Distributed Biomedical Database for Public Health Research

Yavuz Tor\textsuperscript{1} 
yavuztor@sbcglobal.net

Richard H. Fowler\textsuperscript{1} 
fowler@panam.edu

Charles A. Harlow\textsuperscript{2} 
caharlow@panam.edu

Chaitanya Jasti\textsuperscript{2} 
chen@cs.panam.edu

Zhixiang Chen\textsuperscript{1} 
chen@cs.panam.edu

\textsuperscript{1} Department of Computer Science, The University of Texas-Pan American
\textsuperscript{2} Department of Electrical Engineering, The University of Texas-Pan American

Abstract

Results of research in public health highly depend on the validity of data collection and internal research processes. To ensure validity, maintaining records of all assets in the research is very important. Researchers should be able to diagnose problems by tracing back the processes that leads to the problem. In this study, we design and implement BioRio system, a portable, flexible and secure information system for public health research that will reduce human errors, and increase traceability and efficiency.

1. Introduction

The mission of public health is to "fulfill society's interest in assuring conditions in which people can be healthy." [1] This involves a substantial amount of research and surveillance. An essential part of a research study is the validity and quality of the evidence that leads to the result. In public health research, this translates to data and specimens collected from patients and assays that are performed on these specimens. Assuming a sufficient sample size, the clarity of the results will increase as the validity and quality of assays increase. To ensure any level of quality in assays, traceability of assay processes should be achieved.

Despite availability of fully integrated Laboratory Information Management Systems (LIMS), most of the public health schools and public health research institutes prefer to not utilize this software. The following paragraphs will explain the reasons.

Most LIMS vendors have numerous features implemented in their LIMS products, to cover different aspects of laboratory processes in different types of labs. This calls for a customization for the specific research lab/institute. This is where the cost of procurement increases, since the customization process needs a substantial amount of resources.

Secondly, most LIMS products require a long time period to learn how to work with the system, as well as how to maintain system integrity [5]. Research studies that are performed within short and strict timelines cannot afford devoting this much of time into the adoption of such a system.

There are also low-priced, less-featured LIMS products in the market; however they do not offer much customization and configurability. They are mostly targeted to clinical laboratories and do not support proprietary assay protocols to be developed and carried out. This might, of course, be filled up by using proprietary documents to support LIMS; however this leads to loss of visibility and traceability, and hence defeats the whole purpose of having a LIMS in place.

The focus of LIMS products is mainly on the assays and samples. Although this covers requirements of a public health research institute, it does not provide visibility of progress towards the objectives of the research study. In the next section we will focus more on the requirements of public health research for a computerized system.

Our goal in this study is to provide a web-based database system that exploits such features to satisfy these problems of public health research centers.

2. General Requirements of Public Health Research

The main requirement of public health research is to provide traceability in workflow. Researchers should be able to review the procedures of their studies, and have insight into study results. This gives more control over the research study, and leads to more knowledgeable results. In these terms, most
requirements of public health research collide with the attributes of LIMS systems.

McLelland [2] lists attributes expected of a LIMS as:

- To shorten turnaround times for lab tests
- To improve access to the results database
- To improve accuracy of analysis, by eliminating transcription steps
- To count and monitor resource utilization
- To exchange data and information both with analytical equipment and corporate mainframes
- To improve productivity

In addition to those features offered by LIMS products, public health research requires features to enable researchers to evaluate their performance and progress towards having a conclusion. Following paragraphs will address these requirements.

One important requirement is to develop and enforce protocols for handling specimens and performing assays. Most research studies in public health try to develop assays that will help diagnose and investigate a health hazard. To accomplish this, researchers develop certain protocols to be enforced in assays, and also in specimen handling. These protocols are often revised to increase quality and validity. Therefore, enabling revisions of these protocols is also an important feature.

Research studies in public health involve a substantial amount of data collection, in addition to specimens and assays. Data collection is based on questionnaires or forms created by researchers, which are believed to best serve the purpose of the study. A computerized system should provide features to enable researchers to create and maintain questionnaires and forms for data collection. A computerized system should enforce use of these forms and questionnaires in data collection, and audit changes and data collections.

Another feature required for research studies in public health is data import and export utilities. We can refer to this as collaboration requirements. Most of these systems were very targeted and limited in features. The Cancer Laboratory Management System (caLIMS), by National Cancer Institute (NCI), seemed to provide the most extensive features. caLIMS also provides satisfactory documentation on internal procedures and architecture. Features provided include management of assay protocols and processes, quality control and validation tracking, sample management, user, laboratory, equipment/device, and collaborations administration [4]. Although the documentation provides information on the design and architecture of the system, we found it hard to incorporate the current features with the new features that we are trying to implement.

Based on our evaluations of current systems, we decided to implement basic features of public health research information system from scratch and keep the possibility of integration open to incorporate systems like caLIMS in the future.

3. Available Systems

There are numerous LIMS products available in the market, which provide features that, partially, satisfy aforementioned requirements. Avery et al [3] provide a list of general, clinical, and specialized LIMS systems on the market. We have reviewed a few noteworthy systems that could be a part of the solution for the computerized system we are trying to develop. Here, we will discuss our perception of these systems.

One of the most extensive systems is StarLIMS from StarLIMS Inc. StarLIMS offers features, such as Raw data capture and storage, standards inventory, tracking of laboratory instruments and maintenance/calibration schedules, management of laborant training records, tracking of chain of custody and specimen inventory, regulatory management, electronic document management. StarLIMS also provides extendibility through scriptable components and scriptable XML. With these features, StarLIMS seems to be highly customizable and useful; however it is not affordable for most public health organizations to adopt such a big system. In addition, reading the material provided in www.starlims.com, we could not identify a feature of StarLIMS to manage the collection of demographic data from patients.

Analyzing the costs of strong systems such as StarLIMS, we came to a decision that it was not possible to procure such big systems. Therefore we started to review affordable and extensible systems that provide essential features. We have considered RLIMS Pro for this purpose, however we could not find sufficient instructions on how to extend it or integrate it with other systems.

We also evaluated the use of open-source systems. Most of these systems were very targeted and limited in features. The Cancer Laboratory Management System (caLIMS), by National Cancer Institute (NCI), seemed to provide the most extensive features. caLIMS also provides satisfactory documentation on internal procedures and architecture. Features provided include management of assay protocols and processes, quality control and validation tracking, sample management, user, laboratory, equipment/device, and collaborations administration [4]. Although the documentation provides information on the design and architecture of the system, we found it hard to incorporate the current features with the new features that we are trying to implement.

Based on our evaluations of current systems, we decided to implement basic features of public health research information system from scratch and keep the possibility of integration open to incorporate systems like caLIMS in the future.

4. Standards

Public health departments and their clinical partners are moving ahead rapidly to implement systems for early detection of disease outbreaks. In the urgency to
develop useful early detection systems, information systems must adhere to certain standards to facilitate sustainable, real-time delivery of important data and to make data available to the public health partners who verify, investigate, and respond to outbreaks. To ensure this crucial interoperability, all information systems supported by federal funding for state and local preparedness capacity are required to adhere to the Public Health Information Network (PHIN) standards [6].

Standards-based system development is critical for three major reasons.

1) First, the need for real-time information from multiple sources can best be accomplished by standards-based electronic messaging. Although individual custom interfaces can be created with the myriad potentially useful data sources, the cost of development would be prohibitive. The specification for standard Health Level 7 (HL7) [7] messages for early detection data permits health departments to leverage integration-broker technology and health-care delivery site information technology (IT) capacity for creation and processing of these standard HL7 electronic messages.

2) Second, the use of standards enables health departments to leverage previous investments in their IT infrastructures. Systems to support public health capacity for outbreak management, response, alerting, and information dissemination have been under development since investments in the Health Alert Network (Fiscal Year 1999) and the National Electronic Disease Surveillance System (NEDSS, Fiscal Year 2000). A detection system is most valuable when it can communicate with those systems needed to investigate and respond to an epidemic. The availability of standards-based shareable directories, system security, and channels for bidirectional secure communication can support public health agencies’ capacity to respond to outbreaks and provide key elements for early detection systems.

3) Finally, a consistent standards-based approach limits the burden on partners in the clinical-care delivery sector. Health-care providers and hospitals provide information to public health agencies for early detection and routine surveillance as part of their community responsibility. They are not compensated for the cost of providing that information. By using standard formats and electronic reporting, public health agencies can minimize the burden involved in reporting diseases and, ideally, use information that is already available in electronic format within the health-care delivery system.

Nationally, the importance of standards-based, interoperable electronic health records to support objectives for quality and safety within the health-care delivery system has been increasingly recognized. The National Committee on Health and Vital Statistics has recognized standards as an integral part of the National Health Information Infrastructure. The critical role of standards has also been endorsed by the U.S. Department of Health and Human Services and the federal government through the Consolidated Health Informatics. To define how these broad standards can be implemented in surveillance systems that support the specific needs of public health practice, CDC and its state and local health department partners have identified key specifications and functions described as the Public Health Information Network (PHIN) [8].

5. Design Process

The first step in our design process was to identify an intuitive and simple framework to implement a portable, secure and extensible web-based system. We have adopted the model-view-controller (MVC) design pattern for the design of such a framework. Figure 1 depicts this framework and its components.

![Figure 1 Framework for BioRio system](image-url)

The controller works as the system facade. The controller intercepts all client requests, and creates the list of tasks to be carried out by the help of the navigation map. Each task is provided with a connection to the system database and data are updated through this connection. Depending on the status of the tasks, the controller selects appropriate view element to return to the client. Then, each view element selects their template, uses the template engine to create the response to send to the client. When the response is created, the controller sends the response to the client.

The next step was to design an intuitive and simple way to represent specimen data in the system. Looking at the specimen lifecycle, a specimen record is created with the collection of the specimen from the patient.
Next, this specimen is processed, based on the protocols, and separated into different aliquots. So, each aliquot, in this case, are treated as a separate specimen. Of course, new aliquots will have the same root as the specimen they are separated from. In accordance, each aliquot will have a different aliquot number, which defines the path of the specimen in the specimen's family tree.

![Specimen Data Diagram](image)

**Figure 2 Specimen Data**

Regarding every aliquot as a separate entity, gives us the ability to keep a separate inventory for each aliquot. Therefore, we can track every process that an aliquot of a specimen goes through. Since each specimen knows its root, it is easy to link to the corresponding patient's file and the specimen's first collection information. Each specimen has a barcode on its tube. Each barcode reads into the unique identifier of the specimen, which increases reliability and efficiency of specimen inventory. ABC CODABAR barcode technology is used to create the barcodes in barcode labeler within the system to reduce the human errors [14].

Specimen inventory and storage is another issue we addressed in our design. The main storage entities we identified were box and freezer. All other storage units (cabinets, drawers, shelves, racks) are considered as a structure unit in the freezer and described in a storage descriptor XML text within the freezer record. Each box consists of wells. Wells identify the location within the box and referenced only by an index number. Each box record has the information of number of rows and columns. At any time, the box is viewed by first identifying the wells that are occupied. System looks into check-in information, which keeps where the specimen is stored. Each check-in process is recorded with the box-id and well index.

Each specimen’s processing information is designed to have a simple link to the protocol that has been used in the process. A protocol is a documentation of the process that is to be performed, and is created by the primary investigator. A primary investigator can create several versions of the protocol, but only the ones that are marked as ready-to-use can be selected while recording the processing information. Processes that a specimen has gone through are audited in the system.

Patient information has been divided into two parts: identifying information and demographics. Since identifiable data is very sensitive, secrecy is maintained amongst studies. Each study keeps its own unique record of the patient. Although each patient has a unique identifier throughout the system, this identifier is hidden from the study. Therefore, each time a patient is registered to a study, a separate record is created with a link to the patient's public (unidentifiable) data. The patient's identifier in other health organizations, such as clinics and health departments, are also recorded. Therefore if a researcher needs more details, he/she can contact to the health organization with the patient's identifier, with the patient's consent.

Patient demographics information is highly dynamic and very essential for research, to analyze epidemics and draw conclusions. Each research study requires different information from their cases. Usually these demographics information are collected in interviews, via questionnaires and forms. Our system, allows the primary investigator to create these forms and questionnaires online. Each form has an XML descriptor that defines how each field is organized in the page. In turn, each field in the form has two descriptors, one for data input and another for data view. Each response from the patient is tracked by a unique identifier and can be tracked back to the data collection form and fields for the patient. In this way, each patient can enter his/her information into the system directly, eliminating the task of entering information into the system later.

The performance tracking has not yet been designed in the system and is included in our future work. The proposed measures include the constraints on number of patients, number of specimens, number of qualified specimens, number of questionnaires, or number of who responds positively or negatively to a questionnaire field.

Our selection of middleware came natural, as one of our goals was to have portability as well as flexibility. We have selected IBM Websphere Application Server as our middleware, because of its J2EE compliance, availability of other J2EE compliant servers, and portability of Java across different platforms. As the database back-end we have selected IBM DB2, for its robustness and its extensive tools for data warehousing. IBM DB2 also provides an add-on spatial data management, which we project to use in the later steps in our development to incorporate geocoding of patient information. With automatic geo-
coding of patient information, the researchers will be able to see the distribution of a disease over a geographical area, and can get intuitions on the migration of the disease.

6. Security

Security in the data transport is accomplished via use of Secure Sockets Layer (SSL) protocol with 128-bit encryption. This is considered to be strong enough encryption and most commonly used in online banking applications.

Access to the system is limited to those that are registered users with valid passwords. Each user receives a role when registered to the system, and their authorization to do a task is identified by this role. User passwords are hashed with MD5 hashing function, via the method referred to as “salting”. Based on this method, each password is concatenated with the user's login name, and then hashed. Therefore, one needs to know both user name and password to be able to hijack the password, in case of a database breach.

Additional security methods have been proposed but not yet been designed. Some of these include, assigning client certificates to the computers that are used to access the system, using user certificates in USB drives to identify each user, require fingerprint scan together with user name and passwords.

7. Storing hierarchical data in relational database

Specimen information has a hierarchical structure in laboratory environment, since each specimen can be obtained from another one, by simply taking a piece of it into a different aliquot. Although, now considered as two separate specimens, each of these specimens shares a common history. This is especially important to know, because the status of the specimen depends on the processes that it has been through in the past.

Mapping hierarchical data into a relational database is a challenge. Our main consideration is the ease of update, and the cost of path queries. In our design, for each specimen node we have recorded the root node's id number. Therefore, every node in the specimen tree will have the same root node number. This enables us to retrieve the whole tree, with one simple query. Path information is kept in conjunction with the root node's number. Figure 3 depicts how the path information is stored. In our system we call this path information the aliquot number.

8. Guarding patient information

Patient information is protected by regulations such as Health Insurance Portability and Accountability Act (HIPAA). These regulations state that such information should be kept secret, and be revealed only with patient's consent. The nature of a research study, however, implies transparency in the information that is collected and the procedures followed. We try to resolve this dilemma by not keeping such sensitive information in the database, leaving them in paper based forms in locked cabinets. All other information is linked together without personal identifiers (such as name, date of birth, etc.) and only with a patient registration number in the study. Therefore researchers can make detailed study of individuals by looking at the paper documents in locked cabinets.

9. Current Status and Future Work

We have finished the first phase of our design of public health research information system. In this phase, we included the following features:

- Research Management: Start a new research project, specify goals, members of the research, descriptive documentation, protocols, data collection forms
- Specimen Lifecycle: specimen tracking from collection to disposal, bar-coding of all aliquots, keeping track of processes, maintaining specimen history
- Patient Information: collection of demographics information, maintaining patient identifiers, registering patients to research project
• Storage Management: adding freezers, structuring freezers, adding boxes, maintaining checkin and checkout history
• User Management: Maintaining user roles and privileges in research, signing up users, access management

We have not yet finished the implementation of the assay procedures, as of the date of the writing. The design of the assays include the laborant certifications, instrument characteristics, placements of specimens on test plate, test runs, quality control samples, and import of raw results out of the instrument into the system. The system will receive the electronic output from an instrument and parse it into the system using configuration descriptors of that instrument.

There are many features we would like to include in our system. The most important feature in our schedule is standards compliance and regulatory. Enabling regulations such as FDA’s Good Laboratory Practices (GLP) [9] will enable easy validation the laboratory processes. Also, data transfer standards, such as HL7 are very important for electronic transfer of patient information. It will make it very efficient for research centers to receive and communicate patient information with other health organizations.

We also have scheduled automatic geo-coding of the patient information, which will enable researchers to see the geographical distribution of cases in the study. ESRI ArcGIS suite and ArcIMS will enable researchers make queries in the database and view the results geographically, using the system.

10. References


