An Architecture for Streamlining the Implementation of Biomedical Text/Image Databases on the Web

Mike Bopf, Tina Coleman, L. Rodney Long, Sameer Antani, George R. Thoma
Communications Engineering Branch, Lister Hill National Center for Biomedical Communications, National Library of Medicine,
National Institutes of Health, Department of Health and Human Services, Bethesda, MD
{mbopf, rlong, santani, gthoma}@mail.nih.gov

Jose Jeronimo, Mark Schiffman
Hormonal and Reproductive Epidemiology Branch, Division of Cancer Epidemiology and Genetics (DCEG), National Cancer Institute,
National Institutes of Health, Department of Health and Human Services, Bethesda, MD

Abstract

To date the implementation of biomedical text/images databases on the Web has been hindered by the lack of software flexibility to enable new datasets to be "rolled in" to existing database systems without extensive programming modifications. An R&D division of the National Library of Medicine (NLM) is creating a new database system, the Multimedia Database Tool (MDT), built on an architecture that moves the required customization for a particular dataset from the programmer to the database administrator level, thereby reducing potentially high labor costs. The first database application to be supported will be uterine cervix data from the National Cancer Institute, including 100,000 digitized images, and NHANES II and III databases now hosted by the NLM Web-based Medical Information Retrieval System (WebMIRS).

1. Background: biomedical text/image database on the Web

The Lister Hill National Center for Biomedical Communications (LHNCBC), an R&D division of the National Library of Medicine (NLM), has developed the Web-based Medical Information Retrieval System (WebMIRS) as an R&D database system that provides access to data collected from the NHANES II and NHANES III surveys, including 17,000 digitized NHANES II x-rays. WebMIRS is to a large extent custom-designed to support the NHANES databases and image data. To facilitate the deployment of additional databases of text and image data on the Web without extensive software reprogramming, a new system architecture is required. The new prototype system built upon this architecture is called the Multimedia Database Tool (MDT). The MDT has the flexibility to accept new datasets of text and images with the required customization for these new datasets done at the level of the database administrator, rather than the programmer.

The first dataset to be incorporated into the MDT will be uterine cervix data and approximately 100,000 related images collected by the National Cancer Institute (NCI). The image data consists of digitized cervicography images. The database is intended for use in biomedical education and training, and for research toward understanding the natural history of Human Papillomavirus (HPV) infection and cervical cancer. The NCI data is described in

more detail in [1]. The origins and significance of the NHANES data have been addressed previously [2].

2. Overview of MDT functionality

The MDT retains all the functionality of WebMIRS: capability to query a database of text and related images over the Web; to show query results with a multiple image display and associated text data; and capability to export query results for statistical analysis. The MDT will add a new capability for distributed data collection from remote users; users at many geographically dispersed sites will be able to view images and record interpretations for those images into a central database. A system of user privileges is being incorporated into the MDT to restrict database components that are user-writeable and also restrict write privileges to authorized users. The MDT system has been designed to support a broad class of text/image databases.

Of particular interest in the MDT design is the incorporation of capability for retrieval and display of spatial data. For example, the functionality of the Boundary Marking Tool (BMT), developed at the LHNCBC to allow the marking and archiving of the boundaries of irregular spatial regions on the NCI uterine cervix images [1], is being incorporated in the MDT to allow the query, retrieval, and display of these regions, superimposed on the uterine cervix images. Highlighting in other image types, such as spine images, could also be displayed using the same mechanism.

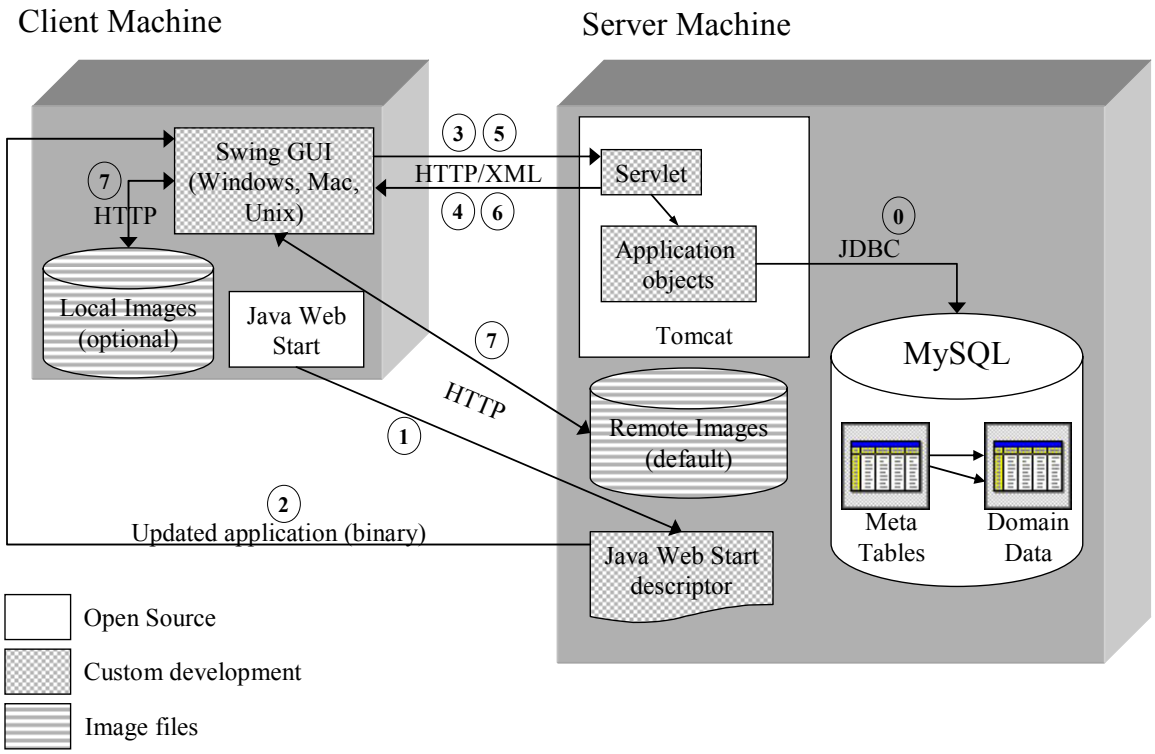

Figure 1. MDT System Architecture

3. MDT system architecture

The MDT is a three-tier Java system consisting of (1) a client that provides a Graphical User Interface (GUI), (2) a servlet running on a remote machine, and (3) a Database Management System (DBMS). The servlet interacts with the DBMS (a MySQL database server) using Java Database Connectivity (JDBC). A user first downloads the client with a Web browser, using Java Web Start Technology (`Figure 1` – flow #1). Each access after the initial download checks for updates to the application and downloads them if necessary (flow #2); after the first download, a Web browser is no longer required. The only additional software required on the client machine is the Java Runtime Environment (JRE); all major architectures (Windows, UNIX, Mac) can function as MDT clients. The client is written using Java Swing, which provides the user robust graphical interaction, especially compared to a thin HTML browser client. All of the components of the system are supplied via open-source systems.

Upon startup, the servlet reads *metadata* tables from the database (flow #0), and caches the information locally. The system distinguishes between two types of data: *domain data* (the medical data of interest) and *metadata,* which describes the domain data within the abstract framework that we have created. `Figure 2` shows the relationships between the primary metadata tables and a subset of the NCI Uterine Cervix domain data.

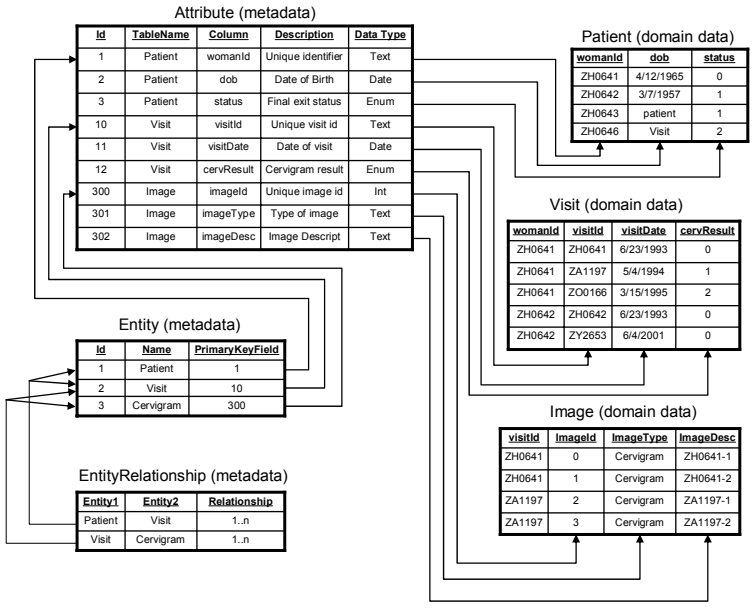

Figure 2 Metadata / Domain Database Table Relationships

For each *data item* in the domain data, the system requires an entry in the Attribute table to define that item. This includes a DataType (Text, Numeric, Date, etc.), and pointers to the tables and columns within those tables that actually hold the values for the data item. The Entity table describes key domain data fields, defined by a content expert who understands the domain data and its interrelationships. For the NCI Uterine Cervix domain, the Entities are *Patient*, *Visit*, and *Cervigram* (*Image*). The Entity_Relationship table describes the relationships between these core Entities. Of primary importance for purposes of data queries is the *cardinality* or many-to-one relationships among the Entities.

At startup, the client communicates with the servlet via XML over HTTP to authenticate the user using an assigned username and password (`Figure 1` - flow #3). HTTP allows the system to avoid firewall issues that were a limitation in WebMIRS since it required a direct

database connection. Many firewalls prohibit direct out-going connections, but all allow HTTP access.

Once logged in, the client requests the *system model*, built from the metadata tables, which describes the domain data interrelationships. Figure 3 shows a UML (Unified Modeling Language) class diagram of the MDT internal class structure of this model. A UML diagram describes Object Oriented software systems and is similar to an Entity-Relationship Diagram (ERD), commonly used in database design. The boxes represent classes, with arrowed lines representing an Association between two Classes, often with a multiplicity (*) symbol showing a many-to-one relationship. A line with a large triangle represents inheritance.

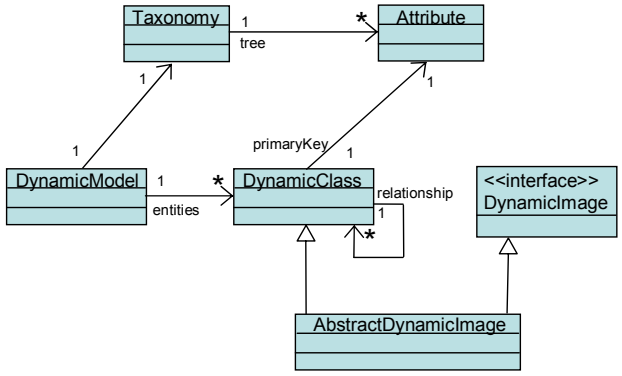

Figure 3 UML Diagram of MDT Model

The core of the MDT Model is the *DynamicClass* which contains the structure of one entity (primary data item). In the NCI Uterine Cervix domain, the entities are Patient, Visit, and Cervigram image. Cervigram is both a DynamicClass and an *AbstractDynamic* image that implements the *DynamicImage* interface. DynamicImages have the ability to be downloaded via HTTP, as they have a URL for the image file location. DynamicClasses contain instances of the *Attribute* class, which encapsulates the information provided by the Attribute table in Figure 2, and describe a particular data field, including its domain (text, numeric, etc) and enumerated type data. The system model groups and orders these attributes in the *Taxonomy* which the client software (and, in turn, the user) uses to determine what fields can be used to construct a query. The GUI presents the Taxonomy in a dynamically-generated tree structure, where the tree nodes are the Entities (DynamicClasses). The *Attribute* data is then used to provide the user with meaningful data with which to build queries.

When the client first requests the system model, the Servlet constructs the model by reading the Attribute, Entity, and EntityRelationship tables and transmits it to the client as a *DynamicModel*. For example, for the NCI Uterine Cervix domain, the DynamicModel contains DynamicClasses for Patient, Visit, and Cervigram, and all the Attributes associated with each of these Entities. The model also knows there can be many Visits for one Patient and multiple Cervigrams for one Visit. None of this information is embedded in the code – it is all constructed *dynamically* from the database.

Abstracting the system data at this high level allows the MDT to incorporate many different datasets without code changes. Once the MDT is fully implemented, it is intended to be used "as is" for a wide range of multimedia datasets. The effort in incorporating new datasets will be shifted from reprogramming software to the initial creation of the required database tables, which currently is largely a manual process. In order to streamline the process of initially constructing the database and subsequently adding new data, additional MDT database administration tools are planned. The goal is to support the incorporation of a wide set of data domains, without reprogramming, and with only database administrators creating and maintaining the system.

Once the system model is received from the servlet by the client (Figure 1 – flow #4), it has enough information for the user to construct a query. As described previously, the system

Attributes are presented in the form of a Taxonomy tree from which the user graphically selects the data items to return. In effect, the user creates an SQL statement by graphical manipulation. However, the client program has no "understanding" of SQL; it works instead with an abstract *Statement* object, which it provides to the servlet; the Statement object contains all the necessary ingredients to query the database. In this way, the client is completely insulated from dependencies on both the database and from dependencies on SQL, and the user thus needs no SQL knowledge.

The client sends the Statement to the servlet, again via XML (`Figure 1` - flow #5). The servlet parses the Statement and creates the SQL syntax required by the DBMS. The results of this query (from the DBMS) are then packaged into XML and returned to the client (`Figure 1` – flow #6), along with references to the image files. The client displays the query results in tabular form and downloads the images via HTTP (`Figure 1` – flow #7). Multiple images are displayed in a horizontal view, and are dynamically linked with their associated query data (`Figure 4`).

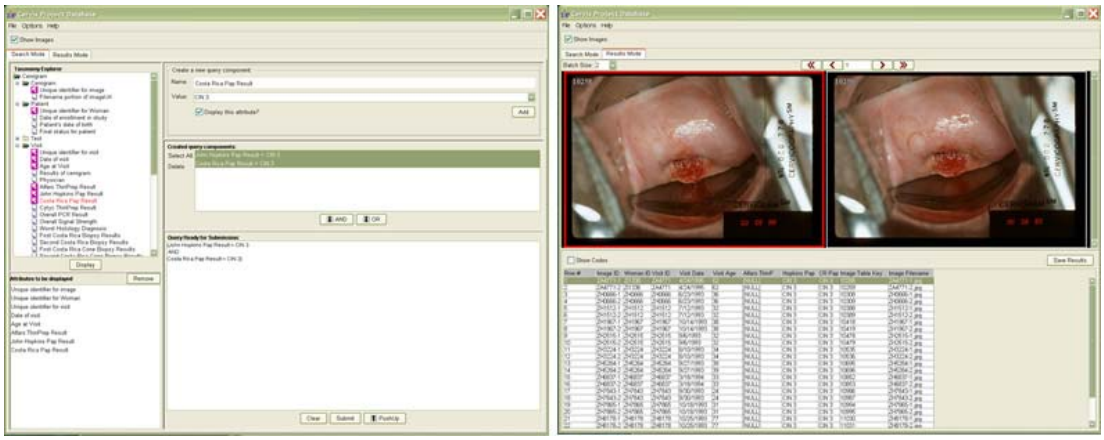

Figure 4 MDT Query Screen (left) and Results Screen (right)

It is important to note that very few assumptions have been made about the images in the code. Our goal is to be able to support a broad class of image types. For some domains, it may be necessary to compress the images in order to produce reasonable retrieval performance, particularly for images retrieved from remote servers. One area we are pursuing is in designing methods to allow domain-specific or user-specific display of different image types. For example, the NCI Uterine Cervix domain will have at least three different images types: cervigrams, colposcopy images, and histology slides. The user may want to see only cervigrams for a particular patient at first, but then switch to a view that displays the three types side-by-side. This type of display may not be appropriate or desired for another domain, however. We are investigating how best to encode these preferences without writing domain-specific code.

Also under investigation is providing different user views into the data results. One example from the NCI Uterine Cervix domain is the capability to *switch to a patient-centric view within query results that were returned for a general, non-patient-centric query*. In this case, the user starts with an initial query and begins to browse through the data and images returned. Once an "interesting" patient is identified, the user then "zooms in" on that particular patient, and easily navigates through data that belongs to that patient and that patient only. We would like to provide a simple GUI mechanism where the user could toggle between this patient-centric view and the general view seamlessly. The challenge is to provide this type of specialized display functionality without incorporating domain-specific code into the MDT.

4. Biomedical significance of the MDT uterine cervix database

This MDT is an exploratory tool that allows researchers to retrieve visual and text data according to specific characteristics such as age, pregnancies, HPV status, viral load, and test results. The MDT identifies the patients who match specified characteristics, and displays their corresponding image and clinical (text) data. Since all of the visual and text data can be accessed through the World Wide Web, this tool enables work among many experts and researchers worldwide, providing researchers in the NCI Division of Cancer Epidemiology and Genetics additional opportunities for collaborative studies.

5. Planned research use of the MDT in advanced imaging

The MDT enables categorization, management, and retrieval of expert-marked biomedical data, including images that are associated with collateral text information. Researchers, educators, and students can retrieve pertinent records with or without images, and study correlations between the text and image data. However, correlation between the pathologies or features of interest in these images is possible only through individual examination, a cumbersome and daunting task. Our planned goal is to take advantage of our continued research in image analysis techniques to augment the image handling capabilities of the MDT, and thereby reduce the manual effort involved in this task.

6. Summary

The Multimedia Database Tool provides important flexibility to text/image database systems to incorporate new datasets with effort required only at the database administrator level, as opposed to potentially high labor costs at the programming level. This new capability is expected to provide a method for streamlining the movement of biomedical image databases to the Web. The first application of the MDT to a biomedical dataset incorporates a large collection of National Cancer Institute uterine cervix images and related longitudinal data that is expected to make a significant contribution to the study of visual precursors of cervical cancer.

References

[1] Jeronimo J, Schiffman M, Long LR, Neve L, Antani SK. A Tool for Collection of Region Based Data from Uterine Cervix Images for Correlation of Visual and Clinical Variables Related to Cervical Neoplasia. Submitted to IEEE CBMS 2004

[2] Long LR, Pillemer SR, Lawrence RC, Goh G-H, Neve L, Thoma GR. World Wide Web platform-independent access to biomedical text/image databases. *Proceedings of SPIE Medical Imaging 1998: PACS Design and Evaluation: Engineering and Clinical Issues,* SPIE Vol. 3339, San Diego, CA, February 21-26, 1998, pp. 52-63.