**Global Rapid Identification Tool System (GRITS)**

**Final Report (July 2014 - January 2015)**

This report covers the work completed during the GRITS extension period from July 18th, 2014 to January 17th, 2015.

**Table of Contents**

[**A. Summary**](#h.vn4gv4un2yr0)

[**B. GRITS milestones**](#h.5r16zamsv2nr)

[1. Enhance the performance of the GRITS Media Diagnostic tool (GRITS.md)](#h.8afjrecofqrn)

[2. Integrate a network of experts (GRITS.net)](#h.xod9wt03dxr)

[3. Build capabilities for the GRITS platform to handle high-volume, real-time data feeds (GRITS.db)](#h.ri6eyvtb9lg7)

[4. Connect GRITS to the EcoHealth Global Repository of Infectious Diseases (GRID)](#h.6r8elpwxxfg9)

**List of Figures**

## 1. Extended Time Extraction

## 2. Key Points

## 3. Ansible Deployment Process

**4. Interaction with GRITS by Email**

**5. GRITS Search Page Enhancements**

**6. GRID Search page**

**7. Report to ProMED button**

**8. BSVE plugin**

# 

# **A. Summary**

This project was originally initiated as the Rapid Identification Tool (RIT) for undiagnosed diseases and evolved into the Global Rapid Identification Tool System (GRITS). This evolution reflects the powerful “system” of tools that were developed to extend the diagnostic capabilities to “global” coverage. With the support of a no-cost extension (through January 2015), the diagnostic capabilities of GRITS were improved. GRITS was expanded making the software scalable and to support high-volume data feeds.

This includes the main web application (GRITS.app), media diagnostics (GRITS.md), and Girder database (GRITS.db). We have also integrated our expert network (GRITS.md) and Global Repository of Infectious Disease (GRID) into the system.

# **B. GRITS milestones**

The second expansion period had four (4) major milestones:

1. Enhance the performance of the GRITS Media Diagnostic tool (GRITS.md)
2. [Integrate a network of experts (GRITS.net)](#h.xod9wt03dxr)
3. [Build capabilities for the GRITS platform to handle high-volume, real-time data feeds (GRITS.db)](#h.ri6eyvtb9lg7)
4. [Connect GRITS to the EcoHealth Global Repository of Infectious Diseases (GRID)](#h.6r8elpwxxfg9)

### 1. Enhance the performance of the GRITS Media Diagnostic tool (GRITS.md)

#### *1.1 Expand the capabilities of the expert annotator to a broad spectrum of disease characteristics (e.g., symptoms, signs, findings, time, location, host, pathogen, drivers, impact)*

The latest version of GRITS is capable of extracting keywords for modes of disease transmission, hosts, host uses, vectors, environmental factors, control measures, demographic information, occupations, and disease categories, symptoms, and risk factors. We manually curated keyword lists for characteristics that were not available from third-party ontologies.

Additionally, the accuracy of location extraction was improved in GRITS, and date and time extraction was expanded to include new information like past and present references, outbreak duration, and ranges. Performance of case count extraction was improved, and case count visualizations were prototyped to determine how these data could be most useful to analysts.

#### 

#### *1.2 Advance the automatic natural language processing tools to accommodate the annotation of expanded disease characteristics*

To incorporate new disease characteristics in GRITS, the annotation format was standardized and migrated our previous annotation code to a new annotation library. A Scala-based tool was developed to use Stanford NLP libraries within GRIT’s annotation library, enabling the new types of date and time extraction. All extracted information includes text offsets, and can be viewed and highlighted from the diagnostic dashboard, or obtained via the GRITS API. A method of grouping disease characteristics was prototyped and outbreak information was summarized to give analysts a quick summary of the events mentioned in a long report.

#### *1.3 Develop a GRITS ontology for infectious disease diagnostics and outbreak detection*

Manually curated keyword lists were combined with information from third-party ontologies to create GRITS ontology for extracting disease and outbreak information. To improve disease diagnostics, the GRITS ontology also includes manually a curated hierarchy of diseases, disease categories, and pathogens.

#### 

#### *1.4 Improve diagnostic accuracy with rigorous statistical approach to machine learning*

The process for training and testing the diagnostic classifier was overhauled. In the redesigned process, the full range of disease information available from HealthMap is aggregated, instead of keeping only one disease and location label for each article. The training script was modified to take advantage of multiple disease labels within a single article. Furthermore, the training and testing scripts were updated to query our database for recent articles instead of using fixed sets of articles. One disadvantage of this update is that these changes prevent the comparison of future f-scores from the previous approach. However, the updated version of the diagnostic classifier has an f-score of 0.49%, with a precision of 56.4%, and a recall of 49.82%.

#### *1.5 Integrate methods from activity models and computer vision classifiers*

The computer vision experts at Kitware suggested that SVM classifiers and cross-validation methods should be used to improve performance.

#### 

#### *1.6 Develop multilingual support for annotation and diagnosis*

The GRITS API processing pipeline is connected to Bing Translator to support annotation and diagnosis of submitted articles in other languages. Bing Translator can translate non-English articles before extracting disease characteristics and diagnosing. Text offsets returned by the API refer to the English translation.

### 2. Integrate a network of experts (GRITS.net)

#### *2.1 Build interface to GRITS web app(s) for experts (e.g., EcoHealth Alliance, Epidemico, and ISID) to collectively diagnose disease reports*

The GRITS web application allows users to create diagnostic dashboards that can be shared with colleagues. Each expert can leave feedback and comments on whether the automated diagnosis from GRITS is correct. The application also includes user profile pages where users can access previous dashboards they have created.

#### *2.2 Build an interface into ProMED-mail and Healthmap editor platforms to allow editors to interact with the GRITS app and submit media for diagnosis*

After multiple consultations with subject matter experts, the team members concluded that email would be the easiest way to incorporate GRITS into their normal workflow. When a subject matter expert receives a disease-related email, they can simply forward the email to GRITS to receive a list of possible diagnoses and a link to a diagnostic dashboard for the article.

#### *2.3 Create data filtering mechanism that uses the diagnostic tools to redirect documents to appropriate experts and/or recommend an expert*

Experts from ProMED-mail strongly prefer to have a moderator manually redirect documents, so instead of automating expert recommendations, GRITS provides key diagnostic information to the top moderators to allow them to quickly determine an appropriate expert to further analyze the article.

#### 

#### *2.4 Create API for receiving requests for expertise from BSVE*

A button was developed that enables an analyst to open an email to the ProMED top moderator with GRITS’ diagnostic information automatically populated. The analyst using GRITS is required to provide the details of their request and send the email, and then the moderator can direct the request to an appropriate expert. The button is currently available in GRITS diagnostic dashboards, but could be incorporated into a standalone BSVE application.

#### *2.5 Develop dialogue mechanisms for discussions with an expert*

The button to send a request to ProMED-mail enables analysts to start a discussion with an expert. As they conduct an email discussion, they can cc GRITS to extract and store information from the discussion.

#### *2.6 Prototype system for multilingual support from regional EcoHealth Alliance, Epidemico, and ISID experts.*

When ProMED-mail moderators receive a request related to a non-English article, they can use English diagnostic information from GRITS to direct the article to a regional expert. When the expert responds, GRITS can also translate the response and extract information. To expand beyond the prototyping stage, GRITS would need to translate requests into regional experts’ languages and internationalize the GRITS user interface.

### 3. Build capabilities for the GRITS platform to handle high-volume, real-time data feeds (GRITS.db)

#### *3.1 Develop capacity of Kitware’s Girder database to handle big data*

Kitware made numerous improvements to the feature set and stability of the core Girder database and released stable versions 1.0 and 1.1, installable via pip. Highlights include support for asynchronous server-side jobs, improved REST endpoint security and fixes for CSRF and XSS vulnerabilities, an admin page in the web interface, and folder-level metadata to support portfolios. Additionally, Kitware restructured the schema used to store HealthMap data in Girder, making it possible to detect changes in data over time.

#### 

#### *3.2 Enhance geospatial capabilities of Girder*

Kitware improved Girder’s geospatial search using MongoDB’s geospatial query extensions, and improved the GeoJS library for client-side visualizations. GeoJS now includes a d3-like API for point, line, graph, and polygon features, a flexible mouse navigation handler supporting animation effects, a renderer-independent mouse-picking API, and UI components such as a zoom control and a legend.

#### *3.3 Develop server stack to process data for filtering and recommendation*

Filtering and recommendations are provided through the API via a Girder plugin and in a search page on the diagnostic dashboard. We use elasticsearch to find recommendations based on keywords present in the text. Users can specify location and date range fields to filter their search, or can pre-populate a search page from a diagnostic dashboard to find related articles.

#### 

#### *3.4 Create infrastructure to diagnose media in a high-volume data stream*

We migrated the GRITS deployment process to an Ansible playbook, which is faster, more reliable, and easier to maintain. Ansible simplifies the deployment process and makes it possible to deploy components of GRITS separately, allowing us to quickly scale our infrastructure as necessary to support higher volumes of data. Ansible also makes it possible to easily retrain the diagnostic classifier or re-index the database independently of deploying code changes. We also load-tested the GRITS API to ensure that performance would be acceptable when processing high volumes of data.

The stability and reliability of our celery task queue was improved to process large amounts of data for our filtering and recommendation API. The celery task queue can potentially scale article diagnosis to multiple machines if there is an extremely high-volume data stream; however, the current volume of articles is not sufficient to justify the extra operating cost, setup time, and maintenance of a distributed system. This improved article-scraping process detects and ignores pages that have been removed or redirected. The diagnostic queue prioritizes new articles submitted through the API, but processes new HealthMap articles in the background. GRITS can preprocess articles before a classifier has been trained, to shorten the time it takes to set up a new GRITS system.

#### 

#### *3.5 Build API for processing high-volume, real-time data to support the BSVE*

The GRITS API provides diagnostic information and extracted disease characteristics for submitted text. It sends submissions to a task queue and prioritizes them. We worked with Digital Infuzion to determine how the GRITS API or its components could fit into the BSVE, and helped Digital Infuzion set up a local GRITS system so they can access their own version of the GRITS API if necessary.

#### *3.6 Support mechanism for BSVE users to connect local datasets to GRITS.db*

The GRITS API was set up so that the BSVE will be capable of running GRITS diagnostic tools on any local datasets users have connected to the BSVE. Additionally, Kitware implemented CORS support in Girder, enabling external applications to use the Girder API. In the future, the BSVE could use the API to upload additional data to Girder, to diagnose and use in training new GRITS classifiers.

#### *3.7 Formalize provenance information (W3C PROV) and standardize metadata*

Kitware implemented a provenance plugin for Girder to standardize provenance information.

#### *3.8 Generalize data hosting capabilities to support capacity to mine, diagnose, filter, prioritize, visualize, connect, and recommend data feeds for the BSVE or other platforms.*

GRITS continually pulls in new links from HealthMap, scrapes relevant text, mines disease characteristics, and diagnoses diseases. The GRITS web application provides diagnostic dashboards with visualizations. GRITS also contains a search page that can automatically recommend articles related to a dashboard and filter the articles based on user input. The search page includes full text search, article aggregations, pagination, and relevance sorting.

Kitware developed a BSVE plugin to provide search results from our HealthMap database to BSVE users, including a map visualization, and Digital Infuzion has set up the GRITS API and can make it accessible to BSVE users in the future.

### 

### 4. Connect GRITS to the EcoHealth Alliance’s Global Repository of Infectious Diseases (GRID)

#### *4.1 Create recommendations to match media with historic disease outbreaks from GRID*

Elasticsearch was used to provide recommendations of GRID events. Recommendations can be filtered by disease, location, and date, and include links to GRID event pages with further information.

#### 4.2 Create portal in GRITS web app to the GRID community data editor

The GRITS web app includes a GRID search page. Like the search page for related media articles, the GRID search page can be pre-populated from a dashboard, or the user can enter diseases or keywords to filter the search. Search results link to event pages in GRID, which include maps and statistics.

#### *4.3 Connect GRID portfolios of historic disease media to GRITS web app*

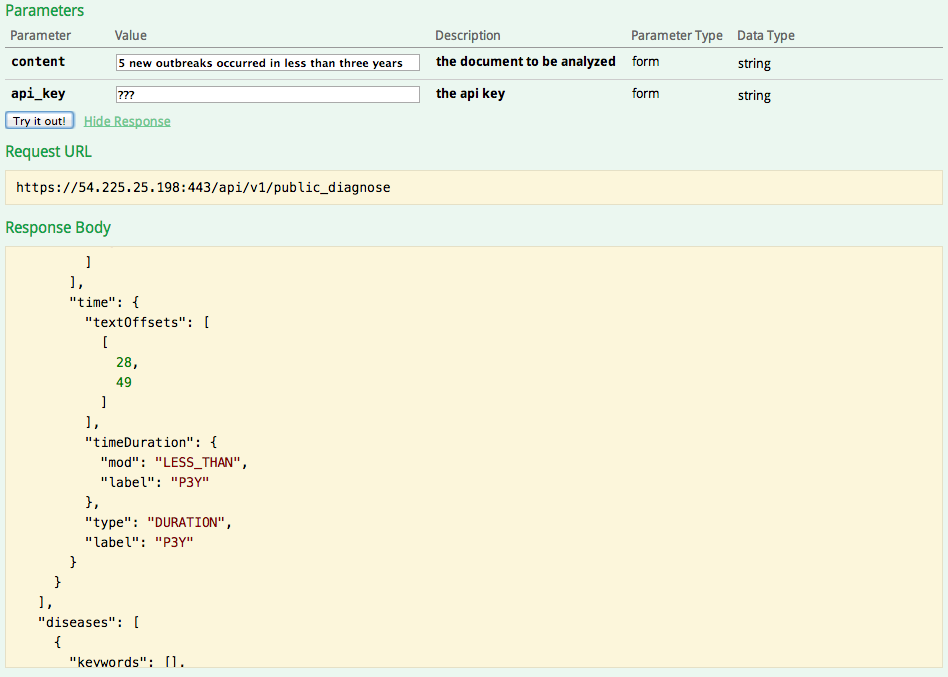
Event pages in GRID include a list of references related to the event, which may include scientific papers, websites, and news media.

#### *4.4 Develop API to recommend and filter historic outbreak data for BSVE.*

We developed a GRID plugin for Girder to provide API access to GRID recommendations.

# 

# 



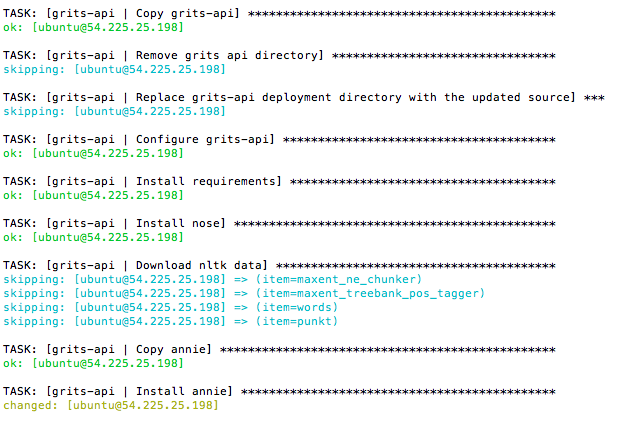
## 

## Figure 1. Extended Time Extraction. GRITS time extraction now captures ranges, durations, and sets. For example, “less than three years” or “from 1999 until late 2001.”

## 

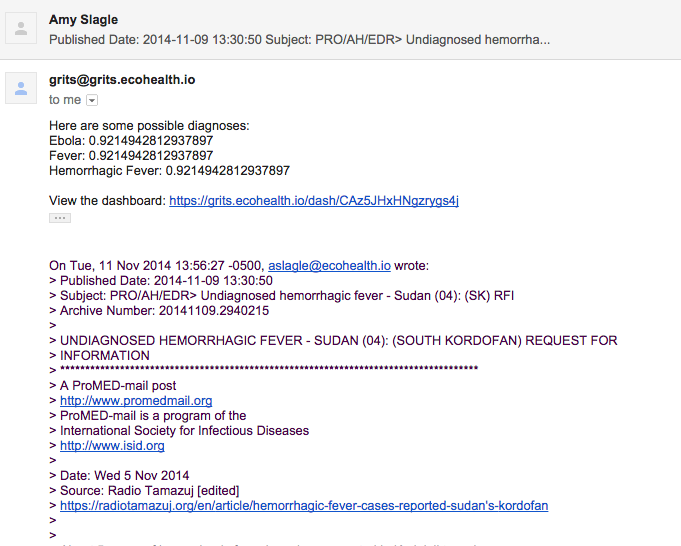


## Figure 2. Key Points. We prototyped a feature that groups extracted information into “key points.” Key points include case counts, locations, occupations, symptoms, risk factors, hosts, and time.



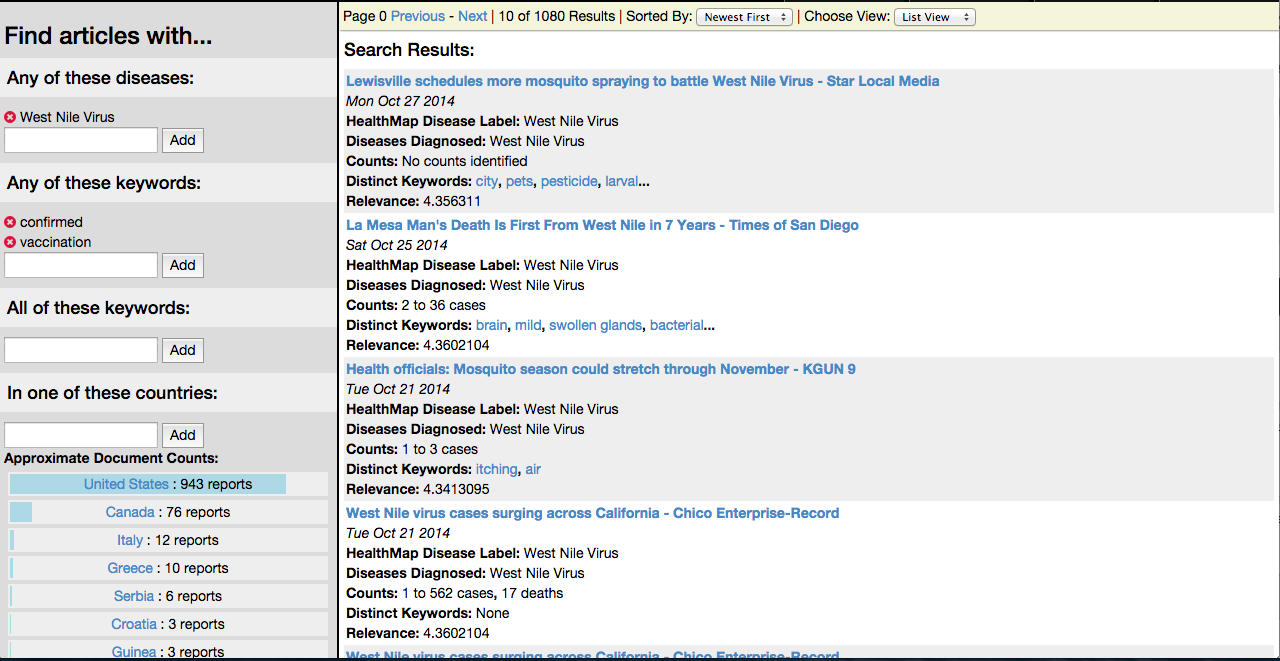
## 

**Figure 3.** Ansible Deployment Process. We migrated the GRITS deployment process to an ansible playbook, which is faster, more reliable, and easier to maintain.



## Figure 4. Interaction with GRITS by Email. Support for receiving and sending email from the GRITS application was added. An expert can send an email to GRITS, and receive a differential diagnosis with probabilities for different diseases, and a link to a GRITS dashboard.

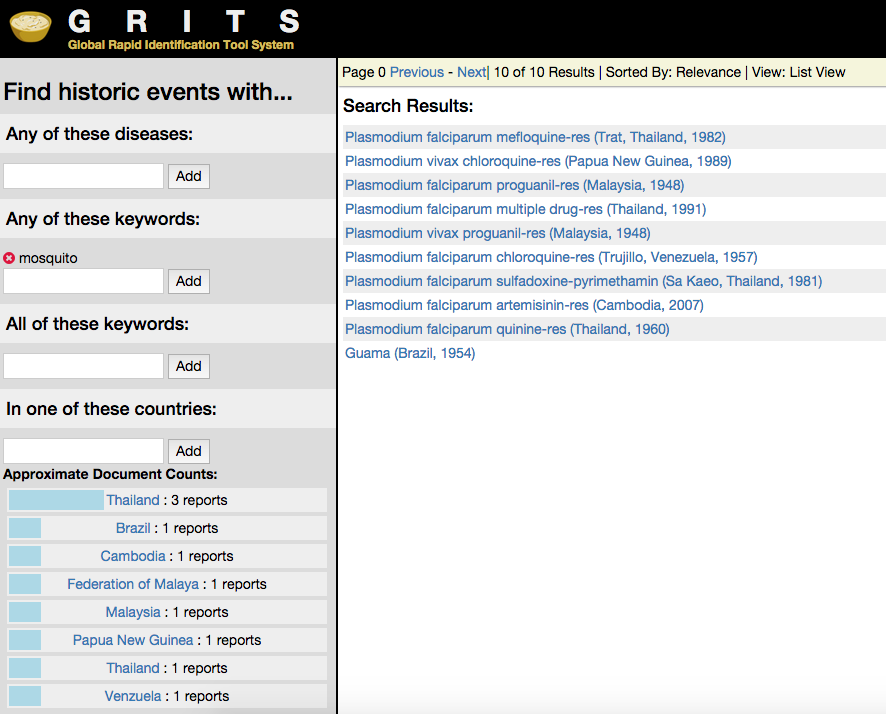
## 



## 

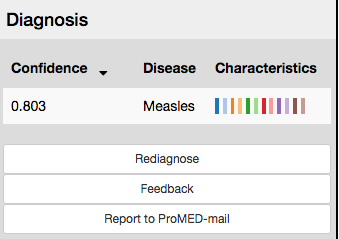
## Figure 5. GRITS Search Page Enhancements. The search page can automatically recommend articles related to a dashboard, and then filter them based on user input. It includes full text search, article aggregations, pagination, and relevance sorting.

## 

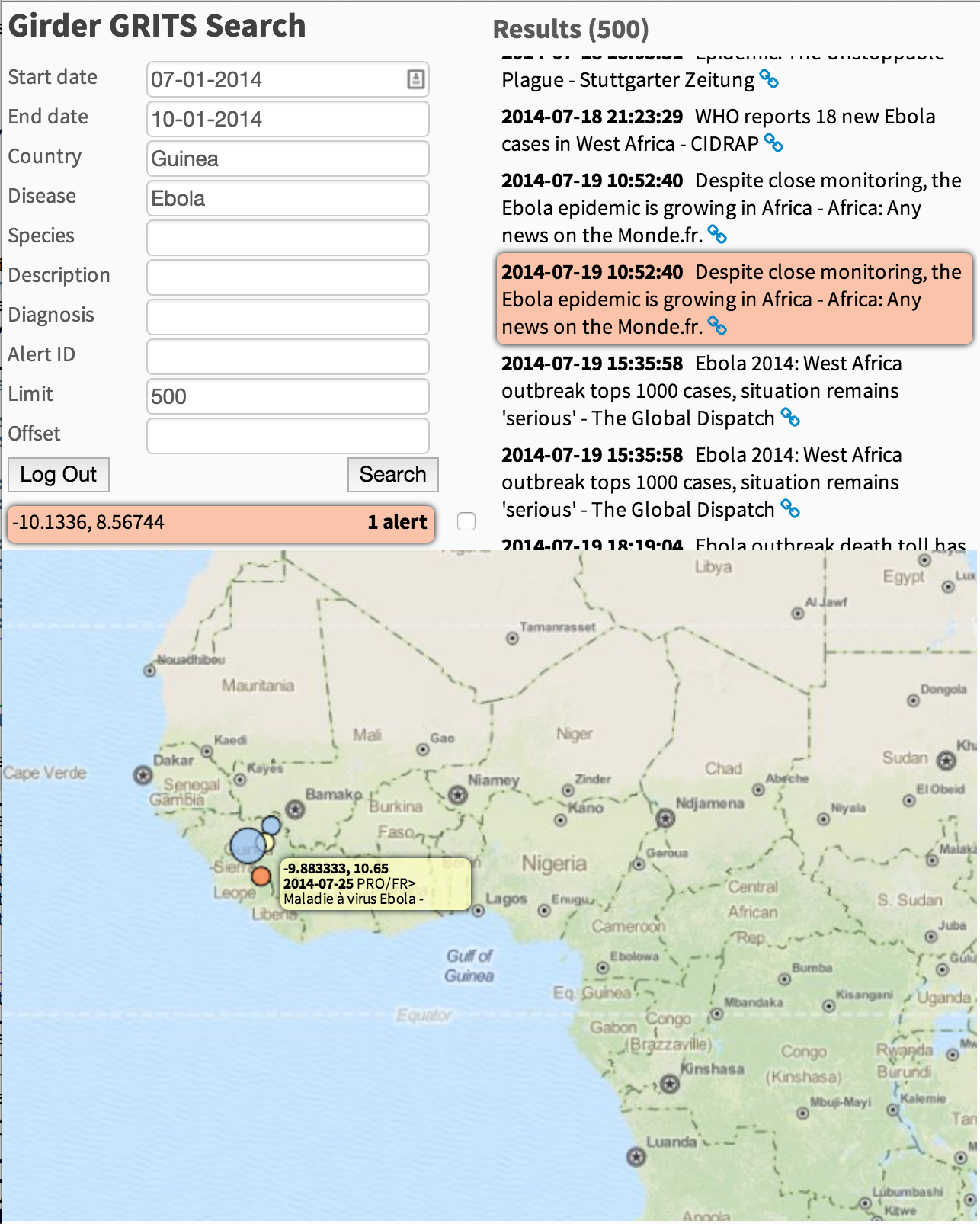


## Figure 6. GRID Search page. The GRID search page provides links to event pages in GRID, and allows searching by any or all diseases and keywords, and filtering by country or date.

## 



## Figure 7. Report to ProMED button. A button to send a request to ProMED from a diagnostic dashboard was added. Clicking the button will open the user’s email client with a email template pre-filled with the diagnosis, article content, and a link to the dashboard in GRITS. The user must fill in the details of their request before sending the email.



## Figure 8. BSVE plugin. Kitware developed a plugin for the BSVE to allow users to search our HealthMap incident database and display results on a map.