

A Web Based Tool For the Detection and Analysis of Avian Influenza Outbreaks From Internet News Sources

Ian Turton¹ and Andrew Murdoch²

May 1, 2008

¹GeoVISTA Center, Pennsylvania State University, University Park, PA 16802, ijt1@psu.edu

²MGIS Program, World Campus, Pennsylvania State University, a_murdoch@hotmail.com

Abstract

An increasing amount of health related data is now available on the Internet (Freifeld *et al.*, 2007; Woodall, 2001, 1997; M'ikanatha *et al.*, 2006), while some of this data is unverified much is produced by medical professionals on the ground (e.g. ProMED, WHO). It is increasingly difficult for interested analysts to keep up with the amount of data available from the many sources but with increased worries about the risk of global pandemics or intentional bio-terrorism incidents it is becoming increasingly important that quick detection and fast response is possible.

This paper describes an experimental project to provide health analysts with a simple web based tool to allow them to quickly and easily view of events in the current outbreak of Avian Influenza (http://www.experimental.geovista.psu.edu/andrew/html/avian_influenza_map.html). The system is constructed using a collection of open source tools that with minor customization could be used to visualize any phenomena that can be accessed using an RSS feed. The system is best characterized as a client server system where the user's web browser provides the execution environment for the client and the server is implemented as a Java middle ware component (GeoServer) over a spatially enabled SQL database (PostGIS).

The client consists of a map and time line that the user can use to zoom into any region of the Earth that is of interest to them and limit the markers displayed by selecting a particular time period using the time bar. This functionality is provided using the OpenLayers mapping client (<http://www.openlayers.org>) which provides Open Geospatial Consortium standard compatible web map server display as well as the ability to display GeoRSS feeds on the map. The timeline is based on the MIT SIMILE project's time line tool (<http://simile.mit.edu/timeline/>). A GeoRSS feed is an RSS feed, such as those used for distributing updates to web sites and news, that has been enhanced by the addition of geospatial coordinates (<http://www.georss.org>).

The server is based on the open source web mapping server GeoServer to serve data from a PostGIS database that stores the georeferenced news items and other background maps. At a set interval a java process is spawned by the cron system on the server that collects the latest items from the RSS news feeds and passes them through a custom geocoder based on FactXtractor. FactXtractor is an information extraction web service for Named Entity Recognition (NER) and Entity Relation Extraction (RE) developed at PSU and available at <http://julian.mine.nu/snedemo.html>. FactXtractor processes a text document using GATE (Cunningham *et al.*, 2002) and identifies entity relations using Stripped Dependency Tree kernels. The named entities are then processed using the GeoNames.org database by adding a set of coordinates for each place entity detected to the item (Turton *et al.*, 2007). These news items are then stored in the spatially enabled database from which they can be served to the map client. By making use of GeoServer's ability to produce output in multiple formats it is possible to send the data to the client encoded as GeoJSON which can be read by both OpenLayers and the SIMILE Timeline.

Currently the system is based on two feeds, one the WHO avian influenza feed from the Epidemic and Pandemic Alert and Response (EPR) section which contains information about outbreaks of avian influenza from around the world. The second feed is a collection of news feeds from news agencies and aggregation services such as Google News, this is more experimental. It suffers from some problems related to being collected from a wider range of sources such as a tendency to collect the same news item from several sources (usually a news agency story that has been widely published by local newspapers) as well as stories that are about avian influenza but not about actual outbreaks. Future work is planned to make use of classification techniques (Konchady, 2006; Segaran, 2007) to reduce the duplication of articles and to classify the news items in to groups that are of more interest to an analyst.

References

- Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. *In: Proceedings of the 40th Annual Meeting of the ACL.*
- Freifeld, Clark C C., Mandl, Kenneth D D., Reis, Ben Y Y., & Brownstein, John S S. 2007. HealthMap: Global infectious disease monitoring through automated classification and visualization of Internet media reports. *J Am Med Inform Assoc*, December.

- Konchady, Manu. 2006. *Text Mining Application Programming (Programming Series)*. Charles River Media.
- M'ikanatha, N. M., Rohn, D. D., Robertson, C., Tan, C. G., Holmes, J. H., Kunselman, A. R., Polachek, C., & Lautenbach, E. 2006. Use of the internet to enhance infectious disease surveillance and outbreak investigation. *Biosecurity and bioterrorism : biodefense strategy, practice, and science*, **4**(3), 293–300.
- Segaran, Toby. 2007. *Programming Collective Intelligence: Building Smart Web 2.0 Applications*. O'Reilly Media, Inc.
- Turton, Ian, Gahegan, Mark, & Jaiswal, Anuj. 2007 (September). Geographic Information Retrieval from Disparate Data Sources. *In: GeoComputation'07*.
- Woodall, J. 1997. Official versus unofficial outbreak reporting through the Internet. *International journal of medical informatics*, **47**(1-2), 31–34.
- Woodall, J. P. 2001. Global surveillance of emerging diseases: the ProMED-mail perspective. *Cadernos de saúde pública / Ministério da Saúde, Fundação Oswaldo Cruz, Escola Nacional de Saúde Pública*, **17 Suppl**, 147–154.