

A Biosurveillance Platform for BioSense Message Analysis Using Integrated Reference Ontologies and Intelligent Agents

Cecil O. Lynch¹, Craig Cunningham¹, Eric Schripsema¹, Tim Morris², Barry Rhodes²

¹ OntoReason, LLC, Salt Lake City, UT

² US Centers for Disease Control and Prevention, Atlanta, GA

ccunningham@ontoreason.com, clynch@ontoreason.com, tom1@CDC.GOV, mbr1@CDC.GOV

Abstract

In this paper we describe a prototype application for real time BioSense message analysis and classification using an intelligent reasoning platform for the distribution of and collaboration between intelligent system components and a common domain knowledgebase provided by the OntoReason Public Health Ontology (OTR Ontology). The purpose of the prototype is to validate the systems' ability to classify diseases and syndromes, correlate and aggregate incoming messages, and provide situational awareness based on the inferred syndrome or disease classification and status. Future work includes planning for the distribution and management of intelligent system components, dynamic integration of distributed reference knowledgebases, and tools for localization of intelligent systems knowledgebases in the field for real-time data collection and analysis.

Introduction

BioSense is a national biosurveillance program initiated by the US Centers for Disease control and Prevention (CDC) as part of the Public Health Information Network for the purpose of early event detection, quantification and spatio-temporal visualization of public health events and risks[1]. Information received is represented in standard Health Level 7 (HL7) format. HL7 is an ANSI standards body which works within the broader health care domain, including Public Health. BioSense currently receives anonymized data in the form of HL7 2.5 messages from more than 600 private and public acute care, Veterans Affairs (VA) and Department of Defense (DOD) hospitals, approximately 1800 VA and DOD ambulatory care centers, the 3 major commercial clinical laboratory systems, and all Poison Control Centers in the US[2]. The system has the capacity to receive up to 72 million messages a day that must be analyzed and posted to analysts within 2 hours, demanding a scalable solution for analysis and routing.

Messages may contain data coded with standard Consolidated Healthcare Informatics (CHI) code systems or may contain free text or local codes in some cases which require conversion to standardized code systems for further analytical processing. The BioSense Messages are of 4 basic

domain types; 1) ADT (Admission, Discharge and Transfer) which captures data about patient presentation and disposition including Chief Complaint, 2) Laboratory Requests and Results, 3) Radiology Orders and Results, and 4) Pharmacy Orders. Messages are analyzed based on content to determine syndromic classification and routing. Messages are correlated to build specific event profiles, refine classification accuracy, provide situational awareness and develop situational assessment.

OTR Ontology

The OTR Ontology was designed using the Protégé Ontology Editor and was modeled to meet the computational reasoning requirements for case classification purposes. Utilizing the Protégé frames structure, the OTR Ontology provides the multi-layered metaclass model necessary for health knowledge representation which is identified by the HL7 message structure and the Case Notification Refined Message Information Model (RMIM).

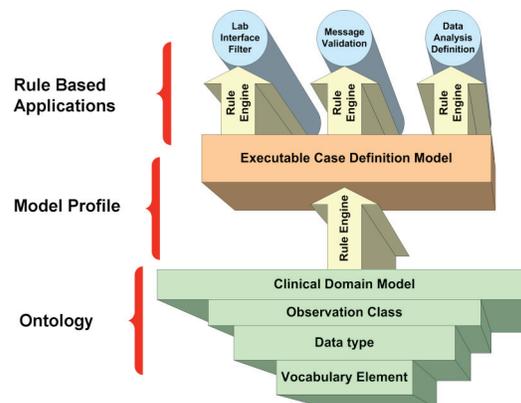


Figure 1 Application Stack

The OTR Ontology contains attributes which cover basic public health case reporting criteria as well as additional knowledge such as incubation period, case frequency, and symptom likelihood for a given disease presentation.

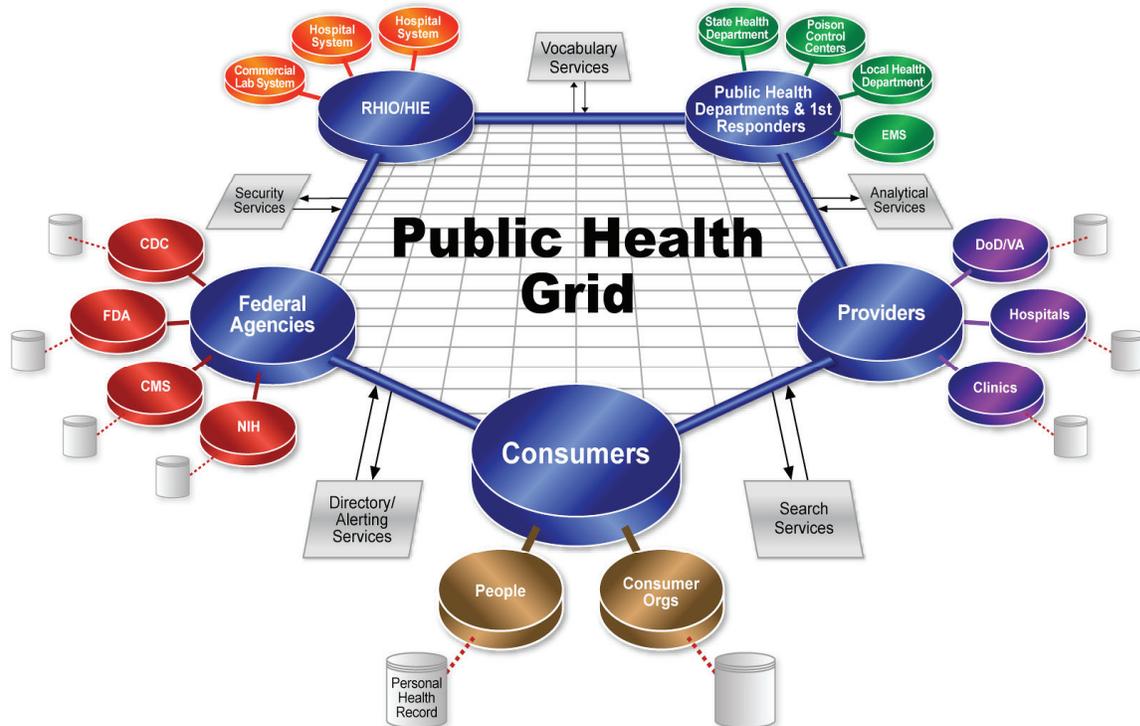


Figure 2 The Public Health Grid

Concept representations in the ontology are structured in a hierarchical relationship native to the code system or in a novel derived hierarchy if required to allow for reasoning at the concept level. The Systematized Nomenclature of Medicine (SNOMED) is used as the core clinical code system and all syndromes and sub-syndromes in International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) (the anticipated healthcare facility diagnosis code system) were mapped to the appropriate synonymous SNOMED ConceptID. Additionally, a lexicon particular to BioSense was also mapped to SNOMED but using a non-synonymous matching slot since the semantics relationship was of an *is_related_type* that was coarsely mapped to a BioSense syndrome.

Where appropriate, HL7 Version 3 attributes in the OTR Ontology were mapped to HL7 2.5 message segments and fields to allow the analysis of the simpler 2.5 messages in the context of the richer semantics of HL7 Version 3 objects.

The ontology allows for the representation of knowledge using the complex coded representation of the standard HL7 complex data types. These complex data types, typically made up of one or more primitive datatypes, are then organized into the metaclass objects that correspond to the HL7 V3 abstract classes and finally organized into an overarching HL7 RMIM metadata model.

This representation then contains the expected artifacts of a patient clinical encounter with a provider given a specific

disease entity. This constellation of clinical acts and entities related to each disease is thus a superset of a CDC

Case Definition for that specific disease, allowing a matching set of necessary and sufficient conditions to be evaluated for confirmation or exclusion as a case.

Platform for Distributed Intelligent Systems

The OntoReason Intelligence and Analytics Workbench is deployed as a platform supporting a series of collaborative expert systems that implement various artificial intelligent constructs each using the ontological content as facts within their knowledgebase. The platform allows for the configuration of various reasoning components to solve complex analysis problems while operating on a consistent, ontology driven knowledge model. The ontological concepts are used to provide traditional pattern matching and differential diagnosis. The expert system implementations leverage correlation factors derived from the ontology in conjunction with domain expertise represented in the knowledgebase (rules and facts). For example, our case classification reasoner uses specific ontology case definition contents structured in JESS facts combined with algorithms that adjust the confidence weighting based on factors such as geographic and seasonal disease occurrence and frequency estimates to instantiate a disease condition, sub-syndrome or syndrome classification. JESS is a java based rule engine which implements the RETE algorithm.

In addition, the ontological model provides the ability to implement pattern matching from inexact concepts based upon the ontological representation that is inherent to the

concept hierarchy structure. The model supports generalized concept matching at the more primitive levels of the hierarchy and specialized matching at the more sophisticated concept levels in the hierarchy.

The evolution of the prototype would involve the distribution of the message filtering and classification within the Public Health Grid, depicted in Figure 2.

Currently providers generate messages based upon the BioSense reporting criteria. These messages are forwarded via a secure routing system (PHIN-MS) to the BioSense message broker. If the message classification reasoners can be implemented at the provider nodes, message traffic on the BioSense network can be greatly reduced by eliminating uninteresting report information that cannot be filtered today.

The knowledgebase for the classification reasoner is very dynamic and will require distribution to potentially thousands of nodes. This distribution will be managed centrally to ensure network consistency and standardization. Management of the dynamic nature of this knowledgebase requires addressing three major factors. The first factor that must be addressed is the update schedule for standard code systems utilized by the knowledge base. These updates will require quarterly modifications to the ontology terminology set. The second factor is the need for frequent updates to the knowledgebase due to the dynamic nature of the public health domain. Emerging disease conditions such as West Nile Virus, SARS, and Avian Influenza drive the need to expand the reference ontology to include these conditions and continually update the domain knowledge, as new clinical information is available. Finally, the network structure will change as well. Additional data sources, regional institutions, and public health interests will be added to the domain, requiring the classification system to be capable of responding to these changes.

Adding to the complexity of predictable changes, additional dynamic knowledge distribution requirements will occur once the classification reasoners are deployed into the network. Based upon the context of regional bio-surveillance activities, knowledge based on event detection and situational awareness will need to be integrated into the distributed classification nodes. This would allow the tuning of operational parameters sensitive to particular condition and syndrome classifications.

Workbench

The OntoReason Intelligence and Analytics Workbench (IAW) product was utilized to provide a functional environment in which to implement the core logic of the application.

The IAW is an application framework that relies on a controlled configuration, and integration structure that allows

the integration of application specific components within a managed environment.

The implementation of the prototype designed within the IAW is described in Figure 3.

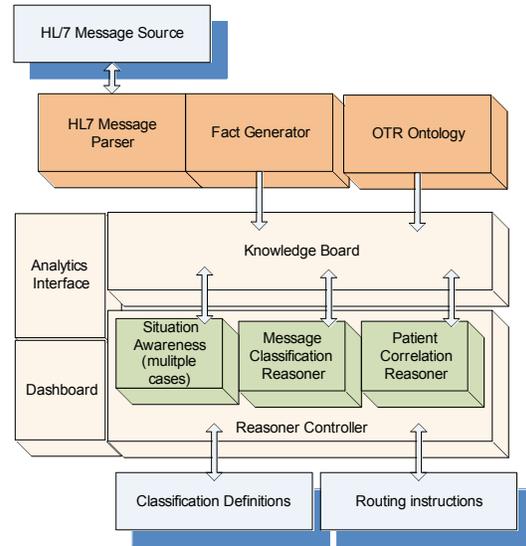


Figure 3 The IAW Workbench™

The workbench allows for the configuration and coordination of a series of intelligent systems each with specific purposes. These intelligent systems called reasoners are configured to operate as distributed applications sharing a common instance of the knowledge board.

The Reasoner controller can operate in two modes: (1) A Data Driven Process Coordinator and (2) A Hypothesis Driven – Process Director. In the Data Driven Process Coordinator mode, the controller enables the individual reasoners to consume and publish knowledge. The controller coordinates collaboration between independent reasoners. The coordination role typically executes in a near real time environment, where the controller is responsible for monitoring current contextual situations and adjusting reasoner priorities. The reasoners react to data inputs, system requests, and execute their individual responsibilities within the current contextual setting provided by the controller.

In the Hypothesis Driven Process Director mode, the controller directs a goal driven application of the reasoner resources with the capacity to configure and execute specific resources in the performance of proving and disproving a given hypothesis. This mode enables the generation and execution of specific analysis based upon available information resources.

As Figure 3 shows, there are various components within the context of the IAW, which have been combined with problem domain specific components to solve the BioSense problem.

One important aspect is the representation of the core knowledge. The OTR Ontology represents the core knowledge base from which the basic inferences can be made. As part of the integration, the ontology is processed to represent facts for the JESS rule engine.

For each given message a process takes place which extracts the core information from the message and translates it into a JESS fact representation. All of this information is presented to the reasoning framework via the Knowledge Board. The Knowledge Board represents a work area for processable data and acts as a two-way access point for reasoners to read and post knowledge to be shared with other aspects of the application.

Once the knowledge has been posted to the knowledge board the reasoners are activated based on their knowledge requirements. The reasoner controller maintains the state of the reasoners, coordinating the processing of individual messages and the correlation of processed messages and reasoning results. As a means for reviewing the processes that the reasoner is following and as a means for visualizing the reasoner results, the IAW provides the Dashboard and Analytics interfaces. These components provide a means for viewing the activity (process counters, activity log) and results (information) on the knowledge board. The dashboard provides a basic tool for understanding the activity, where the analytics interfaces provides a means for viewing knowledge, and a tool for providing custom queries and visualizations.

To provide the basic functionality, the reasoner components define the rule set that is necessary to calculate the results.

Intelligent “Reasoner” Systems

The initial application provides for the implementation of three core expert system components that operate independently. The first component receives HL7 V2.5 messages, from a variety of remote sources. These messages are processed, validated, and then classified. Individual messages tagged with their classifications are placed on the knowledge board for consumption and further refinement. The message classification expert system uses the Public Health Ontology as its core knowledge source, which describes conditions, syndromes, and sub syndromes based upon the expected observations for each. Observations include clinical findings, laboratory findings, radiology findings, pharmacy orders, risk factors and other epidemiologic factors, and demographics. The message classification pattern provides for the generation of multiple hypotheses and uses evidential based reasoning to weight the hypothesis.

The classification reasoner rule patterns have been developed to compare the reference ontology knowledge base to individual BioSense messages using generic pattern match-

ing techniques. Individual observations that are matched for a given classification are represented as evidence supporting or refuting a given classification. The system compiles evidence for a given classification based upon the message observations and then adjusts the belief in a particular classification based upon the strength or weakness of the evidence supporting the classification. The potential message classifications are measured based upon relative strength of their evidential support.

The second intelligent application component correlates messages to further derive classification confidence based upon the combination of evidence and belief. The message correlation reasoner provides consolidation of messages into specific case event profiles; each case event is classified and given a status. Using CDC defined case classifications, individual cases are determined to be suspect, probable, or confirmed.

The third intelligent system component correlates case events based upon case/syndrome classification and spatial-temporal parameters. This reasoner provides specific situation awareness and assessment. Disease specific models defined to include annual instance counts, frequency based upon time of year, reported geographic occurrence, transmission modes and incubation periods are tested with this component.

Additional knowledge will be incorporated into the reasoners based upon context of the biosurveillance mission. As specific alerts are triggered, contextual parameters will adjust the classification parameters or skew the control of the independent reasoners based upon individual alert levels. Regional nodes of the classification reasoner will be dynamically updated based upon the context of events being monitored throughout the system. Critical events must be properly defined based on historical evidence, clinical weighting factors and epidemiologic expertise. Condition alerts are based on higher-level confidence factors. This prevents the creation of an Inhalation Anthrax condition alert solely because of a report of common upper respiratory symptoms. Once a suspect case of Inhalation Anthrax has been identified within the public health network, however, the knowledge of that case would need to heighten the awareness of the distributed classification systems. This may include changing the importance of upper respiratory symptom reports. This sensitivity will be achieved through input of additional data feeds followed by the distribution of additional knowledge based facts, which may be global across the entire network or within regional components. The basic expert system functions will allow for the assertion and retraction of these knowledge base components.

It is anticipated that these intelligent system components will be distributed to the data source nodes for filtering, routing and reporting system feedback. The distribution of these components will allow for better filtering and alert-

ing, but will require the infrastructure to enable dissemination of individual knowledge bases as they evolve. For instance, the emergence of a new disease like SARS or the update of new tests used to diagnose an existing condition requires the dynamic update of these distributed knowledge base facts.

Each expert system is developed on a core ontological knowledgebase for Public Health.

Visualization Platform

We have built an integrated message visualization platform that constructs a directed acyclic graph of the message observations linked to the message segments for viewing by the Biointelligence Center BioSense Monitor. This allows rapid visual analysis and comparison and is updated as additional messages from the same patient are received, parsed and analyzed. A confidence threshold slider control allows the analyst to dynamically reconfigure the graph based on the level of confidence selected (see Figure 4).

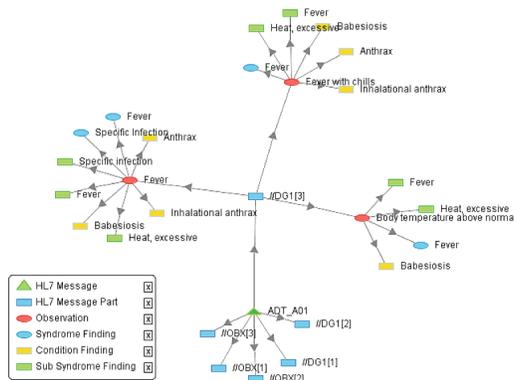


Figure 4 Message Classification Visualization

Prototype Discussion

Using the OntoReason Public Health Ontology as the core reference model, we have produced a demonstration system that evaluates incoming HL7 2.5 BioSense messages, providing parsing, semantic validation, syndromic classification, and case definition in real time as a service add-on to a generic HL7 interface. The ontology classifications can also provide a means to generate public health application value sets linked in context to a particular disease.

The analytical framework is a multithreaded application and scales to high message volumes for runtime analysis as an HL7 interface listener and utilizes the HL7 standard terminology Common Terminology Services (CTS) API for vocabulary maintenance requirements.

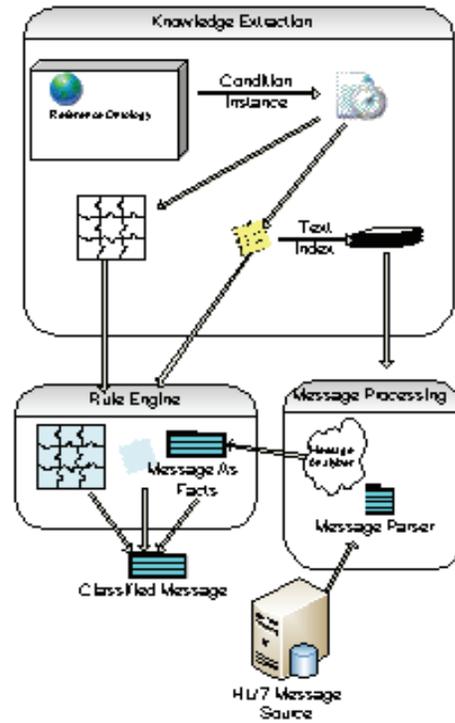


Figure 5 Ontology Driven Knowledge Platform

Buckeridge et al, were one of the first to demonstrate the use of an ontological framework for biosurveillance as part of the BioStorm project[3], defining the benefits of abstracting knowledge from applications as a means of lowering system costs and improving system flexibility. Additional ontological based biosurveillance work has been done by others, notably Mirhaji et al on chief complaint data analysis[4] and modeling of Logical Observation Identifiers Names and Codes (LOINC) for laboratory surveillance[5]. Each of these approaches developed an ad hoc model for building the ontology that suited the purposes of the specific system implementations described.

Our approach differs from the previous work through the instantiation of a standard object model to generalize application functionality and in using both the model and content as a basis for reasoning. This HL7 based ontology approach has the major advantages of standardizing content exchange in a clinical messaging environment to enable standards based Model Driven Architecture software solutions that can be distributed in a Services Oriented Architecture environment as objects and persisted in a more efficient object database therefore maximizing throughput and flexibility for domain processes such as message fragment generation for visualization in context. The object structure and its complex meta data also allows for the generation of documentation associated with an object through supporting literature references to the ontology content embedded as part of the object. This pro-

vides an “audit trail” for the knowledge and gives the end user confidence in the validity of the content.

The major limitations of this approach include the complexity of the model, which requires significant domain expertise both in the ontological modeling and the messaging environment, which increases the time required to instantiate the ontology.

To solve this problem requires the development of simple interface tools for domain experts to maintain and update ontological content and modify or instantiate rules while hiding the complexity of both the model and the rules engine. The development of these tools is in the scope of near term future development.

References

1. Loonsk, J.W., *BioSense--a national initiative for early detection and quantification of public health emergencies*. MMWR Morb Mortal Wkly Rep, 2004. **53 Suppl**: p. 53-5.
2. Steele, L. *BioSense: Integrating Local, Regional, Nationwide Biosurveillance Capabilities*. in *ISDS Annual Conference*. 2006. Baltimore, Maryland.
3. Buckeridge, D.L., et al., *Knowledge-based bioterrorism surveillance*. Proc AMIA Symp, 2002: p. 76-80.
4. Mirhaji, P., et al., *Semantic approach for text understanding of chief complaints data*. AMIA Annu Symp Proc, 2006: p. 1033.
5. Srinivasan, A., et al., *Semantic Web Representation of LOINC: an Ontological Perspective*. AMIA Annu Symp Proc, 2006: p. 1107.