

West Nile Virus and Botulism Portal: A Case Study in Infectious Disease Informatics ¹

Daniel Zeng¹, Hsinchun Chen¹, Chunju Tseng¹, Catherine Larson¹,
Millicent Eidson², Ivan Gotham², Cecil Lynch³ and Michael Ascher⁴

¹Department of Management Information Systems
University of Arizona, Tucson, Arizona

{zeng,hchen,chun-ju,cal}@bpa.arizona.edu

²New York State Department of Health, SUNY, Albany
{mxe04,ijg01}@health.state.ny.us

³California Department of Health Services, UC Davis, Sacramento
clynch@dhs.ca.gov

⁴Lawrence Livermore National Laboratory
ascher1@llnl.gov

Abstract. Information technologies and infectious disease informatics are playing an increasingly important role in preventing, detecting, and managing infectious disease outbreaks. This paper presents a collaborative infectious disease informatics project called the WNV-BOT Portal system. This Portal system provides integrated, Web-enabled access to a variety of distributed data sources related to West Nile Virus and Botulism. It also makes available a preliminary set of data analysis and visualization tools tailored for these two diseases. This system has helped to demonstrate the technological feasibility of developing a cross jurisdiction and cross species infectious disease information infrastructure and identify related technical and policy-related challenges with its national implementation.

1 Introduction

Infectious disease outbreaks are critical threats to public health and national security [1, 4, 12]. With greatly expanded trade and travel, infectious diseases, either naturally occurred or caused by biological terror attacks, can spread at a fast pace within and across country borders, resulting in potentially significant loss of life, major economic crises, and political instability.

Information systems play a central role in developing an effective comprehensive approach to prevent, detect, respond to, and manage infectious disease outbreaks of plants, animals, and humans [2, 5]. Currently, a large amount of infectious disease data is being collected by various laboratories, health care providers, and government agencies at local, state, national, and international levels [11]. However, there exist a number of technical and policy-related challenges hindering the effective use and sharing of infectious disease data, especially datasets across species and across juris-

¹ Research reported in this paper has been supported in part by the NSF through Digital Government Grant #EIA-9983304.

dictions, in regional, national, and global contexts [3]. Several key challenges are summarized below.

- *Existing infectious disease information systems do not fully interoperate.* Most existing systems have been developed in isolation [8]. As such, when disease control agencies need to share information across systems, they may resort to using nonautomated approaches such as e-mail attachments and manual data (re)entry. In addition, much of the search and data analysis function is only accessible to internal users.
- *The information management environment used to analyze large amounts of infectious disease data and develop predictive models needs major improvements.* Current infectious disease information systems provide very limited support to professionals analyzing data and developing predictive models. An integrated environment that offers functionalities such as geocoding, advanced spatio-temporal data analysis and predictive modeling [6], and visualization is critically needed.
- *An efficient reporting and alerting mechanism across organizational boundaries is lacking.* Certain infectious disease information needs to be quickly propagated through the chain of public health agencies and shared with law enforcement and national security agencies in a timely manner. Certain models exist within the human public health community (e.g., the Centers for Disease Control and Prevention (CDC)'s ArboNet and Epi-X) and within certain states (e.g., New York State's Health Information Network (HIN)). However, in general the current reporting and alerting mechanism is far from complete and efficient, and may involve extensive and error-prone human interventions.
- *Data ownership, confidentiality, security, and other legal and policy-related issues need to be closely examined.* When infectious disease datasets are shared across jurisdictions, important access control and security issues need to be resolved between the involved data providers and users. Subsets of such data are also governed by relevant healthcare and patient-related laws and regulations. Negotiating the agreements that must be developed to govern access to and use of disease-related data by agencies and individuals can be labor and time-intensive [9].

This paper summarizes our ongoing research and system development effort motivated to address some of the above challenges. This effort is aimed at developing scalable technologies and related standards and protocols needed by the full implementation of the national infectious disease information infrastructure and at studying related policy issues. The resulting research prototype, called the **WNV-BOT Portal** system, provides integrated, Web-enabled access to a variety of distributed data sources related to West Nile Virus (WNV) and Botulism. It also provides information visualization capabilities as well as predictive modeling support. In this paper, we summarize the background and application context of our project and present the main technical components of WNV-BOT Portal. We also discuss broader technical and policy issues related to the design and development of a scalable national infectious disease infrastructure based on the lessons learned through our prototyping effort.

The rest of the paper is structured as follows. Section 2 discusses WNV and Botulism datasets and the related existing public health systems that WNV-BOT Portal is

designed to integrate and interoperate. In Section 3, we present the overall system design and main technical components of WNV-BOT Portal. We conclude the paper in Section 4 by summarizing our research and discussing our ongoing activities and future plan.

2 West Nile Virus and Botulism Datasets and State Public Health Systems

The emergence of WNV in the Western Hemisphere was reported first in New York State in late summer 1999. This unprecedented event required rapid mobilization and coordination of hundreds of public health workers, expenditure of millions of dollars on an emergency basis, and immediate implementation of massive disease surveillance and vector control measures. The Health Information Network (HIN) system has been used by New York State to enable rapid and effective response to the WNV crisis. The HIN is an enterprise-wide information infrastructure for secure Web-based information interchange between the New York State Department of Health (NYSDOH) and its public health information trading partners, including local health departments and the New York State Department of Agriculture and Markets, New York State Department of Environmental Conservation, and the United States Department of Agriculture's Wildlife Services New York office [5]. This system currently supports 20,000 accounts and 100 mission critical applications, cross-cutting all key public health response partners in the state of New York. It implements sophisticated data access and security rules, allowing for real-time use of the data within the state while protecting confidentiality and scientific integrity of the data. The infrastructure is well suited to public health response, as illustrated by New York's ability to rapidly incorporate it into its plan to respond to the WNV outbreak in NY in 1999-2000 [5]. The system has evolved into an integrated surveillance system containing large quantities of real-time data related to WNV including (a) human cases, (b) dead bird surveillance data, (c) asymptomatic bird surveillance data, (d) mammal cases, and (e) mosquito surveillance data.

WNV has yet to manifest as an indigenous human disease in California, but the historical geographic spread of this disease and the fact that WNV has been detected in sentinel flocks of chickens and mosquito pools in California, would indicate the high likelihood that the state will have to deal with large numbers of cases later this year. Important as a cause of neurological morbidity and death, WNV is also a prototype of an emerging viral infection. The analysis of data collected regarding its occurrence and spread provides a basis for the development of predictive models for other emerging or as yet unidentified diseases. The California Department of Health Services (CADHS) has access to the detailed datasets from California's mosquito control districts and surveillance data on sentinel flock, dead bird, and equine specimens. In collaboration with USGS, we also have datasets concerning domestic and wild animal populations that might be exposed to WNV; some related data is available through CDC.

Botulism is a disease rarely seen in the United States with fewer than 200 cases per year reported to the CDC. Despite the low volume of cases, because of the risks asso-

ciated with the possibility of a terrorist event utilizing botulinum toxin, the importance of having a system in place to identify and manage larger numbers of cases of the disease cannot be overestimated. The transactional data generated in such a system must also be available for post-event analysis in order to improve public health methods and responses. Both New York and California represent a significant part of the world economy and are high risk targets for bioterrorism due to the high level of international traffic into the states. NYSDOH has an internal database for botulism cases that occurred in New York State. In California, no computerized system currently exists that is capable of handling the information gathering, retrieval, and dissemination needs for a bio-terrorist event involving botulinum toxin. However, detailed paper-based information is available on both Botulism cases and antitoxin inventory. In addition, nationwide avian botulism data is maintained and updated by the National Wildlife Health Center.

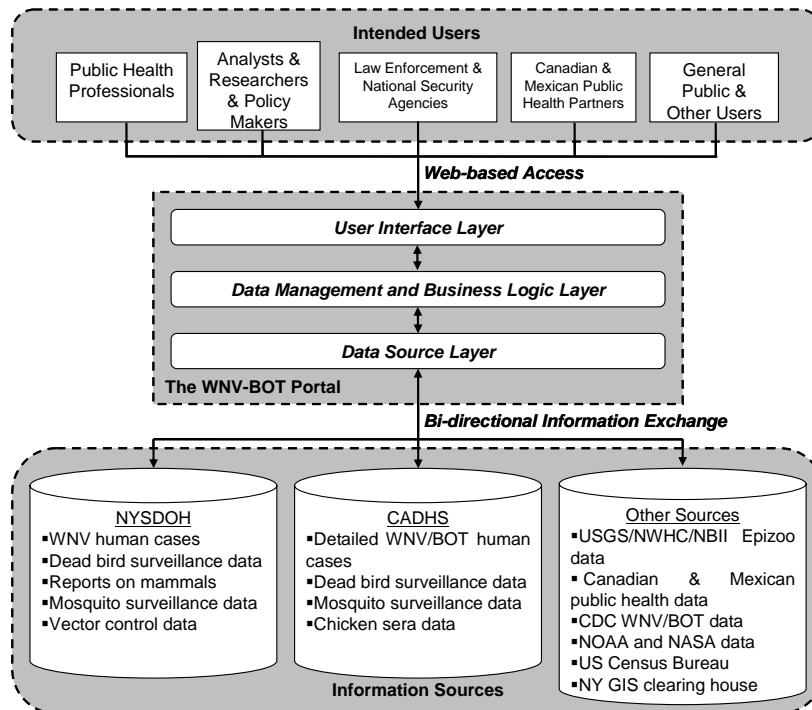


Fig. 1. Data Sources and Intended Users of the WNV-BOT Portal

3 WNV-BOT Portal System Development

The WNV-BOT Portal system has been developed to integrate infectious disease datasets on WNV and Botulism from New York, California, and several federal data sources. It also provides a set of data analysis, predictive modeling, and information

visualization tools tailored for these two diseases. Figure 1 summarizes these datasets and intended users of WNV-BOT Portal.

3.1 WNV-BOT Portal System Design

As illustrated in Figure 2, from a systems perspective, WNV-BOT Portal is loosely-coupled with the state public health information systems in that the state systems will transmit WNV/BOT information through secure links to the portal system using mutually-agreed protocols. Such information, in turn, will be stored in the internal data store maintained by WNV-BOT Portal. The system also automatically retrieves data items from sources such as those from USGS and stores them in the internal data store.

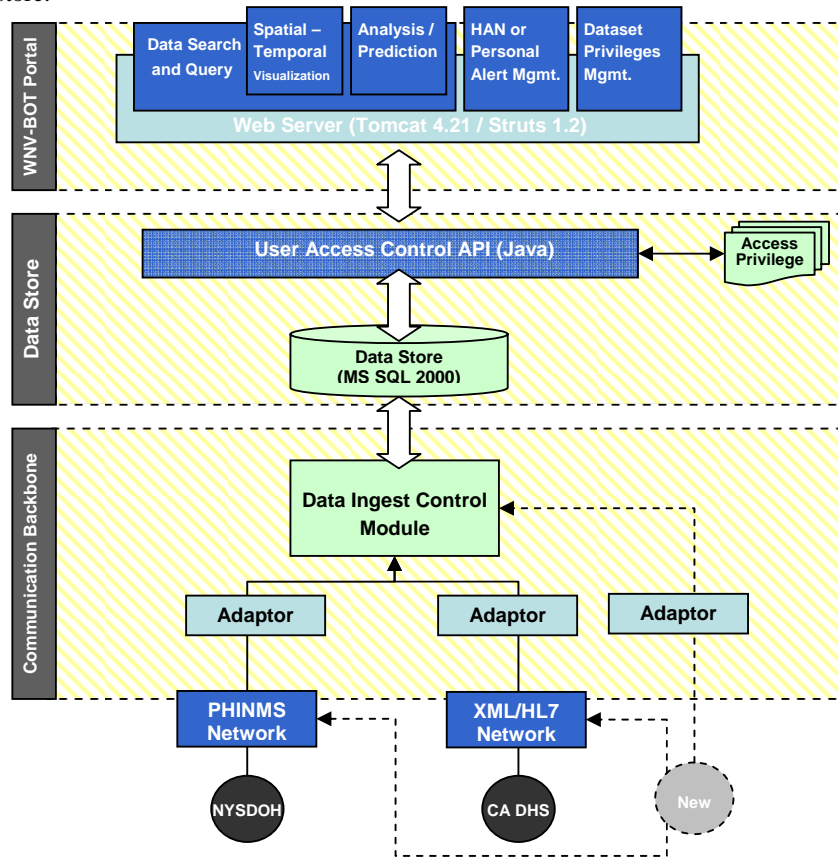


Fig. 2. Overall Architecture of the WNV-BOT Portal System

Architecturally, WNV-BOT Portal consists of three major components: a communication backbone, a data store, and a Web portal. Figure 2 illustrates these compo-

nents and shows the main data flows between them and the underlying WNV/BOT data sources. The communication backbone module implements data transmission protocols. It normalizes data from various participating sources and is responsible for converting incoming data into a data format internal to the Portal system. The data store module receives normalized data from the communication backbone and stores it in a relational database. The data store module also implements a set of Java APIs that serves data requests from the user and imposes data access rules. The Web portal module implements the user interface and provides the following main functionalities: (a) searching and querying available WNV/BOT datasets, (b) visualizing WNV/BOT datasets using spatial-temporal visualization, (c) accessing analysis and prediction functions, and (d) accessing the alerting mechanism. The remainder of this section discusses the design considerations and technical details of these three major modules. A use scenario will be used to demonstrate how these modules work together to satisfy the user's information and analysis needs.

3.2 Communication Backbone

The communication backbone module enables data exchanges between WNV-BOT Portal and the underlying WNV/BOT sources. Several federal programs have been recently created to promote data sharing and system interoperability in the healthcare domain. The CDC's Electronic Disease Surveillance System (NEDSS) initiative is particularly relevant to our research. It builds on a set of recognized national standards such as HL7 for data format and messaging protocols and provides basic modeling and ontological support for data models and vocabularies. NEDSS and HL7 standards are having a major impact on the development of disease information systems. Although these standards have not yet been tested in cross-state sharing scenarios, they provide a solid foundation for data exchange standards in the national and international contexts. WNV-BOT Portal heavily utilizes NEDSS/HL7 standards.

WNV-BOT Portal currently supports two messaging systems: Public Health Information Network Messaging System (PHIN MS) which is developed by the CDC, and XML/HL7 Message Broker Network (MBN) which has been deployed in CADHS's California Public Health Information Network (CALPHIN). For illustration purposes, we use PHIN MS to describe the main components of these messaging systems. PHIN MS uses the Electronic Business Extensible Markup Language, ebXML, infrastructure to securely transmit public health information over the Internet. It is platform-independent and loosely coupled with systems that produce or receive messages. PHIN MS has three major components: Message Sender, Message Receiver, and Message Handler. The Message Sender sends ebXML messages to one or more Message Receivers; the Message Handler takes corresponding actions based on message types. Below is an example ebXML message sent from NYSDOH to the Portal following the PHIN MS system. CALPHIN has a similar architectural design but different applicable data formats:

```

- <Import>
  <Sending_Entity>NY</Sending_Entity>
  <Date_File_Created>2004-02-18</Date_File_Created>
  <Data_End_Date>2004-02-17</Data_End_Date>
  <NETSS_SiteID>36</NETSS_SiteID>
  <Number_Confirmed_Cases>3</Number_Confirmed_Cases>
  <Number_Probable_Cases>1</Number_Probable_Cases>
  <Number_Negative_Cases>1</Number_Negative_Cases>
  <Number_Indeterminate_Cases>0</Number_Indeterminate_Cases>
  <Number_NotDone_Cases>0</Number_NotDone_Cases>
  <Number_InProgress_Cases>0</Number_InProgress_Cases>
  <Number_Birds_Found>5</Number_Birds_Found>
- <Bird_Data>
  <Sending_Entity>NY</Sending_Entity>
  <Bird_ID>8444</Bird_ID>
  <Collection_Date>2002-01-01</Collection_Date>
  <Location_Address>9999 S Swan St</Location_Address>
  <Location_City>Albany</Location_City>
  <Location_Zip>12203</Location_Zip>
  <Location_County>Albany</Location_County>
  <Location_Latitude>42.659142</Location_Latitude>
  <Location_Longitude>-73.772406</Location_Longitude>
  <Bird_Species>WPU OTHER SPECIES</Bird_Species>
  <Number_Birds>1</Number_Birds>
  <WNV_Test_Results>P</WNV_Test_Results>
  <Captive_Or_Free_Ranging>F</Captive_Or_Free_Ranging>
  <Symptoms>Dead, trauma</Symptoms>
  <Necropsy_Diagnosis>WPU diagnosis 1</Necropsy_Diagnosis>
  <CDC_Week>1</CDC_Week>
</Bird_Data>

```

From an implementation perspective, the communication backbone module uses a collection of source-specific “adaptors” to communicate with underlying sources. We use the adaptor linking PHIN MS system and WNV-BOT Portal to illustrate a typical design of such adaptors. The data from NYSDOH to the portal system is transmitted in a “push” manner. NYSDOH sends secure PHIN MS messages to the portal nightly. The PHIN MS adaptor at the portal side runs a data receiver daemon listening for incoming messages. After a message is received, the adaptor will invoke the data ingest control module and stores the verified message in the portal’s internal data store. The adaptor is also responsible for sending error and control messages back to PHIN MS if necessary. Other data sources (e.g., those from USGS) may have “pull” type adaptors which will periodically download information from the source systems and examine and store data in the portal’s internal data store.

Another important function of the communication backbone module is data ingest control, which enforces different types of security checks ranging from account verification to data validation. The ingest control process is invoked by communication adaptors when a message is received. The first step of the process is to perform account authentication and ensure that the account is authorized to ingest data. (The HL7-based message/data payload is protected by public key encryption.) In the second step, the incoming message is examined by normalizing and cleansing subroutines before being saved in the portal data store. In this normalization and cleansing step, predefined dictionaries, taxonomy databases, and related string similarity meas-

ures such as Levenshtein distance function, are used to identify possible typos in the incoming data. When typos are identified, the message is routed back to the data source for confirmation. At the end of the ingest control process, the message is converted into a standard HL7 XML format, ready to be stored in the Portal's internal data store.

3.3 Portal Data Store

A main objective of WNV-BOT Portal is to enable users from partnering states and organizations to share data. Typically data from different organizations has different designs and stored in different formats. To enable data interoperability, we use Health Level Seven (HL7) standards (<http://www.hl7.org/>) as the main storage format. In our approach, contributing data providers transmit data to WNV-BOT Portal as HL7-compliant XML messages through the secure communication backbone. After receiving these XML messages, WNV-BOT Portal will store them directly in its data store, a relational database built in Microsoft SQL server. This HL7 XML-based design provides a key advantage over an alternative design based on a consolidated database. In a consolidated database design, the portal data store has to consolidate and maintain all the data fields for all datasets. Whenever an underlying dataset changes its data structure, the portal data store needs to be redesigned and reloaded to reflect the changes. This severely limits system scalability and extensibility. Our HL7 XML-based approach does not have these limitations.

To alleviate potential computational performance problems associated with this HL7 XML-based approach while searching and querying, an index of stored XML objects is created. We have identified a core set of data fields based on which search will be done frequently and extracted these fields from all XML messages to be stored in separate database tables to enable fast retrieval. A flexible database schema is deployed to facilitate storage of different attributes from different datasets such as bird species from dead bird data and toxin types from Botulism cases. This design enables the Portal to dynamically generate customized search criteria for each dataset in the search process.

The access control API as part of the data store module is responsible for granting and restricting user access to sensitive data. To satisfy important data integrity and confidentiality requirements, we are implementing a central, role-based access control system. In this system, the owners of various WNV/BOT datasets specify explicitly the access privilege on datasets provided by them. To facilitate this access privilege elicitation process, we have developed a predefined set of privileges from which the data owners or providers can choose. Examples of such privileges include (a) "full visibility" indicating that all information is visible to any user; (b) "aggregation only at the county level" indicating that only aggregated information at the county level is visible to the user. There are two models to assign such privileges: the individual model and the trusted model. In the individual mode, the privileges are assigned directly to a specific individual user. In the trusted model, the data source administrator assigns access privileges to user roles as explained below.

In our current design, a user role is defined to have three parts: state, organization, and user type. For example, if user "John Smith" is from organization "AILab" in the

state of Arizona and his type is “COORDINATOR,” then the combination of Arizona, AILab and COORDINATOR is John’s user role. A role can be represented by a string with various parts delimited by dots. For instance, the above role can be specified as “AZ.AILAB.COORINATOR.” One user can have multiple roles in WNV-BOT Portal and these roles are assigned and managed by the administrator of each organization through a Web-based interface. Using the above system, the data owner/provider can easily specify which user roles are allowed to view the data, and which level of access privilege they may have. The access control API will ensure that these specified access control rules apply to all data access requests.

3.4 Web Portal

The WNV-BOT Portal website (<http://wnvbot.eller.arizona.edu>) is the main user interface, through which the user queries WNV/Botulism databases, analyses and visualizes datasets, and accesses alerting messages. Portal and site administrators also manage data access rules and user roles through the website. In this section, we present a use scenario to highlight the key data query and visualization functions provided by the Portal.

3.4.1 Data Query and Search. A typical data query process is comprised of five steps: (1) the user selects the disease of interest; (2) the user selects interested datasets; (3) the user specifies the time and geographic ranges of interest; (4) the system displays the returned query results; and (5) the user uses a visualization tool to summarize and explore interactively the returned results. We use the following example to illustrate how a user can access the Portal databases and perform basic searches and queries. Consider a user who is interested in WNV-positive, dead crow cases in New York State between 2001 and 2003. This example uses a test (not real) data set. After logging onto the Portal, the user first selects the disease of interest, in this case, WNV. The Portal then presents a list of WNV-related datasets from which the user can choose. The metadata associated with the datasets is also shown next to the displayed dataset names. The datasets that the user does not have access privilege to will be listed at the bottom of the page. If the user needs to view the unprivileged datasets, he or she can request the access through the Portal. In the use scenario, the user selects the NY_DEADBIRD dataset to which the access privilege is available. By clicking the “advanced” button, the user accesses the “advanced search” mode. A list of source-specific attributes is presented and the user can submit fine-grained search criteria. In our example, the user selects the “Positive,” “ND,” and “unknown” statuses and the “crow” species. In addition, the user specifies the interested time range as that between “2001-01-01” and “2003-12-31,” and the interested location as “All counties” from New York State.

After receiving all search criteria, the Portal system performs the searches and displays the query results in a table. Results are grouped by the dataset, and the name of the dataset and the number of the returned data items are shown in the title bar. The first row of the table shows the names of the various data fields returned. When the user moves the mouse over this first row, a pop-up tool-tip will be shown to provide a short description of the corresponding data field. (Note that depending on the access

privilege the user or the associated user role has over the underlying datasets, this table may show data in different granularity even given the same search criteria). In our example, we assume the user has full access privilege and 475 data records in total are returned containing details such as case ID, date, state, county, status, and bird species. Figure 3 illustrates these four steps of the data query process discussed above.

The screenshot illustrates the WNV-BOT portal interface, which is used for querying West Nile virus data. The interface is divided into six numbered steps:

- 1) Select disease:** The user selects the disease to search, with options for West Nile Virus and Botulism.
- 2) Select datasets:** The user selects the datasets to search, with options for CA_DEADBIRD, CA_DEADBIRD, CA_WNV_HUMAN, NY_DEADBIRD, and WNV_DEADBIRD.
- 3) Fine tune search criteria:** The user fine-tunes the search criteria, including Case Status (ID, Negative, Unknown) and Bird Species (Quail, Chicken, House Sparrow, Other).
- 4) Select spatial and time period:** The user selects the spatial and time period, including From (YYYY MM DD) and To (YYYY MM DD) dates, and Specify Locations of Interest (CA, NY).
- 5) Aggregation view:** The user views the aggregation view, showing a table of data for various states and counties.
- 6) Detail view:** The user views the detail view, showing a table of data for individual cases, including Case ID, Case Date, State, County, Status, and Bird.

The following table represents the data shown in the 'Detail view' step:

ID	Case Date	State	County	Status	Bird
119-1307	2003-05-10	NY	WESTCHESTER	Unknown	Case
119-1348	2003-05-10	NY	WESTCHESTER	Unknown	Case
087-1198	2003-05-10	NY	ROCKLAND	Unknown	Case
087-1198	2003-05-14	NY	ROCKLAND	Unknown	Case
087-1179	2003-05-21	NY	ROCKLAND	Unknown	Case
087-1173	2003-05-20	NY	ROCKLAND	Unknown	Case
088-1192	2003-03-22	NY	NASSAU	Unknown	Case
089-1180	2003-03-23	NY	NASSAU	Unknown	Case
087-1032	2003-05-18	NY	ROCKLAND	Unknown	Case
119-1028	2003-05-10	NY	WESTCHESTER	Unknown	Case
119-1021	2003-05-04	NY	WESTCHESTER	Unknown	Case
119-1024	2003-05-10	NY	WESTCHESTER	Unknown	Case
119-1082	2003-05-11	NY	WESTCHESTER	Unknown	Case
087-1086	2003-05-08	NY	ROCKLAND	Positive	Case
087-1082	2003-05-21	NY	ROCKLAND	Unknown	Case
071-1038	2003-05-03	NY	ORANIE	Unknown	Case
119-1052	2003-05-05	NY	WESTCHESTER	Unknown	Case

Fig. 3. Query West Nile virus Data through WNVBOT portal

3.4.2 Spatial-Temporal Visualization. This section focuses on Step 5 of the data query and analysis step, i.e., visualization. The role of visualization techniques in the context of large and complex dataset exploration is to organize and characterize the data visually to assist users in overcoming the information overload problem [13]. WNV-BOT Portal makes available an advanced visualization module, called the Spatial Temporal Visualizer (STV) to facilitate exploration of infectious disease case data and to summarize query results. STV is a generic visualization environment that can be used to visualize a number of spatial temporal datasets simultaneously. It allows the user to load and save spatial temporal data in a dynamic manner for exploration and dissemination. STV has three integrated and synchronized views: periodic, timeline, and GIS. The periodic view provides the user with an intuitive display to identify periodic temporal patterns. The timeline view provides a 2D timeline along with a hierarchical display of the data elements organized as a tree. The GIS view displays cases and sightings on a map. Figure 4 illustrates how these three views can be used to explore infectious disease dataset: The top left panel shows the GIS view. The user can select multiple datasets to be shown on the map in a layered manner using the checkboxes. The top right panel corresponds to the timeline view displaying the occurrences of various cases using a Gantt chart-like display. The user can also access case details easily using the tree display located left to the timeline display. Below the timeline view is the periodic view through which the user can identify periodic temporal patterns (e.g., which months have an unusually high number of cases). The bottom portion of the interface allows the user to specify subsets of data to be displayed and analyzed.

We now illustrate how the user can use the visualization module for summarization and analysis purposes. Continuing from the scenario discussed in the previous section, the user clicks on the “visualization details” button to enter the visualization interface. Three background layers are available for the user to select: (1) geographic maps including national and state borders, rivers, and major roads, (2) land information including precipitation, temperature, and vegetation, and (3) demographic data such as population and unemployment rates. In Figure 4, the user has selected “New York State population 2000” and “major rivers” to observe possible correlations between dead crow cases and population/water distribution.

The user first selects to visualize all the cases from 2001-01-01 to 2003-12-31. This example uses a test (not real) data set. The periodic view indicates that in June there has been a surge of dead bird cases. By reducing and moving the time window, the user can further observe case distribution (by month) in each year using the periodic view. For instance, it can be observed that for each calendar year, all cases started in March and reached the climax in June. Using the GIS view, the user can selectively overlay case data on top of major rivers and population maps. It can be observed that many cases have been distributed along the populous areas along Hudson River.

Assume that the user now intends to find out more about case progression in the Long Island area in 2001. The user sets the global data time window to be from 2001-01-01 to 2001-12-31 and then zooms into the Long Island area using the zoom tool provided as part of the GIS view. By reducing the time window down to a two-week period and moving the time slider slowly from the beginning of the year to the end, the user can clearly see the case progression. In this example, the cases started in East

Long Island around March and then gradually moved toward West and upstate. Internally, the STV tool has a scalable and flexible design to support its rich functionality. It has two main components: data preparation and user interface. The data preparation component requires an information source, a map server, and a conversion API to convert data into an XML-based format internal to the STV. The user interface component is implemented using Java WebStart™ technology which enables cross-platform, installation-free, and Web-based execution.

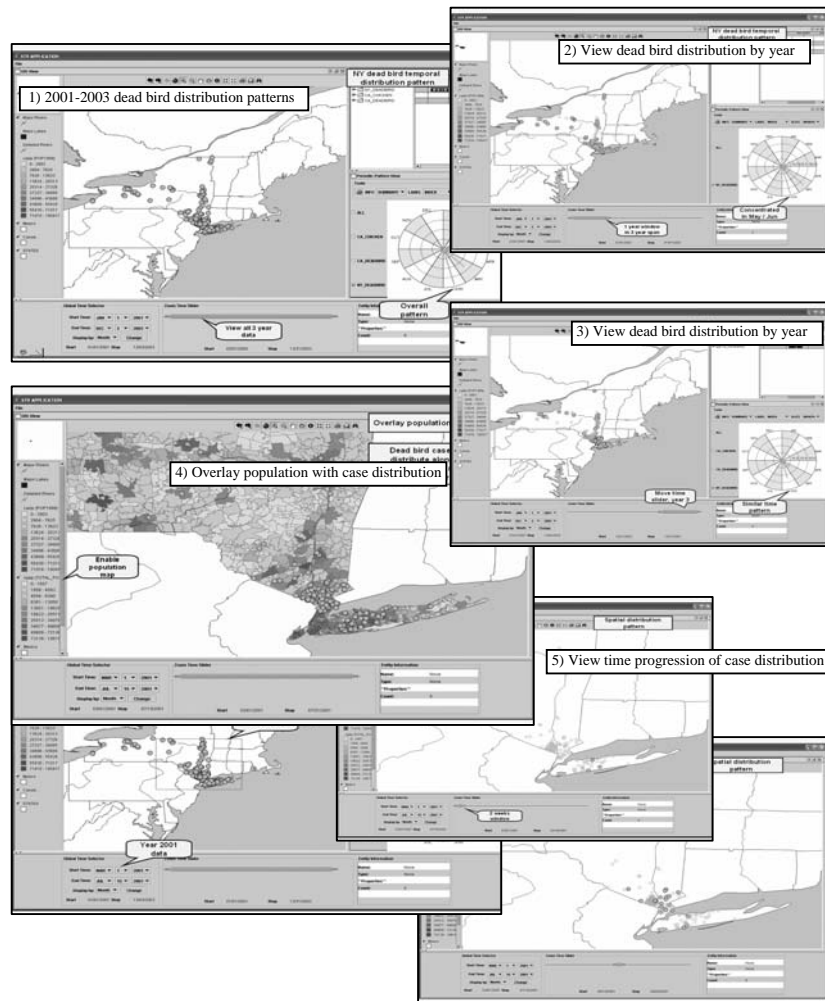


Fig. 4. Using STV to Visualize West Nile virus Data

This component follows the standard Model-View-Control pattern (MVC) and uses a conversion API to convert XML input into an internal data model and an event-

listener bridge to synchronize three views. Because of this flexible design, STV directly supports data feed from various methods including databases, flat files, among others. In WNV-BOT Portal, Microsoft SQL server is used to store the data and the ARCIMS map server is used as a map server. Incorporating new datasets for visualization can be done fairly efficiently. In a recent case, we successfully developed a visualization module for an air pathogen-sensing dataset within a week by one programmer.

3.4.3 Alerting and Hotspot Analysis

Our ongoing effort is focused on two aspects of infectious disease informatics: efficient alerting and hotspot analysis. For the past decades, alert dissemination networks such as Health Alert Network (HAN) systems are being developed and deployed in state public health agencies. However, there is a critical need to create cross-jurisdiction alerts and to automate the dissemination process. We are developing an advanced alerting module as part of WNV-BOT Portal to complement alerting and surveillance systems that already exist in various states. In our current design, alert messages can come from the following three sources: (a) The user can specify personalized triggering conditions (e.g., “notifying me if there are four Botulism cases within the past two days”), (b) The predictive models, as discussed later, will consider datasets from different origins and suggest with high confidence that a disease outbreak is in progress, (c) Public health officials may want to send alerts across organizational and state boundaries. Depending on the nature of the alert messages, some of them may be reviewed by designated personnel. The dissemination module will route the messages through state HAN systems to the public health professionals or send them via registered email address to Portal users when applicable.

In building predictive models for data with spatial and temporal attributes, Hotspot analysis is an approach widely applied in crime analysis and disease informatics applications. Hotspot is a condition indicating some form of clustering in a spatial and temporal distribution. For WNV, localized clusters of dead birds typically identify high risk disease areas. Automatic detection of dead bird clusters using hotspot analysis can help predict disease outbreaks and allocate prevention/control resources effectively. Most of existing disease informatics research uses the spatial scan statistic techniques to perform hotspot analysis. We are currently applying other hotspot analysis techniques (e.g., Risk-Adjusted Nearest Neighbor Hierarchical Clustering) that have been developed and successfully applied in crime analysis to disease informatics [7, 10]. Initial experimental results indicate that these techniques are complementary to the spatial scan techniques in many regards. In a broader context, we are pursuing research in vector borne emerging infection predictive modeling. In particular, we are (a) augmenting existing predictive models by taking additional factors (e.g., weather information, bird migration patterns) into consideration, and (b) tailoring data mining techniques for infectious disease datasets that have prominent temporal features.

4 Summary and Future Research

This paper presents a collaborative effort between IT researchers and public health agencies aimed at developing a scalable information sharing, analysis, and visualization environment in the domain of infectious diseases. The resulting prototype system, WNV-BOT Portal, focuses on two prominent disease types and has successfully demonstrated the technological feasibility of integrating and interoperating infectious disease datasets for multiple diseases and across jurisdictions. Our project has supported exploration of and experimentation with technological infrastructures needed for the full-fledged implementation of a national infectious disease information infrastructure and helped foster information sharing and collaboration among related government agencies at state and federal levels. In addition, we have obtained important insights and hands-on experience with various important policy-related challenges faced by developing a national infrastructure. For example, a nontrivial part of our project activity has been centered around developing data sharing agreements between project partners from different states.

We conclude this paper by discussing the pathway leading to the national infectious disease information infrastructure based on the lessons learned from our WNV-BOT project. Due to the complexity of such an infrastructure from both technical and policy standpoints, we envision that its development path will follow a bottom-up, evolutionary approach. Initially, each individual state will develop its own integrated infectious disease infrastructure for a limited number of diseases. Following successful deployment of such systems, regional nodes linking neighboring states can be established. Such regional nodes will leverage both state sources and data from federal agencies such as CDC, USGS, and USDA. National and international infrastructures will then become a natural extension and integration of these regional nodes, covering most infectious disease types.

References

1. Berndt, D., Hevner, A. and Studnicki, J.: Bioterrorism Surveillance with Real-Time Data Warehousing. *NSF/NIJ Symposium on Intelligence and Security Informatics*, 2003.
2. M.Chang, M. Glynn, and S. Groseclose.: Endemic, notifiable bioterrorism-related diseases, united states, 1992-1999. *Emerging Infectious Diseases*, 9(5):556-564, 2003
3. Chen, H., Zeng, D., Atabakhsh, H., Wyzga, W. and Schroeder, J.: COPLINK: manging law enforcement data and knowledge. *CACM* 46(1), 28-34, 2003.
4. Damianos, L., Ponte, J., Wohlever, S., Reeder, F., Day, D., Wilson, G. and Hirschman, L.: MiTAP for Bio-Security: A Case Study. *AI Magazine*, 23(4), 13-29, 2002.
5. Gotham, I. J., Eidson, M., White, D. J., Wallace, B. J., Chang, H. G., Johnson, G. S., Napoli, J. P., Sottolano, D. L., Birkhead, G. S., Morse, D. L., and Smith, P. F.: West Nile virus: a case study in how NY State Health Information infrastructure

- facilitates preparation and response to disease outbreaks. *J Public Health Manag Pract*, 7(5), 75-86, 2001.
6. Hand, D. J.: *Discrimination and Classification*. Wiley, Chichester, U.K, 1981.
 7. Jain, A. K., Murty, M. N., and Flynn, P. J.: Data clustering: a review. *ACM Computing Surveys*, 31(3), 264-323, 1999.
 8. Kay, B. A., Timperi, R. J., Morse, S. S., Forslund, D., McGowan, J. J. and O'Brien, T.: Innovative Information-Sharing Strategies. *Emerging Infectious Diseases*, 4(3), 1998.
 9. Kargupta, H., Liu, K. and Ryan, J.: Privacy Sensitive Distributed Data Mining from Multi-Party Data. *Proc. of ISI 2003*, 336-342, 2003.
 10. Levine, N.: *CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations (v 2.0)*. Ned Levine & Associates, Houston, TX, and the National Institute of Justice, Washington, DC. May 2002.
 11. Pinner, R. W., Rebmann, C. A., Schuchat, A. and Hughes, J. M.: Disease Surveillance and the Academic, Clinical, and Public Health Communities. *Emerging Infectious Disease*, 9(7), 2003.
 12. Siegrist, D. W.: The Threat of Biological Attack: Why Concern Now? *Emerging Infectious Diseases*, 5(4), 2002.
 13. Zhu, B., Ramsey, M. and Chen, H.: Creating a Large-scale Content-based Air-photo Image Digital Library. *IEEE Transactions on Image Processing, Special Issue on Image and Video Processing for Digital Libraries*, 9(1), 163-167, 2000.