Transboundary and Emerging Diseases

Transboundary and Emerging Diseases

ORIGINAL ARTICLE

Comparison of Web-Based Biosecurity Intelligence Systems: BioCaster, EpiSPIDER and HealthMap

A. Lyon[1,2], M. Nunn[3], G. Grossel[3] and M. Burgman[1]

[1] Australian Centre of Excellence for Risk Analysis, University of Melbourne, Melbourne, Vic., Australia
[2] University of Maryland, College Park, MD, USA
[3] Department of Agriculture, Fisheries and Forestry, Canberra, ACT, Australia

Keywords:
BioCaster; EpiSPIDER; HealthMap; automated biosecurity intelligence; open-source information

Correspondence:
A. Lyon. 1103B Skinner Building, University of Maryland, College Park, MD 20742, USA.
Tel.: +1 301-405-5704;
Fax: +1 301-405-5690;
E-mail: alyon@umd.edu

Received for publication May 6, 2011

doi:10.1111/j.1865-1682.2011.01258.x

Summary

Three web-based biosecurity intelligence systems – BioCaster, EpiSPIDER and HealthMap – are compared with respect to their ability to gather and analyse information relevant to public health. Reports from each system for the period 2–30 August 2010 were studied. The systems were compared to the volume of information that they acquired, their overlaps in this information, their timeliness, their sources, their focus on different languages and their focus on different geographical regions. Main results were as follows: EpiSPIDER obtains the most information and does so mainly through Twitter; no significant difference in systems' timeliness was found; there is a relatively small overlap between the systems (10–20%); the systems have significant differences in their ability to acquire information relevant to different countries, which may be due to the sources they use and the languages they focus on.

Introduction

Open-source and unstructured information concerning disease outbreaks exists in many forms and languages on the Internet. There are a number of automated biosecurity[1] intelligence systems that are trying to gather and analyse this information. Three such systems – BioCaster, EpiSPIDER, and HealthMap – are compared with respect to their ability to do this.

BioCaster (Collier et al., 2006, 2008) collects information from EurekAlert!, European Media Monitor Alerts (EMMA), Google, the CDC's Morbidity and Mortality Weekly Report (MMWR), MeltWater, OIE, ProMED, Reuters, WHO and Vetsweb. It scans for articles in Arabic, Chinese, English, French, Japanese, Korean, Portuguese, Russian, Spanish, Thai and Vietnamese. The system gives a special priority to languages of the Asia-Pacific region and has an open-source, multilingual ontology that relates disease terms, causal agents, symptoms, etc. BioCaster has a mapping feature that allows users to view and filter its reports.

EpiSPIDER (Tolentino et al., 2007; Keller et al., 2009) collects information from Daylife, Google, Humanitarian News, Moreover, ProMED, Twitter and WHO. It scans for articles in English only. The system has a mapping feature that allows users to view and filter reports. It also has a timeline visualization to help users to order events in time, and a word-cloud that helps users to get a sense of what topics are making headlines.

HealthMap (Brownstein et al., 2009, 2010; Keller et al., 2009; Wilson and Brownstein, 2009) collects information Baidu, EuroSurveillance, Google, HealthMap Community

[1] The term 'biosurveillance' is sometimes used to refer to what we mean by 'biosecurity'.

[2] http://www.alchemyapi.com/api/text/.

[3] The term 'country' is used more broadly here than usual. See http://www.geonames.org/countries/ for the list of 'countries' used in this study.

News Reports, OIE, ProMED, SOSO, User Eyewitness Reports, WDIN and WHO. It scans for articles in Arabic, Chinese, English, French, Portuguese, Russian and Spanish. It also has a mapping system that allows users to view reports and apply a number of filters. Users can also comment on articles and rank them for significance.

The purpose of this study is to evaluate comparatively the performance of these three tools and to assess their utility in horizon scanning for biosecurity. A somewhat similar study of HealthMap, EpiSPIDER and another system, GPHIN, has been made (Keller et al., 2009). However, the comparisons made so far have been largely qualitative and have not included BioCaster. To our knowledge, there has not yet been a quantitative assessment of the degree of overlap of these systems, their timeliness, their relative focus on different geographical regions and languages and their reliance on different sources. Understanding the differences and similarities between these systems is important for their improvement, for developing other systems in other areas of biosecurity (e.g. aquatic animal health) and also to public health officials who use the systems and face information overload on a daily basis.

Comparisons of BioCaster, EpiSPIDER and HealthMap

Methods

To compare the systems, it is necessary to count and classify the articles that each system finds. Each system generates at least one report for each article that it collects. Each report has a link to an article, a publish date, the source by which the article was found, location coordinates and some other information, depending on the system. Users can view the reports in various ways – e.g. on a map that can be filtered in various ways. To determine the number of articles a system has found over some period of time, it is not enough to simply count the number of reports published by that system (*Total Original*). This is because often an article will be linked to by a number of different reports, which may have different locations, publish dates, etc.

A better way to determine the number of articles a system has found is to count the number of *unique* links in the systems' reports (*Unique Original*). However, two different links can lead to the same article. So, it is necessary to follow all links to their final URLs and then count how many unique final URLs there are (*Unique Real*). However, a single article can be associated with multiple URLs. For example, http://www.promedmail.org/pls/otn/f?p=2400:1001:4293294425104239::NO::F2400_P1001_BACK_PAGE,F2400_P1001_PUB_MAIL_ID:1050,84399 and http://promedmail.org/pls/otn/f?p=2400:1001:462275334141704::NO::F2400_P1001_BACK_PAGE,F2400_P1001_PUB_MAIL_ID:1055,84399 both link to the same ProMED article. So, it is necessary to compare the pages of each link to see whether they are the same. Pages, however, often vary slightly and in irrelevant ways (as with the previous two links). So to compare the pages of two links to see whether they are links to the same article, the pages' contents must be scraped clean of any irrelevant surrounding material such as headers, sidebars and advertisements [*Unique Content (exact)*]. In this study, this was done using Alchemy's Text Extraction API.[2]

The scraped contents of pages can still vary in irrelevant ways – e.g. when two news media sites share an article, but apply different editing and/or formatting standards. This makes it necessary to compare the scraped contents of pages in a way that allows for some variation. In this study, this pairwise comparison was made using the Python v2.7 SequenceMatcher Class (Python Software Foundation, http://docs.python.org/license.html). If two scraped contents had a high similarity ratio – defined as $2M/T > 0.9$, where M is the number of matches and T the total number of elements – then they were judged to be the same article.

To estimate the number of articles a system has found, the number of scraped contents whose pairwise similarity ratios were below 0.9 was determined [*Unique Content (similarity)*]. This involved following each original link through any redirections and scraping its content (*Unique Real and Scrapable*). These processes were not always possible. Sometimes the original links were broken, and sometimes their contents could not be scraped. Let b be the fraction of links that were broken, m the fraction of pages that were not scrapable and S the number of unique contents (by similarity). The expected number of unique articles is: *Expected Unique* $= S/(1-b)(1-m)$.

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

Unique articles

Table 1 shows the counts at various stages of the estimation of the number of articles for each system. BioCaster had a larger overall reduction (from number of original links to expected unique articles) than HealthMap (49% and 18%, respectively), but it still reported more unique articles than HealthMap – about 28% more. EpiSPIDER had the largest overall reduction with 89% of its links removed as repetitions. However, EpiSPIDER still had more unique articles than BioCaster and HealthMap –

Table 1. Numbers of articles

	BioCaster	EpiSPIDER	HealthMap
Total Original	6860	58 046	3856
Unique Original	4682	8235	3302
Unique Original and Live	4382	7250	3174
Unique Real	4352	7148	3137
Unique Real and Scrapable	4192	6784	3073
Unique Content (exact)	3843	5818	2997
Unique Content (similarity)	3620	5367	2960
Expected Unique	4015	6193	3144
Percentage Reduction = (1−Expected Unique/Total Original)*100%	42%	89%	18%

about 54% more than BioCaster and 96% more than HealthMap.

Overlaps and first reports

To determine the overlaps between the systems, the unique contents (by similarity) of each system were compared. First, each unique content (by similarity) for a given system was combined with all the contents that were similar to it, along with all the final links that linked to it, and the dates of the corresponding original links. Call each combination a story. (There were 3620 stories for BioCaster, 5367 for EpiSPIDER and 2960 for HealthMap.) When two systems had two stories with at least one matching final link, the two stories were defined to be a match. When two systems had two stories that had contents judged to be the same (by similarity), the two stories were defined to be a match. When two stories matched, the earliest dates of the stories were compared and the difference between them was recorded.

About 10% of BioCaster's stories matched a HealthMap story and about 13% *vice versa*. On average, BioCaster's publish date for a story that both it and HealthMap found was later than HealthMap's publish date by 0.4 of a day. However, the two systems operate in different time zones (BioCaster in Japan, HealthMap in the US), so this difference is not important and probably an artefact of the difference in time zones.

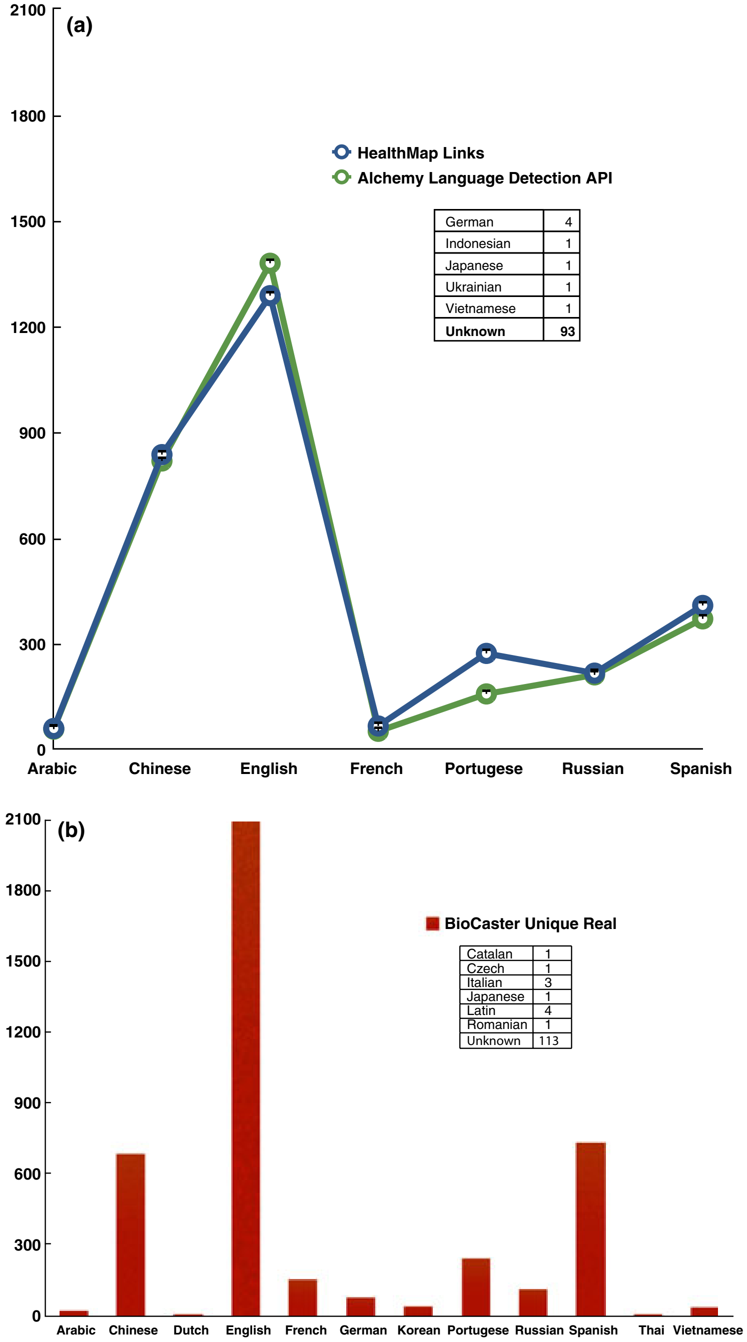
About 10% of EpiSPIDER's stories matched a HealthMap story and about 18% *vice versa*. On average, EpiSPIDER's publish date for a story that both it and HealthMap found was earlier than HealthMap's publish date by 0.2 days. Both systems are based in the United States, so this might mean that EpiSPIDER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

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

Languages

EpiSPIDER finds articles solely written in English, so it was only necessary to compare BioCaster and HealthMap with respect to the systems' focus on different languages. To determine the language of each article reported by the systems, Alchemy's Language Detection API was used. For HealthMap, it was also possible to determine the language of each article directly. The system has a translation schema that it uses in the links of reports to automatically translate the articles it has detected. For example, if 'trto=en&trf=zh' appears in one of HealthMap's links, then when a user clicks on the link, HealthMap gets Google Translate to translate the article that the link points to from Chinese (zh) to English (en). So, it was assumed that whenever 'trto=en&trf=zh' appeared in a link, the original article was in Chinese – and similarly for other languages. Using this translation schema, it was then possible to determine another distribution of HealthMap's reports over languages – Fig. 1a. This second distribution was strongly correlated with the distribution determined using Alchemy's API (Fig. 1a). This strong correlation suggests that the distribution for BioCaster (which was only determined using Alchemy's API) is accurate (Fig. 1b). A notable exception in the correlation between the two distributions for HealthMap's languages was the language Portuguese. Alchemy found 115 fewer pages in Portuguese than those determined using the translation schema in HealthMap's links, and it was unable to obtain the language of 93 of HealthMap's links. It was also unable to obtain the language of 113 of BioCaster's links. It is therefore likely that the actual number of BioCaster's links in Portuguese is significantly higher (by about 100). Incidentally, the Alchemy API also detected a handful of pages in languages not reported by HealthMap (German, Indonesian, Japanese, Ukrainian and Vietnamese) and some for BioCaster (Dutch and German).

Figures 2, 3 and 4 show the numbers of articles in each language for the two systems. English and Chinese are the two most common languages for HealthMap, with 46% and 26% of the expected unique articles, respectively. BioCaster has a similar focus on Chinese and English (18% and 45%, respectively) except that it finds more articles in Spanish than HealthMap (both in absolute



German	4
Indonesian	1
Japanese	1
Ukrainian	1
Vietnamese	1
Unknown	**93**

Catalan	1
Czech	1
Italian	3
Japanese	1
Latin	4
Romanian	1
Unknown	113

Fig. 1. Left (a): HealthMap languages determined by HealthMap and Alchemy for all Unique Real reports. Right (b): BioCaster languages determined by Alchemy for all Unique Real reports.

terms (696 versus 367) and relative terms (17% versus 12%). Given that BioCaster has a priority for finding articles in Asia-Pacific languages – Chinese, Japanese, Korean, Thai and Vietnamese – it finds surprisingly few articles in these languages, with the exception of Chinese, although it finds more articles in these languages than HealthMap (again with the exception of Chinese).

Geographic distributions

The country associated with each report was determined by reverse geocoding each report's latitude and longitude with Geonames' Reverse Geocoding API. In some cases, the coordinates of a report corresponded to a location that was not on land, so a 20-km buffer radius was used.

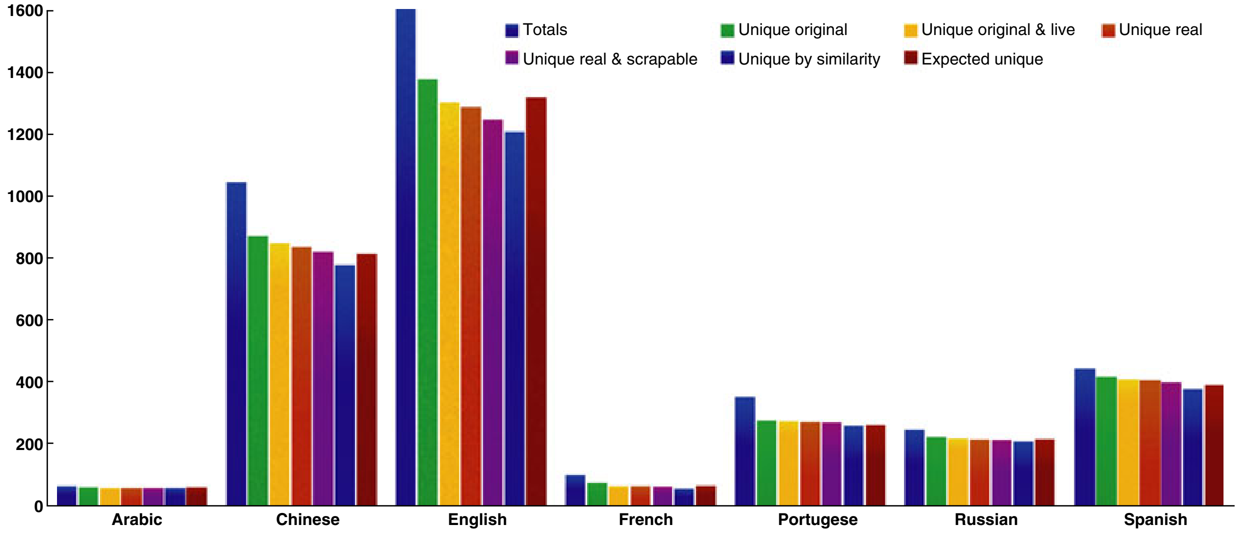


Fig. 2. HealthMap languages determined by URL Translation Schema.

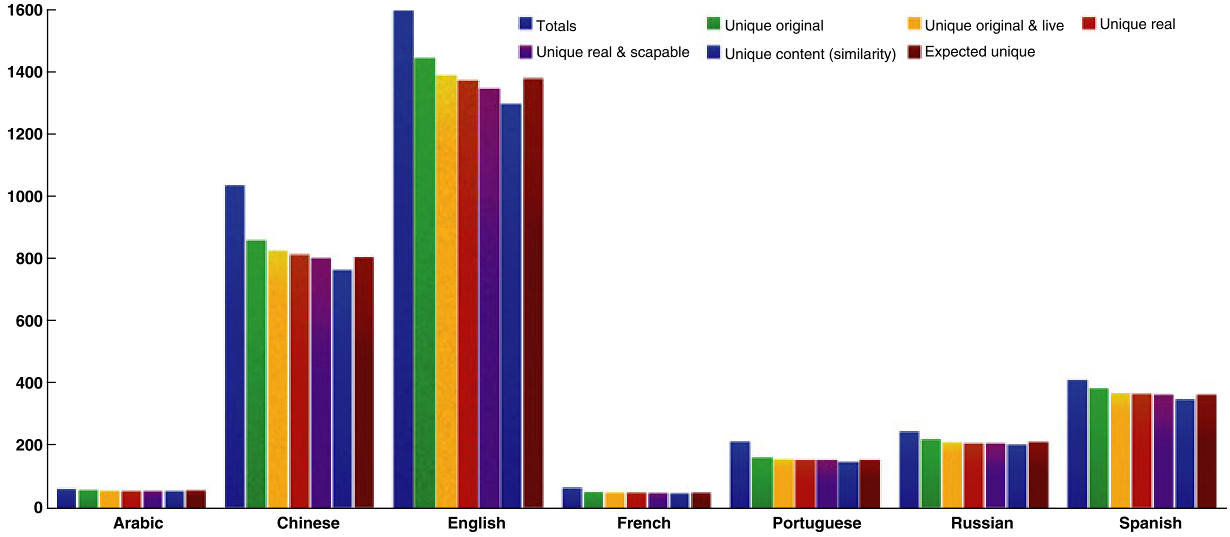


Fig. 3. HealthMap languages determined by Alchemy.

All other reports were treated as errors and ignored (e.g. some reports had 0,0 coordinates). Because a given article can refer to multiple locations, repetitions of links across countries were not removed. All of the systems report an article in multiple locations if they detect that the article makes reference to multiple locations. In what follows, the expected numbers of unique articles, within country categories, are reported.

There were 215 countries[3] that had an article reported by at least one of the three systems. Figures 5–7 contain three choropleth maps representing pairwise comparisons of the three distributions of articles over countries. The colour and intensity of that colour of each country reflect the comparative number of reports by the dominating system for that country. BioCaster's colour is red, HealthMap's is green, EpiSPIDER's is blue and colour intensities come in 10% bands. For example, in Fig. 7, one can see that HealthMap has many more reports than EpiSPIDER for Argentina, and EpiSPIDER has only slightly more for Peru. In general, EpiSPIDER reported the most articles for each country, although there were some notable exceptions: China, France, Brazil, Argentina and Spain. EpiSPIDER sometimes reported a significant number of articles where BioCaster and HealthMap reported none or very few – e.g. Afghanistan, Haiti and Switzerland.

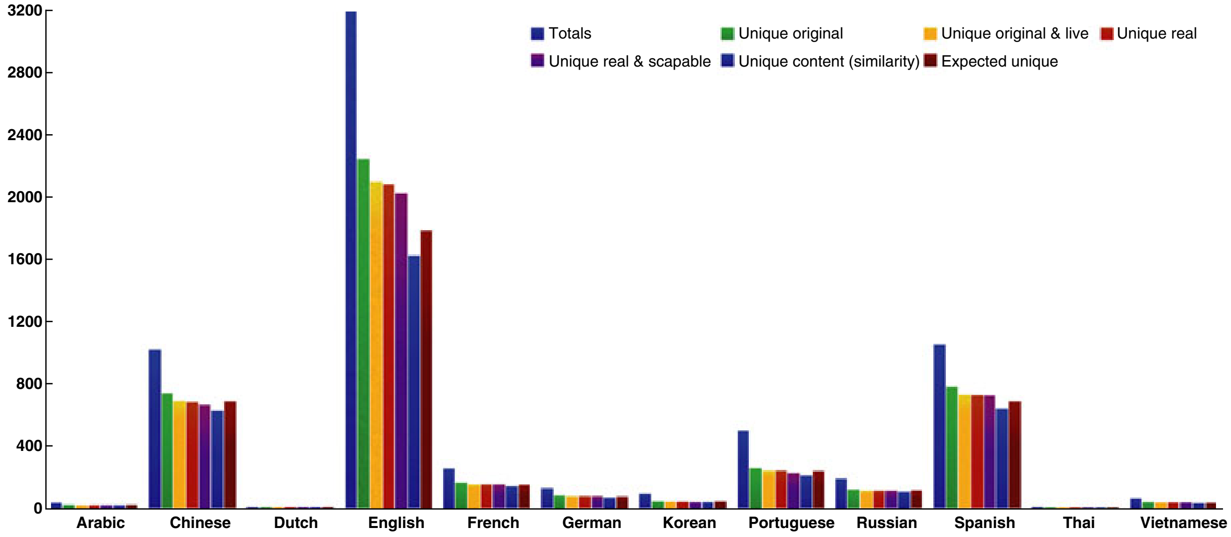


Fig. 4. BioCaster languages determined by Alchemy.

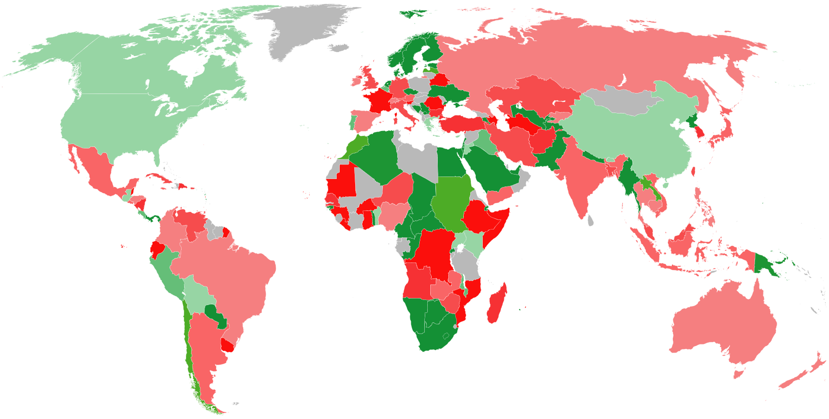


Fig. 5. Comparison of BioCaster and HealthMap. Colour intensities are in 10% bands. BioCaster is red, HealthMap green.

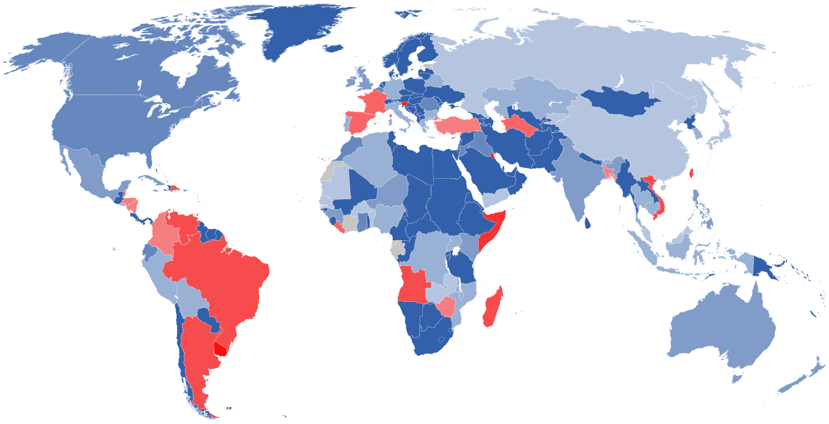


Fig. 6. Comparison of BioCaster and EpiSPIDER. Colour intensities are in 10% bands. BioCaster is red, EpiSPIDER is blue.

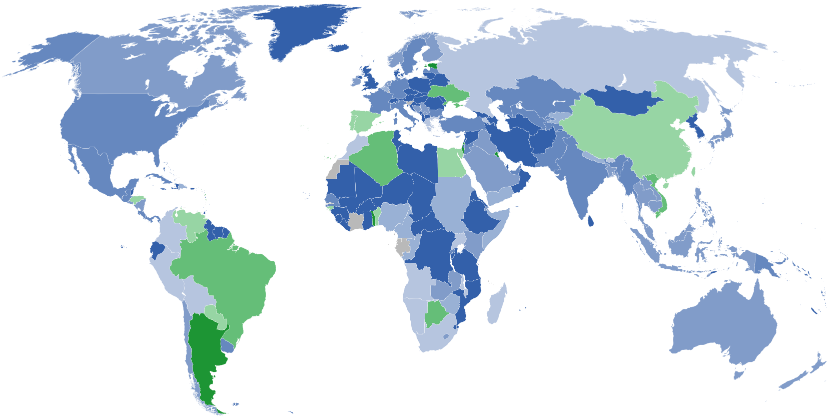


Fig. 7. Comparison of EpiSPIDER and HealthMap. Colour intensities are in 10% bands. EpiSPIDER is blue, HealthMap is green.

There were several interesting differences between BioCaster and HealthMap. HealthMap had more articles in two Asia-Pacific countries: China and Laos. It is not clear what the difference was in Laos, but the difference in China was probably due to the systems' different sources (see next section). HealthMap also had nearly 150 articles for Pakistan; BioCaster had none. However, in the titles of BioCaster's reports, there were 62 instances of 'Pakistan'. This suggests that BioCaster is finding articles about events in (or relating to) Pakistan, but failing to plot them in Pakistan. HealthMap also reported 20–50 articles each for Egypt, Cameroon, Martinique, Nepal, South Africa and Ukraine, whereas BioCaster had at most 3.

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

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

Source distributions

All of the systems acknowledge where they get each of their articles. Using these acknowledgements, it was possible to determine how many articles each system acquired from each source. Figures 8, 9 and 10 show each system's distributions over its sources (with only repetitions within sources were removed).

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

EpiSPIDER is the only system to have a social media platform – viz., Twitter – as a source. Most of the reports from Twitter were plotted in the United States (30%), Pakistan (11%), India (10%), Mexico (4%), UK (3%) and China (3%). However, there were a number of countries for which a large percentage, if not all, of all reports were from Twitter. (This is including BioCaster, HealthMap and the rest of EpiSPIDER's reports.) Interestingly,

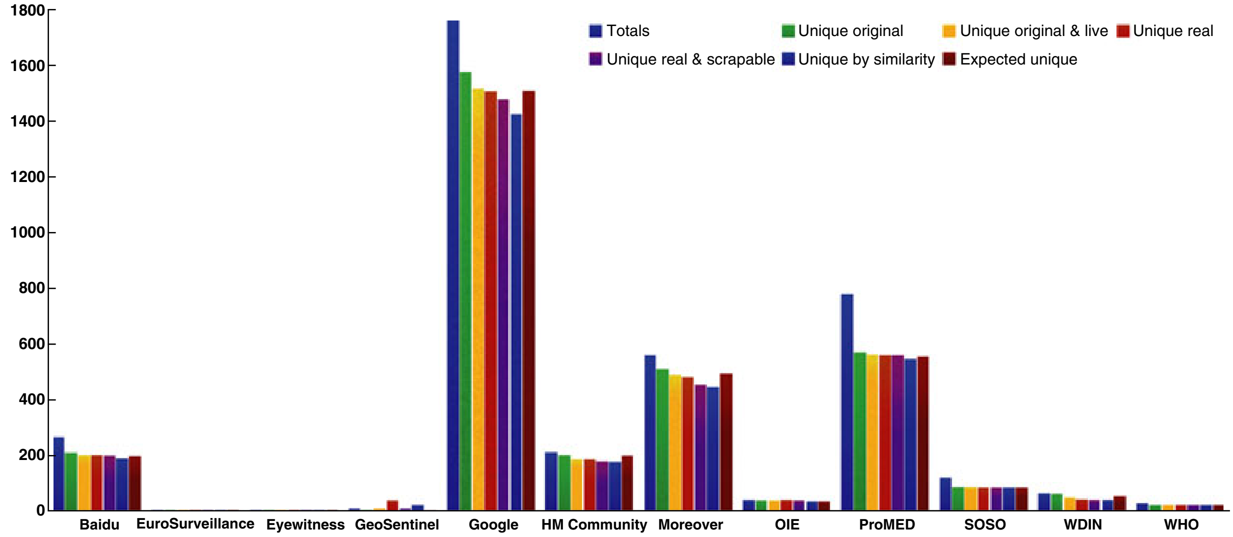


Fig. 8. HealthMap's Sources.

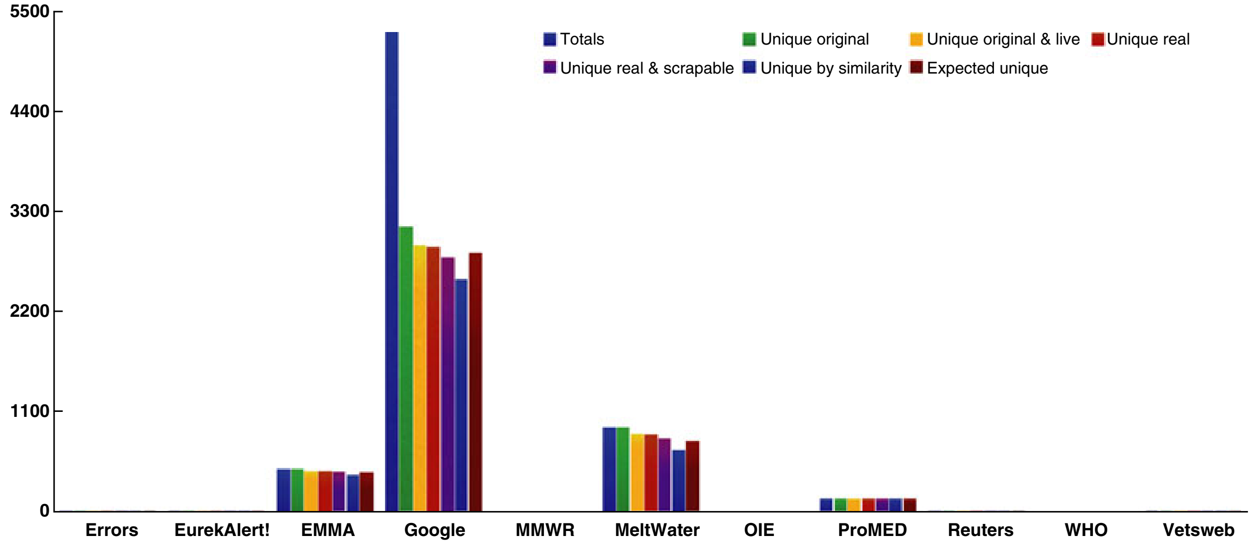


Fig. 9. BioCaster's Sources.

these countries were not the United States, China, India, Pakistan, the UK and Canada. Figure 11 contains a choropleth that shows Twitter's contribution to all of the reports for each country. The countries for which Twitter generated more (unique original) reports than all other reports combined were the following: Faroe Islands, Micronesia, New Caledonia, Seychelles, Suriname, Vanuatu, Northern Mariana Islands, Guam, Latvia, American Samoa, Bermuda, Haiti, Iceland, Trinidad and Tobago, Papua New Guinea, Tonga, Mauritius, Iran, East Timor, Isle of Man, Denmark, Montserrat, Anguilla, Tajikistan, Iraq, Syria, North Korea, Sri Lanka and the Central African Republic. The United States came in at 46%, Pakistan 45%, India 29%, Mexico 32%, UK 35% and China 16%.

Discussion

There are many factors that determine the numbers of articles the systems find. The sources they use appear to play a substantial role. Twitter contains information not found by the systems through any other source, and this can be a significant percentage of the total number of articles for some countries (Fig. 11). Also, although BioCaster has a focus on the Asia-Pacific region, HealthMap

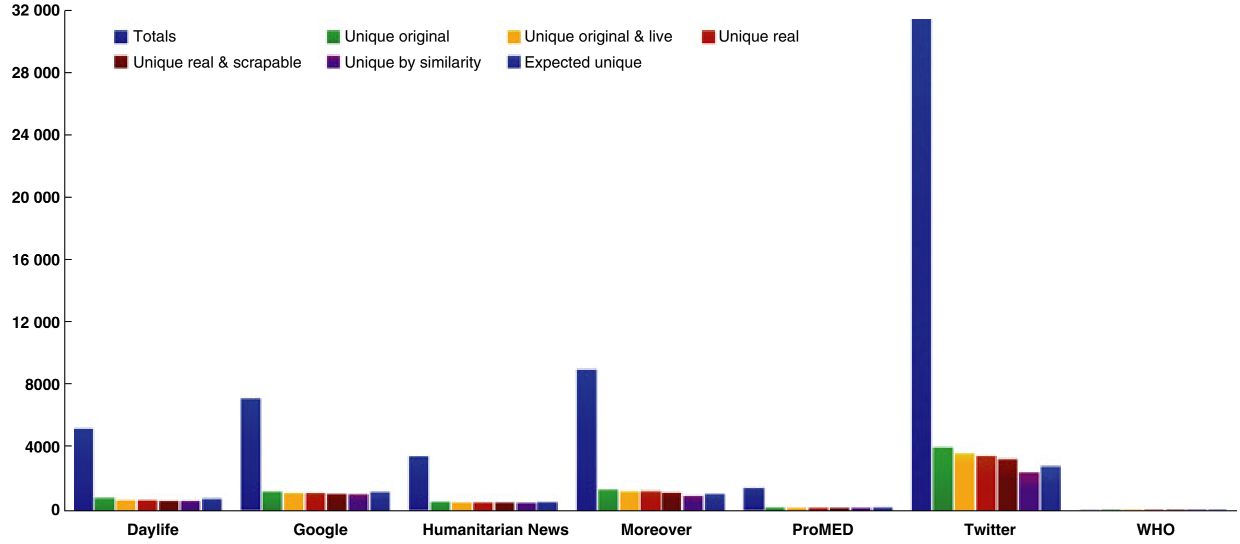


Fig. 10. EpiSPIDER's Sources.

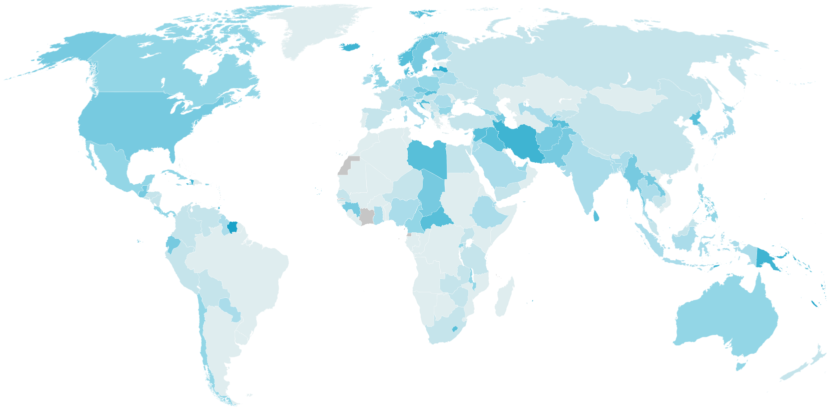


Fig. 11. Twitter's contribution to the reports of each country as a percentage of all reports by BioCaster, HealthMap and EpiSPIDER (without its Twitter reports). Each level of intensity of blue represents a 10% band with the lightest representing 0–10% and the darkest representing 90–100%.

finds more articles in Chinese; this appears to be because HealthMap uses Chinese sources such as Baidu, while BioCaster does not. Another significant factor determining the number of articles that a system finds is the topics it covers. One reason why EpiSPIDER finds more articles than HealthMap and BioCaster appears to be that it has more coverage of natural and man-made disasters than the other two systems. The differences in topics covered may also explain, in large part, why the pairwise overlaps between the systems are only 10–20%.

There does not appear to be an important difference between the timeliness of the systems. If there is any difference, it is that EpiSPIDER is slightly faster than HealthMap. This may be due to the fact that EpiSPIDER uses Twitter as a source, and articles are sometimes tweeted faster than search engines index them. A more detailed study would have to be conducted to know for sure, though.

EpiSPIDER only finds articles in English, while BioCaster and HealthMap find articles in other languages. The most interesting difference between BioCaster and HealthMap with respect to the languages they cover is that HealthMap finds more articles in Chinese than BioCaster. Some articles found by the systems were in languages that HealthMap and BioCaster do not claim to cover. This highlights the fact that automatic language detection is not 100% reliable.

There are also problems with automatic location detection. It was assumed that the number of articles that a system *plotted* for a given country reflected the number of articles it found that were *about* that country. This assumption is problematic. There are clearly errors with automated location detection, and in some cases, these may be quite significant – e.g. BioCaster's lack of reports for Pakistan. No general measure of how accurately the systems detect and plot locations was made.

No measure of the overall quality of the articles found by the systems was made either. This is because overall quality is a vague notion and because this would require reading a large number of the articles of each system and assessing their relevance to biosecurity issues. There are clearly some errors produced by all of the systems, however. Each system at least occasionally reports articles that in no way pertain to biosecurity – but this to be expected of automated textual analysis.

All three systems appear to be improving in a number of ways – by including new sources, refining analysis techniques, adding new ways of visualizing data, etc. One interesting refinement has been the move to make use of social media. HealthMap allows users to submit articles by a number of means as well as eyewitness accounts, and EpiSPIDER scans Twitter. So far, this use of social media has mostly been as a new source of information. However, social media can also be used for analysis purposes – e.g. HealthMap allows users to rank its articles by 'significance' on a five–star system. Using social media to help analyse the articles that the systems find may help reduce the errors that inevitably result from automated textual analysis (Floridi, 2009).

Acknowledgement

This research was supported by the Australian Centre of Excellence for Risk Analysis, Project 1003.

Biography

Aidan Lyon is an Assistant Professor in the Department of Philosophy at the University of Maryland, College Park. He works mainly in Philosophy of Science, Philosophy of Probability and Formal Epistemology. He also works on issues surrounding judgement aggregation/consensus formation and biosecurity intelligence gathering and analysis. This work is done in collaboration with the Australian Department of Agriculture, Fisheries and Forestry (DAFF), and is supported by the Australian Centre of Excellence for Risk Analysis (ACERA).

References

Brownstein, J., C. Freifeld, and L. Madoff, 2009: Digital disease detection – harnessing the web for public health surveillance. *N. Engl. J. Med.* 360, 2153–2157.

Brownstein, J., C. Freifeld, E. Chan, M. Keller, A. Sonricker, S. Mekaru, and D. Buckeridge, 2010: Information technology and global surveillance of cases of 2009 H1N1 influenza. *N. Engl. J. Med.* 362, 1731–1735.

Collier, N., A. Kawazoe, L. Jin, M. Shigematsu, D. Dien, R. A. Barrero, K. Takeuchi, and A. Kawtrakul, 2006: A multilingual ontology for infectious disease surveillance: rationale, design and challenges. *J. Lang. Resour. Eval.* 40, 405–413.

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

Floridi, L., 2009: Web 2.0 vs. the Semantic Web: a philosophical assessment. *Episteme* 6, 25–27.

Keller, M., M. Blench, H. Tolentino, C. C. Freifeld, K. D. Mandl, A. Mawudeku, G. Eysenbach, and J. S. Brownstein, 2009: Use of unstructured event–based reports for global infectious disease surveillance. *Emerg. Infect. Dis.* 15, 689–695.

Tolentino, H., R. Kamadjeu, P. Fontelo, F. Liu, M. Matters, M. Pollack, and L. Madoff, 2007: Scanning the emerging infectious diseases horizon – visualizing ProMED emails using EpiSPIDER. *Adv. Dis. Surveill.* 2, 169.

Wilson, K., and J. Brownstein, 2009: Early detection of disease outbreaks using the internet. *Can. Med. Assoc. J.* 180, 829–831.