



GLOBAL RAPID IDENTIFICATION TOOL SYSTEM

Andrew G. Huff¹,
Zachary S. Gold¹,
Amy M. Slagle¹,
Nathaniel Breit¹,
Russell S. Horton¹,
Jonathan Goley¹,
Karissa A. Whiting¹

¹EcoHealth Alliance
New York, NY, USA
March 2015



EcoHealth Alliance

Partners: International Society for Infectious
Disease, Kitware, ProMED, Epidemico

Project sponsor: Defense Threat Reduction Agency

Executive Summary

Global Rapid Identification Tool Set (GRITS) is a biosurveillance application that enables infectious disease analysts to monitor non-traditional information sources (e.g., online news outlets, ProMED reports, and blogs) for infectious disease threats.

GRITS analyzes these textual data sources by identifying, extracting and succinctly visualizing critical public health information and suggesting possible associated infectious diseases. Via the web-interface, infectious disease analysts can examine dynamic visualizations of GRITS' analyses, perform powerful queries of an index of over 250,000 infectious disease reports, and explore related historical infectious disease emergence events. The GRITS API can be used to continuously analyze information feeds and large collections of data and enables GRITS technology to be easily incorporated into larger surveillance systems. GRITS is a flexible and pluripotent tool that contains robust Natural Language Processing (NLP) and machine learning algorithms that can be modified to conduct sophisticated report triaging, expanded to include customized alert systems, or tailored to address other surveillance needs. In conjunction with human expertise, GRITS is a valuable tool for infectious disease surveillance.

Background

Infectious diseases pose a significant threat to global health, and economic stability.^{1,2} Due to extensive globalization and urbanization, infectious diseases can spread at unprecedented rates.³ Small and localized infectious disease threats can rapidly become international catastrophes, as demonstrated by Severe Acute Respiratory Syndrome (SARS) in 2003, influenza (H1N1A) in 2009, and Ebola Virus Disease in 2014.^{4,5,6} Early detection of emerging threats is critical to implementing effective responses, and is achievable through robust global disease surveillance.⁷ Disease surveillance is defined

by the U.S. Centers for Disease Control and Prevention (CDC) as, "the ongoing, systematic collection, analysis, interpretation, and dissemination of data regarding a health-related event for use in public health action to reduce morbidity and mortality and to improve health".⁸ Despite international awareness of the essential role disease surveillance can play in mitigating infectious disease threats, there are currently gaps in global biosurveillance infrastructure.

Problem

Gaps in Traditional Disease Surveillance

Traditional disease surveillance relies predominantly on local clinicians, laboratory technicians, and public health practitioners to detect infectious disease outbreaks. Once detected, these outbreaks are communicated to regional, national and international public health organizations for evaluation and response. However, traditional disease surveillance faces several challenges. The systems' ability to detect and collect information is dependent on regional healthcare infrastructure that is heterogeneous in quality. Disease threats that emerge in regions with poor healthcare infrastructure may be detected slowly or missed completely.⁹ This problem was illustrated recently in the ongoing Ebola Virus Disease epidemic in West Africa, in which the etiological agent was not identified until 85 days after the first case.¹⁰ Additionally, traditional disease surveillance typically uses official communication channels that are often controlled by governments. However, governments are often concerned that increased awareness of emerging infectious disease events will adversely affect their economy, and will pressure government agencies to restrict communication to the public.¹¹ The over protection of critical infectious disease outbreak data is highly problematic to global health.

Digital Disease Detection: Promising and challenging

The burgeoning field of digital disease detection (DDD) attempts to complement traditional disease surveillance by using software to expand data collection to non-tradition sources, bypass official communication routes and automate epidemiological analyses. Prominent examples of DDD tools include the Global Public Health Intelligence Network (GPHEN), PulseNet , HealthMap, Argus, BioCaster, and the Global Early Warning System for Major Animal Diseases Including Zoonoses (GLEWS).^{12,13,14,15,16,17}

DDD is an innovative, but innately challenging field. Although, the quantity of digital information is immense, the signal to noise ratio is very low.

To curate data sources for epidemiologically significant information many DDD tools require substantial human capital. This dependency may be rate limiting and makes it difficult to provide disease detection that is superior (faster and more accurate) to traditional surveillance systems. Conversely, some DDD tools make extensive use of NLP or machine learning algorithms to predict disease risk, but do not incorporate sufficient public health, or medical expertise. This has led to inaccurate predictions of disease risk. In some cases, like early experiences with Google Flu Trends, this has led to inaccurate forecasts of infectious disease.¹⁸ DDD tools that enable analysts to examine non-traditional information more efficiently, and do not circumvent health analysts, are critically needed.

Solution

GRITS is a biosurveillance tool created to aid infectious disease threat detection efforts by monitoring and analyzing textual data sources (e.g., news articles, medical reports). GRITS extracts critical epidemiologic information (e.g., case-counts, symptoms, pathogens, transmission types, hosts, dates, locations) from text and delivers a list of possible diseases associated with submitted text ranked by probability. GRITS visualizes infectious disease occurrence for analysts on a timeline and map, and provides links to similar and potentially related infectious disease reports using a customizable search feature. Prior to analysis, non-English articles are translated using Bing Translator. GRITS is available through a web-application for detailed analysis of specific text, and through an API for rapid analysis of large collections of text, and integration with other applications.

Intended Audience

GRITS is intended to serve as a tool for public health and military analysts in their daily infectious disease surveillance efforts. GRITS will serve both individual users through the web interface, and developers of other surveillance applications through a comprehensive diagnostic API.

Components & Features

GRITS Web-Application

The GRITS web-application is a dynamic web-interface for performing detailed analysis of a text sample. Through the web-application, users can submit a text sample to GRITS for analysis and exploration, and view the previous GRITS analyses they have conducted. Users can view a dynamic visualization of GRITS analyses through the Diagnostic Dashboard, use the Find Similar Articles Feature to conduct customizable searches of a pre-analyzed article index for related reports, and search for relevant infectious disease emergence events using the Find Similar Disease Emergence Events feature.

Diagnostic Dashboard

The GRITS Diagnostic Dashboard provides users with a ranked diagnosis of potential infectious diseases associated with submitted text. Diagnoses are determined using a MaxEnt BoW classifier trained on disease labels HealthMap assigned to articles over a recent 2-year interval. Possible diagnoses are ranked based on confidence score. The keywords that contribute to each diagnosis and their relative weights can be inspected from within the diagnostic dashboard (Fig. 1).

The GRITS Diagnostic Dashboard identifies, and succinctly visualizes the important information in a text sample (Fig. 2). This includes disease related keywords (disease, pathogens, symptoms, modes of transmission, and hosts), dates, locations and case counts. Extracted information is highlighted in an annotated article view that can be filtered by category. Locations are plotted in a map view (Fig. 3), and a histogram of temporal information is shown in a timeline view (Fig 3). Date extraction is done via the Stanford SUTime library. The Geoname resolution uses a custom algorithm with geonames.org data. Case count extraction uses the CLiPS Pattern library's



figure 2

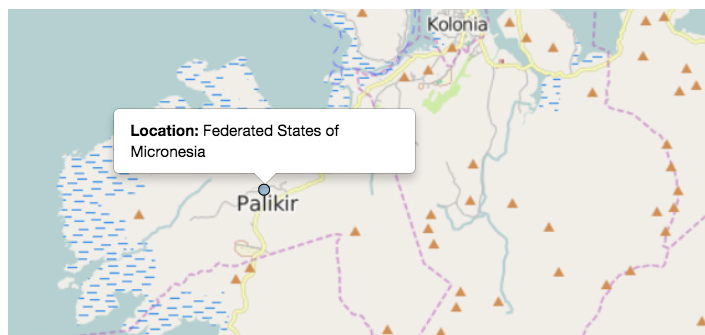


figure 3

search module with a number of search phrases tailored to meet GRITS requirements.

Disease *	animal b	ite	bleeding	contact with infected	crimean congo hemorrhagic fever	endemic	exposure	hemorrhagic fever	infected animal	rash	risk	surveillance	tick	tickborne	ticks	turkey	veterinarians	viral	virus
Crimean-Congo Hemorrhagic Fever	0.059	0.168	0.179	0.032	0.44	0.161	0.035	0.046	0.293	0.037	0.074	0.034	0.427	0.022	0.194	0.057	0.063	0.2110	.009

figure 1

Find Similar Articles Feature

The GRITS Find Similar Articles Feature (Fig. 4) provides a simple, customizable interface to query over 250,000 disease related articles collected by HealthMap over a 2-3 year interval. These articles have all been subjected to GRITS analysis. Articles can be searched by any combination of the following information obtained from their completed GRITS analysis; diagnosed disease, extracted keywords, publication date range, and/or country. A query of articles similar to the text sample is automatically conducted based on the results of the text sample's GRITS analysis. Search results can be viewed spatially using Kitware's geojis library, or in a list view. The list view includes links to the original articles, and selected meta-data, including diagnosed disease, case counts, distinct keywords and more. Aggregations of publication dates and countries associated with search results are visualized in histograms.

The screenshot displays the 'Find articles with...' interface. It includes three main filter sections: 'Any of these diseases:', 'Any of these keywords:', and 'All of these keywords:'. Each section has a list of items with radio buttons for selection and an 'Add' button. The 'Any of these diseases:' section lists 'Fever' and 'Dengue'. The 'Any of these keywords:' section lists various medical terms like 'leukopenia', 'fever', 'dengue', etc. The 'All of these keywords:' section lists 'mosquito' and 'village'. Below these filters is a section titled 'In one of these countries:' with a list of countries and their approximate document counts: India (32 reports), Philippines (14 reports), Brazil (13 reports), Pakistan (9 reports), China (6 reports), DR Congo (6 reports), Fiji (6 reports), Greece (5 reports), Botswana (2 reports), and Collectivité spéciale de New Calédonie, France (2 reports). To the right of the filters, a 'Page 0' header shows '10 of 111 Results' and sorting options. Below this, 'Search Results:' are listed, showing snippets of search results with titles like 'Fiji dengue case count now 68: Report - Outbreak News Today' and 'Spike in dengue cases alarm health officials - Catanduanes Tribune', along with their respective dates, health map labels, diseases diagnosed, counts, distinct keywords, and relevance scores.

figure 4

Find Similar Disease Emergence Events Feature

This feature allows users to query the Emerging Infectious Disease Repository (EIDR) developed by EcoHealth Alliance for infectious disease emergence events that may be related to a text sample. Development of EIDR is ongoing. As of March 2015, EIDR contained information on 369 infectious disease emergence events occurring between 1940 and 2013 (Z.S. Gold, A.G. Huff, Tilchin C. et al., unpublished). GRITS automatically generates a list of emergence events associated with the diagnosed disease(s) for a text sample. EIDR events can also be queried by host, pathogen, disease, transmission mode, country and date. Aggregations of emergence event start dates and associated countries are visualized in histograms. Users can examine emergence events in detail through the EIDR web-application, which can be accessed by clicking on any emergence event search result.

GRITS API

The GRITS API allows users to apply GRITS diagnostics continuously on large collections of data and to use GRITS intelligence in any application. API users can submit documents for analysis through a REST API and receive comprehensive results in a JSON format, including the disease diagnosis, identified pathogens, hosts, symptoms, locations, dates and more.



Extracted keywords are annotated with their occurrences in the original text, so they may be easily highlighted in a user interface, indexed for search, used as features for statistical evaluation or machine learning, or subjected to further analysis. This open and flexible API allows developers to easily integrate GRITS analysis capabilities into their own application with simple calls to a central web service with no requirement to set up or maintain their own installation of GRITS, although that is supported as well.

In the hands of the astute public health analyst, GRITS is a powerful tool for infectious disease surveillance. GRITS allows users to more efficiently monitor non-traditional data sources for infectious disease threats. GRITS can be expanded to incorporate a customizable triaging system that curates text sources temporally, spatially, by diagnosed disease, or by public health keyword. Additional ontologies can be created to train GRITS to make educated conclusions on additional complex variables besides disease, like pathogen class, report risk level, or the emergence of a novel pathogen. An alert system could be built into GRITS to notify users of particular report clusters that may signal the emergence of potentially dangerous situations. Additionally, through the GRITS API, the tool can be incorporated into larger surveillance systems, like the Defense Threat Reduction Agency's prototype Biosurveillance Ecosystem, and run continuously on those systems data feeds. Through the web-interface users can evaluate the functionality and public health foundations of GRITS, increasing the transparency of the tool. This will lead to scrutiny and the refinement, and continued development of GRITS. GRITS, and surveillance systems like it, may be able to fill gaps in traditional global biosurveillance systems.

Applications beyond

disease surveillance

GRITS is fundamentally sophisticated NLP and machine learning software that has been tailored to address disease surveillance needs. The GRITS technology could be applied to a plethora of topics. For instance, GRITS could be of value to the financial sector. Textual sources are rich with indicators of investor sentiment, and are often monitored by NLP tools. GRITS could be tailored to detect clusters of investor sentiment indicative of an emerging financial crisis, or market shift.

¹Morens DM, Folkers GK, Fauci AS. 2004. The challenge of emerging and re-emerging infectious diseases. *Nature*430:242-249

²Fonkwo PN. 2008. Pricing infectious disease. *EMBO Reports*9:S13-S17. doi: 10.1038/embor.2008.110.

³Hosseini P, Sokolow SH, Vandegrift KJ, Kilpatrick AM, Daszak P. 2010. Predictive power of air travel and Socio-Economic data for early pandemic spread. *PLoS One*5: e12763. doi: 10.1371/journal.pone.0012763

⁴Centers for Disease Control and Prevention. 2003. Outbreak of severe acute respiratory syndrome – worldwide. *MMWR*52:226-228. url: <http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5211a5.htm>

⁵Centers for Disease Control and Prevention. 2010. The 2009 H1N1 pandemic: summary highlights, April 2009-April 2010. url: <http://www.cdc.gov/h1n1flu/cdcresponse.htm>.

⁶Schar D, Daszak P. 2015. Ebola economics: the case for an upstream approach to disease emergence. *EcoHealth*11:451-452.

⁷Ferguson NM, Cummings DAT, Cauchemez S, Fraser C, Riley S, Meeyai A, Iamsirithaworn S, Burke DS. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*437:209-214.

⁸German RR, Lee LM, Horan JM, Milstein RI, Pertowski CA, Waller MN. 2001. Guidelines working group Centers for Disease Control and Prevention (CDC). Updated guidelines for evaluation public health surveillance systems: recommendations from the guidelines working group. *MMWR Recomm Rep* 2001 July 27;50:1-35.

⁹Hosseini P, Sokolow SH, Vandegrift KJ, Kilpatrick AM, Daszak P. 2010. Predictive power of air travel and Socio-Economic data for early pandemic spread. *PLoS One*5: e12763. doi: 10.1371/journal.pone.0012763

¹⁰World Health Organization. 2015. Global alert and response. One year into the Ebola epidemic: A deadly, tenacious and unforgiving virus. url:<http://www.who.int/csr/disease/ebola/one-year-report/introduction/en/>.

¹¹Sturevant JL, Aranka A, Brownstein JS. 2007. The new international health regulations: Considerations for global public health surveillance. *Disaster Med and Public Health Prep*1:117-121.

¹²Mykhalovskiy E, Weir L. 2006. The Global Public Health Intelligence Network and Early Warning Outbreak Detection: A Canadian Contribution to Global Public Health. *Can J of Public Health* 97:42-44.

¹³Pulsenet. 1996. Centers for Disease Control and Prevention. <http://www.cdc.gov/pulsenet>

¹⁴Brownstein JS, Freifeld CC, Reis BY, Mandl KD. 2008. Surveillance Sans Frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS medicine*5:e151. doi:10.1371/journal.pmed.0050151.

¹⁵Hartley DM, Nelson NP, Walters R, Arthur R, Yangarber R, Madoff L, Linge JP, Mawudeku A, Collier N, Brownstein JS, Thinus G, Lightfoot N. 2010. Landscape of international event-based biosurveillance. *Emerg Health Threats J*3:e3. doi:10.3134/ehj.10.003.

¹⁶Collier N, Doan S, Kawazoe A, Goodwin RM, Conway M, Tateno Y, Ngo Q, Dien D, Kawtrakul A, Takeuchi K, Shigematsu M, Taniguchi K. 2008. BioCaster: Detecting public health rumors with a Web-based text mining system. *Bioinformatics*24:2940-2941

¹⁷FAO, OIE, WHO. 2011. GLEWS. Global early warning system for major animal diseases, including zoonoses. <http://www.glews.net/> (cited April 2015).

¹⁸Olson RD, Konty KJ, Paladini M, Viboud C. 2013. Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales. *PLoS Comput Biol* 9:e1003256. doi:10.1371/journal.pcbi.1003256.

¹⁹Cleary ME, Antelman ET, Markuzon N, Miller SM, Postlethwaite TA, Prasov Z. 2014. COMBS: A Biosurveillance Ecosystem (BSVE) Prototype. *Online J of Public Health Inform*6. doi: <http://ojphi.org/ojs/index.php/ojphi/article/view/5060>.

²⁰Lugmayr A, Friedrichsen M. 2013. Predicting the future of investor sentiment with social media in stock exchange investments: A basic framework for the DAX performance index. In Friedrichsen M, Muhl-Benninghaus W (ed), *Handbook of social media management value chain and business models in changing media markets*. Springer Berlin Heidelberg.



EcoHealth Alliance