Reverse Site Visit
Summary Report
National Library of Medicine

CONTRACT#:	N01-LM-3-3506
PROJECT TITLE:	Applications of Advanced Network Infrastructure in Health and Disaster Management: *Project Sentinel Collaboratory*
CONTRACTOR:	Georgetown University 37th and O Streets, NW Washington, DC 20057
PRINCIPAL INVESTIGATOR:	Seong K. Mun, PhD
DATE:	August 28, 2007

Table of Contents

1 Introduction

When anthrax struck the nation's capital in October 2001, Washington Hospital Center (WHC), the largest emergency department in the capital city, became the epicenter of medical response to a major sustained bio-terrorism that shocked the nation. This first hand experience of responding to a sustained biological attack revealed the nation's alarming lack of integrated information management infrastructure to (1) assess the nature and the scale of the threat, (2) seek and coordinate external assistance, and (3) share necessary information in a timely manner with concerned healthcare organizations and government. This research project deals with the information technology aspect of the deficiencies identified though our first-hand experience. The project is named the *Project Sentinel Collaboratory* as we aimed to serve as an ever-present watchful guardian of patient health and public safety. It is a collaborative project between Georgetown University Medical Center (Principal Investigator: Seong K. Mun, PhD) and Washington Hospital Center (Principal Investigators: Craig Feied, MD and Mark Smith, MD)

The WHC developed a state-of-the-art medical information system called Azyxxi which has been in use since 1996. Using this system as the cornerstone of our research, we built an information network test bed that links the emergency departments of two major medical centers and the public health department in the Washington DC metropolitan area. Internet2 class networking, wireless, biometrics and the Global Information Systems (GIS) were implemented to investigate (1) the scalability of key applications, (2) the ability to share essential information across multiple organizations and (3) the ability to access rich sets of clinical, public health and surveillance data to gain new knowledge for improving responses to future threats. We integrated new technologies such as geographic information system, biometrics technology and visualization in our test bed and applied emerging technologies from the "Core Middleware" and "Grid Services" concepts to investigate how well these new tools work supporting our functional requirement of scalability, and the ability to communicate accurate, relevant and complete information and knowledge in a secure manner across the virtual organization.

The overall goal of Project Sentinel has been to build and deploy a data-centric collaboratory that supports real-time sharing of clinical data from all patients at two hospitals. The users of the system include public health authorities at the Washington, DC Department of Health (DC DOH), researchers at two institutions, and clinicians from the two hospitals. Patient-identified data is available to authorized clinicians who are involved with patient care and who have been authenticated. De-identified data is available to researchers while pseudo de-identified data is made available to public health authorities, who are able to request access to patient identity when they normally have access to patient-identified data by traditional means.

The collaboratory is a network-dependent set of applications making use of middleware from the National Science Foundation Middleware Initiative (NMI), including the Shibboleth package to support biometric identification and authorization in a federated environment, and the NMI releases for Grid Services for better clinical data sharing and interoperability support. A GIS is used for data visualization and analysis. Internet2 class networking has been deployed to support advanced network functionality. The intent of the public health component of this project is to understand how access to the collaboratory data can provide the ability to track and manage evolving large-scale events and automate surveillance for sentinel events that may herald a bioterrorist attack or an emerging illness.

Project Sentinel has been divided into six sub-projects which are as follows:

- **Project 1**: Instantiating the Collaboratory and Linking WHC-GUH-DCDOH
- **Project 2**: Public Health Functions and Surveillance with GIS
- **Project 3:** Data Visualization beyond GIS

- **Project 4**: Federated Organizations and Biometric Authentication and Authorization using the Shibboleth middleware components
- **Project 5:** Grid Service for Data Sharing and Interoperability dealing with Grid and Web Services and using the Globus toolkit and related tools
- **Project 6**: Knowledge Dissemination and Sustainability

2 Project 1: Instantiating the Collaboratory and Linking WHC-GUH-DCDOH

The intent of Project 1 was to establish the technical software and hardware collaborative foundation upon which Project Sentinel research could be accomplished and is comprised of three components: (1) provide the required software infrastructure (2) establish robust and secure network connectivity between the collaboratory participants (3) develop tools to enhance collaboration between participants

2.1 Software Infrastructure

Azyxxi is a modularly built data scaffolding system that exists as a superstructure around hospital legacy systems. It interfaces with legacy systems and accepts all data from them as well as enabling direct data input. Azyxxi treats all data equally ("data are data"), and does not distinguish between clinical, demographic, financial, image, or other types of data. It enables the look-up of specific data items, answers particular questions about individual data items, discovers patterns in the data, and exposes problems in ongoing data-generating processes in real time. Azyxxi has been in operation at the WHC since 1996 and is also installed at five sister hospitals.

The bio-surveillance aspects of Project Sentinel required access to the raw clinical data, already available from the WHC Azyxxi system (and five sister hospitals). To obtain the data from GUMC, cooperative agreements with GUMC were secured, and Azyxxi servers and software for GUMC data were installed. Interfaces and parsing scripts for GUMC data sources were developed. Nine interface data streams from GUMC were made live and real-time available across the collaboratory to include: patient demographics, laboratory results, pathology results, dictations, images (X-ray, CT, ultrasound, nuclear medicine, MRI data), and blood bank data.

At the core of Azyxxi is the concept of data centric models for data storage. Incoming data streams are broken down to the smallest "data atom." All structure and formatting is converted to meta-data associated with that respective data atom. The data centric approach of Azyxxi allowed the Project Sentinel collaboratory to unite data across seven hospitals for use by the collaboratory participants for bio-surveillance research.

The Azyxxi electronic medical record (EMR) was configured with base-views for each class of user across Project Sentinel. Users across the collaboratory were trained for how to customize views and develop parameters for improving research efforts.

The end-user tool configure efforts helped spur Azyxxi's use across GUMC thus adding additional data for bio-surveillance research from a variety of hospital users and locations. From a base of zero users at GUMC at the beginning of Project Sentinel, by March of 2006, there were 3081 user accounts and 548 different machines using Azyxxi at GUMC. Usage covers the following units: Endoscopy, Limb Center, Surgery Center (Same Day Surgery), Conference Rooms and Auditoria, Neurology, Cardiology, Lombardi Cancer Center, Surgical and Medical Intensive Care Units and Neurosurgical Intensive Care Unit (C61, C62, C63), Pediatric Intensive and Critical Care Units (C52, C53), Critical Care and Step-Down Units (C41, C42, C43), Medical Units (6 Main, 2 North, 4 East, 5 West), Neonatal Intensive Care

Unit, Medical and Oncology Units (2, 3, 4, 5 Bles Building) Emergency Department, Radiology, Laboratory, Physician Billing, Physician Lounges, and Registration.

2.2 Establishing the network

Access to the Azyxxi software and data platform for the collaboratory was established by linking researchers at the Imaging and Information Systems (ISIS) Center with patient data at WHC, GUMC, five other hospitals part of MedStar hospital system, and the DC DOH.

Our previous experience with 100 Mbps connections has shown that frame rates for certain video conferencing applications drop to 15-20 frames per second due to bandwidth demands. As a Collaboratory could be expected to have many simultaneous video conferencing streams, gigabit Ethernet fiber connections were established between WHC and GUMC with standard user-level switches, endpoint routers, and fiber optics. The fiber allowed high-speed, high-bandwidth network and VPN connections to be made across the Collaboratory between WHC, GUMC, and ISIS.

Site-to-site virtual private network (VPN) connections were established over the dark fiber network. In accordance with the National Institute of Standards and Technology (NIST) recommendations, encryption has been established using AES-256. NIST selected AES as a Federal Information Processing Standard in November 2001 for protecting sensitive electronic data. In June 2003, the National Security Agency announced that AES-256 was sufficient for protecting classified information up to the TOP SECRET level, which is the highest security level. This high level of security standard will help ensure that the Collaboratory will be a viable platform for research when collaborating with other government agencies. Top Secret security capacity for the network could be particularly valuable for future Collaboratory work given the national capital region's unique risk of bioterrorism attacks.
The hashing algorithm applied is secure hash algorithm (SHA). The client to site VPN was established as an IP Security Protocol (IPSec) tunnel for use for access from Internet2. Encryption was established using the triple data encryption standard (DES) and the hashing algorithm applies is SHA. The Collaboratory is protected at Internet connection points via firewalls. There is direct, encrypted access between Georgetown University and Washington Hospital Center.

Because Internet traffic can surge unreliably during disasters (which are the very times when connectivity between the DC Department of Health and the collaboratory) a direct T1 line was established to the DC Department of Health from MedStar's data center for WHC (the data hub for the Collaboratory).

2.3 Development of Collaboration Tools

The Azyxxi team at WHC developed several special projects for users at GUMC, including an interactive, electronic white board for the emergency department and quality improvement views for monitoring and analyzing hospital-acquired conditions such as urinary tract infections and deep vein thromboses. Azyxxi was modified to allow multi-institutional viewing of patients so that patients at both institutions could be seen and explored in a single view. Vast amounts of current and historical data was provided to project participants. User baseviews were created for each class of users. Through Azyxxi's core data exploration capabilities, users were able to see data from individual patients, examine cohorts of patients across hospitals, investigate trends over time, aggregate data for reporting or statistical access, and extract data.

The ProMonitor system (a.k.a. MonitorMan)is a standalone system that queries existing databases, executes scripts using the acquired data, and produces results displays that include a monitoring console and graphical displays, as well as alerts. The ProMonitor system was also instantiated at GUMC and project participants were given access to the system. Nearly all ASCII data at both institutions (GUMC and WHC) were made available in a replicated form for access by ProMonitor. Users were given role-specific access to ProMonitor's data. ProMonitor allows data access and exploration in both discrete and aggregated form across institutions.

In addition, the WHC team:

- Explored de-identification methods and pseudo-de-identification methods. Those methods included
  - Parsing pseudonyms and alternate patient identifiers directly into tables at the time of parsing.
  - Removing certain patient identifier fields from the available columns in a user's Azyxxi baseview.
  - Hiding the patient identifier data entirely from the title bar of viewer components.
  - Limiting access to patient data for some to ProMonitor where only aggregated data could be accessed.
  - Creating a "Pseudonymized" mode in Azyxxi in which the parsed pseudo-identifiers would be displayed in the viewer component title bars rather than the true identifiers.
- Developed multimedia teleconferencing and messaging:
  - Successful and robust teleconferencing, chat, whiteboarding, and application sharing with real-world utility was easily achieved when commonly and freely available external applications were deployed and used for this purpose. "Embedding" was achieved through the "Shortcut" function of Azyxxi that allows users to launch any external program from within Azyxxi .
  - Ghostly teleconferencing was achieved using FaceTop, an open-source package that supports face-to-face communications within a shared full screen experience. Semi-transparent images of dual participants are placed over the screen. In this manner, content can be seen through these "ghosted" images. The advantage of this model is that video conferencing participants can use natural gestures such as pointing on the screen which could facilitate discussion about medical cases or complex data.
- Built an infrastructure for creating user scripted alerts and alarms. System-generated alerts and alarms can deliver messages by page, email, automated printing at a specified printer, broadcast message across the Azyxxi interface, automated file FTP, generation of an outgoing message, an automated fax, or a call to a web service. User-scripted alerts and alarms can deliver messages by page or email.
- Fully embedded MapPoint into the Azyxxi client and integrated it into the data environment so that any cohort of patients can be instantly geo-mapped. Azyxxi's embedded report structure allows any user to generate surveillance graphs. A browser for the ProMonitor surveillance graphing system was also directly embedded into Azyxxi.
- Provided integrated access to online resources such as AHRQ guidelines and PubMed

2.4 Lessons Learned from Project 1

- Data atomic approach of Azyxxi allowed rapid instantiation of the data feed infrastructure at GUMC
- Primary threats and challenges were not technological. Rather, they involved unanticipated cultural shifts, new legislation and changing processes across a complex environment. Managing separate institution cultures proved very challenging.
- Mapping across terminologies is greatly facilitated by the use of the UMLS Metathesaurus. Therefore, tagging data with a Concept Unique Identifiers (CUI) is a highly efficient surrogate for

tagging data with dozens of different terminology codes. Natural language tools are useful for mapping short free text strings to CUIs, but the MetaMap Transfer tool (MMTx) is severely limited in its utility for longer strings as it does not include the concept of negation

- For de-identification/pseudoanonymization:
  - Full de-identification through the removal of all 18 HIPAA identifiers was both exceptionally difficult for a dataset of this breadth and scope as well as highly detrimental to the real-world utility of the dataset. Limited de-identification was far more readily achievable and yielded truly useful datasets
  - Automated de-identification of a comprehensive data set requires natural language processing research for embedded patient identifiers
- Peer-peer teleconferencing is not feasible in a strict firewall environment (hospital) – a solution using a central server intermediary is optimal in this environment
- For ghostly teleconferencing:
  - New users need a small amount of practice to reliably point areas on the screen. Automatic adjustment to location or GestureCentering would be a valuable feature
  - Stationary background objects obscure content – automated background elimination feature would be useful
- Some of the online resources to which access was provided proved to be “moving targets”. Even government-sponsored web service resources utilizing an SGML-based standard format (such as PubMed utilities) changed during the course of the project in a non-backward compatible way. Reliance on external resources cannot be “set it and forget it”. Rather, we learned that success depends with minimal ongoing effort depends upon setting up automated monitoring and periodic attention in order to succeed.

3 Project 2: Public Health Functions and Surveillance with GIS

Rapid access to hospital data (and especially to emergency department data) is useful for any city, county, or state health department that wishes to conduct effective infectious disease surveillance and to do effective planning for managing medical consequences. Existing efforts to transmit such data electronically have focused on data transmission during special periods of high concern and on the batch transmission of small amounts of subselected data. The Sentinel collaboratory is different, in that it uses a real world data system to deliver "always-on" real-time access to the clinical data in the context of a collaborative working environment. The concept is that this capability could then enable automated surveillance of sentinel events within the test bed and thus, facilitate the tracking and management of evolving large-scale events.

3.1 Surveillance efforts using Sentinel data

Although the Sentinel data were readily available electronically, the status of syndromic surveillance in combination with the limitations of DC DOH made it difficult to conduct surveillance. Within the United States, the status of syndomic surveillance reveals a nascent science with poorly defined and validated data requirements and standards (Table 1). The progression of annual Center for Disease Control (CDC)-sponsored Syndromic Surveillance Conferences from 2002-4 (www.syndromic.org) have revealed a significant concern regarding the validity of using chief complaints and ICD-9 codes to passively detect human disease events. In attempting to develop geotemporal baselines for infectious disease, we have noted these issues as well. We were unable to confirm the reported sensitivity and specificity in using chief complaints, as reported by ESSENCE and RODS, in the Azyxxi system used by the Sentinel

Collaboratory. This result reflects a need for more basic statistical and epidemiological research in defining the requirements for syndromic surveillance.

Potential Benefits
▪ Helps to define scope of outbreak, providing reassurance that a large scale outbreak has not occurred ▪ May provide more rapid characterization of the population potentially at risk following a bioterrorist event ▪ May provide information on noninfectious public health problems ▪ May provide more efficient targeting of prevention/control activities ▪ May provide an increase in knowledge concerning naturally occurring infectious diseases ▪ May help strengthen the public health infrastructure at the local and state levels and its ability to deal with other diseases of public health significance ▪ May help define the at-risk population in need of preventive measures (cost effectiveness after an event) ▪ May provide better public health data systems
Potential Limitations

▪ Low sensitivity e.g., the limited ability to detect a few number of events (Anthrax 2001)
▪ Poor timeliness of event detection
▪ High false positives
▪ Low positive predictive value (low number of events)
▪ Low utility of alternative data sources
▪ Difficulty in identifying individual patients and diagnostic confirmation
▪ High cost of surveillance systems
▪ HIPAA regulations - confidentiality of health records may limit sharing of health data
▪ Lack of reliable specific detection algorithms to account for:
    ▪ Fast spreading agents and slow spreading agents
    ▪ Common and uncommon events
    ▪ Varying geographic patterns
    ▪ Seasonality

Table 1 – Potential benefits and limitation of syndromic surveillance

To partially address these concerns, we examined other, more objective test markers for both passive and active surveillance. Specifically, a statistical retrospective analysis of objective markers for pneumonia was performed to define the utility of using objective parameters for pneumonia (i.e., body temperature and oxygen saturation) that are simple and cost-effective to use. The current model suggests a combination of body temperature of 100.4 or higher and an oxygen saturation of 97% or lower predicts an eventual diagnosis of pneumonia (ICD-9 codes 480-8) 78% of the time. The utility of using objective markers offers a more specific approach to detecting high threat agents that are aerosol or large-droplet nuclei-transmitted. Such agents have tremendous implications for a healthcare facility. Our model offers a statistically defined, simple method to screen for patients that may pose a serious health risk, either at an international Port of Entry or at a healthcare facility. Efficient screening for these patients can enable more targeted medical intervention, as well as protect critical healthcare provider staff and other patients from exposure. This has implications not only for emerging infectious diseases introduced via air transportation, but also for bioterrorism. We propose consideration of objective markers as a component to current syndromic surveillance methodology. Use of the Azyxxi system greatly facilitated our research, as it enabled efficient, cost-effective investigative evaluation of over 100,000 patients for these marker analyses.

We have found that baselining infectious disease should follow several additional principles that are not currently considered by the CDC in a broadly-applicable approach. Because of the need for rapid event detection for outbreaks of concern, current syndromic surveillance requirements fall within the most imprecise area of diagnostic information (i.e., chief complaints) due to the need for rapid identification of a problem. A major requirement of an IT system to enable full data visualization is more efficient statistical evaluation of the full spectrum of disease markers. It was outside the scope of work for the Sentinel Collaboratory to do this for all disease given the magnitude of markers available for such analysis (>8,000 data elements per patient).

From the public health agency perspective, DC DOH found itself in a precarious position. Current syndromic surveillance produces a tremendous number of false positive alerts that required hiring additional personnel to "ground truth" and rule out outbreaks. The vast majority of the alerts coming out of Sentinel were false alarms, and the effort as a whole became an unwelcome drain on their infrastructure and resources.

As part of this project, we also analyzed the utility of SatScan. We have evaluated SatScan's geotemporal anomaly detection algorithm and found it to produce confusing outputs, from an operational perspective. DC DOH agreed with this assessment, as clusters revealed by the algorithm vary depending on which model is selected and what temporal and spatial window is selected. This is a problem because there are no standards for how to use this product properly. For this reason, DC DOH indicated a reluctance to use this tool. Again, this poses a problem for approach validation both in the research and operational setting.

3.1.1 Project Argus – Born out of Project Sentinel

The surveillance research done under Project Sentinel made it apparent that even with access to real-time clinical data, the current limitations of syndromic surveillance severely compromise the ability to detect a biological threat in a timely manner, leaving little to no time to manage any catastrophic bioevent. Instead, the biodefense team concluded that a system of Indications and Warnings (I&Ws) would be far more effective in alerting responders of an imminent bioevent weeks to months in advance; thus, priming the national response infrastructure by alerting agencies of an evolving threat that could ultimately be catastrophic. Retrospective analyses of major bioevents have demonstrated the presence of multiple I&Ws were present in multiple data sources weeks to months in advance, which were not recognized and utilized properly by the national response community. To meet present and future biothreats, an integrative strategy for information discovery, exploitation, and effective proactive use by the response community is critical. I&Ws provide a key component for integration within the U.S. bio-surveillance portfolio, enabling earlier warning potential. Project Argus is funded separately by the Department of Defense and Department of Homeland Security. It is the first attempt to integrate I&Ws in an effort to detect catastrophic bioevents on an international scale and Project Sentinel is considered its precursor.

3.1.2 Emergence of bio-surveillance IT standards

When Project Sentinel began, there existed no standard IT requirements for interoperability for information exchange in the Bio-surveillance domain. However, since the launch of the National Health Information Network (NHIN), interoperability specifications have emerged. The American Health Information Community (AHIC) is a federal advisory body chartered to make recommendations on how to accelerate the development and adoption of health information technology. It was formed to help advance efforts to achieve President Bush's goal for most Americans to have access to secure electronic health records by 2014. Since its formation, the AHIC has identified four initial areas with potential for early breakthroughs in the advancement of standards that will lead to interoperability. Bio-surveillance is one the breakthrough areas. In 2006, AHIC produced a bio-surveillance use case and subsequently, the Health Care Information Technology Standards Panel (HITSP) has created the Bio-surveillance Interoperability Specification which defines standards that promote the exchange of bio-surveillance information among health care providers and public health authorities. HITSP specifies the use of the IHE IT Infrastructure Technical Framework as the foundation for information exchange (see Figure 1).

Although making the Azyxxi system IHE-compliant is outside the scope of this project, it is important to understand the implications of these specifications. Hence, we have created a software architecture document for an IHE ITI reference implementation of the Web Services version Cross Document Sharing (XDSb) and Cross User Authentication (XUA) profiles. It is the intent to present this information at the RSNA 2007 and HIMSS 2008.

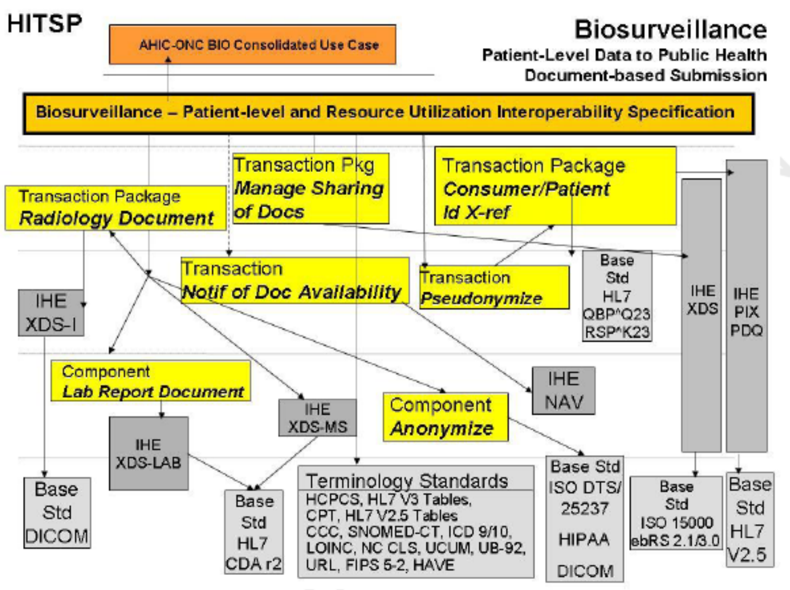


Figure 1 – HITSP Document Map for Bio-surveillance Document Submissions

3.2 Lessons Learned from Project 2

- Although syndromic surveillance offers a broad scope of potential benefits there exist numerous limitations that make it difficult to perform successfully.
- Additional basic statistical and epidemiological research in defining the requirements for syndromic surveillance is a critical need. This includes consideration of other clinical markers for disease as revealed by the Sentinel Collaboratory.
- Information technology can greatly facilitate advancing the research and operationalization of both passive and active disease surveillance. Rapid, cost-effective access to complete patient records, as shown in the Azyxxi system, is critical.
- Currently, syndromic surveillance itself cannot detect a biological threat in a timely manner, leaving little to no time to manage any catastrophic bioevent. Instead, a system of Indications and Warnings (I&Ws) would be far more effective and is being implemented through Project Argus.

4 Project 3: Advanced Visualization

There are several gaps that have existed in the field of advanced data visualization in medicine. Approaches to integrating end user tools for "nth dimensional" (4^{th} dimensional and above) data visualization have not been extensively studied with medical datasets. Few dedicated toolsets for advanced data visualization are available in the open-source community. Lastly, few advanced data visualization tools have been put into the hands of practicing clinicians. The intent of this aspect of the project was several fold: (1) to explore data visualization options beyond typical 2D and 3D charting tools; (2) to develop and release advanced data visualization tools to the open-source community; (3) to investigate platforms for geospatial visualization; (4) to put advanced data visualization tools into the hands of clinicians.

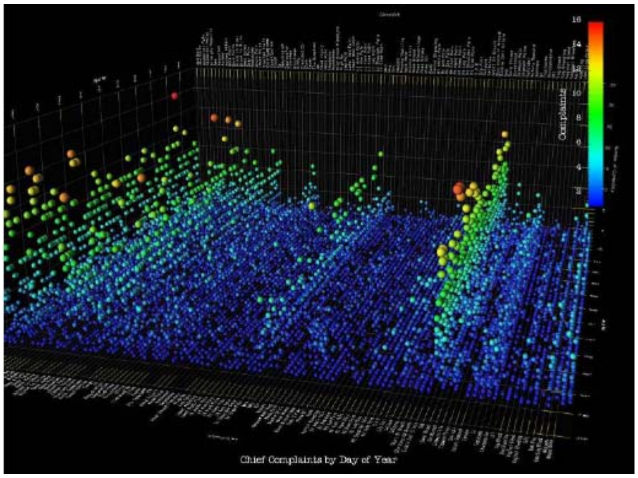
Data from the Project Sentinel collaboratory were expected to be of a bio-surveillance type in which geospatial analysis is often a component; thus, the goal was to investigate the use of software that could serve to unify a broad variety of heterogeneous data into a single geospatial platform. One challenge facing the effort is that as graphical complexity increases, high-end graphics cards are typically needed. For technology to diffuse widely across standard computers in hospital environments, software approaches that have low CPU demands and can run on low-end graphic cards are needed. The goal of the collaboratory was to identify a platform which could scale over time but would also have a low CPU footprint today. Approximately twenty different commercial and open source software packages were reviewed for their capabilities and applicability, ease of use, cost, and extensibility for advanced data visualization (ADV). The following software packages were identified as candidates for use in the ADV piece of Project Sentinel: OpenDX, Miner3D and GeoFusion.

4.1 OpenDX

OpenDX is the open source software version of IBM's Visualization Data Explorer software that runs on Linux. OpenDx uses an object-oriented GUI for creating visualizations and gives users the ability to apply advanced visualization and analysis techniques to their data. The system offers tools for "manipulating, transforming, processing, realizing, rendering and animating data and allow for visualization and analysis methods based on points, lines, areas, volumes, images or geometric primitives in any combination." (www.ibm.com/dx)

A client-server model enables data to be posted to the OpenDX server via CGI (Common Gateway Interface) over HTTP and the OpenDX server returns an image. An Apache web server was installed on the server to allow this interaction. Importantly, each of these graphing template scripts can be used by other researchers to plot their own data via OpenDx. Figure 2a represents a 5^{th} dimensional tool. Apart from the X, Y and Z axes, the color of the objects can change along with the size – thus adding two dimensions. In this particular case, the data plotted shows one year of visits to the ED. Across the X axis is the day of the year, the Y axis categorizes the complaint, and the Z axis is the count. The count has been emphasized by replicating the count to the color and size of the graphing icon. Three distinct walls of "bubbles" emerge from the plot representing the dominant complaints in emergency medicine: chest pain, shortness of breath, and abdominal pain. The dominant role these complaints play in emergency medicine is not always realized from the "ground level" by practitioners but is brought out clearly in this plot.

Figure 2b illustrates cylindrical or bar charting that was instantiated into the suite of tools made available in the open source collection. In the first instantiation of this graph, count values for some diseases could go into the thousands making comparisons difficult without performing camera angle adjustments. The rendering time for each re-rendered frame could take over a minute for some complex graphs. To overcome this hurdle, the values are scaled to a percentage of 100%. This minimizes the number of needed camera angle changes, potentially saving researchers time while analyzing data.



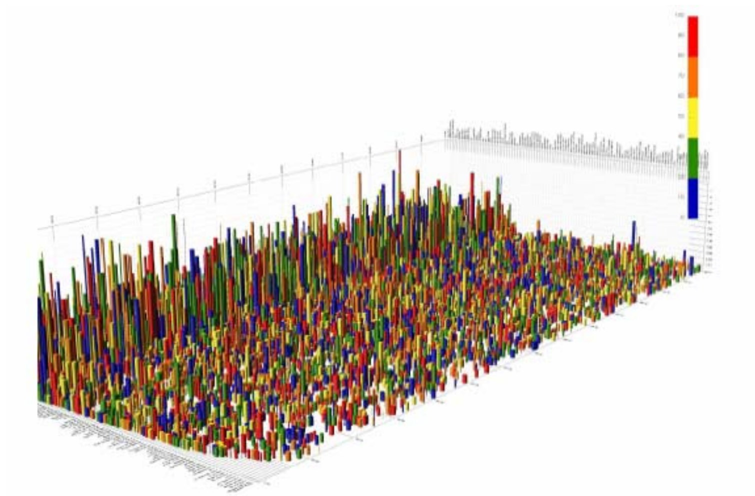


Figure 2a – OpenDX – 5th dimension tool

Figure 2b- OpenDX Cylindrical Charting

4.2 Miner3D

It was determined that OpenDX had some significant limitations with regard to real-time data-plotting and camera angle changes. Miner3D was identified as another software tool which could be fully integrated into the Azyxxi EMR infrastructure. It is an ActiveX component that allows up to fifteen dimensions of data to be plotted into a single 3D space. Dimensionality of the 3D space is increased through using such attributes as color, shape, vibration, size, audio characteristics, labels, and other features. The Miner3D package was instantiated into the Project Sentinel Collaboratory software platform and supports a variety of graphing modalities. Clinicians or researchers across the enterprise can utilize the tool to plot clinical or bio-surveillance data.

During disease outbreaks and high volume seasons, every bed will fill in hospital and patients will be forced to board in the emergency department. Figure 3a represents a 4-dimensional graph showing time of registration on the X axis, date of registration on the Y axis, throughput time on the Z axis, and disposition through color. The figure shows the vast clusters of patients at the bottom that are either discharged (green) or admitted in a few hours. “Bubbling up” from the mass at the bottom are a number of patients who are kept for up to 8 days before being either discharged or admitted to the hospital. The importance of this 4-dimensional graph is three-fold: first, a sense of the proportion of patients in the cohort is immediately discerned; second, up to 11 other dimensions can be added to this graph if needed; third, this tool is accessible with just a few clicks through the EMR across all seven hospitals as well as the DC DOH.

In Figure 3b, each year is represented by a different color. The height represents the number of patients with flu like symptoms. January is on the left with December on the right. A steady and regular flu-season spike is seen in the winter of every year, except in 2001, there is a new unusually high spike. That spike represents all the patients seen after the anthrax attack

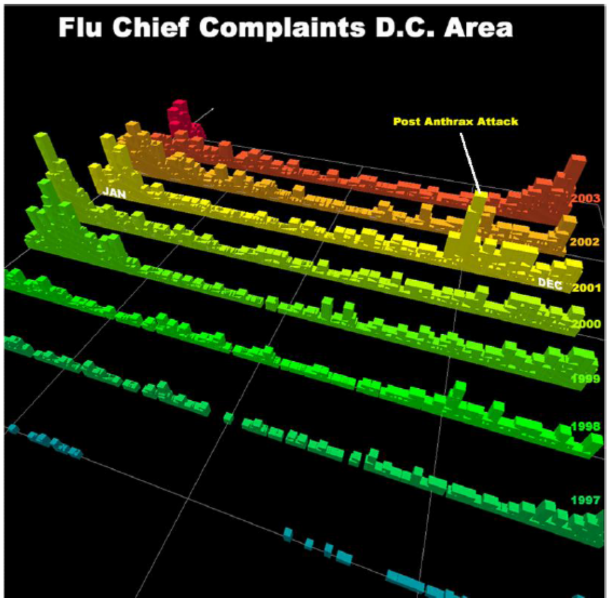


Figure 3a - Miner3D

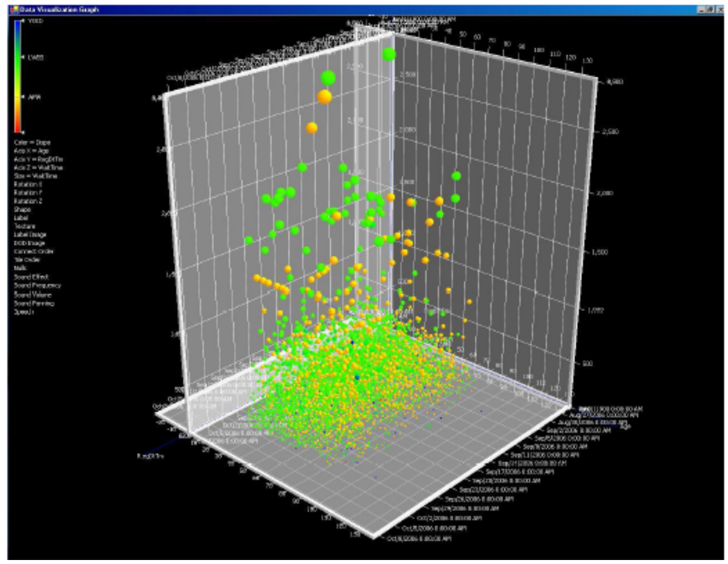


Figure 3b- Miner3D

4.3 GeoFusion

The GeoFusion™ software platform was identified as a flexible, inexpensive software platform on which geospatial visualizations could be built. The company, though commercial, releases its viewer for end-users for free allowing any visualization created by the collaboratory to be viewed by end-users across the world without cost. The developer license is affordable (about $8K). The system supports not only importing satellite level data, but also integration of terrain feature sets with resolutions at the sub-centimeter range. The CPU demands of the system are very low. The standard system will run high resolution animations and fly-ins that pan from space and continue to zoom to staggering sub-centimeter level resolutions even on a commodity Pentium laptop. It also supports IMAX™ quality animations.

The GeoFusion geospatial rendering engine was interfaced to data from the collaboratory and researchers designed interfaces to determine what principles could be drawn on how one might create a "Medical Weather Service Muse." The system supports the ability to map patient cases onto any point in a global, real-time navigable map. In Figure 4a, disease cases by zip code are counted and plotted for the national capital region. One can see an extreme close up view of the zip code borders and the counts of cases for any particular zip code region. Hovering over the red bars gives more detailed information about the outbreak. Most significantly, the dark zip code borders represent a unique type of technology: layering. Layers allow other forms of data to be superimposed on baseline satellite data. Figure 4b illustrates how terrain can be potentially useful for tracing the origin of an event or estimating impact. Terrain scalability was added to the system allowing users to interactively and dynamically change the relative height of the terrain to draw out terrain features more clearly.

Figure 4a – GeoFusion

Figure 4b - GeoFusion

4.4 Lessons Learned from Project 3

- OpenDX is not recommended for on-the-fly image or video creation and analysis. A second limitation of OpenDX is in the area of real-time 3D manipulation. Substantial reworking of a Java based viewer would have been necessary to view the data in 3D and manipulate it in real-time. An even greater limitation of the Java viewer is that it could not utilize client side computational resources for rendering. All rendering would be performed by the OpenDX server and the Java client would just view the video results. The responsiveness of the system would be poor because large video streams would be necessary. It was for these reasons that the OpenDX platform was ultimately replaced by the Miner3D commercial application.
- The Azyxxi clinical repository for Project Sentinel is a rich source of raw data, and extensive data transformation is often necessary before plotting data. Our experience shows that future system designers should make adequate plans for data transformation layers to be added to their systems. These layers should be arbitrarily stackable to allow data to go through several stages of processing.
- One of the consistent challenges of charted data is interpretation. We found that isolated glyphs (such as the “bubbles” one can see in previous graphs) are difficult to place into three dimensional context using ordinary two dimensional displays. As one can see from some of the previous charts, the problem can be compounded when there are tens of thousands of data points.
- GeoFusion’s use of XML syntact for images, points, lines, polygons, 3D objects, symbols, etc. are effective for displaying data with bio-surveillance implications.
- Useful features of the GeoFusion system include the ability to:
  - create multiple layers with variable transparency
  - scale terrain real-time
  - use sub-centimeter scaleable resolutions tosupport any resolution of satellite to local imagery
  - support a low CPU footprint for navigation and viewing
  - plot disease data by region, zip code, or individually.

5 Project 4: Federated Organizations and Biometric Authentication and Authorization using the Shibboleth middleware components

The Sentinel Collaboratory provided a use case in which the each hospital had its own instantiation of a medical information system, and each is administered separately. This means that each hospital handles the identity proofing of medical personnel who will require access to patient data, and each will be responsible for the issuance and maintenance of electronic identity for those personnel taking part in the project. Given these conditions, Project 4 focused on exploring the infrastructure required for the sharing and accessing of medical data between multiple organizations. Specifically the goal is to allow healthcare workers from one organization to access pertinent medical information about a patient from other healthcare organizations that have agreed to share the data.

To accomplish this goal the project focused on three general areas:

- push vs. pull and message vs. document sharing models
- mechanisms for controlling access to shared information
- federated identity management systems which could be used to identify users

The requisite criteria of all the possible solutions investigated are the ability to interoperate across vendor products and integrate with existing organizational infrastructure. A note of interest is that HITSP has begun to recognize the significance of these areas as well.

5.1 Proposed Technical Approach

We propose the use of the Integrating the Healthcare Environment (IHE) IT Infrastructure (ITI) Cross-enterprise Document Sharing, version b (XDSb). XDSb is the most recent implementation choice for the XDS Integration Profile based on a use of the Web Services and ebXML Registry/Repository standards that is consistent with the current developments and best practices in the industry. The use of Simple Object Access Protocol (SOAP) within the XDSb specification provides the basis for creating a network capable of publishing, discovering, and retrieving patient medical documents from participating organizations. This specification supports numerous types of documents (e.g. continuity of care records (CCRs) and radiology images). An extensible set of metadata is associated with the document and subsequently used to query for the document.

We selected XDS.b for two primary reasons. First, it uses the Message Transmission Optimization encoding mechanism (MTOM), which enables arbitrarily large documents to be published and retrieved, given an appropriate implementation of the service. Moving to MTOM addresses a technical limitation that the previous version of the specification (XDS.a) had with documents larger than several megabytes. The ability to handle large documents is significant since medical data can include medical images, which can be multiple gigabytes. Second, because of the wide adoption of IHE specifications and their choice to use simple SOAP constructs, we believe many vendors will choose to support this service directly in their products, greatly reducing the burden to deployers.

In order to control access to the shared information we recommend the use of eXtensible Access Control Markup Language (XACML) within the XDS.b registry and repository components. XACML is an OASIS standard that defines a general policy language used to protect resources as well as an access decision language. It allows administrators to define the access control requirements for their application resources. Specifically we recommend that the registry/repository components act as XACML Policy Enforcement Points (PEPs). These PEPs would communicate with a XACML Policy Decision Point (PDP) located at the organization that is sharing the information. Using this approach allows the organization to maintain control of who access their shared documents.

In order to provide the information necessary for policy decisions to be rendered we recommend that use of both the IHE Audit Trail and Node Authentication (ATNA) and Cross-enterprise User Assertion (XUA) profiles. The ATNA profile will provide the system with verifiable information about the organization making a request (and even machine within the organization) through the use of mutual (both the client and server) cryptographic authentication. The XUA profile provides verifiable information about the individual making the request by defining a mechanism for binding an OASIS Security Assertion Markup Language (SAML), version 2.0, assertion to an XDS.b message. These assertions identify the individual and can contain additional attributes about the user (e.g. medical license number, professional specialty, phone number, HR department, method the user employed to authenticate themselves, etc.). Together, ATNA and XUA, allow access control policies to be based on not only which machine is making the request, as is the case with the current XDS.a specification, but also who is making the request. For example, a policy could be made indicating that only doctors in the radiology department of Georgetown Hospital could view a patient's radiology images but that anyone from Georgetown Hospital or the CDC could view a patient's discharge record.

While this system should provide a strong, secure, means for sharing data it does not deal with the political, organizational, and inter-organizational issues that would need to be addressed. Our final report, and supporting documents, will discuss this further.

5.2 Project Contributions

Perhaps the most important contribution that Project 4 has made to date has been the harmonization of the IHE specification with SAML, version 2.0. The technical lead of this project, Georgetown University's Chad La Joie, is a world-respected expert in system security and SAML. He participated in the IHE ITI meetings, throughout 2006 and 2007 in order to provide nearly all of the XUA specification and offer guidance on the XDS.b specification. His work resulted in an XUA that should be immediately compatible with the federal eAuthentication, version 2.0, initiative and most SAML 2.0 implementations.

In addition to his main work for Georgetown University on Project Sentinel, Chad also works as the technical lead for the Java-based components for Internet2's Shibboleth system. In this role he and his team were able to incorporate, in Shibboleth 2.0, essential work for creating globally scalable trust models and non-browser based application uses for SAML 2.0. Because of this work, as schools transition to Shibboleth 2.0 their researchers would automatically be able to participate in the information sharing system outlined above.

Lastly, and not yet complete, Project 4 will produce a set of documents that further detail this proposed system. These documents will give guidance to software developers looking to implement some, or all, of the software components discussed here. It will offer deploying organizations information on the pre-requisites, integrating the software with their current systems, and an analysis of the risks and benefits of using the system. Another document will discuss the thorny issues surrounding the creation and administration of such a sharing environment as a whole. It will discuss legal issues such as liability and privacy, operational issues such as problem reporting and resolution, and technical issues such as the metadata standards and registry maintenance.

5.3 Observations and Lessons Learned

We have learned a great deal from the work performed in Project 4. First, current medical record systems are not up to the task of sharing data. Most such systems are incapable of sharing any useful data, those that do offer this capability use a proprietary method and as such the data is only available to other product lines from the same company. Perhaps because of their inability to share information, current record systems cannot provide meaningful structured patient records or record metadata nor can they provide access to individuals outside the healthcare organization. Taken together this results in an environment where multi-organizational data analysis is extraordinarily difficult, time intensive and the results are rarely worth the effort.

Next, systems currently being developed to allow access to some of patient information are based on old models and technology that does not scale well. For example, the current IHE XDS specification (XDS.a) uses the same audit-based security model generally used in hospitals. This approach, however, is not suitable for a multi-organizational model where the data owners likely do not have the ability to take disciplinary actions against members of other organizations that may be abusing the system. Other systems, such as caBIG, have chosen to implement their technical security mechanisms and models on, again, systems that were designed for single organizations. This has led to highly brittle security that in some cases is trivial to circumvent and/or monstrously difficult to maintain. In addition, many of these systems assume data traversing the infrastructure is de-identified or anonymized while the systems producing the data make no such assumptions.

Finally, during our research, we observed that push-based message systems did not provide enough contextual data in order to perform surveillance-related analysis without the analytical systems having to store up and maintain a history of medical data. This limitation poses two problems. First, from a security standpoint, the risk of data exposure increases dramatically with each additional system that maintains a copy of the patient data. Even if this data is de-identified the collection as a whole is still subject to correlation attacks. Second, the types of analysis that may need to be done on the data will change over time. Because of this, such analytical systems would need to keep a significantly large timeframe of data available in order to account for analysis over long time periods. For example, a change in an analysis process from looking at data from the most recent month to the most recent year.

Despite the current limitations, it is important to recognize that existing systems represent the first foray into this field and a significant amount of information has been gained from the effort and is informing the latest round of specifications like SAML 2 and XDS.b. The results of which, as suggested here, allow for the creation of a strong, useful, information sharing system that could be used not only for bio-surveillance but also to support clinical practice as well as clinical/translational research in a multi-organizational setting.

6 Project 5: Grid Service for Data Sharing and Interoperability dealing with Grid and Web Services and using the Globus toolkit and related tools

The initial focus of Project 5 was to pioneer Grid technology in clinical environments in a dual capacity: routine and emergency (biodefense mode). At the time the contract was awarded the Globus Toolkit (GT) transitioned from version 2 to version 3. Version 3 was the first time the GT made extensive use of Web Services.

Healthcare delivery in general and biodefense in particular is inherently a decentralized activity where participants cross many institutional boundaries. However, current healthcare systems are not flexible in supporting clinical workflow, cannot easily work with other systems, and impose onerous solutions whenever integration or modification is required. Thus a key problem with Electronic Health Record (EHR) is interoperability, which is the subject of vast amounts of ongoing research. Interoperability can be improved through judicious use and introduction of standards, thus helping overcome proprietary formats.

At the same time, however, it is recognized that present standards are not sufficient to solve the EHR interoperability issues. Any serious EHR effort clearly needs interoperable management of persons' identities: they need to be handled in a consistent way across the healthcare continuum, and identifiers must be coordinated and correlated among different organizations. The only standard that addresses these needs is the Person Identification Service (PIDS). It provides rich and well-researched semantics, and has been adopted both by government and commercial users in several countries. The problem is that PIDS is specified using CORBA semantics, and CORBA is no longer considered to be a mainstream distributed environment. This issue motivated the focus of the project to bring PIDS within the IT infrastructure of Web Services and Grid environments.

There are other clinical services and related standards (such as HL7 3X, RIM) which can help the EHR interoperability in Grid and Web Services environments. However, after a brief assessment in 2003, it was decided that while these standards hold great promise, they were still not at the maturity level to justify the considerable resources required for their use in Grid and Web Services environments.

Thus, our refocused efforts were to:

1. Extend PIDS to Web Services and then Grid environments. We call the resulting specification WS/PIDS.
2. Insure WS/PIDS compatibility with mainstream data grid standards (such as Web Services Resource Framework – WSRF – which provides a standardized way of handling state in Web/Grid Environments)
3. Validate WS/PIDS in research and clinical domains

6.1 Technology Description

Grid and Web Services Convergence

The (data) grid technology pioneered by Globus Toolkit focused on offering standardized services relevant to mainstream computing. This focus motivated the use of Web Services within the GT3 and offering technology (such as Grid Services, included in GT3) as an extension to Web Services. Following critical reviews from the user community a re-factoring of the toolkit code resulted in the discarding of some of the initial GT3 offerings (such as Grid Services) and the creation of standards and associated implementations amenable to widespread use. Thus, the latest GT version 4 introduces WSRF, a well received and widely accepted standard for handling state in Web Services environments. The WSRF implementation became a standard layer used by a number of more advanced GT4 capabilities.

Technology Description WS/PIDS

The WS/PIDS specification is WS-I compliant and it is compatible with GT4 WSRF. It is a very close rendition of the original CORBA PIDS semantics. The open source implementation of the WS/PIDS specification depicted in Figure 5 is platform agnostic because it is used equally well with .NET or J2EE environments. The implementation is database independent and works equally well with MySQL open source database and with proprietary databases such as Oracle.

A number of significant additions and extensions to WS/PIDS were necessary to make use of the WS/PIDS in the live clinical research environment provide by the collaboration with Washington University in St. Louis (WUSTL).

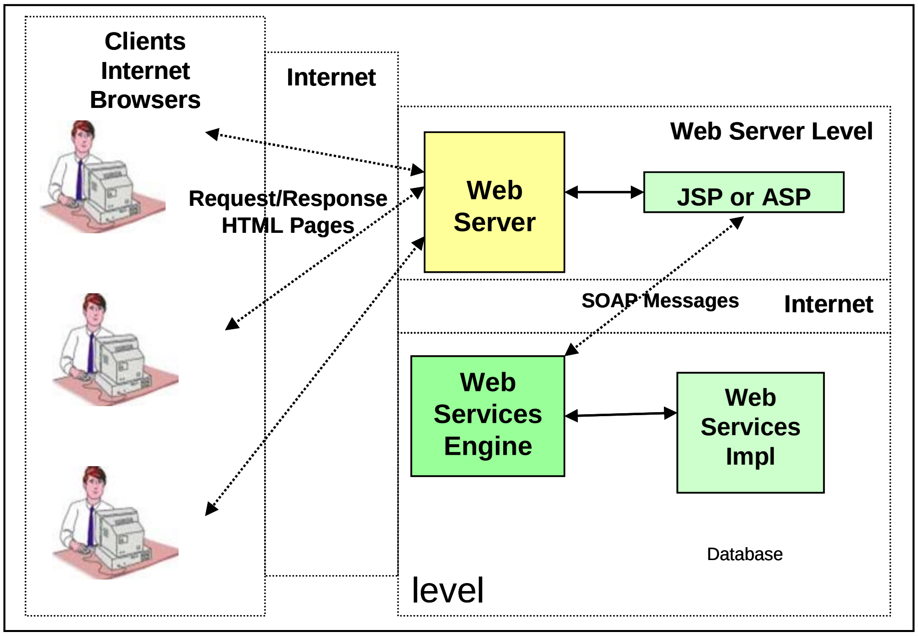


Figure 5 - Current implemented general architecture for WS/PIDS:
Web Services using .NET and J2EE clients

6.2 Accomplishments and Potential Impact

- The open source WS/PIDS specification and implementation, backed up by comprehensive and validated use cases is currently in use in a clinical research environment at the Clinical Imaging Research (CCIR) at WUSTL. WS/PIDS is the core component of the Research Master Participant Index (RMPI) (Figure 6).
- The WS/PIDS specification is compatible with IHE profiles such as PIX, PDQ, insuring its use IHE compliant environments
- There is an ongoing standardization effort called Entity Identification Service (EIS) sponsored jointly by OMG and HL7. The RFP was issued in February 2007 and call for respondents to propose a (Web Service) based specification and make commercial implementation commitments. The initial submission deadline to the RFP is in 9/07 with revisions in 11/07. To date there are at least five respondents. Most functional capabilities of EIS are openly inherited from CORBA PIDS in the following main areas:
    - Metadata Interface
    - Entity Management Interface (manages entity information)
    - Query Interface
- It is expected the final standard EIS specification will overlap to a large extend with WS/PIDS positioning WS/PIDS to become the open source compliant implementation of EIS.

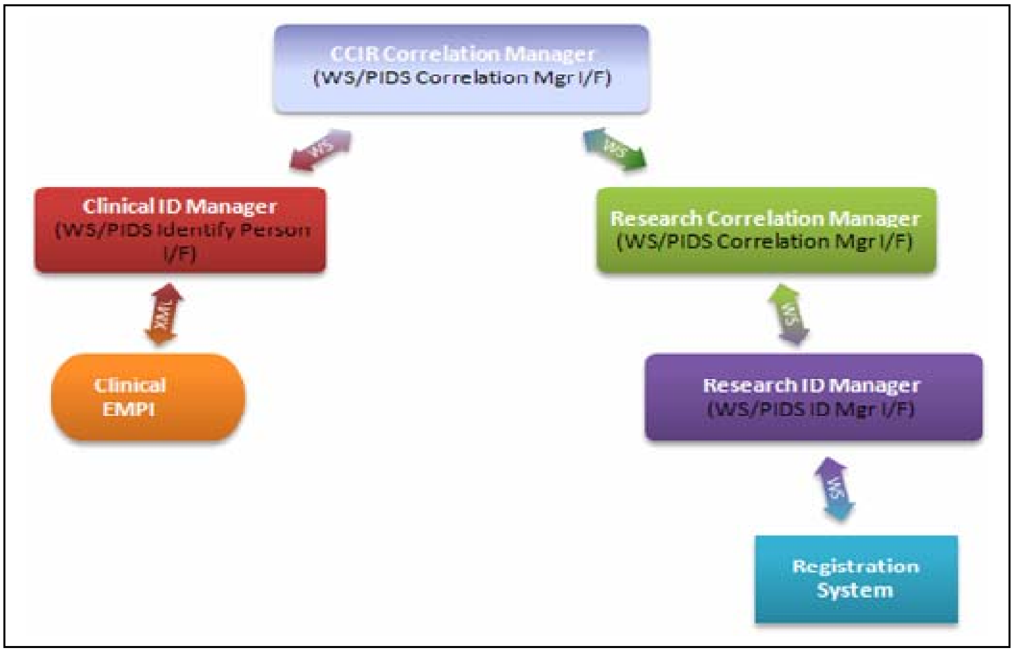


Figure 6 - WS/PIDS-RMPI System Design Concept
The research components (left) have been implemented

6.3 Lessons Learned

- Agility and dynamism of the web environment participants may create divergence with established directions because of the speed of adopting new mainstream priorities and associated standards. Thus Globus Toolkit iteration lifecycle may fall behind new developments/standards.
- Lack of standards for (patient) identification within web environments is a central issue in clinical environments which is clearly recognized and addressed by standard bodies (EIS). WS/PIDS can become a viable contender in the EIS based integration efforts because its judicious choice of clinical semantics and technology.
- Real live implementations of WS/PIDS uncovered new valid requirements that may enhance standards usability such as:
    - Need for domain management capabilities
    - Need for flexible "linking" of ids

6.4 New Direction and Future Activities

- Finalize the use of WS/PIDS in conjunction with WUSTL partners of a real world RMPI application and achieve a stable and consistent support of a hierarchy of research domains in medical imaging
- Enhance and Validate WS/PIDS by:
    - adding appropriate features (linking, domain management, insuring its applicability in a particular clinical domain
    - providing interoperability features (such as PIX or PDQ front-end wrappers)
- Assess the final EIS specification
- Align WS/PIDS with EIS

7 Project 6: Knowledge Dissemination and Sustainability

The knowledge dissemination component of Project Sentinel is multi-faceted. We have produced a large number of publications posters and oral presentations throughout the duration of the project that reflect the achievements of the project. The knowledge gained from a bio-surveillance perspective is foundational to the launching of Project Argus which is separately funded. However, to truly sustain the knowledge into the future, dissemination should occur outside the narrow scope of the project. In March 2006, Georgetown organized a workshop on Multi-center Image Management (MCIM) to explore open source strategies in support of flexible access to biomedical data. As a result, the group recognized the technology gaps between commercial information systems that focus on efficient clinical operations within a single institution and the urgent needs of the research community that require flexible access to multimedia data from multiple institutions. Since these gaps are not likely to be addressed by the commercial sector, an effort by the research community to develop robust software tools that can mend the gap is essential. The group concurred that an open source approach could support efficient software development, incorporate the rigor required to meet the highest standards of software quality and encourage collaboration between government, industry and academia. To support this effort, imaging and informatics experts at Georgetown University, Washington University in St. Louis, the Northwestern University Feinberg School of Medicine, and the University of Geneva (Geneva, Switzerland) have formed the Information Management Toolkit (ImTK™) Consortium. The ImTK™ Consortium aims to expedite translational biomedical research through the development of software tools that enable efficient exchange, sharing, management, and analysis of multimedia medical information such as clinical information, images, and bioinformatics data. ImTK™ is based on an open source and open architecture approach to allow scientists, engineers and physicians throughout the world to participate in this initiative.

Much of the work accomplished in Project Sentinel is directly related to biomedical information management across institutions. There significant overlap of the IT requirements for bio-surveillance, clinical practice and clinical research/multi-center clinical trials. Thus, in an effort to support this effort we have accomplished the following:

- In June of 2006, formed the ImTK™ Consortium
- In April of 2007, organized the MCIM 2007 Workshop which focused on clinical trial management and determined there is a clear need and compelling business case for this effort.
- Validated WS/PIDS in a clinical research environment with the WUSTL RMPI project
- Created an open source development environment based on the environment and process implemented by Kitware Inc.
- Developing an IHE XDSb/XUA architecture and implementation guide. It will be presented at the RSNA 2007 and HIMSS 2008.
- Developing a bio-surveillance implementation prototype based on the HITSP interoperability specification, concept of federated authorization using SAML 2 and use of WS/PIDS. This infrastructure created to support this effort will have direct applicability to the clinical research/multi-center clinical trials domain.
- Planning to hold an ImTK-based session at Computer Assisted Radiology and Surgery (CARS) in June 2008.