

# Data Integration and Analytics within the National Bio-Surveillance Integration System

Nancy Grady<sup>2</sup>, PhD, Jim Jones<sup>2</sup>, MS, Lowell Vizenor<sup>3</sup>, PhD, Joe Dudley<sup>2</sup>, PhD,  
Jeanne Marin<sup>2</sup>, PhD, Anne Havey<sup>2</sup>, BS, Tom Sabo<sup>4</sup>, MS, Ray Champoux<sup>4</sup>, BA,  
Robert Hogan<sup>2</sup>, MS, Dennis Beal<sup>2</sup>, MS, Laura Peitersen<sup>2</sup>, PhD, Jean Orr<sup>2</sup>, BA,  
Karl Kropp<sup>2</sup>, PhD, Kimothy Smith<sup>1</sup>, PhD

<sup>1</sup>Department of Homeland Security, Office of the Chief Medical Officer,  
<sup>2</sup>SAIC, <sup>3</sup>Ontology Works, <sup>4</sup>Attensity

## OBJECTIVE

This presentation introduces the U.S. Department of Homeland Security (DHS) National Bio-Surveillance Integration System (NBIS) and the analytics functionality within the NBIS that integrates and analyzes structured and unstructured data streams across domains to provide inter-agency analysts with an integrated view of threat scenarios. The integration of Human and Animal incidences of Avian Influenza will be used to demonstrate initial capability.

## BACKGROUND

The National Bio-surveillance Integration System[1][2] began initial development in the fall of 2006 with the goal of integrating data from federal, state, private and international sources (including food, agricultural, public health, environmental monitoring and intelligence communities). This integration and subsequent analysis assists inter-agency analysts in developing continuous situational awareness, early recognition of an attack, and decision support for outbreak and event response from an incident whether intentional or naturally occurring.

## METHODS

NBIS currently acquires data on the incidence of avian influenza (AI) primarily from a variety of web-accessible open sources, from international health organizations, non-governmental organizations, and from multiple media sources. The AI data must be integrated while maintaining its provenance.

Unstructured case-based data currently undergoes manual extraction before being stored in dimensional tables. Natural Language Processing (NLP) is being evaluated for the automatic identification of noteworthy events.

The now structured AI data is semantically integrated through an ontological framework and stored in a knowledge base, from which data sets are extracted for analytics. The ontology encodes subject matter expertise and provides a representation of organisms {humans, animals, viruses,...}, events {exposure, infection, hospitalization, death, reporting,...}, attributes {age, gender-quality, latitude, longitude,...}, and relationships {died\_on, located\_in...}.

The reporting events have been further identified in specific sequences for several countries, a “chain of denial”, that relate the official admission or denial of outbreaks to the actual presence of the AI virus.

The case-based AI data is being processed by two Bayesian Belief Network (BBN) models to determine (i) the likelihood of human to human transmission in the actual human and animal cases, and (ii) the likelihood of an outbreak given a series of potentially untrue denial statements in the chain of denial corpus.

## RESULTS

While manual integration of a restricted number of data streams is possible, integrating the content of numerous disparate data streams with collateral contextual data such as weather and transportation routes is not scalable without the use of semantic methods. We have developed an ontological framework to support this large scale integration effort.

The latest metrics will be presented in the talk on the tuning of the NLP identification of specific events as measured against manual extraction, and on the comparison of BBN models to the evaluation of subject matter experts.

## CONCLUSIONS

While this is a work in progress, the integration of human and animal AI data through the semantic concepts expressed by subject matter experts and the performance of the initial NLP and BBN models demonstrate the efficacy of this approach. The principal challenges to ensuring the scalability of the system lie in ensuring the completeness of the ontology, identifying important contextual datasets, and encoding the surety and provenance of the analysis.

## REFERENCES

- [1] Smith, Kimothy. Statement to the House, Subcommittee on Prevention of Nuclear and Biological Attack, Creating a Nationwide, Integrated Biosurveillance Network, Hearing, May 11, 2006. Available at: <http://chs.clientapp2.com/hearings/viewhearing.aspx?id=30>
- [2] Smith, K., Keynote Presentation at the NSF Biosurveillance workshop, New Brunswick, NJ, May 22, 2007.

Further Information:  
Nancy Grady, [gradyn@saic.com](mailto:gradyn@saic.com)