The RODS Open Source Project: Removing a Barrier to Syndromic Surveillance

Jeremy U. Espino MD, Michael M. Wagner MD, PhD, Fu-Chang Tsui PhD, Hoah-Der Su, Robert T. Olszewski PhD, Zhen Liu, MS, Wendy Chapman PhD, Xiaoming Zeng MD PhD, Lili Ma, Zhong Wei Lu, Jagan Dara

Real-time Outbreak and Disease Surveillance Laboratory, University of Pittsburgh, Pittsburgh, PA, USA

Abstract

The goal of the Real-time Outbreak and Disease Surveillance (RODS) Open Source Project is to accelerate deployment of computer-based syndromic surveillance. To this end, the project has released the RODS software under the GNU General Public License and created an organizational structure to catalyze its development. This paper describes the design of the software, requested extensions, and the structure of the development effort.

Keywords: bioterrorism, public health informatics, population surveillance, computer systems, disease outbreaks

Introduction

Syndromic surveillance has been defined as the systematic and ongoing collection, analysis, and interpretation of data that precede diagnosis.[1] Surveillance methods that can detect disease at a pre-diagnostic stage are generally referred to as syndromic because they have the goal of recognition of outbreaks based on the symptoms and signs of infection—and even its effects on human behavior—prior to first contact with the healthcare system.[2, 3]

There are different forms of syndromic surveillance including for example drop-in surveillance, which is the stationing of public health workers in emergency rooms and special clinics during high-profile events such as the Super Bowl to capture data on patients presenting with symptoms potentially indicative of bioterrorism. The major disadvantage of this approach is the cost of round-the-clock staffing for manual data collection.

A less expensive approach—and the one taken in the Real-time Outbreak and Disease Surveillance (RODS) system [4]—is detection based on data collected routinely for other purposes. Examples of such data include absenteeism data, sales of over-the-counter healthcare products, and chief complaints from emergency departments (EDs).[5] This approach is emerging as the dominant approach.

Since 1999, the design objective of the RODS system has been to develop the explicit capability of detecting a large-scale, outdoor, surreptitious, aerosol release of anthrax. A focus of our research and development has been syndromic surveillance from free-text chief complaints routinely collected by triage nurses in EDs and acute care clinics during patient registration. A second focus has been on collection and analysis of sales of over-the-counter healthcare products [6]. These choices have been based primarily on feasibility issues. We are gradually adding requirements and functions to detect other types of threats such as a water contamination.

In the initial years of the project, the main barriers to dissemination and deployment of this technology seemed to be doubts about its efficacy, the cost of the technology, concerns about the cost and effect of false alarms on the practice of public health, legal and administrative issues [7]. Research has addressed concerns about efficacy [8-12], HIPAA and state laws have addressed legal issues [13], and growing experience with these systems suggest that the cost of deployments and false alarms is much lower than expected [14].

At this time, a key barrier to dissemination and deployment—and the one addressed by the RODS Open Source Project discussed in this paper—is the availability of high quality, well-supported, and affordable syndromic surveillance software. The rate of adoption of syndromic technology, although accelerating rapidly, is still not commensurate with the threat posed by bioterrorism, emerging infections, and common outbreaks of disease. This threat warrants even faster deployment, which in turn suggests the need for identifying and removing remaining barriers to adoption.

In December 2002, the University of Pittsburgh, recognizing the national urgency, first released RODS as compiled byte code for non commercial use. This step led to several new deployments. However, the lack of availability of source code was identified as a concern by some state health departments. Recently, the University—responding to these concerns—took the additional step of releasing the source code under the GNU GPL license.
The RODS software for syndromic surveillance has been under continuous development and use since 1999, first in Pennsylvania and then for the 2002 Winter Olympics in Salt Lake City, Utah where it still operates.[14] Additional deployments are underway in Georgia, New Jersey, and Ohio. This software provides a solid starting point for an open source project.

The RODS Open Source Project makes available source code (and compiled modules) for HL7 listeners, parsers, and natural language processing software that state that local health departments can use to automatically collect surveillance data from hospitals in real time. A health department may use a subset of these modules to create a data collection system that populates an existing state integrated surveillance database with real-time hospital data, or it may use all of them (with the RODS database, analytic modules, and user interface) to create an end-to-end syndromic surveillance solution.

The RODS Laboratory, recognizing the additional need for an effective development community to support the code then organized the “The RODS Open Source Project” under the direction of Dr. Espino. With funding from the Commonwealth of Pennsylvania, the RODS Open Source Project is providing resources to stimulate and support open source development of this syndromic surveillance software. In the same way that open source methods created Linux as a mainstream operating system, it is the combination of open source availability and creation of a development community that together can catalyze rapid adoption.

This paper discusses the particular GNU open source license under which the software is distributed, the organization of the development effort, the function and modular architecture of the software, and our suggestions for future development by the community that the project hopes to foster.

**Materials and Methods**

The GNU GPL allows anyone to freely use, copy, and modify RODS. Furthermore, the license permits users to redistribute modified versions of RODS as long as the source code for those versions is redistributed under the GPL. [15]

The online resources in support of the RODS Open Source Project include the existing RODS Laboratory website (http://www.health.pitt.edu/rods) and a project website hosted on Sourceforge (http://openrods.sourceforge.net). The Sourceforge site provides standard software project management tools—a Concurrent Versions System (CVS) server so that developers can collaborate on code modifications, email lists so that developers can communicate with each other, a software bug reporting system, and source code for the stable versions of the system. (Figure 2)

To further facilitate and guide the software development effort, we have identified qualified individuals—listed on the website—to serve as coordinators for the further development of various modules of the RODS system and they have inventoried features that should be a priority for development because users or the coordinators have requested them.

**Results**

**Software**

In this section, we describe the software architecture of the RODS system. The latest version of RODS consists of over 42,000 lines of Java code contributed by a team of eight programmers. RODS is a highly modular system that adheres to the CDC’s National Electronic Disease Surveillance System (NEDSS) standards and any of the components may be incorporated into a foreign surveillance system, or they may all be used to create a native end-to-end RODS system.

RODS follows NEDSS recommendations and is architected as a multi-tiered application, employing Enterprise Java Beans (EJBs) to encapsulate the database. Using EJBs for database encapsulation allows us to make changes to underlying database structure without needing to modify surrounding modules, such as the web-based user interface.

The RODS system has modules that serve six major surveillance functions: data collection, syndromic classification, data warehousing, database encapsulation, outbreak detection, and the user interface. (Figure 3) In the following descriptions of these functional areas, we also identify extensions for future open source development that have been requested by RODS users (Table 1).

The data collection modules consist of: (1) an HL7 listener that accepts and maintains connections from a hospital’s HL7 integration engine, (2) an HL7 parser that extracts patient visit data from HL7 messages, and (3) a text file parser that extracts patient visit data from text files uploaded in batches by non-HL7 capable hospitals. HL7 is the standard format for exchanging health-related data between hospital information systems. The development team would like to see the development of modules capable of handling other data types—beyond the emergency room visit data and retail data that are already supported. (Figure 1)

The syndromic classification module is called CoCo (Complaint Classifier).[16] CoCo uses a simple Bayesian classifier to assign a free-text chief complaint into a syndromic category. These syndrome categories are user-specifiable and the mappings are created automatically through Bayesian analysis of a user-provided training set. Other classification methods such as keywords,[17] ICD-9 codes,[18] and additional NLP modules would be useful development projects.

The software includes a database schema and data warehousing modules that function to store and provide efficient access to syndromic data. The database schema is defined by a set of Oracle data definition statements required to create the data warehouse. The data warehousing module consists of a cache table updater that keeps running counts of the number of visits for each syndrome. The data encapsulation modules have functions that efficiently retrieve this preprocessed data from the
tables. We have received multiple requests from health departments to support a wider range of databases such as Microsoft SQL Server and mySQL. The project developers would like to see RODS successfully deployed and integrated by these health departments on their existing database systems and with other surveillance projects.

**Detection algorithms** provided in the current open source release include the recursively least squared algorithm. We designed this algorithm to detect sudden increases in the number of visits—for example, an increase in the number of respiratory type visits that would accompany a large-scale covert release of Anthrax. Users requested that there be a way for them to import syndromic data into more familiar statistical tools, such as Excel and SAS. To support this request we created a way for them to download data directly from the web site. We recognize that there are some analyses that health departments will want to do automatically on a regular basis. Many statistical packages can import data and run automatically using scripts. With the source code available, other developers can customize RODS to automatically perform these analyses. There are already other algorithms that have been adapted to run off of the RODS database externally as a separate program. For example, another RODS Laboratory project is the development of the Patient and Population-based ANomaly Detection and Assessment or PANDA algorithm and our collaborators at Carnegie Mellon University created What’s Strange About Recent Events (WSARE).[19] Data from the RODS database has been successfully imported into PANDA and WSARE for automatic analysis.

The final functional area are **user interfaces** and include modules that (1) display syndromic data as time-series graphs, and (2) work with a geographic information system (GIS) to view the data spatially and (3) authenticate the user. The coordinators have a list of requested analytic extensions. State health departments have requested LDAP support so that seamless links can be created from existing state surveillance front ends to the syndromic surveillance functions provided by RODS. The project leaders would also like to see the development of animated images that could better show the relationship between spatial and temporal information. Additionally, the developers would like to rewrite the code used to interact with the presently supported GIS system (ArcIMS) using open source libraries.

**Development Community**

There are currently eight active members of the development community arranged around the six functional areas. Three of the developers, Drs. Wagner, Espino and Tsui, are the founding members of the RODS project and created the original RODS prototype (in PERL and C). The current lead developers are:

- Overall - Jeremy Espino
- Data Collection – Rich Tsui
- Datawarehousing – Fu-Chang Tsui
- Syndromic Classification and NLP - Wendy Chapman
- Outbreak Detection - Rich Tsui
- Database Encapsulation - Howard Su
- User Interface - Jeremy Espino

The project has released version 1.6 of the RODS software under the GPL, and the developers expect version 1.7, which is undergoing Beta testing, to be available in October.

**Discussion**

The goal of the RODS Open Source Project is to accelerate the development and deployment of syndromic surveillance by providing free software and catalyzing the formation of a development/consulting community. This community can continuously improve the software and serve as a resource for IT managers who require customized versions of the software.

The first open source release of RODS contains modules that can be used separately in foreign systems or as an end-to-end solution for real-time syndromic surveillance of hospital visits and retail data sales. The code base is a resource for creating, implementing, and testing new syndromic surveillance methods.

The GPL license has the desirable property of encouraging improvement of the software. Modifications to the software are allowed and encouraged. Under the GPL, the modifications can only be redistributed for free; thus, programmers are assured that their “work” will not be turned into a proprietary product and will be further developed and used. Open Source projects can create a community of like-minded individuals that have the vision of creating innovative, well-supported syndromic software. We hope that other syndromic surveillance software developers will join this cause.

The importance of catalyzing the formation of a larger community of scientists, programmers, consultants, and users cannot be overstated. Syndromic surveillance is a highly automated, computer-based activity that involves a great deal of software and system development. The RODS Open Source Project hopes to provide a focus around which a critical mass of developers can coalesce to create and support the needed software. The existence of such a community will strengthen the position of information technology managers and public health officials who wish to deploy syndromic surveillance in their planning deliberations. The existence of the Open Source Project will provide assurances to their supervisors that source code is available, that there is a community of developers and consultants supporting the code and that there are ongoing exemplary projects that use the software. Open source also allows the user to fix and further customize the features in the program without restriction.

We note that the RODS Open Source Project may benefit from the existing base of GPL-compatible source code (i.e., source code that could be combined with the source code of the RODS project without violating software licenses). Freshmeat.net, a database of software popular with open source developers, reports that 69.98% of the 30,001 projects in its database are licensed under the GPL. A majority of the yet-to-be
implemented features in Table 1 are available as existing GPL-compatible software.

**Conclusion**

The RODS Open Source Project is making software modules available that cover the spectrum of processing tasks involved in syndromic surveillance. The software complies with NEDSS standards and therefore should integrate well with surveillance systems already deployed in state health departments. Open sourcing software is expected to accelerate software development and more importantly deployment by lowering cost, increasing reliability, providing IT managers with increased assurance of sustainability (by creating a community of developers and consultants) and ensuring customizability of the software. The RODS Open Source project is providing a leadership nucleus to catalyze the formation of a community of syndromic surveillance developers. As amply proven by the example of Linux, this approach can result in an extremely high quality software product that achieves mainstream acceptance.

![Figure 1 – The Epiplot user interface of the RODS system displays syndromic data as time-series graphs. Here, graphs for the constitutional, gastrointestinal, and respiratory emergency department visits are shown alongside the unit sales of antidiarrheals, electrolytes, and pediatric cough syrup.](image)

![Figure 2: The RODS Open Source Development Project website is located at http://openrods.sourceforge.net.](image)

![Figure 3 - The Real-time Outbreak and Disease Surveillance Software architecture consists of six different types of modules—data collection (red), syndromic classification (blue), data warehousing (green), database encapsulation (orange), outbreak detection (yellow), and user interface (pink).](image)

**Table 1 - Existing features and suggested features for development in the RODS Open Source Project. The features are shaded according to the different module types highlighted in Figure 3.**
<table>
<thead>
<tr>
<th>RODS Feature</th>
<th>Already exists in RODS</th>
<th>Exists as GPL-compatible source code</th>
<th>Needs to be developed or tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>HL7 Listener</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HL7 Parser</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text File Parser</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XML Parser</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple Bayes Syndromic Classifier</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syntactic/Semantic NLP Classifier</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>ICD-9 Classifier</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Multiple Datatype Classifier</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diverse Database Options</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Integrated Data Warehouse Engine</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Integrates with external statistical analysis tools</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>RLS Detection Algorithm</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manual Data Entry Interface</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diverse GIS Software Options</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDAP Interface</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time Series Graphing</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Email Notifier</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Database Encapsulation</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Acknowledgments

The RODS Open Source Project is supported by Pennsylvania Bioinformatics Grant ME-107.

Address for correspondence

Jeremy Espino
RODS Laboratory, University of Pittsburgh
Suite 500 Cellomics Building
500 Technology Drive
Pittsburgh, PA 15219
jue@cbmi.pitt.edu

References