THE LISTER HILL NATIONAL CENTER
FOR BIOMEDICAL COMMUNICATIONS

A research division of the U.S. National Library of Medicine

TECHNICAL REPORT
LHNCBC-TR-2008-004

The Lister Hill National Center for Biomedical Communications Annual Report FY2008

Clement J. McDonald, M.D.
Director

U.S. National Library of Medicine, LHNCBC
8600 Rockville Pike, Building 38A
Bethesda, MD 20894

LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS FY2008 ANNUAL REPORT

Clement J. McDonald, M.D.
Director

The Lister Hill National Center for Biomedical Communications (LHNCBC), established by a joint resolution of the United States Congress in 1968, is a research and development division of the NLM. The Center continues its active research and development, seeking to improve access to high quality biomedical information for individuals around the world. It leads a research and development program aimed at creating and improving biomedical communications systems, methods, technologies, and networks and enhancing information dissemination and utilization among health professionals, patients, and the general public. An important new focus of the LHNCBC is the development of Next Generation electronic health records to facilitate patient-centric care, clinical research, and public health, an area of emphasis in the new NLM Long Range Plan 2006-2016.

The Lister Hill Center research staff is drawn from a variety of disciplines including medicine, computer science, library and information science, linguistics, engineering, and education. Research projects are generally conducted by teams of individuals of varying backgrounds and often involve collaboration with other divisions of the NLM, other institutes at the NIH, other organizations within the Department of Health and Human Services, and academic and industry partners. Staff regularly publish their research results in the medical informatics, computer and information science, and engineering communities. The Center is visited by researchers from around the world.

The Lister Hill Center is organized into five major components: Cognitive Science Branch (CgSB), Communications Engineering Branch (CEB), Computer Science Branch (CSB), Audiovisual Program Development Branch (APDB), and the Office of High Performance Computing and Communications (OHPCC).

An external Board of Scientific Counselors meets biannually to review the Center's research projects and priorities. The most current information about the Lister Hill Center research activities can be found at http://lhncbc.nlm.nih.gov/. The Center's principal research activities and accomplishments are described in the remainder of this chapter.

NEXT GENERATION ELECTRONIC HEALTH RECORDS TO FACILITATE PATIENT-CENTRIC CARE, CLINICAL RESEARCH, AND PUBLIC HEALTH

These projects are efforts to target the overall recommendations of the NLM Long Range Plan (LRP) Goal 3: *Integrated Biomedical, Clinical, and Public Health Information Systems that Promote Scientific Discovery and Speed the Translation of Research into Practice.*

NLM Personal Health Record (PHR)

A Personal Health Record (PHR) is an electronic medical record whose contents are controlled and managed by the person whose data it carries. Having attracted much press attention in recent years, the US IT industry, health industry, and federal government envision the PHR as a possible solution to the information sharing and efficiency problems in health care.

The NLM has embarked on the development and deployment of a PHR in order to study and improve their utility, reduce barriers to their adoption, identify best practices, and provide a platform and test bed for advanced PHR applications. The development of the NLM PHR is based on a set of existing health care message and vocabulary standards that have either been developed by, or are supported by, the NLM and for the most part are also part of the accepted standards of the Secretary of HHS.

The PHR is a pure web application. It is based on a forms generator (developed by NLM that produces web input forms on the fly. These forms include built-in skip logic (e.g. if the person is male, it does not show the question about pregnancy history), performs edit-checks and auto-completion of user entries. It uses AJAX server access techniques, so the response times are fast. The PHR uses Ruby on Rails (an open source web management system) as the software for the application server and MySQL or Oracle as the database server. The PHR makes heavy use of JavaScript on the client in order to approximate the speed and capabilities of a desk top application. It also borrows heavily from the JavaScript open source world using Scriptaculous and Dojo.

The PHR provides tools for managing clinical information by different family members, so, e.g., a mother can maintain immunization records for each of her children and/or keep track of her ailing father's medications. It provides for the recording of medications, medical problems, allergies, surgeries and implants, immunizations (vaccines), measurements such as blood pressure and laboratory results (e.g. serum glucose), and questions to remember to ask one's care provider.

A key feature of the PHR is its emphasis on the coding of the names of medications, drugs, problems, surgeries, immunizations, and allergies that a user enters. When a user types in one of these names, the PHR presents a menu of the names that "match" with what the user has typed so far. The menu gets shorter with each additional key stroke. The user can either select from this short menu or keep typing until there is only one choice. The computer matches the input string to any word in its list of vocabulary entries so it can find heart failure whether the user types in "heart" or "failure." It also has synonyms for its internal vocabulary to account for the many ways that people name things. So it will find "heart attack" whether one types in "heart attack" or "Myocardial infarction."

The computer encodes the user's entries into NLM-supported coded vocabularies: RxNorm for drugs, LOINC for measurements and laboratory tests, CDC's vaccine codes for vaccines and SNOMED CT for problems. To ease the entry of medication records, we developed a special subset of RxNorm, called RxTerms. This vocabulary is used by the Centers for Medicare and Medicaid Services (CMS) for their post-acute care project and by the PHR. The PHR borrows its data type and much of its data structure from HL7.

The orientation toward automatic coding enables the PHR to provide two special capabilities: decision support and one-click access to information recorded in the record. Decision support in the PHR is based on predefined rules that the patient conditions to care recommendations. The PHR compares the patient's data against rules for preventive care and reminds the individual about interventions such as mammograms or colonoscopy which are due.

The PHR provides one-click links to information about most of the items (drugs, problems, medications, allergies surgeries) that the user records on the form. This content comes from Medline plus, CDC, the American Academy of Allergy Asthma and Immunology, and AHQR, depending upon the topic. When the PHR suggest preventive or other care, it also provides one-click context-sensitive information from the US preventive Task Force web site.

The user can print a paper copy or download an electronic copy of their PHR (to their local computer or to a thumb drive). Currently, the electronic version is delivered as a spreadsheet in which separate tabs are dedicated problems, medications, allergies, tests, etc. The spreadsheet will provide text guidance about how to best keep the information private. The spreadsheet format provides an easy way for the users to review the contents on their own machine in a familiar media. We have plans to provide other export formats in the future.

We have invested substantial work in the development of mechanisms for user registration, management of passwords and user Ids and logging on. We do this without asking for user name, address or the other usual identifying information. We have also expanded the database support to both Oracle (new support) and MySQL. We moved to Oracle in order to obtain the levels of encryption and security required by NIH rules.

This young project addresses the longstanding NLM interest of facilitating health care management and is part of the NLM strategic plan. It will help tune the message and vocabulary standards that NLM has supported and also provide another consumer entry point to a rich trove of patient-oriented data. Early research projects will focus on user needs, usability, and usage patterns to guide the next round of development and research.

De-identification Tools

De-identification can unlock the research potential of long term clinical records. No well-supported and freely available de-identification tools exist. Taking advantage of past efforts and experience with de-identified procedures (NCI Shared Pathology Informatics Network (SPIN) grant) and existing Lister Hill Center tools that can recognize sensitive content such as dates, person names, locations, and numeric identifiers, LHC researchers initiated an effort to develop an open source text de-identification tool.

The system uses more than 700,000 clinical records from the Clinical Center for testing and validating current work (under IRB exemptions). Currently, developers are identifying and scrubbing sensitive information in the clinical text and labeling these items by type (e.g. personal name and postal address, etc.). To accomplish the task, the system utilizes several data source, including databases of first and last names along with frequency counts derived from the Social Security database of 480 million persons. U.S. street names and information about cities, states, and zip codes were obtained from the U.S. Postal Service and 2000 U.S. Census databases. Statistical information about the usage of clinical words, common English words, and word co-occurrences have been extracted from multi-billion word corpora such as Wikipedia and Core clinical journal article abstracts.

Developers presented preliminary results to the LHNCBC Board of Scientific Counselors in April 2008. They plan to release a beta version of the software package for evaluation in the next few months and a complete version of the software in 2009.

Clinic Database Research and Development

As part of our medical record research and development we have embarked on the creation of a general purpose longitudinal database structure that could be used for many purposes, e.g.:

a) as a major part of the data base for an open source EHR as has been proposed in some pending legislation
b) as the structure for the clinical data that might be needed for an HER oriented to disasters
c) as part of a systems for statistical analysis of de-identified longitudinal research data

We have used a number of sources to inform this data structure. We have looked at the structure of a number of longitudinal data models. These have been especially instructive:

a) the data model employed by the HL7 version 2 message segments
b) the Regenstrief SPIN data model (1 Billion results)
c) the Women's Health Initiative (WHI) database for the data collected over decades (500+ million results)

We have also obtained a de-identified clinical data set from a University (under a restricted use MOU). This data set carries clinical data for over 15,000 ICU encounters, including laboratory, data, radiology reports, discharge summaries and electrophysiological data – more than 100 million individual records, in toto. This data set has the complexity and size we need to test and tune the database design and it includes clinical content rich enough to test the utility of applications that we might attach to the database statistical analysis tool.

We have successfully loaded his de-identified clinical data into our early stage database system – achieving a major goal for this year. We have also developed a linkage to the R statistical system. R is an open source analysis system that encapsulates very sophisticated statistical knowledge. The current interface will perform simple statistical analysis and data manipulation. R has powerful facilities for analyzing and graphing data. One of our goals is to experiment with innovative ways to present and analyze an individual patients' data. But it may also provide a greater window into research databases.

Clinical Data Entry Tools

The initial goal of this project is to develop a tool that can generate data entry forms dynamically based on specifications stored in a database. The development platform is Ruby on Rails, an open-source web application framework. Developers are using this tool in the data capture function of personal health records. They are also using several terminology resources from the UMLS (e.g. RxNORM, ICD9-CM) in data entry fields that require a set of controlled terms. Further development will involve work with very large databases of de-identified patient data. The goal is to create additional reusable software tools, some of which will involve biostatistical analysis with the "R" package.

Collaboration with Centers for Medicare and Medicaid Services (CMS)

LHC has assisted CMS in the development of many aspects of Medicare's Post Acute Care data collection project demonstrated in the spring of 2008. One of Medicare's goals in this project is to standardize and meld different data collection forms from four different post-acute care settings. LHC proposed and demonstrated a Web approach to auto-completion of entered text, much of which has been adopted by Medicare for this demonstration project. LHC has proposed and delivered the full content of many of the look-up tables that Medicare will use in this project, including the RxTerms database, now in the third release, and shaped their data conceptualization to fit a LOINC/HL7 model. A Memorandum of Understanding (MOU) to formalize this collaboration is now in final review stages within CMS.

Concept Recognition in Narrative Clinical Reports

Through collaborations with a major university and the NIH Clinical Center the LHC Natural Language Processing, researchers have gained access to over 400,000 progress notes, 15,380 discharge summaries, and 178,126 radiology reports which are de-identified. These resources are instrumental in the ongoing effort to identify the clinical concepts in a narrative clinical report and map them into UMLS Concept Unique Identifiers (CUIs). Clinical text presents the additional challenge of dealing with grammatically incomplete sentences, significant numbers of acronyms and abbreviations (often idiosyncratic to a particular clinical unit), and distinguishing positive from negative statements about a finding, disease or symptom. Tools for finding important clinical concepts in narrative records would have widespread use in quality assurance, clinical research, and decision support, with the right level of sensitivity and specificity. The NLM has existing tools for converting strings found in text into concepts that can facilitate this effort.

Standards for Identifying Clinical Observations, Forms and Panels

As part of an effort to ease the mapping of local laboratory systems to a universal standard (LOINC) in collaboration with other NLM divisions and our contractors, we have identified a sample of the 800+ most commonly reported test results and embedded them in an application that lets users map entries in their lab to the LOINC database. We have also produced a guide for mapping the tests that HL7_eLINCs chose as the most important. They have used that guide to shape their standard but did not incorporate it into their standard.

To make LOINC terms more accessible and directly understandable to users in the same collaboration, we have developed tools that systematically convert the formal names to short names and to long common names. The long common names are a brand new development that produces names that replace the formal names with conventional names, e.g. Prothrombin time instead of the more formal "Coagulation surface induced." It also eliminates parts of the formal name that are not distinguishing in context and not part of common usage, and includes prepositions and connectors to make them more readable, e.g. Prothrombin time in blood by coagulation method. These will be distributed in the next release and encompass at least 90% of the laboratory tests.

In collaboration with Office of the National Coordinator for Health Information Technology (ONC) and the Healthcare Information Technology Standards Panel (HITSP), we engaged in an effort to create LOINC terms and standardization for two clinical areas of high priority to the secretary of HHS. The first is prenatal screening with a major focus on the reporting of newborn screen laboratory results, especially those produced by mass spectrophotometry testing. We produced a set of LOINC terms and panels to accommodate the reporting of quantitative values for all newborn screening laboratory tests. We have also begun work on newborn screening for hearing loss.

The second of these efforts was the development of an approach to the reporting of clinical genetics studies. This was a collaborative effort with HTSP, Partners of Boston, Intermountain Healthcare, HL7, and NCBI. It resulted in the creation of 12 LOINC panels and about 100 new LOINC terms and a balloted HL7 implementation guide that describes exactly how to report the results of such studies. The approach is a general one that uses a repeating series of panels, one per region of interest, one per genetic marker noted, and a few others to carry the full information in the report.

The LOINC vocabulary system continues to grow in usage and size. More than 3000 new terms have been added in the last year. The LOINC mapping tool now supports both search and display of those non-English languages that have been submitted (Spanish, Chinese and French).

BIOMEDICAL IMAGING AND MULTIMEDIA

The overall goal of this major research area is to address fundamental questions that arise in the handling, organization, storage, access and transmission of very large electronic files in general and digitized biomedical images in particular. A special focus is research into these topics as applied to heterogeneous multimedia databases consisting of both images and text. Projects in this area have benefited from collaborators in several universities as well as at agencies such as the National Center for Health Statistics (NCHS) and the National Institute of Arthritis, Musculoskeletal and Skin Diseases (NIAMS), and a continuing partnership with the National Cancer Institute (NCI) in their research in cervical cancer caused by the Human Papillomavirus (HPV).

Interactive Publications Research (IPR)

The IPR project is in line with *Recommendation 13: Illuminate the value of multimedia-rich interactive publications ... to a broad readership that would reuse the media content for analysis.* This project intends to demonstrate a type of highly interactive multimedia document that could serve as a model for next-generation publishing in biomedicine. The project focuses on the standards, formats, authoring and reading tools necessary for the creation and use of such interactive publications (IP) containing many media objects relevant to the biomedical literature: text, video, audio, bitmapped images, interactive tables and graphs, and clinical images such as x-rays, CT, MRI, and ultrasound. These objects are in a wide range of file formats: e.g., text in MS Word or PDF, animations in Flash, spreadsheets in Excel and clinical images in the DICOM format. Following a definition of authoring procedures, we have created and demonstrated two prototype documents. One focuses mainly on still and moving clinical images and animations, and the other focuses on large stores of tabular research data subject to statistical analysis.

The LHC has developed the first version of an IP viewer named Panorama, an Eclipse-based Java module, which can present different kinds of multimedia, tables, graphs, and image modalities in multiple panes. A reader may easily move among them, or perform analysis on a graph in one pane and display results in tabular form in another, for example. A component of Panorama is **iTAG**, a tool for selecting data subsets, creating graphs and charts from tables (and the reverse), performing basic statistical analysis, and exporting data in formats required by SAS or R for further analysis. In 2009, developers will enhance Panorama with additional features, such as volume rendering from 2D clinical images.

Research in FY2009 will also address a serious challenge to the network delivery of IP arising from their large file size, conceivably in the hundreds of megabytes. This research will seek strategies to overcome the barriers posed by network bandwidth limitations in the delivery of IPs, e.g., progressive transmission in which small text files are delivered for immediate reading, while larger media files are transmitted in the background. In addition, we will develop **Forge**, an authoring tool in Java. Forge will enable authors to create multimedia-rich interactive articles using a set of wizards while permitting them to write content in conventional word processors, e.g., MS Word or LaTeX.

Developers are also working with the Optical Society of America (OSA) to produce four interactive issues of regularly published OSA journals during the next year. The articles in each of the special interactive issues will contain image data sets with which the reader can interact. The journals will be accompanied by a free interactive publication reader similar to the free Adobe PDF Reader.

Multimedia Database R&D

Goals of this project are: (1) to research latest technological approaches for information retrieval and delivery for biomedical databases that include non-text data, with an emphasis on biomedical images, and (2) to develop prototype systems for the retrieval and delivery of such information for use by the research and, potentially, the clinical communities.

These tools are under development as part of the collaboration with NCI. This suite of tools earned the Internet2 IDEA award in Spring 2008. The tools include the following:

Multimedia Database Tool (MDT)

The **Multimedia Database Tool (MDT)** accommodates the new text/image database currently being created for the collection of 100,000 uterine cervix images from NCI, but also existing databases of x-rays and associated text from NHANES. It has a flexible database schema and GUI that allows new databases to be incorporated easily by a database administrator without software modification. The system allows not only data dissemination but also distributed (remote) data collection.

Developers plan to add capabilities to the **Multimedia Database Tool (MDT)** for users to write to the database, add levels of user privileges for user access, and add support for flexible schema and for simultaneous display of multiple images related to a single subject.

In FY2009, research will be extended beyond the uterine cervix images to other NCI image sets, e.g., histology images. Techniques are to be developed to support the dissemination of these very large files using image "tiling" approaches.

Boundary Marking Tool (BMT)

The **Boundary Marking Tool (BMT)** is a Web-accessible system for the accurate and efficient collection of descriptive data and boundary data for specific structures of biomedical interest in digitized uterine cervix images. It is in active use by NCI for about a dozen studies, e.g., to select biopsy sites in colposcopic images. The BMT Study Administration Tool (BSAT) is a recent addition that allows users to create their own studies by uploading their own images and configuring the screens to collect boundary data particular to their studies.

Virtual Microscope (VM)

The **Virtual Microscope (VM)**, currently at an intermediate level of development, will provide Web capability to view and collect information on histology images from expert observers. There are both a simple demo system of basic histology image and data collection capability and a fully-functional system, currently being used to support a multiple-observer study of lung histology images, in collaboration with the NCI Genetic Epidemiology Branch, the NCI Cell and Cancer Biology Branch, and their medical collaborators in Italy.

V**irtual Microscope (VM) and Virtual Slides (VS)** are an archive of virtual slides has been developed from the teaching set of glass slides from the Department of Pathology of the Uniformed Services University and other collaborating institutions. An entire slide is digitized, segmented and processed to simulate an examination of a glass slide under the microscope but with a Web browser. The collection preserves the specimen for posterity and allows viewing by users worldwide anytime. Annotations and automatic linking to Medline/PubMed is planned. A related collection of images from the AFIP fascicles allows users to search images and automatically link to Medline citations.

Teaching Tool (TT)

This system is for training medical personnel in cervix anatomy/pathology. It displays uterine cervix images and quizzes an observer in the categories of medical knowledge, pattern recognition, and patient management, and enables a medical expert to tailor exams by specifying images and questions to use on an examination. A prototype system is available for experimentation by NCI and American Society for Cervical Pathology and Colposcopy (ASCCP) experts, which includes capability to administer and score the ASCCP "Resident's Online Exam."

Visual Triage Study (VTS) Software

The NCI Visual Triage Study concluded in FY2008. This study had the objective of estimating the effectiveness of “screen-and-treat” cervical cancer prevention programs using HPV testing and cryotherapy. The study used image and diagnostic data collected from 552 HPV-positive women in the NCI Guanacaste Project, and had three phases.

First, using both image and diagnostic data collected during the Guanacaste work, two expert gynecologists independently determined, whether the women were treatable (at the time of enrollment as participants in the Guanacaste Project) for HPV infection by cryotherapy, or whether they required referral for advanced treatment; these expert opinions were then reconciled into the final “truth” standard for the Visual Triage Study. Second, 12 midwives in Peru performed a simulated visual triage on each of the 552 women by viewing images for each of them and judging, based on visual criteria, whether a woman was treatable by cryotherapy or required referral for advanced treatment. The work of the midwives was carried out at Internet cafes in the Amazon region of Peru. Finally, 5 gynecologists, located in Costa Rica, Ghana, Peru, and Thailand, repeated the triage carried out by the midwives. The cumulative data was used to compare the effectiveness of the midwives, and of the gynecologists, in assessing treatability by visual methods. The conclusion of the study was that the performance of the visual triage was suboptimal, and that such screen-and-treat programs might be ineffective. The study was used for a major part of the successful Ph.D. dissertation in epidemiology of Julia C. Gage of NCI, and has been submitted for publication to the *International Journal of Gynecology and Cancer*, with the title, “An Evaluation by Midwives and Gynecologists of Treatability of Cervical Lesions by Cryotherapy Among HPV-Positive Women.”

The Visual Triage Study software developed by LHNCBC staff used a Web-browser interface to display both images and related text, and to collect responses from the study observers (midwives and gynecologists). Images were provided in a range of pixel sizes so that the physical images appeared on all monitors at a uniform size, regardless of the monitor’s screen size. This was done to simulate the appearance of the cervix area when viewed through a colposcope. Secondary, larger versions of the images were also provided to simulate the appearance of the image after magnification. For most of the study participants, images were accessed in real-time over the Internet. For two of the very-low-bandwidth participants, images were provided on DVD; the Visual Triage Study Software was written to accommodate both options.

Other tools include the following:

WebMIRS (Web-based Medical Information Retrieval System)

Developed some years ago and still in active use, WebMIRS continues to provide access to spinal x-ray images and associated text from nationwide surveys conducted by the National Center for Health Statistics.

At present, there are 568 users of WebMIRS in 58 countries. 43% of the users are from academia, the rest from government agencies, corporate and medical organizations. More than half the users are in the U.S., but relatively high numbers are from Canada, U.K., India and China. This Java application allows remote users to access data from the National Health and Nutrition Examination Surveys II and III (NHANES II and III). The NHANES II database contains records for about 20,000 individuals, with about 2,000 fields per record; the NHANES III database contains records for about 30,000 individuals, with more than 3,000 fields per record. In addition, 550 of the 17,000 x-ray images collected in NHANES II contain vertebral boundary data collected by a board-certified radiologist. Users may do queries for both radiological and/or health survey data.

LHNCBC
FY 2008 ANNUAL REPORT

Digital Atlas of the Cervical and Lumbar Spine

The Digital Atlas remains available for the public from the CEB Web site either as a Java applet or a downloaded Java application or as a CD version of the Java application. The Java application version allows the user to add images (either grayscale or color) in a special “My Images” section, and to annotate and title those images for later use.

Content Based Image Retrieval (CBIR)

We continue our research in **Content Based Image Retrieval (CBIR)**, using text descriptors, image examples, sketches, or combinations of these to create queries. As part of this work, we will incorporate user-relevant feedback and image indexing trees to enhance the accuracy of image retrieval in CBIR systems. Based on this research, a next-generation prototype CBIR system will be developed with capability of demonstrating *image retrieval by content* to professional biomedical groups for initiating collaborative work with biomedical subject matter experts. The goal is to validate the image data and to acquire technical critiques of the usefulness of our design approaches for biomedical research and/or clinical practice. CBIR techniques are also used in the **Image Text Indexing** project in which the illustrations in medical articles are indexed by processing figure captions and mentions in the text using natural language processing techniques, as well as by image features in the illustrations.

In FY2009, research will continue into computer-assisted image segmentation using efficient and accurate manual boundary point marking, Active Contours, Active Shape Modeling, the Generalized Hough Transform, Active Appearance Models, Live Wire, Level Sets, and general deformable template methods. We will test these methods focusing on the NHANES II x-ray images and on other image collections and modalities, such as digitized color images of the uterine cervix. Developers are working to improve the accuracy of these methods, integrating them into a practical system for shape segmentation, and making them useful for efficient production-level segmentation of large collections of images by shape. These functions are being incorporated into the new Web-based **Spine Pathology and Image Retrieval System (SPIRS)**.

In FY2009, research will continue toward establishing a gold standard for evaluating segmentation algorithms, as they relate to both the spine x-rays and NCI uterine cervix images, that will be useful to the biomedical community.

The Visible Human Project

The Visible Human Project image data sets are designed to serve as a common reference for the study of human anatomy, as a set of common public domain data for testing medical imaging algorithms, and as a test bed and model for the construction of image libraries that can be accessed through networks. The Visible Human data sets are available through a free license agreement with the NLM. They are distributed to licensees over the Internet at no cost; and on DAT tape for a duplication fee. The data sets are being applied to a wide range of educational, diagnostic, treatment planning, virtual reality, and virtual surgeries, in addition to artistic, mathematical, legal, and industrial uses by over 2,450 licensees in 49 countries. The Visible Human Project has been featured in more than 900 newspaper articles, news and science magazines, and radio and television programs worldwide.

saw the continued maintenance of two databases to record information about Visible Human Project use. The first, to log information about the license holders and record statements of their intended use of the images; and the second, to record information about the products the licensees are providing NLM in compliance with the Visible Human Dataset License Agreement.

In FY2007, a planning workshop, “VHP: Scope and Scale for the Future,” assembled an expert panel of radiologists, anatomists, pathologists, computer scientists, and engineers from across the country to advise NLM on future directions for the Visible Human Project. Topics such as human variation, community data annotation, algorithm validation, and multiscale anatomy emerged as leading areas of interest. Based on these findings, a project entitled **A Knowledgebase of Human Variation** was started during . The goal is to build a database of the parameters and variances which would define the normal range of all human anatomical structures and the dependencies and covariances between them. The initial data will be gleaned from the existing anatomical literature. The database will then be confirmed using data obtained from radiological scans of normal human anatomy. A Web 2.0 paradigm, inviting the participation of the interested community, will be used to glean and collect the needed data.

3D Informatics

The 3D Informatics Program has expanded research efforts concerning problems encountered in the world of 3-dimensional and higher-dimensional, time-varying imaging. Among its many projects, the 3D Informatics (TDI) Group has continued work on image databases, including ongoing support for the National Online Volumetric Archive (NOVA), an archive of volume image data. This collection contains 3D data from across medicine. Contributors to the collection include the Mayo Clinic Biomedical Imaging Resource and the Walter Reed Army Medical Center Radiology Department. Integrated and multimodal data such as virtual colonoscopy matched with recorded video from endoscopic interventions, time-varying 3D cardiac motion, and 4D MRI of a human hand appear in the archive.

The 3D Informatics group continues its partnership with the NLM Specialized Information Systems Division and the U.S. Veterans Administration to study content-based retrieval methods for medical image databases. In the pharmaceutical identification project, we are assisting in the acquisition of imagery through digital macro-photography of the thousands of prescription pharmaceuticals dispensed routinely by the VA Centralized Mail-Order Pharmacies. Together we are creating a new, updated, visual database of all these products and developing techniques for automatically identifying any product in the inventory from a representative photograph. New OHPCC research has developed computer vision approaches for the automatic segmentation, measurement, and analysis of solid-dose medications. In particular, recent focus has been on robust color classification tools to help identify prescription drugs.

Beyond data collection, this group is engaged in aggressive image analysis programs in partnership with the High Resolution Microscopy Laboratory at the National Cancer Institute. 3D Informatics staff members are developing new techniques using the **Insight Toolkit** from the Visible Human Project to study sub-cellular structure from 3D dual-beam scanning electron micrographs and 3D transmission-electron tomograms of cultured cells exhibiting critical pathologies such as melanoma and HIV-AIDS. This effort includes analysis as well as visualization and rendering of complex microscopy data, integrating the HPCC facilities in 3D printing to support intramural, trans-NIH research.

3D Telepresence for Medical Consultation

Completed in 2008, this project involved testing the efficacy of 2D versus 3D representations of video data transmitted in real time in remote clinical consultations. The technology infrastructure continues to be developed at the University of North Carolina and its efficacy continues to be researched there with help from colleagues at other institutions. The research team made substantial progress in implementing the technology infrastructure. A prototype portable camera unit was added to the stationary one and calibrated. The PDA application was completed and all the basic components of the system proposed are in place. Of special concern were optimizing camera and sensor placement, refining calibration and rendering algorithms, and dealing with problems when perspective changes from different points of view, such as occlusion when an intervening object obstructs the view of interest. Real time 3D video is a very difficult problem to solve and, while the team did not solve the problem completely, substantial advances were made. Initial programs rendered the video as computer graphics while the later renderings were of video quality. Moreover, the team completed research comparing 2D video with a 3D proxy condition where the consulting ER physicians advised first responders at a distance via high quality 2D video or in the same room, but in constrained conditions mimicking 3D video communication. The consulting doctors could move to get different views and could point with a laser pen, but could otherwise only talk to the responder. Responders made 11 critical life threatening errors when unassisted, 7 when given advice via 2D video and only one in the proxy condition. In addition, they had greater self-efficacy and confidence in performing the airway task. Consultants had to ask fewer questions because the view was less obstructed, allowing responders to concentrate more on the task. The higher level of questioning in the 2D context made many responders feel the consultant lacked confidence in their performance and they had less self-efficacy as a result.

Advanced Network Infrastructure for Distributed Learning and Collaborative Research

This project built on previous work with HAVnet (Haptic Audio Visual Network for Educational Technology) and was collaboration between Stanford University and the University of Wisconsin at La Cross. The project was completed in 2008 and focused on developing visual and haptic applications for anatomy and surgical training and included aspects of self scaling technology, self-optimizing end-to-end, network aware, real time middleware, wireless technology, and GIS. The technology was developed and refined in the context of teaching anatomy and surgical skills and addressed issues concerning network bandwidth and latency and the integration of 3D visualization, haptic, and real time online collaboration tools. The project delivered enhancement and integration of two existing middleware applications, Information Channels and Weather Stations, allowing correlations to be made between network metrics and actual application performance; addition of self-optimizing features to the six applications using the core middleware; development of a new application, Anatomy Window, that uses a handheld computer to map a cadaver and present corresponding images derived from the Visible Human data set; development of a Remote Tactile Sensor, capable of capture and transmission of tactile dermatology information over a network; implementation of the anatomy teaching suite over local, national and global networks for use in early, laboratory based and actual field teaching; and implementation of the clinical skills test bed, primarily in early phase and laboratory testing.

Work on the remote stereo viewer and haptic probe was completed that suggested videoconferencing is essential for dermatologists to see and communicate with patients while using the haptic device. Research was conducted on sense of touch and ability to detect thickness and resistance of membranes as part of the effort and to compare this feedback to haptic feedback generated by computer. The SPRING surgical simulator engine and its Remote Tactile Sensor component were made open source. The engine allows building of software modules providing haptic feedback for simulated surgical tools. The testing on many of the visualization tools was done as part of an iAnatomy collaboration with the Northern Ontario School of Medicine involving the use of the stereo viewer for anatomy teaching and distance learning.

Insight Tool Kit (ITK)

The Insight Toolkit, a research and development initiative under the Visible Human Project, is now in its seventh year with a recent official software release of ITK 3.8 in July 2008. ITK makes available a variety of open source image processing algorithms for computing segmentation and registration of high dimensional medical data on a variety of hardware platforms. Platforms currently supported are PCs running Visual C++, Sun Workstations running the GNU C++ compiler, SGI workstations, Linux based systems and Mac OS-X. Support, development, and maintenance of the software are managed by a community of university and commercial groups, including OHPCC intramural research staff. The Insight Toolkit continues to have an impact on the medical imaging research community. Researchers are testing, developing, and contributing to ITK in more than 40 countries, with more than 1500 active subscribers to the global mailing list for the project. ITK is an essential part of the software infrastructure of such projects as Osirix, an open-source diagnostic radiological image viewing system available from a research partnership between UCLA and the University of Geneva. ITK is also having an impact in other scientific fields, appearing in the Orfeo Toolbox (OTB) from the Centre Nationale D'Etudes Spatiales, the French National Space Administration. ORFEO Toolbox (OTB) is distributed as an open source library of image processing algorithms. OTB is based on the medical image processing library ITK and offers particular functionalities for remote sensing image processing in general and for high spatial resolution images in particular.

Across NIH, ITK is providing a foundation for new imaging investigations. The National Alliance of Medical Image Computing (NA-MIC), an NIH Roadmap National Center for Biomedical Computing (NCBC), has adopted ITK and its software engineering practices as part of its engineering infrastructure. Staff members participate as science officers and lead science officer for the NIH-Roadmap for the NA-MIC consortium.

ITK also serves as the software foundation for the Image Guided Surgery Toolkit (IGSTK), a research and development program sponsored by the NIH National Institute for Biomedical Imaging and Bioengineering (NIBIB) and executed by Georgetown University's Imaging Science and Information Systems (ISIS) Center. IGSTK is pioneering an open API for integrating robotics, image-guidance, image analysis, and surgical intervention. The external advisory board for IGSTK includes members of the Lister Hill staff.

From 2002 to 2008, approximately 20 purchase orders were awarded for reference data sets and enhanced algorithms to support the further development of ITK. This effort supported the integration of ITK into research platforms such as the Analyze from the Mayo Clinic, SCIRun from the University of Utah's Scientific Computing and Imaging Institute, and the development of a new release of VolView, free software for medical volume image viewing and analysis. At the current time, the Optical Society of America is adopting VolView as their free 3D data viewing software as part of a joint NLM/OSA project in Interactive Publication, intended to distribute open access journals accompanied by open data. Among the data acquisitions for NLM, the Mayo Clinic Biomedial Imaging Resource has provided over 100 datasets collected across dozens of animals and clinical cases representing a wide cross section of anatomy, pathology, modality, and pre- and post-operative clinical conditions.

Image and Text Indexing for Clinical Decision Support

The title of a publication is not always sufficient in determining the Evidence-Based Practice (EBP) relevance of a publication. Given that medical illustrations often convey essential information in compact form, this project seeks to automatically identify illustrations from the articles that could help clinicians evaluate the potential usefulness of a publication in a clinical situation. We explored feasibility of automatic image annotation by utility for EBP, and if such images can be reliably extracted from the original articles.

Our study showed that images presented in clinical journals can be successfully annotated by their usefulness in finding evidence to assist a clinical decision. The feasibility of automatic image classification with respect to its utility in finding clinical decision support demonstrated in this study provides several venues for further exploration. We plan to study the influence of augmenting bibliographic references retrieved from a database search with images; new ways of organizing and presenting retrieval results using annotated images; and further improvement in the automatic single and multi-panel image extraction, annotation, and complementary text extraction.

InfoBot (formerly Medline on Tap – MDoT)

The **InfoBot** project is in line with *Recommendation 27: Promote the development of just-in-time, patient-relevant knowledge bases that link scientific and clinical information within personal health records.* Its goal is a system that will enrich an institution's existing EMR system with useful information from NLM resources. The InfoBot software would run as background agents, both at the institution and at NLM. APIs would be supplied to the institution to allow them to integrate the search setup and to display and store results in their existing EMR system, in accordance with their own preferences. Part of this project is to automatically generate a repository of key facts extracted from the biomedical literature to support informed clinical decision making. This subproject is RIDeM (Repository for Informed Decision Making). Practitioners of Evidence Based Medicine (EBM) advocate informed decision making that combines the clinician's expertise and judicious use of current best evidence. EBM provides guidance in both how to best find needed information, and appraise information found in literature. The three basic components of EBM are [1] Clinical Task (e.g., prevention, therapy, etiology, etc.) [2] The PICO framework for question formulation and document analysis (PICO = **P**atient/**P**opulation, **I**ntervention, **C**omparison, and **O**utcome); [3] Strength of Evidence. Tools will extract these items automatically from MEDLINE records to populate the RIDeM database.

Turning The Pages Information Systems

In line with *Recommendation 12: Educate the public about the historical development of biomedical sciences and technology*, the Turning The Pages project aims to provide the lay public a compelling experience of historically significant and normally inaccessible books, e.g., medieval volumes on anatomy and zoology. In a photorealistic manner, this project allows users to turn and view page images on touch-sensitive monitors in kiosks, as well as 'click and turn' in an online version. In the kiosk version of TTP, techniques and tools are explored to optimally capture the original pages, enhance these page images to improve quality, animate the page images, and import them into a software environment that allows user interaction. We developed a 3D wireframe model in Maya, a commercial modeling and animation system to replace the labor-intensive manual process used earlier in collaboration with the British Library. Within Maya, each pair of page images is texture-mapped to both sides of the wireframe model of a turning page, with a multisource lighting model that provides realistic highlights and shadows. Our technique allows the automated generation and rendering of 15 intermediate animation frames which are then imported into Macromedia Director for viewing and interaction by library patrons.

While retaining the eye-catching appearance and easy touch-and-turn characteristics of a virtual book, the TTP project has also prototyped an extension of this *exhibit* product to an *information system* (TTP+) by linking the displayed pages to relevant biomedical information sources such as databases of clinical trials information, anatomic images and biomedical literature citations. Blackwell's Herbal and Vesalius' Anatomy, both rare books normally inaccessible to a library patron, has been extended to the TTP+ metaphor.

Two significant developments in 2008 included the creation of a 3D model for flat scrolls leading to the construction of the Edwin Smith Papyrus in TTP form ('touch and scroll'), and the release of a French language version of TTP Online ('Tournez les Pages'). Plans are under way to increase the TTP collection with other rare books, selected in collaboration with NLM's historians.

LHNCBC
FY 2008 ANNUAL REPORT

Video Retrieval and Reuse Project

APDB includes four major project areas: core resources (COREAPDB), NLM support (AVSNLM), LHNCBC research support (AVRES), and NLM Media Assets (AVSNLM/Media Assets). The NLM media assets and the NLM support project contribute to the NLM-wide AV support area in the NLM Long Range Plan. The LHC research support and the core resources contribute to ongoing LHC information services projects.

The core resources continue to provide the equipment, software, support and training required for the videographics and animation facility, the interactive multimedia development projects, and the video production capability of the NLM. These video and graphics communications areas depend on technologies which are changing rapidly and are driven by major market requirements not influenced significantly by biomedical needs. Therefore, technology transformation and adaptation are major branch activities pursued along with the development of high quality facilities and support resources required to produce the educational and informational materials essential to a major emphasis by NLM of providing health information to the consumer and to encourage the use of high quality information by health professionals and the public.

In FY2008, the Media Assets project area continued the upgrade of the high definition (HD) editing systems and the transition to digital video and high definition production workflow. The changeover to a completely high definition video production facility remains a high priority for FY2009. Much of the post production area of the facility has been converted to the DTV standard, including accommodation of the HD portion of the standard. The network integration of the four existing nonlinear editing systems and the development of a media asset management system within a networked content creation environment, allows more efficient manipulation of video for faster editing turn around, and the ability to include revisions and updates much more quickly. In FY2009, staff will continue the transition to the HD standard and provide the IT (information technology) convergence for both the production and postproduction functions. This will allow the efficient use of HD still graphics, animation, and full motion HD video in post production editing. Network access to branch-created high definition graphic and animation materials will assist the expanding role of high quality digital images in support of media development.

The NLM still image, graphics, and video support team provides ongoing capability in these areas to all of the NLM and includes the production, post-production and authoring services for the development of Internet video, kiosk interactive multimedia and DVDs. The number of requests for content creation continues to increase and has exceeded the in-house resources available. In addition to meeting the requests through the application of advanced technology, support from contract sources is essential. This area of the budget also contains some equipment funding to maintain the audio, video and multimedia capability in the NLM board room, auditorium and other conference areas supported through branch project management and technical resources.

The fourth project area is LHNCBC research support (AVSRES). A number of LHNCBC development projects require videographics, interactive multimedia development, imaging or video production as part of the overall project objectives. A major effort for FY2009 in this project area is the improvement of rendering times for videographics, and 3D visuals and animations for DVD and other interactive multimedia productions. These animations are rendered on state of the art workstations enhanced with specialized graphics engines and require continual development and upgrading of current modeling and rendering software modules. Some of these 3D animations have been integrated into Web-based materials and into kiosks equipped with DVD systems and touch screen monitors. Resources are required to continue to improve image creation, graphic design, rendering times, and interactive multimedia application development.

Biomedical Image Transmission via Advanced Networks (BITA)

In FY2009, the **BITA** project will continue to have both an R&D as well as a technical support role in NLM's Next Generation Internet (NGI) activities. It is in line with *Recommendation 10: Pilot test and evaluate new ways to use the digital infrastructure to enhance access to online health information by minority and underserved persons and communities*. Technical support for the NIH's Malaria Initiative program in Africa will continue, with a focus on characterizing end-to-end performance measurements over the MIMCom network, 802.11a and 802.11b wireless network implementations, and networks exhibiting narrow bandwidths, high latency and high jitter.

In addition, the BITA project will continue to deploy and test basic connectivity over advanced implementations of Internet2. Developers are collecting and examining data for speed, error, and QOS performance and analysis of advanced network infrastructures including Packet-Over-Sonet, Wavelength Division Multiplexing, and ATM using different techniques, e.g., FTP, multisocket, IP over ATM, native ATM, as well as client server systems such as WebMIRS.

With the worldwide increase in the Internet Protocol addresses needed to network devices of all kinds, research will focus on IPv6, in light of its emergence as a viable architecture for wide area networking. The project aims to investigate the speed, error and QoS performance of this protocol in the transmission of medical images, of fundamental importance in telemedicine and digital libraries.

AUTOMATED CONCEPT EXTRACTION FROM DOCUMENTS

Research in this area is directed toward developing techniques and algorithms to extract bibliographic data from biomedical journal articles, both digitized and Web documents, to build MEDLINE citations. The projects in this category are MARS and its various spin-offs and the Indexing Initiative. These systems address the NLM Goal 1: *Seamless, Uninterrupted Access to Expanding Collections of Biomedical Data, Medical Knowledge, and Health Information.*

Medical Article Records System (MARS)

The **MARS** project and its several spinoffs address NLM's Long Range Plan *Recommendation 4: Support the development of data mining techniques to integrate access and repackage information*. The projects in this group aim to develop and operate systems that efficiently extract bibliographic information from the paper-based or online medical journal literature to build MEDLINE citations. This is done by a combination of document scanning, optical character recognition (OCR), and rule-based or machine learning algorithms. The MARS production system currently in operation extracts and organizes bibliographic data by using advanced algorithms for automated zoning, field identification, and syntax reformatting. In addition, biomedical lexicons are used to implement pattern matching algorithms to correct errors in OCR-detected affiliation information, and reducing incorrectly highlighted words for increased operator productivity.

WebMARS

Developed as part of the MARS project, WebMARS is a system to automatically extract data from Web-based online journals. Besides serving this purpose, in FY2009, the basic WebMARS technology will be used to help achieve goals of NLM's Indexing 2015 Initiative, by the development of the Publisher Data Review (PDR) system. This system will provide operators data missing from the XML citations sent in directly by publishers such as databank accession numbers and NIH grant numbers, thereby lowering the manual effort in completing citations for MEDLINE. In addition, incorrect data sent in by the publishers can be corrected by PDR. The current manual effort in filling in missing data and correcting wrong data from the publishers is considerable since the operators generally have to look through an entire article to find this information, and then key them in.

In FY2009, while the PDR system will be completed and placed into daily production for NLM indexers, we will continue all engineering support for the offsite production facility: installation of upgraded modules, testing, maintenance and operation of all hardware and software for servers, clients and networks, and the necessary system administration.

Analysis of Images for Data Extraction (AIDE)

These projects support the goals of the NLM *Indexing 2015 Initiative* and the *NLM Long Range Plan Recommendation 1: NLM Resources and Infrastructure for the 21st Century*. The goal of this research is to increase the efficiency of creating citations for MEDLINE in order to accommodate the expected doubling of the citation rate within the next few years.

ACORN

Conducted within the **AIDE** project is the underlying research in image analysis and lexical analysis that contributes to the continual improvements in the MARS system, as well as the creation of new initiatives in which these techniques could find application. An example is the ACORN initiative (Automatically Creating OldMedline Records for NLM) which aims to capture bibliographic records from pre-1960 printed indexes (e.g., IM, QCIM, QCICL, etc.) for inclusion in NLM's OldMedline database, thereby creating a complete record of citations to the biomedical literature since Index Medicus appeared in the late 19th century. Manual data entry has proven to be too slow and costly. In FY2009 we will continue our investigation of scanning, image enhancement, OCR, image analysis, pattern matching, and related techniques to extract unique records from the printed indexes, and develop a prototype system for real-world testing and as a precursor to a production system.

Validated Test Set for Document Image Analysis

As part of the AIDE project, a ground truth database, Medical Article Records Groundtruth (**MARG**), has been released for use by the international computer science and informatics communities for research into advanced algorithms for data mining. It has attracted more than 16,500 visits from 96 countries. The MARG database consists of document images and the corresponding OCR data, zones, labels and verified data obtained from the normal operation of the MARS production system. In FY2009 MARG will be expanded and tools provided for easy use.

Indexing Initiative

The **Indexing Initiative** project investigates language-based and machine learning methods for the automatic selection of subject headings for use in both semi-automated and fully automated indexing environments at NLM. Its major goal is to facilitate the retrieval of biomedical information from textual databases such as MEDLINE. Team members have developed an indexing system, Medical Text Indexer (MTI), based on two fundamental indexing methodologies. The first of these calls on the MetaMap program to map citation text to concepts in the UMLS Metathesaurus which are then restricted to MeSH headings. The second approach, a variant of the PubMed related articles algorithm, statistically locates previously indexed MEDLINE articles that are textually related to the input and then recommends MeSH headings used to index those related articles. Results from the two basic methods are combined into a ranked list of recommended indexing terms, incorporating aspects of MEDLINE indexing policy in the process. Image Indexing Initiative (I3) has used interactive MetaMap successfully for automated mapping of terms suggested by subject matter experts to Metathesaurus concepts. I3 will continue efforts to improve the precision and recall of these mappings with additional features of the MetaMap API.

The second approach, a variant of the PubMed related articles algorithm, statistically locates previously indexed MEDLINE articles that are textually related to the input and then recommends MeSH headings used to index those related articles. Results from the two basic methods are combined into a ranked list of recommended indexing terms, incorporating aspects of MEDLINE indexing policy in the process.

The MTI system is in regular, increasing use by NLM indexers to index MEDLINE. MTI recommendations are available to them as an additional resource through the Data Creation and Maintenance System (DCMS). This year MTI recommendations are being augmented by the attachment of subheadings to some of the MeSH headings it recommends. Indexers will now have the option of accepting MTI heading/subheading pairs in addition to unadorned headings. In addition, indexing terms automatically produced by stricter version of MTI are being used as keywords to access collections of meeting abstracts via the NLM Gateway. These collections include abstracts in the areas of AIDS/HIV, health sciences research, and space life sciences.

Indexing Initiative (II) development focuses on testing and updating recently added functionality to the Medical Text Indexer (MTI) system such as the inclusion of subheading attachment recommendations and the addition of an explanation facility to inform indexers how MTI arrived at specific MeSH recommendations. Improvements to MTI will benefit the Indexing 2015 project through its MTI subproject. A secondary focus will be to test the application of MTI to NLM Cataloging and to make any necessary modifications to the cataloging version of MTI. System-related objectives for II include completing the final testing of the migration of our systems to a Linux environment.

The major objectives for the Indexing Initiative in FY2009 are the development of the Basic Medical Text Indexer (MTI):

- Test and update MTI’s subheading attachment function and continue to coordinate with NLM’s Document Creation and Management System (DCMS) team.
- Test and update MTI’s explanation facility and coordinate its inclusion in DCMS.
- Assess catalogers’ acceptance of MTI’s adaptation for NLM Cataloging and to modify the cataloging version of MTI as needed.
- Improve MTI’s ability to handle citations without an abstract, i.e., title-only citations.
- Develop MetaMap 3D, a tool for colorizing and otherwise highlighting features of biomedical text, accounting for both literature and clinical views.
- Explore further Word Sense Disambiguation (WSD) algorithms for improving MTI’s accuracy.
- Complete the code modifications that have been necessitated by the migration to new machine and secure network architecture.

Automatic Extraction of Outcomes from Published Documents

Originally part of the MDoT project, research was conducted toward automatically finding patient outcomes (e.g., the population under study) from MEDLINE citations using knowledge extractors that rely upon NLM Unified Medical Language System and tools. Our Extractor system identifies an outcome and determines whether a found outcome pertains to the topic of interest, the type of treatment studied, and the quality of the study. We evaluated the ability of the Extractor both to find outcomes in general, and to find high quality outcomes that answer specific clinical questions. Possible application areas might include clinical trials design, EMR, and a patient-oriented service. Developed to provide access to the repository, a server accepts requests containing information about a patient (at present, current problems, age and medications) and searches MEDLINE via any of three search engines (Essie, PubMed, or the RIDeM database). The extracted information is sent to the client. The repository will be evaluated in a planned pilot study of supporting Evidence Based Nursing Practice at the NIH Clinical Center.

Digital Preservation Research (DPR)

This project is in line with *Recommendation 1: Build a repository that employs advanced technology for the storage and preservation of large quantities of print and digital material.* It addresses an important problem for libraries and archives, viz., to retain electronic files for posterity, both documents in multiple formats (e.g., TIFF, PDF, HTML) as well as video and audio resources. For document preservation, our objective is to develop a prototype **System for Preservation of Electronic Resources (SPER)** that builds on open source systems (e.g., DSpace from MIT) while incorporating in-house developed modules that implement key preservation functions: ingesting, automated metadata extraction and file migration. One role of SPER is as a testbed to evaluate alternative approaches to these functions and to select the optimum ones. Another role is to use it to preserve a collection of historic medico-legal documents that NLM has acquired from the FDA. This effort will continue in FY2009.

While the accuracy of the automated metadata extraction (AME) module is reasonably high, further work is planned for FY2009: enhancing AME's layout classification and pattern recognition modules; investigating other SVM implementations for use in SPER; modularizing AME to extend its application to other NLM collections.

In FY2009 we will continue to collaborate with OCCS and LO in an effort to create a production-level NLM archive, drawing upon our design experience with SPER. The target collections to be preserved will be selected in collaboration with NLM's preservation managers and curators, and could include Profiles in Science, medical pamphlets, or some other collection.

Also, in FY2009, as the production system is developed, the SPER prototype will be used as a testbed system to investigate promising approaches to implement critical preservation stages, for eventual inclusion in the operational system. Issues to be researched: selecting optimal file formats for the long term; maintaining unique and reliable references to migrated files; feature selection, quantization and clustering in a Bayesian Learning approach to automatically classify digital documents; extending the learning algorithms to Web and video resources; intelligent decision making for file migration.

INFORMATION RESOURCE DELIVERY FOR CARE PROVIDERS AND THE PUBLIC

The Lister Hill Center performs extensive research in developing advanced computer technologies to facilitate the access, storage, and retrieval of biomedical information.

Clinical Research Information Systems

ClinicalTrials.gov provides the public with comprehensive information about all types of clinical research studies, both interventional and observational. The site has over 63,000 protocol records sponsored by the U.S. Federal government, pharmaceutical industry, academic and international organizations from all 50 States and in 158 countries. Some 41% of the trials listed are open to recruitment, and the remaining 59% are closed to recruitment or completed. ClinicalTrials.gov receives over 52 million page views per month and hosts approximately 800,000 unique visitors per month. Data are submitted by over 5,800 study sponsors through a Web-based Protocol Registration System, which allows providers to maintain and validate information about their studies.

ClinicalTrials.gov was established by the National Library of Medicine (NLM) in 2000 in response to the Food and Drug Administration Modernization Act of 1997 and to support NLM's mission of disseminating biomedical knowledge and advancing public health. ClinicalTrials.gov was enhanced in 2007 - 2008 to implement initial requirements of Section 801 of the Food and Drug Administration Amendments Act of 2007 [Public Law 110-85]. The law required the expansion of the registry and the addition of a results database. In response to the law, the ClinicalTrials.gov registry was expanded in November 2007 to provide members of the public, health care professionals, and researchers with additional descriptive, recruitment, location, contact, and administrative information about ongoing and completed applicable drug and device clinical trials. As a consequence of this law, new registrations have from December 2007 to September 2008 increased by 44% (average 360 per week) compared to an average of 250 per week from December 2006 to November 2007) and modifications to existing registrations over the same time periods increased by 300% (average rate of 2,400 per week compared to an average of 800 per week). Also in response to the law, links were added from registration records to related results on the FDA Website (e.g., Drugs@FDA) and links were enhanced to NLM's Medline and DailyMed Websites. ClinicalTrials.gov also researched, designed, tested and implemented a results database which is complementary to the registry. The results database is required by law and is the first-of-its-kind. It includes results information on primary and secondary outcomes of registered trials, as well as information on the patient populations studied. The results database also includes a module for reporting serious and frequent adverse events observed in a clinical trial, although this component is not required by law until September FY2009. Preliminary versions of the database were made available for public testing and comment and the final version was implemented on September 22, 2008. The expanded registration requirements as well as the results database will be further implemented through rulemaking and NLM is working with the Food and Drug Administration (FDA) on drafting of the registration rule. ClinicalTrials.gov continues to work on other aspects of the law, including but not limited to expansion of the results database, a pilot quality control study, hosting a public meeting, and consulting with risk communication experts. When fully implemented, the registry and results database will become a unique resource for scientific and clinical information that can assist in providing patients, healthcare providers, and researchers more comprehensive information about ongoing and completed research

ClinicalTrials.gov was actively involved in educating the public on the new law and continuing to promote standards of transparency in clinical research through trial registration. This information was communicated to a broad range of U.S. and international stakeholders via presentations and peer-reviewed publications. As a result of increasing awareness of the law and the importance of trial registration, nearly 27,000 new registrations were received over the last calendar year. ClinicalTrials.gov continues to collaborate with other registries and professional organizations, working towards developing global standards of trial registration.

Genetics Home Reference

Genetics Home Reference (GHR) is an online resource that offers basic information about genetic conditions and the genes and chromosomes related to those conditions. This resource provides a bridge between the public's questions about human genetics and the rich technical data that has emerged from the Human Genome Project and other genomic research. Created for the general public, particularly individuals with genetic conditions and their families, the site currently includes summaries of more than 325 genetic conditions, more than 500 genes, all the human chromosomes, and information about disorders caused by mutations in mitochondrial DNA. The web site also includes a handbook, *Help Me Understand Genetics*, which introduces users to fundamental topics in human genetics including mutations, inheritance, genetic testing, gene therapy, and genomic research. Usage of the GHR web site, as measured by the daily unique visitors, increased almost 50 percent in the past year, and the site continues to be recognized as an important health resource.

GHR specifically targets NLM's goal of advancing scientific knowledge in molecular biology by providing information about hereditary conditions and their underlying genetic causes. The LHNCBC continues to investigate a variety of ways to make the results of the Human Genome Project more readily available to the public through the Genetics Home Reference (GHR) Web site. Because GHR is designed for patients, families, and the general public, the genetic information is written in a consumer-friendly format. Brief summaries of genetic conditions and genes will continue to be added in FY2009 and existing topics will be reviewed and updated. Strategies for creating "just-in-time" links to salient resources for additional consumer information (e.g., MedlinePlus, support/advocacy groups, and recent literature) will continue to be developed. Additionally, new types of content are under development, including pages that introduce families of related genes.

The GHR web site celebrated its fifth anniversary in FY2008 with a combined total of more than 800 user-friendly summaries of human genetic conditions, genes, and chromosomes. In the past year, the project expanded its genetics content for consumers, added new features, and continued with outreach activities to increase public awareness of the web site. Specifically, GHR staff added new summaries of 94 genetic conditions and 121 genes to the web site. On average, about 20 new summaries were added per month, which is more than twice the monthly average from the previous year. Staff intend to continue this rate of production in FY2009, with a goal of covering as many Mendelian genetic disorders as possible. In the past year, staff members developed a new feature about gene families that allows users to find out how genes are related to one another. The GHR web site currently includes almost 20 gene family summaries, with more under development.

GHR continues to support the Information Rx initiative, a free program that enables doctors and nurses to write "prescriptions" directing patients to the GHR web site for an explanation of genetic disorders and related topics. In other outreach activities, GHR staff presented the web site to several visiting groups, including educators and students, and represented the project at local and national meetings. Staff members will continue to educate others about this important resource in FY2009.

Profiles in Science Digital Library

The Profiles in Science® Web site provides researchers, educators, and potential future scientists worldwide access to extraordinary, unique biomedical information previously accessible only to patrons able to make an in person visit to the institutions holding the physical manuscript collections. "Profiles" also serves as a tool to attract scientists to donate their collections to archives or repositories in order to preserve their papers for future generations. Profiles in Science decreases the need for handling the original materials by making available high quality digital surrogates of the items. Standardized, in-depth descriptions of each item make the materials widely accessible, even to individuals with disabilities. The growing Profiles in Science digital library provides ongoing opportunities for future experimentation in digitization, optical character recognition, handwriting recognition, automated image identification, item description, and search and retrieval.

The Profiles in Science® Web site showcases digital reproductions of items selected from the personal manuscript collections of prominent biomedical researchers, medical practitioners, and those fostering science and health. Profiles in Science provides researchers, educators, and potential future scientists worldwide access to extraordinary, unique biomedical information previously accessible only to patrons able to make an in person visit to the institutions holding the physical manuscript collections. "Profiles" also serves as a tool to attract scientists to donate their collections to archives or repositories in order to preserve their papers for future generations. Profiles in Science decreases the need for handling the original materials by making available high quality digital surrogates of the items. Standardized, in-depth descriptions of each item make the materials widely accessible, even to individuals with disabilities. The growing Profiles in Science digital library provides ongoing opportunities for future experimentation in digitization, optical character recognition, handwriting recognition, automated image identification, item description, digital preservation, emerging standards, digital library tools, and search and retrieval.

The content of Profiles in Science is created in collaboration with the History of Medicine Division of NLM, which processes and stores the physical collections. Several collections have been donated to NLM and contain published and unpublished materials, including manuscripts, diaries, laboratory notebooks, correspondence, photographs, poems, drawings and audiovisual resources. The collections of Arthur Kornberg, Maxine Singer, Alan Gregg and Paul Berg were added this year. An additional 6,192 digital items composed of 12,052 image pages were also added to the twenty-six existing Profiles in Science collections. Presently the Web site features the archives of twenty-seven prominent individuals:

Christian B. Anfinsen
Virginia Apgar
Oswald T. Avery
Julius Axelrod
Paul Berg
Francis Crick
Rosalind Franklin
Donald S. Fredrickson
Edward D. Freis
Alan Gregg
Michael Heidelberger
C. Everett Koop
Arthur Kornberg
Mary Lasker
Joshua Lederberg
Salvador E. Luria
Barbara McClintock
Marshall W. Nirenberg
Linus Pauling
Martin Rodbell
Florence R. Sabin
Wilbur A. Sawyer
Maxine Singer
Fred L. Soper
Sol Spiegelman
Albert Szent-Györgyi
Harold Varmus

The 1964–2000 Reports of the Surgeon General, the history of the Regional Medical Programs, and Visual Culture and Health Posters are also available on Profiles in Science.

In addition to releasing new Profiles in Science collections during , LHNCBC staff made several enhancements to the Profiles in Science systems. Among these was the addition of the ForeSee Results American Customer Satisfaction Index (ACSI) survey. Survey results are expected to identify who is using Profiles in Science as well as what collections users want NLM to add. Staff also increased the accessibility of the Profiles in Science Web site. They extracted OCR text from the Web site's PDF documents and made the OCR available as an alternative format to PDF, developed algorithms to identify documents whose OCR is particularly bad or good, and made more noticeable the transcripts of the Web site's digitized files. LHNCBC staff developed software to automate the review and addition of items to existing Profiles in Science collections. Developers also experimented with various conversion software and created mockups of the Data Entry program in Java Swing and ASP.NET. They completed most of the new Linux-based Profiles in Science, including synchronizing data with the current Solaris-based system and displaying search results. Developers also migrated machines to new networks and migrated applications to new machines.

Nursing Home Screener

In line with *Recommendation 9: Work to reduce and eliminate health disparities by providing underserved populations ... with access to high quality health information that is understandable..*, the **Nursing Home Screener** is a tool for the public to judge the quality of nursing homes in keeping with the NLM goal for customized personalized health information. This Web 2.0 tool will deliver the Quality Indicators for such facilities, derived from publicly available data from the Center for Medicare and Medicaid Services, in an easily navigable geographic graphical interface. The Nursing Home Screener locates homes on a Google map. It allows users to survey nursing home quality, indicated by map icons, in any of four major categories: staffing, fire safety deficiencies, healthcare deficiencies, and quality of care inferred from residents' health. Options can be tailored within each category and other filters may be set to selectively hide home markers of less interest.

Based Medicine - PubMed for Handhelds

PubMed for Handhelds was publicly released in FY2003. Developed to facilitate evidenced-based medical practice with Medline access at the point of need via smartphones, wireless PDA's or portable laptops, PubMed for Handhelds requires no proprietary software and reformats the screen display as appropriate for the wireless handheld device being used. In support of evidence-based clinical practice, clinical filters feature easy access to relevant clinical literature. Newly developed resources allow searching Medline through text-messaging. An algorithm to derive "the bottom line" (TBL) of published was recently added for a clinician's quick reading at the point of care.

User Focused Portals

NLM Gateway

The **NLM Gateway** is an ongoing production system that provides results from 23 NLM information resources with a single search. Since these resources are "moving targets" frequently updated, improved, and otherwise modified, the Gateway must change with them. Periodic changes to the NLM DTD (Document Type Definition), to MeSH and the MeSH Mapping File, and to the UMLS Metathesaurus are accommodated each year. More than 100,000 meeting abstracts are indexed using the tools of the Indexing Initiative. The most recently added resource to which the Gateway provides access is NLM's Profiles in Science.

A formal usability study of the NLM Gateway was performed in-house in FY2007, leading to a comprehensive evidence-based redesign of the system. The redesigned Gateway was placed into production late in February 2008. We are pleased to report that usage is up substantially (more than 50%) since that time.

COMMUNICATION INFRASTRUCTURE RESEARCH AND TOOLS

The Lister Hill Center performs and supports research to develop and advance infrastructure capabilities such as high-speed networks, nomadic computing, network management, and wireless access. Other aspects that are also investigated include security and privacy.

Advanced Biomedical Tele-Collaboration Testbed

The **Advanced Biomedical Tele-Collaboration Testbed (ABC Testbed)** project was completed in 2008. It involved the use of open source, cross-platform technologies based primarily on grid technologies in general and the Access Grid (AG) in particular. The research was a collaborative effort with the University of Chicago, Argonne National Laboratory, the University of Illinois at Chicago, Northwestern University, the University of Rhode Island, and other institutions. Among the scenarios that have been identified to test technologies: using the AG to link different patient safety and medical simulation; using AG with the daVinci surgical robot for distance education; using AG for wireless communication from mobile ambulances for patient treatment prior to arriving in the ER; the use of AG with handheld devices so residents can communicate more effectively; using the AG for 3D teleradiology; and using AG for volume rendering of patient image data in the operating room with wearable (e.g., eye-glass-like) environment. The latter allows surgeons to view the 3D data and to share it with colleagues and consultants while working on a patient.

In , the research team completed testing the scenarios, including those implementing color algorithms for real time volume rendering of CT and MRI data and stereo display in the AG environment as well as the use of the technology in surgical education and planning. Virtual reality methods were employed in a haptic environment allowing surgeons to rehearse liver operations. In addition, the technology was tested in simulation labs for team training and in classroom settings for anatomy teaching. Several additional successful wide area wireless demonstrations of transmitting video and other patient data from ambulances using 3G and mesh cellular technology were completed. The University of Chicago has patented the imaging algorithms.

Scalable Information Infrastructure Initiative

The Scalable Information Infrastructure (SII) Project, which ended during , encouraged the development of relevant health applications that are network aware and able to automatically adjust to changing network conditions and resources. Public next generation networks with SII capabilities hold the promise of adding advanced networking capacity to the tools available to healthcare professionals. Virtual reality and home health care may become realizable at reasonable costs based on next generation networking technology. Applications included wireless and geographic information system (GIS) techniques.

Videoconferencing and Collaboration

A new initiative was undertaken to experiment with uncompressed video over IP as well as high definition television. Compressed HD videoconferencing codecs were investigated using the H.264 technology that is compatible with and part of the revised H.323 standard. Digital Video Transport System (DVTS) technology was implemented, both as a standalone technology and as a component of the Access Grid. DVTS was developed by the WIDE consortium in Japan and is used by various Internet2 members to send uncompressed digital video at 30 megabits per second over IP. In addition, the Collaboratory became part of the Research Channel Working Group within the Internet2 and started acquiring components to implement uncompressed HD video at 1.5 gigabits per second (iHDTV). In 2008, staff successfully implemented the technology within the Collab and are now exploring options for transmitting the video to distant end points. These options include implementing multiplexing and optical network switching and upgrading NLM’s bandwidth from its current 1 gbps. Staff are also exploring UltraGrid technology as a method of transmitting uncompressed video as well as ways to adjust the bandwidth usage of iHDTV. This work is being undertaken with an eye toward testing the technology in telemedicine settings and, possibly, performing a clinical trial. Collaboratory staff developed major enhancements for the Access Grid’s (AG) shared browser and presentation tools. The use of open source browsers and presentation software as the basis for making the enhancements is being considered.

A distance learning program in collaboration with SIS, coordinator of the NLM Adopt-A-School Program, continued to provide on-site and distance education about varied health science topics and information sources to students at the King Drew Medical Magnet High School, affiliated with the Charles R. Drew University of Medicine and Science in Los Angeles. The NIH Office of Science Education participated again in the program and conducted several sessions on health science careers. In 2007-2008, the program was expanded to include a high school serving Native Americans in Kotzebue, Alaska. Each session was assessed as in previous years. As in the past, a statistical analysis of student ratings of teaching showed students rated the presentations quite highly. Over the years the ratings of the distant presentations are sometimes higher than the face to face ones at King Drew. These differences are only two tenths of a point (5 point scale) apart. The Alaska students, who receive all presentations at a distance, rated the presentations higher overall than the students at King Drew.

Methods for providing application sharing and image manipulation with low latency were identified and methods developed enabling the instructor at NLM to view each remote student's desktop. Successful pilot training sessions have been done with the University of Puerto Rico using the application sharing methods in conjunction with H.323 videoconferencing and with the University of Michigan with Access Grid (AG) technology to offer NCBI's biotechnology training at a distance. A follow up implementation was done with the Charles R. Drew University of Medicine and Science. Unfortunately, budget constraints have forced NCBI to cut back all training, but the methodology was validated.

A study of collocation as a factor in synchronous learning was completed with the University of Alabama at Birmingham in which students were tested on lectures delivered by videoconference and asked to collaborate on search tasks before being tested. They also were asked to rate teaching effectiveness of the lectures. Students were either physically collocated in a computer lab or meeting virtually in a multipoint videoconference. The data indicate that there are no significant differences between groups except for perceived interaction, which was much higher for the dispersed group. Observations confirmed that those groups experienced higher levels of actual interaction because the technology required all to work together. The collocated students simply interacted with the person next to them, if at all. The results are being written for possible publication.

The Center for Public Service Communication (CPSC) completed a successful pilot test of the use of video over IP to provide remote medical interpretation services at public health clinics in Duval County Florida. Valuable information about how the technology was and should be used was obtained, and the CPSC will move the technology to another public health environment in Florida. A more formal assessment has been undertaken with CPSC and the Medical University of South Carolina and data is being collected of patient, provider, and interpreter judgments of the quality of communication in clinical encounter when interpretation is provided by video, by phone or in person. Depending on the outcome, the use of video technology beyond the medical center to remote rural clinics may be explored, given a) interest and b) in place network infrastructure.

Both the Web casts of the bi-monthly Washington Area Computer Assisted Surgery Special Interest Group and videoconferencing added last year continued. There is now two-way interaction between those attending the meeting in the Lister Hill auditorium, where the presentations are made, and those in an auditorium at the Allegheny Hospital System in Pittsburgh. Attendees are able to obtain continuing medical education credits because of this linkage.

OHPCC Collaboratory for High Performance Computing and Communication

The "Collab" was established primarily as a resource for researching, testing, and demonstrating imaging, collaboration, communications and networking technologies related to NLM's Next Generation Network initiatives. This infrastructure is also used by staff to keep abreast of and test new technologies of possible interest to NLM (and others in biomedical informatics) and to conduct ongoing imaging, collaboration and distance learning research within OHPCC. The technology infrastructure is used to collaborate with researchers outside the NLM and, when appropriate, it is leveraged to support other activities and programs of the NLM. The facility can be configured to support a range of technologies, including 3D interactive imaging (with stereoscopic projection), the use of haptics for surgical planning and distance education, and interactive imaging and communications protocols applicable to telemedicine and distance education involving a range of interactive video and applications sharing tools. The latter enable staff to collaborate with others at a distance and, at the same time, demonstrate much of the internal and external work being done as part of NLM's Visible Human and advanced networking initiatives. The collaboration technologies include a complement of tools built around the H.323 and MPEG standards for transmitting video over IP and open source technologies such as the Access Grid.

BabelMeSH

BabelMeSH is a multilanguage and cross-language search tool for healthcare personnel who prefer to search MEDLINE/PubMed in their native languages. Journals' language of publications can be selected. Through international collaborations, including WHO Eastern Mediterranean Regional Office in Cairo, users can now search in Arabic, Chinese, Dutch, French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Swedish and English. Some specialty organizations are using BabelMeSH as a tool to search their collection of images.

PICO Linguist

PICO (Patient, Intervention, Comparison, and Outcome) **Linguist** is an application available through BabelMeSH that allows users to search Medline/PubMed in a more clinical and evidence-based manner. This work is significant because it is the only cross-language search portal on the Internet that allows the input in more than two languages. It is also unique because it allows the user to search in character-based languages (non-Latin alphabet), transform it to an English language search, and retrieve citations published in any language or language combination. Full-text articles may be linked to the result available online without subscription requirements.

Computing Resources Projects

The Computing Resources (CR) has a variety of core projects to build, administer, support, and maintain an integrated secure infrastructure that facilitates the LHC research activities and thereby leverages the overall effectiveness of research staff members. The integrated secure infrastructure contains network management, security management, facility management, and system administration support for large numbers of individual workstations and shared servers.

The network management team plans, implements, deploys, and tests high speed network connectivity over Internet and Internet-2. One core project is studying the implementation of OMB requirements in the next generation Internet version 6 (IPv6). The security management team incorporates firewall administration, patch management, anti-virus management, intrusion monitoring, security scanning, and vulnerability remediation to ensure a safe working environment in overall security perspectives. The facility management team facilitates the product deployments, including power planning, network connection, cabling and space allocation in the computer room as well as the NCCS co-location. The system administration team provides center IT services, such as DNS, NIS, backup, printing, and remote access to ensure an efficient operation across the NIH campus.

Core projects also provide critical project planning and support for FISMA security mandates, such as yearly Certification and Accreditation (C&A) audit. Additionally, core projects provide operation assistance and troubleshooting functions for shared communication resources for public access to production systems.

DocView Project: Tools for Using and Exchanging Library Information

This research area applies communications engineering and digital imaging techniques to document delivery and management, thereby addressing the NLM mission of providing document and information delivery to end users and libraries. An additional focus is to contribute to the bulk migration of documents for purposes of digital preservation, also part of the NLM mission. The active projects in this area are DocView, MyDelivery, DocMorph, and MyMorph.

DocView

The goal of the **DocView** project is to conduct R&D on advanced tools allowing libraries and users to access biomedical information. Originally released in 1998, this Windows-based client software is widely used to facilitate delivery of TIFF documents for interlibrary loan services. More than 18,600 users in 195 countries have downloaded it since it was released.

MyDelivery

In FY2009, research will primarily focus on the development of **MyDelivery**, a new Internet communications system designed to solve an important biomedical and health sciences communications problem. Health science applications often require the use and exchange of information contained in very large electronic files (e.g., digitized x-ray images, sonographic images, digital video files, MRI, CT scans, PET scans, and scanned document images). The delivery of this information should be fast, easy, reliable, safe, and secure (HIPAA-compliant). Because no current Internet communications technology (email, ftp, instant-messaging and Web delivery) fully meets these criteria for large file communication, especially over potentially unreliable wireless networks used by an increasingly mobile user population, we are designing and developing a new communications system that will provide these capabilities to the biomedical and health science communities. Targeted for use in clinical, research, administration, and library environments, the **MyDelivery** system will be capable of reliably communicating biomedical information contained in files of any size over networks of all types, including potentially unreliable ones.

DocMorph

As part of the DocView project, research and system engineering design will continue to maintain and improve the operation of **DocMorph**, a Web-based server providing users remote image and information processing capabilities via the Internet. DocMorph serves as a test bed for evaluating new image and library information processing algorithms, as well as a public service for document format conversion. This system now accepts more than fifty file formats, including black and white images, grayscale and color images, text and word processing files, to produce four outputs: PDF files, TIFF files, text, and synthesized speech. DocMorph averages 1,000 conversions daily, and 1,000 unique users monthly. It is used by several hundred libraries, including NLM, mostly in their interlibrary loan operations.

MyMorph

While DocMorph is generally accessed via a Web browser, the **MyMorph** client software allows users to perform *large scale* conversion of thousands of files at a time. MyMorph has more than 9,800 registered users, many of whom are document delivery librarians in small libraries around the country, using MyMorph as part of their daily document delivery operation. In 2009, we will expand MyMorph's role to digital preservation, both for bulk migration of archived files as well as conversion of ingested files to canonical forms such as PDF/A. PDF/A is an ISO-proposed standard file format for long-term electronic document preservation that promises to be of importance in preserving part of NLM's collection. Initial investigation has started in modifying both MyMorph and DocMorph to produce PDF/A files from document images.

Image Storage and Transmission Optimization (ISTO)

The **ISTO** project *(Recommendation 28: Develop software for acquisition, organization and access to ... digitized biomedical images)* addresses the problems of efficiently storing and delivering large biomedical image collections through research into advanced compression and transmission techniques. The focus is on Visible Human images, spinal x-rays from the NHANES surveys, and digitized color images of the uterine cervix from NCI. In FY2009, we will research the best approaches for using advanced compression techniques, such as Wavelet Transform compression, for the storage of the NHANES images in our multimedia database, as well as efficient decompression algorithms. The result of this effort will be executable modules for Wavelet decompression and multiscale display to be incorporated into the dissemination system for the NHANES x-rays. This work will position NLM to consider large file size image databases as service components for telemedicine as well as future Health Networks.

LANGUAGE AND KNOWLEDGE PROCESSING

Terminology Research and Services

LHNCBC research staff build and maintain the **SPECIALIST** Lexicon, a large syntactic lexicon of medical and general English that is released annually as one of the Unified Medical Language System (UMLS) Knowledge Sources. The SPECIALIST Lexicon consists of over 360,000 records describing the syntactic, morphological and orthographic properties of words. These records are released in a unit record format as a set of relational tables. New lexical items are continually added by a team of lexicon builders using LexBuild a lexicon building tool developed by LHNCBC staff. LexBuild is an evolving lexicon building tool designed to aid the lexicon building team by facilitating entry of lexical information and providing real time quality control. The SPECIALIST lexicon release tables are annually generated using the LexBuild tool. The SPECIALIST lexicon and tools are UTF-8 compliant and capable of dealing with non-ASCII characters.

The UMLS Lexical tools, including lexical variant generator (LVG), wordind, and norm are distributed with the UMLS as are text processing tools which analyze documents into sections, sentences, and phrases. The SPECIALIST Lexicon, lexical tools, and text processing tools are released as open source resources and available under an unrestrictive set of terms and conditions for their use. LHNCBC researchers have also released a text classification system which is a JAVA port of the Journal Descriptor Indexing tool originally developed by LHNCBC staff in LISP. The JDI algorithm is an unsupervised text statistical classification method that can provide context for word sense disambiguation and other natural language processing tasks.

LHNCBC research staff also develop and maintain the UMLS Knowledge Source Server (UMLSKS), an application that provides Internet access to the UMLS knowledge sources. UMLSKS is updated quarterly to accommodate quarterly UMLS releases. A beta version of the Grid/Web services implementation of the UMLSKS backend and portlet-based user interface has been released and is undergoing usability testing.

Medical Ontology Research (MOR)

While existing knowledge sources in the biomedical domain may be sufficient for information retrieval purposes, the organization of information in these resources is generally not suitable for reasoning. Automated inferencing requires the principled and consistent organization provided by ontologies. The objective of the **Medical Ontology Research** project is to develop methods whereby ontologies can be acquired from existing resources and validated against other knowledge sources, including the Unified Medical language System (UMLS).

This year, the research team focused on biomedical information integration from the perspective of translational research. Effective data integration of data repositories created by different communities (e.g., basic research and clinical care) is often realized through the integration of the terminologies used for the annotation of data in these repositories. We evaluated semantic integration among such terminologies, more specifically between SNOMED CT and the NCI Thesaurus, and between LOINC and

SNOMED CT. We also evaluated methods based on Semantic Web technologies for data integration, with application to nicotine dependence.

RxNav, the standalone browser for **RxNorm**, NLM's drug terminology integration database, was extended in two different directions. First, we created an Application Programming Interface, making it possible for developers to integrate RxNorm information in their programs. Second, we broadened the scope of RxNav to include clinical information about drugs.

The research team continues to work on assessing the quality of biomedical terminologies and ontologies. Investigated this year were RxNorm and the UMLS (categorization of polysemous concepts, alignment of relationships between Metathesaurus and Semantic Network). Recommendations were made for evaluating the quality of vocabularies for use in NCI's caBIG.

We continue to collaborate with leading ontology and terminology centers, including the National Center for Biomedical Ontology and the International Health Terminology Standards Development Organization.

The major objective of the Medical Ontology Research project is to develop methods whereby ontologies could be acquired from existing resources (including the Unified Medical Language System), as well as validated against other knowledge sources.

Specific objectives for FY2009:

- To integrate biomedical information from various knowledge sources using Semantic Web technologies.
- To format biomedical terminologies (e.g., MeSH) for use in the Semantic Web.
- To evaluate the use of UMLS concept identifiers as a source of permanent identifiers (Uniform Resource Identifiers) for the Semantic Web.
- To assess similarities and differences between the UMLS Semantic Network and other top-level biomedical ontologies.
- To integrate publicly available drug information sources through RxNav, including RxNorm, NDF-RT and MedlinePlus.
- To continue providing the service of mapping between vocabularies to client projects such as ClinicalTrial.gov and The Indexing Initiative.

Semantic Knowledge Representation (SKR)

The **Semantic Knowledge Representation** project provides a context for basic research in natural language processing based on the UMLS knowledge sources. Research focuses on development of SemRep and MetaMap to extract semantic predications from text to support innovative information management applications in biomedicine. We are currently developing Semantic MEDLINE, a Web application which exploits semantic predications to help users manage the results of PubMed searches. Research is being conducted to adapt the application to support clinical practice guideline development (in cooperation with NHLBI) and scientific portfolio management (in cooperation with OD/OPASI). Further collaboration with NLM/SIS is extending the technology to medical aspects of disaster information management.

The context of the SKR project is articulated in the NLM Long Range Plan, especially 1.6.1 (Discovery initiative: Facilitate scientific discovery through computational methods which identify relevant linkages across a variety of information resources) and 1.6.3 (Advanced literature search tools for finding articles and facts for targeted purposes, including decision support and guideline development).

Major objectives for the planning year include:

- Continue to expand SemRep effectiveness, concentrating on recall
- Develop an algorithm to efficiently process large amounts of biomedical text (MEDLINE citations, grant applications, and Web documents) to accommodate Semantic MEDLINE
- Collaborate with the Disaster Information Management Research Center in the Division of Specialized Information Services to expand SemRep to medical aspects of disaster information management, focusing initially on influenza epidemics, burns management, and post-traumatic stress disorder
- In continued collaboration with NHLBI and OPASI enhance Semantic MEDLINE as an adjunct to traditional information retrieval systems for helping biomedical researchers managing large amounts of text
- Expand the use of Semantic MEDLINE for literature-based discovery
- Conduct research to combine SemRep processing with MedLEE for effective processing of clinical text

UMLS AND CLINICAL VOCABULARY STANDARDS

Unified Medical Language System (UMLS)

The most recent release of the UMLS Metathesaurus contains over 1.5 million concepts and 7.7 million concept names. After the successful transition of the production of the Metathesaurus to NLM's Office of Computer and Communications Systems (OCCS), staff are focusing on the research and development aspects of the Metathesaurus. A NLM-wide UMLS Priorities and Services Working Group is convened to formulate proposals for priorities of development in the coming 3 -5 years. The group solicits input from recent surveys of UMLS users, questions and comments received via the UMLS listserv and at professional meetings, customer inquiries; other NLM staff; knowledgeable people in other Federal agencies and in standards organizations; and known heavy users. Staff continue to provide assistance to the Library Operations (LO) team in providing user support; and in various terminology related projects initiated or supported by NLM nationally and internationally.

UMLS-CORE Project

The UMLS-CORE (Clinical Observations Recording and Encoding) Project has finished the data collection phase and data analysis is ongoing. The goal of this project is to identify a clinical subset of the UMLS to support consistent high level encoding of clinical information (e.g. in discharge summaries or problem lists). Lists of terms and their actual frequencies of usage in real-life clinical systems are collected from large healthcare providers including: Kaiser Permanente, Mayo Clinic, Intermountain Health Care, Regenstrief Institute, Nebraska University Medical center and the Hong Kong Hospital Authority. These terms are mapped to the UMLS and their pattern of usage and overlap are studied. The CORE subset will promote and facilitate the use of standard clinical terminologies by helping users to identify the most frequently used portion of these terminologies. It will also enhance data interoperability by reducing coding variability. For terminology developers, the subset will help them identify gaps in coverage and focus their qualify improvement efforts on the most frequently used terms.

Terminology Representation and Exchange Format (TREF)

The specification of the Terminology Representation and Exchange Format (TREF) has been finalized. The purpose of TREF is to serve as a standard publishing format for single-sourced terminologies. Its use will facilitate the exchange of terminologies and the sharing of terminology related tools. TREF will simplify the task of inversion of source terminologies into the UMLS editing environment. There is an ongoing collaborate with the National Center for Health Statistics to produce a TREF version of ICD9CM. This will be the first major terminology published in the TREF format and will fill the need for a machine-readable version of ICD9CM.

UMLS User and Usage Pattern Study

Ongoing collection and analysis of UMLS user and usage information through the web-based annual report application. This information will help guide the NLM-wide UMLS Products and Services Working Group which will recommend priorities for UMLS development in the next 3 - 5 years.

RxTerms

RxTerms is an innovative solution to a common problem in the development of clinical applications that capture medication or prescription information. The lack of a publicly available interface terminology for drugs has meant that application developers must either use proprietary terminologies or build from scratch. RxTerms fills this gap by providing a free, user-friendly, and efficient drug interface terminology that links directly to RxNorm, the national terminology standard for clinical drugs which is also developed by NLM. Efficiency of data entry is enhanced by systematic segmentation of RxNorm clinical drug names and aggressive pruning of drugs that are not available in the US. Additional synonyms from sources outside RxNorm further enhance the user-friendliness of RxTerms. RxTerms is currently being used in one of CMS's applications in the post-acute care environment (CARE). It will also be used in the NLM Personal Health Record. RxTerms will be freely available for download from NLM's website for wider testing and feedback.

DISASTER INFORMATION MANAGEMENT

RxHub Medication Reconciliation Project

Funded by the Bethesda Hospitals Emergency Preparedness Partnership, a collaboration formed by three Bethesda area healthcare facilities (National Naval Medical Center, NIH Clinical Center, and Suburban Hospital), this project will study the feasibility and value of using prescription history information from RxHub in a disaster situation. RxHub access points will be set up in these three local Bethesda hospitals and RxHub data will be compared with that obtained by the traditional manual medication reconciliation process in terms of coverage, accuracy and completeness. This external source of medication information that can be obtained automatically could be both time and life saving in disaster circumstances.

Our hypothesis is that when a patient has medication information in the consortium of Pharmacy Benefit Managers (PBM) databases, that information will be more complete and more precise than the corresponding information collected manually from the patient. We also hypothesize that the PBM medication history will be obtained more quickly and with less effort than the manual history. The value of such data for a population will depend upon the proportion of patients in the population who have medication information within the consortium database. The PBM consortium will have no information about patients without insurance or those with insurance who have not been processed by the PBM consortium. We will set a binary variable to identify patients who had no data in the PBM consortium and will collect demographic, administrative variables (arrival mode), and insurance class on all patients. Then we will model the existence of PBM consortium data on these attributes. We will use this model to predict the proportion of patients with data in the PBM and to assess policies that might eliminate this gap.

Lost Person Finder (LPF)

The **Lost Person Finder (LPF)** is included as a part of BHEPP program and funded by interagency agreement. This partnership seeks to create systems that would be used in the event of a disaster, either natural or manmade. The LPF system addresses the problem of missing people, a common problem in the chaos of a disaster event. At registration, pictures may be taken of patients/disaster victims by registrars or volunteers using digital cameras or cell phones. These pictures are uploaded to the database along with other information. The LPF matches these with pictures and descriptive information from the general public searching for lost children, spouses, or friends. The LPF displays pictures and some descriptive information on large screens situated at the hospitals, both indoors and at key outside locations. In addition, through remote computers or handheld devices the public may access LPF to search for this information.

TRAINING OPPORTUNITIES

Working towards the future of biomedical informatics research and development, the Lister Hill Center provides training and mentorship for individuals at various stages in their careers. The LHNCBC Informatics Training Program (ITP), ranging from a few months to more than a year, is available for visiting scientists and students. Each fellow is matched with a mentor from the research staff. At the end of the fellowship period, fellows prepare a final paper and make a formal presentation which is open to all interested members of the NLM and NIH community.

In , the Center provided training to 37 participants from 13 states and 5 countries. Participants worked on research projects including 3-D informatics research, personal health record research, medical image processing, image & text retrieval, InfoBot research, interactive publication research, information retrieval, document analysis, natural language processing, ontology research, question answering research, grid computing, medical terminology research, medical ontology research, telemedicine, and ubiquitous computing. The program maintains its focus on diversity through participation in programs supporting minority students, including the Hispanic Association of Colleges and Universities and the National Association for Equal Opportunity in Higher Education summer internship programs. Participants in the program have been authors on 24 of the 64 (38%) manuscripts published by LHC researchers in the period from June 2007 through June 2008.

In , we started a new Clinical Informatics Postdoctoral Fellowship Program to attract young physicians to NIH to pursue research in informatics. This program is run jointly by the Lister Hill Center and the Clinical Center to bring postdoctoral fellows to labs throughout NIH. Funding is from the LHC. Our first Clinical Informatics Fellow arrived in May.

The Center continues to offer an NIH Clinical Elective in Medical Informatics for third and fourth year medical and dental students. The elective offers students the opportunity for independent research under the mentorship of expert NIH researchers. The Center also hosts the eight-week NLM Rotation Program which provides trainees from NLM funded Medical Informatics programs with an opportunity to learn about NLM programs and current Lister Hill Center research. The rotation includes a series of lectures covering research being conducted at NLM and the opportunity for students to work closely with established scientists and meet fellows from other NLM funded programs.