

# Open Source Health Intelligence (OSHINT) for Foodborne Illness Event Characterization

Catherine Ordun, Jane W. Blake\*, Nathanael Rosidi, Vahan Grigoryan, Christopher Reffett, Sadia Aslam, Anastasia Gentilcore, Marek Cyran, Matthew Shelton and Juergen Klenk

Booz Allen Hamilton, McLean, VA, USA

## Objective

We propose a cloud-based Open Source Health Intelligence (OSHINT) system that uses open source media outlets, such as Twitter and RSS feeds, to automatically characterize foodborne illness events in real-time. OSHINT also forecasts response requirements, through predictive models, to allow more efficient use of resources, personnel, and countermeasures in biological event response.

## Introduction

An increasing amount of global discourse reporting has migrated to the online space, in the form of publicly accessible social media outlets, blogs, wikis, and news feeds. Social media also presents publicly available and highly accessible information about individual, real-time activity that can be leveraged to detect, monitor, and more efficiently respond to biological events.

## Methods

Salmonella and Escherichia Coli (E. coli) events were selected based on the magnitude and number of reported outbreaks to the Centers for Disease Control (CDC) in the last ten years (1). These events affect multiple states and were large enough to ensure appropriate confidence levels when developing response metrics obtained from our prediction models. We collected social media data between 2006 – 2012 due to the emergence of Twitter, Facebook, and other social media utilization during this time period.

Characterization is defined as the process of identifying specific event features that inform overall situational awareness. The number hospitalized, dead, or injured, in addition to patient demographics and symptoms were determined to be useful for our characterization and forecast event metrics. Analytical methods, such as term-frequency-inverse document frequency (TF-IDF), natural language processing (NLP), and information extraction, were used to characterize events according to our metrics. Lexicon development, during NLP implementation, was generated from online news articles used to describe the events. Lastly, forecasting algorithms were developed to predict the potential response based on similar historical events that were initially characterized by our information extraction algorithms.

## Results

The OSHINT system was developed in Amazon Web Services and includes real-time social media collection for event characterization (see Figure 1). OSHINT currently characterizes number of victims ill, hospitalized, and dead due to foodborne illness events.

OSHINT was used to characterize the recent national 2012 Salmonella event related to cantaloupes, during which OSHINT characterized social media posts related to the event, as news articles and Twitter tweets streamed into the system (Figure 2). On August 17, 2012 the OSHINT system identified a large increase in Twitter tweets mentioning salmonella. Social media data found absent (victims missing work or school day), death, hospital, and sick events to involve 2, 4, 17, 283 media mentions, respectively. Our TF-IDF algo-

rithm characterized the salmonella event impact as two dead and 150 sickened by salmonella-tainted cantaloupe. Retrospective analysis of CDC reported data on August 30, 2012 indicated the salmonella event involved two deaths in 204 cases (2).

## Conclusions

The OSHINT team is continually developing and refining characterization and forecasting algorithms used in the system. Upon completion, OSHINT will characterize symptoms, geography, and demographics for E. coli and Salmonella events. The system will also forecast number sick, dead, and hospitalized for an effective and quick response. We will refine our algorithms and evaluate the system against past and future events to provide confidence in our results.



Figure 1. OSHINT System in Amazon Web Services.



Figure 2: 2012 Salmonella Outbreak in Cantaloupe

## Keywords

Open Source; Forecasting; Social Media; Response; Food Safety

## Acknowledgments

Frederika Conrey, Kenneth Decker, Willam Lei, Dania Shor, Misha Zhurkin

## References

- (1) CDC. Retrieved September 7, 2012 from <http://www.cdc.gov/outbreaknet/investigations>.
- (2) CDC. Retrieved August 30, 2012, from <http://www.cdc.gov/salmonella/typhimurium-cantaloupe-08-12/index.html>.

\*Jane W. Blake  
 E-mail: [blake\\_jane@bah.com](mailto:blake_jane@bah.com)

