles are distributed in a weekly e-mail to all PPQ employees, and other selected groups and individuals. The e-mail notification is also available to anyone else upon request. Each e-mail contains 3–7 articles, depending on the number and relevance of information gathered, and the capacity of the analyst team. A searchable archive of all articles is also available to users.

PestLens also includes functionality for coordinating decisions made by various parts of the agency in response to articles, and for tracking the status of action items.

Usage status PestLens remains operational.

5.9 IBIS

The following information has been collated from published and unpublished literature [54, 38, 55], from the IBIS website (http://biointel.org/home), and from personal communication with the system developer.

Introduction The International Biosecurity Intelligence System (IBIS) gathers publicly available information on the internet about the emergence and spread of pests and diseases, in order to inform biosecurity policy. IBIS is the successor system to AquaticHealth.net, which first went online in January 2010. The system was administered by CEBRA, but was handed over to DAWR in mid-2016.

Data acquisition The system searches selected online sources for plant and animal disease reports, industry reports, articles from relevant journals, and any other articles or comments that may be relevant to biosecurity. Sources include search engines, RSS feeds, blogs, social network sites, specialist news sites, newsletter subscriptions and Twitter. The sources are searched each day, week or month, as appropriate to each source.

Automatic processing The system identifies and deals with duplicates. Third party web services including AlchemyAPI (http://www.alchemyapi.com/) and GeoNames (http://www.geonames.org/) are used to extract information from the article such as the title, text, author, language and locations. If necessary, articles are translated to English using Bing Translator (http://www.bing.com/translator) or Google Translate (http://translate.google.com/). The quality of translations is deemed sufficient for intelligence-gathering purposes. Articles are automatically assigned categorical tags that are appropriate to the article, based on the search criteria that resulted in the match, and metadata retrieved from the match. This approach obviates the need for natural language processing. All processed reports are uploaded to a MySQL database, which can be viewed via the system's website. The website is built with the open source content management system Drupal (https://www.drupal.org/).

Human processing A large number of active users are subject matter experts with technical backgrounds in aquatic, plant or terrestrial/animal diseases. Instead of using automatic algorithms, IBIS draws on users' collective expertise with the aim of promoting an active and collaborative user community with control over scanning, gathering and content. The subject matter experts

a. Evaluate the relevance of articles
b. Edit the articles
c. Ensure that automatically extracted dates and locations are correct, and that automatically assigned categories and tags are correct; and
d. Maintain the search terms and lists of tag words.

The relevance of each article affects how it is treated in the system. All relevant articles are retained in the system; irrelevant articles are deleted. Of the relevant articles, some

are deemed important enough to be promoted to the website's home page and be included in the daily email digest. The most serious articles warrant an immediate alert email to users.

A small number of users create and maintain search queries, and can add and edit search sources, pests, diseases, hosts, qualifiers and blocked sites. All users are able to add comments to articles, and submit articles directly for consideration.

Dissemination All users can receive the daily email digest, any ad hoc alert emails, and can access and search the database of collected articles via the system's website. The daily digest and alert emails are customisable to suit each user's interests in terms of information sources, geographical location, pests/diseases, and pest/disease hosts. The user preferences also affect which articles are displayed on the user's home page, and in the user's optional RSS feed. The home page displays recent and/or filtered articles as a list and via a clickable map.

In addition to the customised article views shown on the user's home page, there are three pre-defined sites for broad subject areas, namely plant health, aquatic animal health and terrestrial animal health. The articles shown on these pages override the user's custom preferences.

Usage status The system is in active use, but is due to be deactivated pending successful implementation of the replacement system (see Section 5.10). Deactivation is expected by the end of 2017.

5.10 IBIS Replacement

The following information has been collated from personal communication with the IBIS system developer and personal contacts within DAWR.

In light of limitations in IBIS and changes to its funding arrangements, DAWR decided to commission a replacement system for IBIS. The new system provides improved functionality, and can accommodate multiple user communities.

The search component has been separated from other parts of the system, allowing it to operate with any search criteria supplied by the user. Translation capability has been retained and the web-scraping capability improved. An automated synonym builder has been added, and the natural language processing capability deals with large volumes of false search matches and duplicates. This reduces the importance of query string specificity and removes the need for much of the human processing of articles required by IBIS.

Rather than being limited to a few pre-defined subject areas, as with IBIS, the new system allows for as many sub-sites as desired. The owners of each sub-site control their own membership, add and maintain their own search parameters (query strings, target URLs etc.), and can choose which of the other available functions to include, such as mapping and advanced statistics. This makes the system more suitable for various user communities who want different things from the system.

Other new features include advanced filters for searching articles, information dashboards, and simplified user privilege systems.

The system architecture is provided by a third party (the developer, YesWeCan https://yeswecan.digital/), so clients other than DAWR should be able to purchase access to the system for their own purposes. The architecture is yet to be formally named, but DAWR's deployment of it will reportedly continue to be called IBIS.

The system is expected to begin formal operation in early 2017, and will run concurrently with IBIS for approximately six months. At the end of this period, assuming the new system proves successful, IBIS will be deactivated.

5.11 GLDIATR

Information in this section has been collated from published literature [56], a page on the GLC website (http://glc.org/projects/invasive/internet-trade-ais/), and personal communication with the GLDIATR project manager.

Introduction The Great Lakes Commission (GLC) has developed software called the Great Lakes Detector of Invasive Aquatics in Trade (GLDIATR). This software searches the internet to identify online sales and sellers of aquatic invasive species of concern to Great Lakes stakeholders, and gives users access to the collated information. The project was initiated in 2012 with a grant through the U.S. EPA-led Great Lakes Restoration Initiative. The system was formally deployed in August 2014.

Data acquisition The system searches the online retailers Amazon and eBay, and the search engine Bing, via their respective APIs. Query strings currently use the names of target species as the primary search terms, but in future may include generic terms such as 'for sale' or 'aquarium'. The search specificity is improved by excluding domains known to be of no interest, such as .gov websites, Wikipedia.org and YouTube.com. Searches are carefully scheduled to maximise the APIs' service within their respective terms of use (e.g. some APIs restrict the number of searches that can be done free of charge per month). Pages matching the search are automatically downloaded.

The GLC had intended to search relevant regional Craigslist sites (such as https://detroit.craigslist.org/), but Craigslist does not provide a suitable API, so automated searching and scraping of the web-pages would be required. However, this is against their terms of use (https://www.craigslist.org/about/terms.of.use), and may result in being blocked permanently.

Selected users can also perform one-off searches of specific websites. This sees all pages within the selected URL automatically scanned for the target species.

Automatic processing The content of each downloaded page is scanned for target terms, and a match score of 0–1 assigned. A classifier algorithm, originally trained on approximately 3200 webpages, then determines whether the page is a 'sale page'. The seller's location information is collected automatically for identified eBay sale pages, and would also be collected for Craigslist matches.

Human processing Nil described.

Translation Nil described.

Dissemination Information is available to authenticated users via a web-based dashboard. General users can access, search and filter the list of target species and the reports of identified sale page URLs.

Usage status The system is in active use. The U.S. EPA awarded a second grant to the GLC in May 2016 for continued use of and improvements to the system.

5.12 ISIMS

The Invasive Species Internet Monitoring System (ISIMS) was developed by NSF CIPM as joint project with USDA–APHIS. Unless otherwise specified, the following information has been collated from published literature [57, 58], from a web-page describing an earlier version of the software (https://ncsu.pure.elsevier.com/en/projects/agriculture-internet-monitoring-system-project-aims-fast-data-sea), and from personal communication with a USDA–APHIS officer. Requests for additional information from the system designers were unsuccessful.

Introduction ISIMS was developed to automate searches of internet sites involved in the sale or trade of certain species. The software was first developed in 2002, and was previously known as AIMS (Agricultural Internet Monitoring System).

Data acquisition The system uses the FAST Search and Transfer[2] engine to automatically crawl the web to identify sites that match search parameters. Web documents, news feeds, blogs and publicly accessible chat rooms are searched, and more than sixty languages can be queried. The search results and metadata, including URIs and owner details, are downloaded to a database.

Automatic processing Nil described.

Human processing Subject matter experts evaluate the relevance of the documents returned by the search. Potential violators are sent correspondence by enforcement officers.

Dissemination Nil described.

[2]FAST was acquired by Microsoft in 2008 (http://www.wsj.com/articles/SB119978543120974587, retrieved 15 Feb 2016).

Usage status The system was used until 2007 or 2008 by USDA–APHIS SITC to search websites for goods of concern based on keywords. The search terms were not specific enough, so it returned too many false positives. Work began on improving the search parameters, but the decision was made to stop using the software completely due performance problems and high costs. As described in Section 4.5, the search function has since been outsourced to a specialist company, which reportedly provides a more cost-effective service than could be provided in-house. The estimated cost of running this system was quoted in an unrelated paper as US$75 000 p.a.[3]

In 2002 ISIMS was also adapted to detect internet sales of endangered species. The results demonstrated proof-of-concept well, but there were too many false positives returned by the query parameters, and the results had to be constantly reviewed manually to determine their relevance. This labour-intensive aspect limited the system's usefulness.[47]

According to a USDA–APHIS representative, the system is no longer in use due to its high cost and inadequate performance.

5.13 ETH Zürich

Information in this section is drawn from personal communication with the authors of the paper in question.

The authors of a recent paper [13] developed software to automatically trawl online auction sites for sale offers of invasive plants. The software is written in Java and relies on JSoup (http://jsoup.org/) for scraping website data. Given a predefined set of plant identifiers it autonomously performs daily searches for matches on different e-commerce sites. For each item matched they retrieved the name of the seller, the item location, and the countries available for shipping. The retrieved data are stored in a MySQL database, and subsequent processing performed with Java, SQL and Python. The software was developed solely to collect data for research purposes and it would need additional development to make it suitable for regulatory use.

5.14 IRIS

Information in this section is collated from http://www.wcoomd.org/en/media/newsroom/2014/december/wco-launches-iris.aspx and http://www.wcoomd.org/en/media/wco-news-magazine/previous/~/media/7C878D64C5664C7CAFDFFEFD804216B9.pdf.

The Intelligence Resources Information System (IRIS) was developed by the World Customs Organisation (WCO), and launched in December 2014. It collates Customs-related news from public websites, and makes it available on a webpage and via a daily digest. It is also used by the WCO to publish information about major Customs seizures, as advised by WCO member organisations.

The system automatically retrieves Customs-related news items from public news websites by means of web crawlers. News items are drawn from a wide variety of sources in English and French. Results from the web crawlers are stored on a database. Authenticated users and WCO member organisations can also submit news items for potential publication.

Recent results are presented on an interactive world map on the public webpage https://iris.wcoomd.org/. Authenticated users can subscribe to daily digests and set up customised alert emails, and can perform searches on the database and download results as Excel files.

The system is in active use.

6

Discussion

6.1 Elements of intelligence-gathering

6.1.1 Data sources

The data sources chosen for interrogation will depend on the type of information being sought. For example, PHS and APH systems monitor news sites, aggregators, social media, and relevant academic publications and official organisations. Crowd-sourced content is also gaining in importance for PHS [59]. The sources of primary interest to BRT monitoring would be e-commerce sites, enthusiast forums, general search engines and possibly social media. The experience and knowledge of inspectorate staff and advocacy groups would be a valuable source in this regard.

E-commerce platforms Identifying specific e-commerce platforms for monitoring is outside the scope of this project, but those commonly encountered in the literature include eBay, TradeMe, Amazon, Gumtree and Alibaba. Apps are an emerging e-commerce platform, which do not require any 'traditional' web presence. Trade is conducted within the app itself, often on a direct user-to-user basis, but they may also have themed user groups.

Aggregators Aggregators are a useful way to source data from potentially thousands of disparate sources. Most PHS and APH systems rely heavily on news aggregators, such as Factiva, Al Bawaba, GoogleNews and Moreover [40]. There are no known general-purpose aggregators for e-commerce, although there are increasing numbers of specialist aggregators such as Skyscanner (travel: http://www.skyscanner.com/), Lyst (fashion: http://www.lyst.com/), and Hotels.com (accommodation: https://www.hotels.com/). If a general-purpose e-commerce aggregator becomes available, it may be a useful source of information for BRT monitoring.

Social media A recent study — the first of its kind — demonstrated that orchids, including potentially endangered species, are being openly traded on multiple specialist forums on a large social media platform [31]. This indicates that social media sites should be considered as sources of information on trade in regulated goods.

There are significant challenges to the monitoring of social media by regulators. Firstly, access to a social media circle is often limited by membership, which may be authenticated, and may only be available by referral. Secondly, dealers may monitor the activity of their followers, and revoke membership to those who make no purchases, who ask uncomfortable questions, or who have a government IP address. TRAFFIC is working on methods to overcome such challenges when monitoring IWT [20], but these have not been investigated for this report. The legal issues associated with monitoring social media — such as privacy requirements, methods for masking government IP addresses etc. — have also not been considered for this report.

6.1.2 Language of sources

When the performance of BioCaster, EpiSPIDER and HealthMap were compared, it was found that performance is significantly affected by the sources accessed and the terms, languages, and regions covered [37].

The reliance on English language sources for obtaining PHS information leads to gaps in coverage due to inherent bias; sourcing additional information from purely local media outlets, particularly those publishing in languages other than English can address some of these gaps [60]. Another study showed that 97% of detected PHS signals were taken from English, Spanish, French, Russian, Portuguese, Arabic and Chinese sources [61].

Monitoring e-commerce trade is likely to have similar challenges. The commodities that are of biosecurity concern will vary between jurisdictions, as will the domestic market segments that are seeking to purchase these commodities online. The languages spoken by the various market segments may help decide which languages to cover in a surveillance program, and therefore which data sources to monitor.

6.1.3 Data acquisition

RSS feeds push information to the subscriber in XML (Extensible Markup Language) format, but the content is often limited to a brief headline and a URL link. This is a simple way to streamline manual searches of target websites. Ingesting RSS data is relatively straightforward via a newsfeeder or other software.

Some e-commerce sites provide RSS feeds, including eBay[1] and TradeMe[2]. E-commerce sites often allow for user accounts that include 'watch list' terms. Using this functionality, in tandem with an RSS feed, would exploit the in-built search features of the e-commerce sites themselves.

If a target website doesn't provide an RSS feed, one can be constructed by means of third-party services. These services can be found via a web search, or via review sites such as http://techpp.com/2009/04/27/top-10-free-tools-to-create-rss-for-any-website-2/.

Automatically retrieving information from a web-page is more complex than from an RSS feed. It can be performed with a web-scraper, but this may be against some websites'

[1] http://www.ebay.com.au/gds/eBay-RSS-FEEDS-FOR-SELLERS-AND-BUYERS-/10000000000756357/g.html

[2] http://www.trademe.co.nz/Help/RSS.aspx

terms of use, and some websites actively block such methods. A regulator may be able to negotiate with website owners for access.

Scraping each webpage will usually require a script specific to that page, which is needed to isolate informative content from non-relevant content such as advertising, unrelated headlines and hyperlinks. These scripts must be updated whenever the layout or functions of each target webpage change, which can occur without warning and may cause down-time. The number of custom scripts and their constant manual revision could become overly burdensome, and limit scalability. [4, 35]

Instead of scraping, it would be preferable to retrieve data via an API (application programming interface) or even direct database access, wherever these options are available. This is the primary approach used by the GLDIATR and iTrade systems. Large e-commerce sites usually have APIs available, such as eBay (https://go.developer.ebay.com/what-ebay-api), Amazon (https://developer.amazon.com/public/apis), TradeMe (http://developer.trademe.co.nz/) and AliExpress (http://portals.aliexpress.com/help/help_center_API.html). One advantage of APIs, other than removing the need for scraping, is that search results are not filtered by the algorithms that otherwise tailor results to the user's perceived preferences. Another advantage is that the metadata provided by the API may be sufficient for identifying items of interest, obviating the need for analysis of raw unstructured text.

The software requirements for this project include the capture of contact details of buyers wherever possible. If the details appear on the webpage matched by the search, they will be retrieved along with the rest of the text on the page. Details that aren't publicly available, such as those behind passwords and CAPTCHAs, normally cannot be retrieved. However, the system used by TRAFFIC China — Train Browser — is described as being able to log into websites automatically and deal with user credentials.

While sale offers are relatively easy to monitor, it is more difficult to determine the buyers and destinations of actual sales, because such information is usually not publicly available. Unless a buyer states publicly on a social media platform that he/she has bought a commodity of concern, and that the buyer's name and location are public, then a subpoena would normally be required to obtain such information from the social media platform.

6.1.4 Query strings

Performing searches on target websites usually requires a suite of query strings, which must be carefully crafted to balance specificity and sensitivity. In developing query strings the input of subject matter experts is essential [58], and inspectorate staff would be invaluable in this regard. The query strings can also be informed by historical interceptions data, reports of incursions, and general browsing [47]. Developing the suite of query strings is an iterative process of trial and error.

One example of a query string from the IWT realm is a search for 'poaching'. This term in isolation returns a large number of irrelevant matches (e.g. from cooking websites), so qualifiers are added to improve the specificity, such as 'in title: poaching wildlife' and 'in title: poaching –bank –egg.' [47]

In the APH realm, PestLens users found that queries on specific pests are not very effective due to the unwieldy number of pests and synonyms, the need to continually update queries as the target organism lists change, and the large number of irrelevant matches. PestLens has found more success with generic search terms, such as 'first report', 'new detection', 'new host' etc., combined with modifiers to exclude irrelevant results. [49]

Query strings must be continually monitored for sensitivity and specificity — including in response to changes in the market and trader behaviour — and adjusted accordingly. This will be a crucial function no matter which system is chosen, but the above examples suggest that the scale of the job will be reduced by using generic queries with well-crafted excluding modifiers.

6.1.5 Text analysis

Once the full text has been retrieved from a page of potential interest it needs to be processed to determine its relevance, category, urgency etc. The number of pages returned by a search may be very large, making manual processing too time-consuming. Text mining algorithms can process large amounts of text from many documents, and can assist with tasks including

- *Data cleaning*, including the removal of irrelevant content such as advertisements, and the restoration of broken sentences
- *Data triage*, including the identification and removal of duplicate and irrelevant documents
- *Fact extraction*, such as names, locations, dates and other key words; and
- *Ranking* to determine relative relevance of each document. [35]

One example of useful automation was the recent development of an algorithm for identifying likely elephant ivory items from over 1000 potential sale items. The potential sale items were obtained via a simple keyword search on eBay UK, and the algorithm was trained with the input of subject matter experts. The algorithm subsequently achieved prediction accuracy comparable to that of the experts, despite using less data than those experts had available to them. [48]

The GLDIATR system includes a classifier algorithm trained on approximately 3200 webpages that automatically determines whether a page is a 'sale page'. The system also includes machine learning components intended to iteratively improve the classifier algorithm, but initial attempts at improvement were not successful. The published report on the system [56] discusses this issue at pages 6, 7–8, and 20–21.

In order for text mining techniques to interpret the text well, the properties and relationships between entities in the subject area may need to be formally defined in one or more ontologies. BioCaster relies heavily on an advanced custom-built ontology, which enables the system to know, for example, that 'A(H1N1) influenza', 'swine flu' and 'swine flu A' all refer to the same disease [35]. The HealthMap system also makes use of an ontology, which at last published count in 2008 consisted of over 2300 location and 1100 disease patterns [46]. The collective knowledge of inspectorate staff would be a valuable source when developing ontologies for e-commerce surveillance.

Text mining algorithms and ontologies have the potential to improve the performance of a large-scale e-commerce monitoring system, but they require significant investment to develop and maintain. The algorithms and ontologies would need to be custom-built for the subject matter of interest. This study has not identified any existing algorithms or ontologies that are immediately ready for use in an e-commerce biosurveillance system.

Depending on the data sources used, translation to or from English may be required. Machine translation would normally be applied before any text analysis steps, to avoid the need to develop and maintain rule books for each language [35]. Various machine translation services are available from commercial companies such as Google Translate or Microsoft's Bing Translate. An open source option is Moses (http://www.statmt.org/moses/). Locally installed translators, including Moses, have the advantage of being able to be customised with specialist vocabulary [35], but this comes with a development and maintenance overhead. No one translator need be selected: the GPHIN system uses a variety of translation services, based on which performs best for a given language [40].

6.1.6 Detecting implicit risks

Of the preferred software requirements, the desire to detect unstated risks such as hitchhiker pests and other contaminants is particularly difficult. None of the candidate software identified have this sort of functionality built-in. They all rely on query strings and text analysis to detect key words explicitly stated in text or metadata.

It may be possible to identify key words that can adequately predict the presence of certain contaminants. For example 'used tyres' may be a reasonable predictor of soil contamination. Complex analysis of observational data will be required to identify potential predictor terms and their sensitivity and specificity. This analysis is not within the scope for the present study.

6.1.7 Human vs. automated processing

A key difference between several systems identified is the degree to which they rely on automatic software to perform various functions, as opposed to human beings. The BioCaster, EpiSPIDER and GPHIN systems use automated methods for all or most stages of data collection, processing and dissemination [35, 42]. In contrast, IBIS and PestLens rely more heavily on a human users.

While automated approaches are useful for dealing with large volumes of data, they often don't perform as well as humans when it comes to assessing the quality of information or adding value to it, and are prone to error particularly when dealing with ambiguous terms [38]. As an example, the ability of PHS systems to detect previously undetected disease outbreaks is 53% better with human moderation of raw search results than without [61].

IBIS and its predecessor, AquaticHealth.net, were designed to capitalise on the respective strengths of humans and automation. Finding and organising the large quantities of information is a tedious task that is performed well by software, whereas humans are better at (and more interested in) understanding and making judgements about the information collected. [38, 62]

6.2 Promoting and enforcing compliance

Targeted information campaigns can be informed by the intelligence gathered through an e-commerce monitoring program. One example of outreach activity, described in an official audit of DAWR, could be expanded to include such intelligence:

> *In 2013, [DAWR] contacted 5165 Australians who had received non-permitted plant or seed imports through international mail. They received a brochure informing them of Australia's biosecurity import requirements, with links to the department's website. Of the recipients, 345 sought additional information from the website and a further 41 contacted the department directly.* [26]

Active engagement with the managers of online marketplaces can also yield results, and the activities of DAWR and ISC, described previously, are one example. TRAFFIC and IFAW have also worked successfully with marketplace managers and regulators to reduce IWT in those forums ([20], http://www.ifaw.org/australia/news/click-delete-%E2%80%93-website-selling-endangered-wildlife-update and http://www.ifaw.org/australia/news/progress-making-websites-endangered-species-free-zones).

USDA–APHIS' SITC team has had success in closing down non-compliant e-commerce sites, removing prohibited items from sale, and in bringing vendors into compliance. Some non-domestic sellers have responded to SITC's engagement by including disclaimers in their posts that items cannot be shipped to the USA due to biosecurity regulations. However, SITC's scope for enforcement activity is constrained by factors such as jurisdictional reach and the anonymity and global nature of the online market. [27]

In its first month of operation GLDIATR identified over 200 unique websites and sellers offering 58 target species for sale. The GLC sent information about invasive species regulations and best practices to 162 sellers. Eleven direct (non-automated) responses were received, and in 27 cases the GLC observed subsequent changes to seller inventory and shipping restrictions. [56]

Targeted banner advertising may be an option where the owners of online marketplaces are unwilling to restrict sales directly. The costs and benefits of such advertising would need careful analysis, particularly given the wide availability of software that blocks advertising.

The Australian Department of Human Services (DHS) has an arrangement with eBay to obtain the names, addresses and birth dates of sellers with sales believed to be more than AU$20 000 p.a. DHS then cross-matches these names against lists of welfare claimants to detect potentially fraudulent claims. This arrangement, which does not require any in-house capability for monitoring e-commerce, may be instructive for other enforcement activities.

6.2.1 Understanding the market

The characteristics of the market and the behaviours of its participants must be considered when setting up and maintaining a monitoring program, and specifically when determining target sites, languages and query strings, and the frequency of searches. Inspectorate staff, advocacy groups and historical inteceptions data would be valuable sources of such information, along with literature such as that listed in Table 4.1.

By way of example, the following issues were identified by researchers investigating how to control e-commerce trade in goji berry plants (*Lycium barbarum*) and Japanese maple plants (*Acer palmatum*) destined for the UK:

1. *In addition to the larger traders who have their own websites, smaller nurseries are also advertising their products on third-party websites*
2. *Plant sharing/trading websites facilitate the direct exchange or sale of plants between members of the public*
3. *Some suppliers are traders rather than growers, buying in the plants and selling them on quickly, often in very large quantities*
4. *The supply chain may be complex, e.g.* L. barbarum *plants supplied to the U.K. from China via intermediaries in the Netherlands and Guernsey.* [3]

The use of trading apps on smartphones etc. is an emerging phenomenon. As with social media sites, transactions made on such platforms would only become apparent if a package were subsequently intercepted via inspectorate activities, or if the buyer/seller discusses the trade publicly in an online forum that is being effectively monitored.

The experience of IWT enforcement programs can provide lessons for the nascent regulation of online BRT. Some programs have found that as certain product names and code words are targeted by regulators and website managers, the traders respond with new code words or descriptions [20].

While the number of IWT products monitored by TRAFFIC has grown over time, the number of keywords for those products has grown at a faster rate. In January 2012 there were 9 keywords across 5 product types (1.8:1), but in September 2014 this had risen to 64 keywords across 8 product types (8:1). TRAFFIC's published longitudinal data on sales of new illegal wildlife goods on traditional e-commerce sites show sharp drops in the first half of 2012, remaining roughly steady thereafter [20]. However, this may not necessarily correspond to a reduction in trade.

As regulators take measures to limit sales of regulated goods on traditional e-commerce sites, trade may increasingly shift onto alternative platforms. The use of social media and instant messaging services has already been observed in the IWT sphere [19, 20]. IWT on the dark web (that part of the internet that is not indexed, and therefore not discoverable via traditional search engines) appears at present to be negligible [21, 22]. This may be due to ineffective regulation of trade on the surface web.

6.3 Potential solutions

The main options for addressing the needs listed in Section 3 are

1. Procure a suitable existing system
2. Collaborate on further development of a suitable nascent system
3. Outsource the surveillance function to a third party service provider
4. Develop a new system in-house
5. Have a new system built by a third party.

Options 1–4 are discussed in further detail in the sections that follow.

Several papers that compare PHS systems point out that those systems are complementary to one another: each system has different strengths and weaknesses, and no one system clearly outperforms the others [40, 34, 37, 61]. This situation may be repeated as the maturity of BRT- and IWT-surveillance systems improves, in which case their complementary strengths could be pooled through inter-agency collaboration to provide a more comprehensive picture than a single system would allow.

Several contacts in the IWT and BRT realms expressed interest in the advancement of automated e-commerce monitoring capabilities. This may present opportunities for collaboration, be it on software development or procurement, sharing knowledge and experience of marketplaces and trader behaviours, identifying data sources, formulating effective and efficient query strings, or building text mining algorithms and ontologies.

6.3.1 Summary of existing systems

The PHS systems identified — BioCaster, EpiSPIDER, GPHIN and HealthMap — may not be readily adaptable to BRT surveillance, but they are instructive in terms of the technologies they use, the sources they monitor and their approaches to analysing and interpreting search results. They are well-advanced in terms of automated translation, relevance ranking and named entity recognition.

Most of the PHS systems identified have or had an operational life of more than five years, with GPHIN being the longest running at almost twenty years. The maturity of the PHS realm is reflected in the number and technical capabilities of the systems, the depth and breadth of their user communities, and the quantity of peer-reviewed literature describing and comparing them.

The APH and CNM systems identified — PestLens, IBIS and IRIS — use similar sources and methods to the PHS systems, because their primary purpose is to provide situational awareness of relevant news items and other publications. All three systems are in active use, although IBIS will soon be replaced by a system that will make it suitable for conducting a wider variety of searches, including the monitoring of BRT and IWT via e-commerce.

Existing systems for monitoring IWT are under various stages of development. The Wildlife Trade sub-site to HealthMap is well-developed, but it doesn't monitor the trade directly; it is focused on situational awareness much like the PHS, APH and CNM systems. In contrast, the other two IWT systems identified — Train Browser (as used by TRAFFIC China) and iTrade — do focus on detecting online sales of regulated goods. The iTrade system is being developed in consultation with expected IWT regulatory users. TRAFFIC China has been using Train Browser to monitor online trade, but the system is based in China, and available only in Chinese, which limits its direct utility to non-Chinese clients. These two systems are instructive in terms of their data sources, query strings and data acquisition processes, since they deal with the structure and content of e-commerce platforms.

Existing systems that monitor BRT are in various stages of maturity. The IBIS Replacement is in the final stages of development, and the generic nature of its architecture should make it suitable to monitoring BRT. GLDIATR is in active use, and funding has been secured for further development. ISIMS appears to meet the requirements, but its use

was abandoned in 2007 or 2008 due to issues with performance and cost. The software built by ETH Zürich researchers is currently suited only to research purposes, and would require further development before being suitable for use by a regulator. It can also be instructive for e-commerce monitoring because it retrieves important details including the name of the seller, the item location, and the countries available for shipping.

6.3.2 Evaluation of existing systems

When procuring software or an outsourced service any organisation should consider how well its needs will be met over the longer term. By way of illustration for the case at hand, the core functions of IBIS (search, retrieve, process, disseminate) were developed to address the specific requirements of a discrete project within DAWR. Those core functions would be useful to many other intelligence-gathering programs in DAWR and elsewhere, but the architecture of IBIS limits its use to its original purpose. This limitation, in part, led to the decision to completely redevelop the system. Another system, ISIMS, was also built for purpose, but the costs associated with performance improvement and maintenance led to the decision to abandon the software and outsource the intelligence gathering function.

A business analyst or similar IT professional should be engaged to advise and/or conduct a formal software evaluation process against a fully developed set of functional requirements. As part of such an evaluation it may be instructive to consider the criteria developed in a recent study [23] that assessed published evaluations of several PHS systems. Those criteria were developed in reference to the Centers for Disease Control and Prevention surveillance system evaluation guidelines [24], which are very comprehensive.

For the present study it is only possible to provide a preliminary assessment of the software against the requirements given in Section 3. Some of those requirements allude to specific technologies, which are better rendered in 'technology-agnostic' terms. This enables a fairer assessment of each system by allowing for the possibility of adapting non-BRT systems to a BRT-monitoring role, and for systems still under development. The assessment criteria and associated scoring values used to assess the systems identified in this study are as follows.

- Initial triage criteria
  - A. Currency: is the system in active use (2), in development (1), or defunct (0)
  - B. Adaptability: can the system be used for BRT monitoring without architecture changes (5), with minor changes (3), or with major changes (1)
- Specific criteria adapted from original software requirements
  - C. Does the system automatically interrogate selected sources for matches against a set of search terms: yes (1) or no (0)
  - D. Does the system retrieve sellers' contact details automatically (2), potentially automatically after some development (1), manually via the retrieved text (1), or not at all (0)
  - E. Does the system prioritise or filter results based on relevance: yes (1) or no (0)
  - F. Can the search parameters (sources and terms) be changed easily (3), with a little difficulty (2), or with greater difficulty (1) such as via a formal change to source code

G. Can the system identify new sources from the internet at-large automatically (3), indirectly via routine matches in a generic search engine (2), potentially automatically after some development (1), unknown (1), or not at all (0)

H. Does the system include capability, such as machine learning, to automatically improve the performance of its search function: yes (1) or no (0)

I. Can the system detect tacit threats, such as hitch-hiker pests: yes (1) or no(0).

Assessing each system against these criteria was hampered in some cases by the lack of available information, but scores have been assigned based on the information at hand. It is hoped that the descriptions provided in this paper will be sufficient to determine which systems and technologies warrant further investigation or investment. The scores assigned to each system are summarised in Table 6.1.

	Triage			Mandatory				Desired			
System	A	B	Subtotal	C	D	E	F	G	H	I	Total
BioCaster	0	1	1	1	0	0	1	2	0	0	5
EpiSPIDER	0	1	1	1	0	0	1	2	0	0	5
GPHIN	2	1	3	1	0	1	1	0	0	0	6
HealthMap	2	1	3	1	0	1	1	2	0	0	8
Wildlife Trade on HealthMap	2	1	3	1	0	1	1	2	0	0	8
Train Browser	2	1	3	1	2	0	1	1	0	0	8
iTrade	1	3	4	1	1	0	2	1	0	0	9
PestLens	2	1	3	1	0	0	2	2	0	0	8
IBIS	2	1	3	1	0	1	1	2	0	0	8
IBIS Replacement	1	5	6	1	1	1	3	3	0	0	15
GLDIATR	2	5	7	1	2	1	1	1	1	0	14
ISIMS	0	1	1	1	1	0	1	1	0	0	5
ETH Zürich	1	3	4	1	1	0	2	1	0	0	9
IRIS	2	1	3	1	0	0	1	0	0	0	5

Table 6.1: Assessment scores for software suitability to BRT monitoring.

The triage subtotal is an initial indicator of whether the system may be a viable option for BRT monitoring. Systems with a score of three or less are either defunct, difficult to adapt to BRT monitoring, or both. Based on these initial scores the four systems that warrant further consideration (highlighted in the table) are the IBIS Replacement, GLDIATR, iTrade and the ETH Zürich system.

IBIS Replacement The system with the highest preliminary score of 15 is the IBIS Replacement. At the time of writing, this system is in the final stages of development. It has been designed as a generic tool to perform automated searches of any online resource for any set of search terms. It therefore should be able to perform BRT monitoring, as well as other intelligence gathering functions such as news aggregation. It is the only system identified with a ready ability to search the internet at-large for potential sites of interest.

GLDIATR GLDIATR has a preliminary score of 14, only marginally less than the IBIS Replacement. This system is purpose-built for detecting sales of biosecurity risk organisms

on e-commerce platforms. It is already operational, and funding has been secured for further development. Its scope is currently limited to detecting sales of invasive aquatic species of concern to the Laurentian Great Lakes region, but there may be potential to adapt its use to other goods and jurisdictions. This is the only system known to include a machine learning component designed to incrementally improve search performance, although that goal has not yet been realised.

iTrade and ETH Zürich These two systems each have a preliminary score of 9. The lower score is due to their still being in relatively early stages of development, and the absence of any known relevance scoring component. The developers of these systems have advised that further development is required before being suitable for operational deployment. This may represent an opportunity for collaboration with BRT and/or IWT regulators.

6.3.3 Alternative: in-house development

Developing new software in-house may save on procurement costs, particularly if making use of open source technology, but the additional development and maintenance costs would need to be considered carefully.

There are many commercial and open-source tools available that perform at least one of the functions listed in the software requirements, some of which are listed below. The relative performance of these tools has not been assessed.

- Google Alerts (https://www.google.com.au/alerts) is a free service that allows users to schedule potentially complex Google queries for regular execution and to receive the results by email or RSS feed.
- Tools that can retrieve, parse and analyse text data from the internet include the open source Python (https://www.python.org/) and R (https://www.r-project.org/), and the commercial systems SAS (http://www.sas.com/) and SPSS (http://www.ibm.com/software/analytics/spss/).
- Hyphe (http://hyphe.medialab.sciences-po.fr/) is free software for web crawling and corpus curation.
- JSoup (http://jsoup.org/) is an open source tool for web-scraping.
- Sheffield University's GATE project (https://gate.ac.uk/) offers open-source text analysis tools.
- Moses (http://www.statmt.org/moses/) provides free machine translation.

EpiSPIDER is one example of a system that was successfully built using only open-source components. Another is Scrapy (http://scrapy.org/), a Python application which extracts data from websites either with or without APIs. There are a number of companies using Scrapy, listed at http://scrapy.org/companies/. One of these companies is Parse.ly (http://www.parsely.com/); they provide a presentation about the technologies they used, the challenges they faced, and some demonstrations at https://speakerdeck.com/amontalenti/web-crawling-and-metadata-extraction-in-python.

The classification algorithm for sales of ivory items mentioned previously used the Orange data mining toolbox in Python (http://jmlr.org/papers/v14/demsar13a.html) [48].

6.3.4 Alternative: outsource the function

Engaging a third party has reportedly worked well for USDA–APHIS. The service is responsive and flexible, and means that the agency doesn't need to concern itself with hiring IT professionals, working through technical details and administering the system.

Appendices

Appendix A

Abbreviations

Abbreviation	In full
CDC	US Center for Disease Control
CEBRA	Centre of Excellence for Biosecurity Risk Analysis
CFIA	Canada Food Inspection Agency
CIA	US Central Intelligence Agency
CITES	Convention on International Trade in Endangered Species
DAWR	Australian Department of Agriculture and Water Resources
DEFRA	UK Department for Environment, Food and Rural Affairs
DHS	Australian Department of Human Services
EPA	US Environmental Protection Agency
EPPO	European Plant Protection Organisation
ETH Zürich	Swiss Federal Institute of Technology in Zurich (German: *Eidgenössische Technische Hochschule Zürich*)
FAO	Food and Agriculture Organization
GLC	Great Lakes Commission
IFAW	International Fund for Animal Welfare
IPPC	International Plant Protection Convention
ISC	Invasive Species Council
NSF CIPM	National Science Foundation Center for Integrated Pest Management
NZ MPI	New Zealand Ministry for Primary Industries
OIE	World Organisation for Animal Health (French: *Office International des Epizooties*)
SITC	Smuggling Interdiction and Trade Compliance
USDA–APHIS	US Department of Agriculture – Animal & Plant Health Inspection Service
USFWS	US Fish and Wildlife Service
WHO	World Health Organization

Table A.1: Abbreviations of organisations

Abbreviation	In full
EPICA	Exotic Pest Information Collection and Analysis
EpiSPIDER	Semantic Processing and Integration of Distributed Electronic Resources for Epidemiology
GLDIATR	Great Lakes Detector of Invasive Aquatics in Trade
GPHIN	Global Public Health Intelligence Network
IBIS	International Biosecurity Intelligence System
IRIS	Intelligence Resources Information System
ISIMS	Invasive Species Internet Monitoring System
OPIS	Offshore Pest Information System
ProMED	Program for Monitoring Emerging Diseases
UMLS	Unified Medical Language System

Table A.2: Abbreviations of systems

Abbreviation	In full
API	Application Programming Interface
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
NLP	Natural Language Processing
RSS	Really Simple Syndication
SQL	Structured Query Language
URL	Uniform Resource Locator
XML	Extensible Markup Language

Table A.3: Abbreviations of technical terms

Abbreviation	In full
APH	animal and plant health
BRT	biosecurity-regulated trade
CNM	Customs news monitoring
IWT	illegal wildlife trade
PHS	public health surveillance

Table A.4: Abbreviations of convenience for this document

Appendix B

Bibliography

[1] L. J. Walters, K. R. Brown, W. T. Stam, and J. L. Olsen, "E-commerce and Caulerpa: unregulated dispersal of invasive species," *Frontiers in Ecology and the Environment*, vol. 4, no. 2, pp. 75–79, 2006.

[2] R. P. Keller and D. M. Lodge, "Species invasions from commerce in live aquatic organisms: Problems and possible solutions," *BioScience*, vol. 57, no. 5, pp. 428–436, 2007.

[3] N. Giltrap, D. Eyre, and P. Reed, "Internet sales of plants for planting – an increasing trend and threat?" *EPPO Bulletin*, vol. 39, no. 2, pp. 168–170, 2009.

[4] D. Parrott and S. Roy, "A preliminary assessment of non-native species pathway-the UK internet pet trade," *Central Science Laboratory Final Report*, 2009.

[5] J. G. Derraik and S. Phillips, "Online trade poses a threat to biosecurity in New Zealand," *Biological invasions*, vol. 12, no. 6, pp. 1477–1480, 2010.

[6] A. L. Barroso de Magalhães and C. M. Jacobi, "E-commerce of freshwater aquarium fishes: potential disseminator of exotic species in Brazil," *Acta Scientiarum. Biological Sciences*, vol. 32, no. 3, pp. 243–248, 2010.

[7] D. Morrisey, G. Inglis, K. Neil, A. Bradley, and I. Fitridge, "Characterization of the marine aquarium trade and management of associated marine pests in Australia, a country with stringent import biosecurity regulation," *Environmental Conservation*, vol. 38, no. 01, pp. 89–100, 2011.

[8] G. D. Martin and J. A. Coetzee, "Pet stores, aquarists and the internet trade as modes of introduction and spread of invasive macrophytes in South Africa," *Water SA*, vol. 37, no. 3, 2011.

[9] National Invasive Species Council, *Invasive Species and E-commerce*, U.S. Department of the Interior, 2012, https://www.doi.gov/invasivespecies/isac/isac_whitepapers. Retrieved 1 Dec 2015.

[10] K. Kaminski, F. Beckers, and J.-G. Unger, "Global internet trade of plants – legality and risks," *EPPO Bulletin*, vol. 42, no. 2, pp. 171–175, 2012.

[11] A. L. Barroso de Magalhães and V. A. São-Pedro, "Illegal trade on non-native amphibians and reptiles in southeast Brazil: the status of e-commerce," *Phyllomedusa: Journal of Herpetology*, vol. 11, no. 2, pp. 155–160, 2012.

[12] H. G. Pearson and R. L. Hill, "Risks for New Zealand's biosecurity from internet trading of plants," in *19th Australasian Weeds Conference*. Tasmanian Weed Society, 2014, pp. 182–185, http://www.caws.org.au/awc/2014/awc201411821.pdf. Retrieved 23 Nov 2015.

[13] F. Humair, L. Humair, F. Kuhn, and C. Kueffer, "E-commerce trade in invasive plants," *Conservation Biology*, vol. 29, no. 6, pp. 1658–1665, 2015.

[14] G. Mazza, L. Aquiloni, A. F. Inghilesi, C. Giuliani, L. Lazzaro, G. Ferretti, L. Lastrucci, B. Foggi, and E. Tricarico, "Aliens just a click away: the online aquarium trade in Italy," *Management of Biological Invasions*, vol. 6, no. 3, pp. 253–261, 2015.

[15] International Plant Protection Convention, "Internet Trade (eCommerce) in Plants and other Regulated Articles," https://www.ippc.int/en/core-activities/governance/cpm-recommendations/ippc-aquatic-plants-trade-commerce/, 2014, retrieved 30 Nov 2015.

[16] International Fund for Animal Welfare, "Click to delete – Endangered wildlife for sale in New Zealand," http://www.ifaw.org/australia/resource-centre/click-delete-endangered-wildlife-sale-new-zealand, 2014, retrieved 25 Nov 2015.

[17] Y. Gao and S. G. Clark, "Elephant ivory trade in China: Trends and drivers," *Biological Conservation*, vol. 180, pp. 23–30, 2014.

[18] TRAFFIC, "Wildlife Trade in the European Union," http://www.traffic.org/general-reports/traffic_pub_gen56.pdf, 2014, retrieved 15 Dec 2015.

[19] J. Guan and L. Xu, "Deadly Messaging – Illegal ivory trade in China's social media," http://www.trafficchina.org/node/252, 2015, retrieved 25 Jan 2015.

[20] Y. Xiao and J. Wang, "Moving targets: Tracking online sales of illegal wildlife products in China," http://www.traffic.org/home/2015/3/3/traffics-e-commerce-monitoring-reveals-shifting-illegal-wild.html, 2015, retrieved 13 Dec 2015.

[21] K. Soska and N. Christin, "Measuring the longitudinal evolution of the online anonymous marketplace ecosystem," in *24th USENIX Security Symposium (USENIX Security 15)*. Washington, D.C.: USENIX Association, 2015, pp. 33–48. [Online]. Available: https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/soska

[22] J. R. Harrison, D. L. Roberts, and J. C. Hernandez-Castro, "Assessing the extent and nature of wildlife trade on the darkweb," *Conservation Biology*, 2016. [Online]. Available: http://dx.doi.org/10.1111/cobi.12707

[23] K. N. Gajewski, A. E. Peterson, R. A. Chitale, J. A. Pavlin, K. L. Russell, and J.-P. Chretien, "A review of evaluations of electronic event-based biosurveillance systems." *PLoS ONE*, vol. 9, no. 10, pp. 1–4, 2014.

[24] Center for Disease Control, "Updated guidelines for evaluating public health surveillance systems: recommendations from the guidelines working group," *Morbidity & Mortality Weekly Report*, vol. 50, no. RR-13, pp. 1–35, 2001.

[25] Centers for Disease Control and Prevention, "Framework for evaluating public health surveillance systems for early detection of outbreaks: recommendations from the CDC Working Group." *Morbidity & Mortality Weekly Report*, vol. 53, no. RR-5, 2004, http://www.cdc.gov/mmwr/pdf/rr/rr5305.pdf. Retrieved 28 Feb 2016.

[26] Interim Inspector-General of Biosecurity, *Managing biosecurity risks associated with international online purchases*. Commonwealth of Australia, 2015, no. 2013–14/2, http://www.igb.gov.au/Pages/completed-audits-and-reviews.aspx. Retrieved 23 Nov 2015.

[27] USDA-APHIS, "Smuggling Interdiction and Trade Compliance (SITC): Overview of SITC and Internet Activities," 2015, via pers. comm.

[28] S. C. Chng and J. Bouhuys, "Indian star tortoises," *TRAFFIC Bulletin*, vol. 27, no. 2, p. 73, 2015.

[29] P. D. Shirey, B. N. Kunycky, D. T. Chaloner, M. A. Brueseke, and G. A. Lamberti, "Commercial trade of federally listed threatened and endangered plants in the United States," *Conservation Letters*, vol. 6, no. 5, pp. 300–316, 2013.

[30] M. Sajeva, C. Augugliaro, M. J. Smith, and E. Oddo, "Regulating internet trade in CITES species," *Conservation Biology: The Journal Of The Society For Conservation Biology*, vol. 27, no. 2, pp. 429–430, 2013.

[31] A. Hinsley, T. E. Lee, J. R. Harrison, and D. L. Roberts, "Estimating the extent and structure of trade in horticultural orchids via social media," *Conservation Biology*, vol. 30, no. 5, pp. 1038–1047, 2016.

[32] N. Collier, S. Doan, A. Kawazoe, R. M. Goodwin, M. Conway, Y. Tateno, Q.-H. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, M. Shigematsu, and K. Taniguchi, "BioCaster: detecting public health rumors with a Web-based text mining system," *Bioinformatics*, vol. 24, no. 24, pp. 2940–2941, 2008.

[33] N. Collier, R. M. Goodwin, J. McCrae, S. Doan, A. Kawazoe, M. Conway, A. Kawtrakul, K. Takeuchi, and D. Dien, "An ontology-driven system for detecting global health events," in *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 215–222.

[34] D. Hartley, N. Nelson, R. Walters, R. Arthur, R. Yangarber, L. Madoff, J. Linge, A. Mawudeku, N. Collier, J. Brownstein, G. Thinus, and N. Lightfoot, "Landscape of international event-based biosurveillance," *Emerging Health Threats Journal*, vol. 3, no. e3, 2010.

[35] N. Collier, "Uncovering text mining: A survey of current work on web-based epidemic intelligence," *Global Public Health*, vol. 7, no. 7-8, pp. 731–749, 2012.

[36] N. Collier and S. Doan, "GENI-DB: a database of global events for epidemic intelligence," *Bioinformatics*, vol. 28, no. 8, pp. 1186–1188, 2012.

[37] A. Lyon, M. Nunn, G. Grossel, and M. Burgman, "Comparison of web-based biosecurity intelligence systems: BioCaster, EpiSPIDER and HealthMap." *Transboundary And Emerging Diseases*, vol. 59, no. 3, pp. 223–232, 2012.

[38] A. Lyon, G. Grossel, M. Burgman, and M. Nunn, "Using internet intelligence to manage biosecurity risks: a case study for aquatic animal health," *Diversity and Distributions*, vol. 19, no. 5-6, pp. 640–650, 2013.

[39] H. Tolentino, R. Kamadjeu, P. Fontelo, F. Liu, M. Matters, M. Pollack, and L. Madoff, "Scanning the emerging infectious diseases Horizon-visualizing ProMED emails using EpiSPIDER," *Advances in Disease Surveillance*, vol. 2, p. 169, 2007.

[40] M. Keller, M. Blench, H. Tolentino, C. C. Freifeld, K. D. Mandl, A. Mawudeku, G. Eysenbach, and J. S. Brownstein, "Use of unstructured event-based reports for global infectious disease surveillance," *Emerging Infectious Diseases*, vol. 15, no. 5, pp. 689–695, 2009.

[41] A. Mawudeku and M. Blench, "Global public health intelligence network (GPHIN)," in *7th Conference of the Association for Machine Translation in the Americas*, 2006, pp. 8–12.

[42] A. Mawudeku, M. Blench, L. Boily, R. St. John, R. Andraghetti, and M. Ruben, "The Global Public Health Intelligence Network," in *Infectious Disease Surveillance*, N. M. M'ikanatha, R. Lynfield, C. A. Van Beneden, and H. de Valk, Eds. John Wiley & Sons Ltd, 2013, pp. 457–469.

[43] M. Dion, P. AbdelMalik, and A. Mawudeku, "Big data and the Global Public Health Intelligence Network (GPHIN)," *Canada Communicable Disease Report*, vol. 41, no. 9, p. 209, 2015.

[44] J. Brownstein and C. Freifeld, "HealthMap: the development of automated real-time internet surveillance for epidemic intelligence," *Eurosurveillance*, vol. 12, no. 48, p. 3322, 2007.

[45] J. S. Brownstein, C. C. Freifeld, B. Y. Reis, and K. D. Mandl, "Surveillance Sans Frontières: Internet-based emerging infectious disease intelligence and the HealthMap project," *PLoS Med*, vol. 5, no. 7, p. e151, 2008.

[46] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein, "HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports," *Journal of the American Medical Informatics Association*, vol. 15, no. 2, pp. 150–157, 2008.

[47] A. L. Sonricker Hansen, A. Li, D. Joly, S. Mekaru, and J. S. Brownstein, "Digital surveillance: A novel approach to monitoring the illegal wildlife trade," *PLoS ONE*, vol. 7, p. e51156, 12 2012.

[48] J. C. Hernandez-Castro and D. L. Roberts, "Automatic detection of potentially illegal online sales of elephant ivory via data mining," *PeerJ Computer Science*, vol. 1, p. e10, 2015.

[49] H. Meissner, J. Fritz, L. Kohl, H. Moylett, M. Moan, S. Emerine, and A. Kaye, "Pestlens: an early-warning system supporting u.s. safeguarding against exotic plant pests," *EPPO Bulletin*, vol. 45, no. 2, pp. 304–310, 2015.

[50] R. J. Balaam, "USDA APHIS Offshore Pest Information System," in *Proceedings, XV U.S. Department of Agriculture interagency research forum on gypsy moth and other invasive species 2004*, K. W. Gottschalk, Ed., vol. Gen. Tech. Rep. NE-332,

U.S. Department of Agriculture, Forest Service, Northeastern Research Station, 2005, pp. 1–2.

[51] J. M. Van Dersal, "Managing plant diseases offshore," *Australasian Plant Pathology*, vol. 36, no. 6, pp. 520–523, 2007.

[52] M. Bateman, C. Brammer, C. Thayer, H. Meissner, and W. Bailey, "Exotic Pest Information Collection and Analysis (EPICA) – gathering information on exotic pests from the World Wide Web," *EPPO Bulletin*, no. 2, 2007.

[53] J. M. Fritz, M. J. Moan, S. E. Emerine, and H. E. Meissner, "Exotic Pest Information Collection and Analysis (EPICA): plant health biosurveillance for the United States," in *Proceedings of the Caribbean Food Crops Society*, vol. 46, 2010, pp. 182–188.

[54] G. Grossel, A. Lyon, and M. Nunn, "Open-Intelligence Gathering and Analysis for Biosecurity," in *Risk-Based Decisions for Biological Threats*, T. Walshe, Ed. Cambridge University Press, forthcoming, draft from http://aidanlyon.com/media/publications/GG-draft.pdf Retrieved 14 Feb 2016.

[55] S. Moss, "International Biosecurity Intelligence System: System Overview," 2014, via pers. comm.

[56] Great Lakes Commission, "Protecting the Great Lakes from internet trade of aquatic invasive species," http://glc.org/files/projects/ais/GLDIATR-Report-Web-Aug-2016.pdf and http://glc.org/files/projects/ais/GLDIATR-Report-Appendices-Web-Aug-2016.pdf, 2016, retrieved 28 Nov 2016.

[57] M. J. Rotstein, S. G. Lloyd, A. Fowler, and R. E. Stinner, "Internet surveillance–new web search and response application is a valuable biosecurity tool," in *Thirteenth Australian Weeds Conference Proceedings*, 2002, pp. 8–13.

[58] K. Suiter and S. Sferrazza, "Monitoring the sale and trafficking of invasive vertebrate species using automated internet search and surveillance tools," in *Managing vertebrate invasive species: proceedings of an international symposium, Fort Collins, CO*, G. W. Witmer, W. C. Pitt, and K. A. Fagerstone, Eds., 2007, pp. 90–93.

[59] D. M. Hartley, N. P. Nelson, R. R. Arthur, P. Barboza, N. Collier, N. Lightfoot, J. P. Linge, E. Goot, A. Mawudeku, L. C. Madoff, L. Vaillant, R. Walters, R. Yangarber, J. Mantero, C. D. Corley, and J. S. Brownstein, "An overview of internet biosurveillance," *Clinical Microbiology and Infection*, no. 11, p. 1006, 2013.

[60] J. S. Schwind, D. J. Wolking, J. S. Brownstein, J. A. K. Mazet, and W. A. Smith, "Evaluation of local media surveillance for improved disease recognition and monitoring in global hotspot regions," *PLoS ONE*, vol. 9, no. 10, pp. 1–9, 2014.

[61] P. Barboza, L. Vaillant, Y. Le Strat, D. M. Hartley, N. P. Nelson, A. Mawudeku, L. C. Madoff, J. P. Linge, N. Collier, J. S. Brownstein, and P. Astagneau, "Factors influencing performance of internet-based biosurveillance systems used in epidemic intelligence for early detection of infectious diseases outbreaks," *PLoS ONE*, vol. 9, no. 3, pp. 1–9, 2014.

[62] A. Lyon, A. Mooney, and G. Grossel, "Using AquaticHealth.net to detect emerging trends in aquatic animal health," *Agriculture*, vol. 3, no. 2, p. 299, 2013.