LISTER HILL NATIONAL CENTER
FOR BIOMEDICAL COMMUNICATIONS

An Intramural Research Division of the U.S. National Library of Medicine

TECHNICAL REPORT
LHNCBC-TR-2011-003

The Lister Hill National Center for Biomedical Communications Annual Report FY2011

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
FY2011 ANNUAL REPORT

Clement J. McDonald, MD
Director

LHNCBC
FY2011 ANNUAL REPORT

LHNCBC
FY2011 ANNUAL REPORT

The Lister Hill National Center for Biomedical Communications (LHNCBC), established by a joint resolution of the United States Congress in 1968, is an intramural research and development division of the US National Library of Medicine (NLM). LHNCBC seeks to improve access to high quality biomedical information for individuals around the world. It leads programs 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 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 NLM Long Range Plan 2006-2016.

LHNCBC 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 members regularly publish their research results in the medical informatics, computer and information sciences, and engineering communities.

LHNCBC 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 semi-annually to review LHNCBC's research projects and priorities. LHNCBC research activities can be found at http://lhncbc.nlm.nih.gov/.

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 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

The goal of the NLM Personal Health Record (PHR) project is to help individuals manage health care for themselves and/or their relatives. The PHR is intended to serve as a test-bed for patient-specific and reminder-based consumer education information, validating and improving NLM clinical vocabularies, studying consumers' use of PHR systems, studying the potential of PHR-based educational reminder systems to improve prevention, and as a potential vehicle for gathering patient information during clinical trials.

The PHR supports the entry and tracking of key measurements, test results, prescriptions, problems, immunizations, and future health appointments. Patients can produce digital and paper copies of PHR contents in various formats. Patients can record questions they want to ask their doctors, maintain lists of their current medications, view educational material that pertains to their specific health information, and monitor trends in their weight, blood pressure or other items of interest. Users can get access to MedlinePlus information resources by clicking the icon adjacent to the name of any prescription drug, medical condition, or surgery that they enter into the system. The PHR automatically assigns codes to the medications, observations, and problems as users enter them. These codes come from national vocabulary standards that are supported or developed by NLM, e.g., Logical Observation Identifiers Names and Codes (LOINC), RxNorm, and Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT). The strong use of coding in the NLM PHR vocabulary standards enables many computer-generated features such as personalized reminders about preventive care and/or healthy behaviors and automatic calculations based on other values in the PHR (for example, calculation of body mass index based on height and weight) and direct links to information sources such as Medline plus.

In FY2011, LHNCBC researchers and developers continued to improve the capabilities of the PHR including expanded and enhanced vocabulary and reminder rules, sophisticated graphing capabilities and auto-save features. Staff continued to work with NLM and others to develop and implement policies concerning the PHR. We worked with the NIH Office of General Counsel to write a software license agreement. We conducted an analysis of PHR user agreements and wrote a user agreement template. We began negotiations with a local hospital that may license the PHR software and provide NLM with de-identified data for usability research testing.

This young project addresses the long-standing NLM interest of facilitating health care management and is closely aligned with the NLM strategic plan. It will help refine the message and vocabulary standards that NLM supports and will provide another consumer entry point to a rich trove of patient-oriented data.

Use of Surescripts Prescription Data in Direct Patient Care

Studies have shown that a significant proportion of Emergency Department (ED) visits are related to adverse events of drugs. It is vitally important that the ED physicians have access to a full and accurate medication history. However, gathering such information from the patient is time-consuming, expensive, and sometimes unfeasible when patients are unconscious. Patient-provided medication histories are also known to be incomplete. The ED of Suburban Hospital has employed the service of Surescripts, a consortium of major Pharmacy Benefit Managers (PBM) and the largest e-prescribing network in the U.S, to provide an electronic summary of a patient's full year prescription filling history. Before the system went live, Suburban Hospital collected both Surescripts data and patient-provided history for quality assurance. We obtained this information in a de-identified form. We compared the two sources of information and found that Surescripts information covered a high proportion (almost 90%) of a patient's current medication and added significantly to the manual history.

The Surescripts reports are now a routine information source for patients attending the Suburban ED. The nurses, pharmacists and physicians using the system have provided positive feedback about its use. In a high proportion of cases, drugs that are missing from the patient-provided medication report, but are present in the Surescripts report, are drugs that the patient is in fact taking. The full year prescription history is also helpful in identifying potential problems of drug compliance, drug seeking behavior, and duplicative prescriptions.

EMR Database Research and Natural Language System Development

Developed to investigate secondary use of data collected in electronic medical records, this general purpose longitudinal database structure proved to be robust when we added microbiology results obtained with the sixth update of the MIMIC II EMR dataset.

To facilitate the use of this database by researchers, in FY2011 we mapped the majority of the MIMIC-II local laboratory test codes to standard LOINC codes and incorporated them into MIMIC-II public releases starting with the sixth version. We are currently mapping medications from multiple tables to RxNorm nursing observations from MIMIC-II local codes to LOINC, RxNorm, and/or SNOMED CT as appropriate.

We also have used the MIMIC-II de-identified data under a restricted-use Memorandum of Understanding to conduct retrospective clinical studies on the significance of obesity and metabolic syndrome, glucose control, and vitamin levels in ICU mortality and post-discharge survival. In the process of using the data for our research, we found problems that required correction and normalizing the clinical data within MIMIC-II.

We also use the MIMIC-II as a test bed for comparing speed and ease of use of a no-SQL database (SOLR) with a traditional SQL database (Oracle). For some kinds of queries, SOLR is 10-50 times faster than fully optimized Oracle. Because SOLR requires flattening of the database, it may provide an easier platform for the typical researcher to search: the researchers might no longer be burdened with understanding the relations between data elements and formulating complex SQL queries. We are developing an interface to test this hypothesis.

In line with the NLM mission to facilitate access to health information resources, this year we became a mirror site for PhysioNet, a very large database of physiologic wave form tracings gathered from health care institutions world-wide by the MIT researchers who also developed MIMIC-II. We have recently updated the MIMIC-II waveform data with the latest, 3-terabyte collection.

We developed natural language processing techniques to extract important clinical variables from the rich narrative text in MIMIC-II, e.g., smoking status and discharge destinations. For NLP experiments we acquired over 200,000 reports that were made available to participants of the 2011 Text REtrieval Conference (TREC). The UMLS lexicographers group uses these reports to augment the Specialist Lexicon with entries specific to clinical narrative. In addition, our research team participates in and regularly scores among the top finishers in the annual TREC Genomics track competition. The LHNCBC team was this year's top-scoring team among 29 international participants from industry and academia.

Biomedical Imaging, Multimedia, and 3D Imaging

This research area has several objectives: build advanced imaging tools for biomedical research; create image-based tools for medical training and assessment; investigate design principles for, and develop multimedia image/text databases with particular focus on database organization, indexing and retrieval; develop Content-Based Image Retrieval (CBIR) techniques for automated indexing of medical images by image features.

Imaging Tools for Biomedical Research

LHNCBC and the American Society for Colposcopy and Cervical Pathology (ASCCP) launched one of our imaging systems, the Teaching Tool, for operational use in the assessment of professional knowledge and skills in the field of colposcopy. As of October 2011, the Teaching Tool is being used by 88 resident programs nationwide in Obstetrics/Gynecology and Family Practice (at 80 universities and other premier institutions such as the Mayo Clinic).To date, the tool has been used to administer 760 individual online exams.

In addition, the National Cancer Institute (NCI) used another of our imaging programs, the Boundary Marking Tool, at the University of Oklahoma Health Sciences Center and other sites in Costa Rica, Nigeria, the Netherlands, and Spain to collect and annotate colposcopy images for the creation of a worldwide database for cervical cancer research. We incorporated new capability in our Multimedia Database Tool to retrieve and display histology images from the NCI ASCUS/LSIL (atypical squamous cells of undetermined significance / low-grade squamous intraepithelial lesion) Triage Study (ALTS), and we are currently working with NCI in a study of visual precursors of pre-cancer. This year, pathologists at the University of Oklahoma Health Sciences Center used the CEB Virtual Microscope to make visual diagnoses and graphically annotate histology images of the uterine cervix, and we are using this annotation data to develop computer-assisted diagnosis methods for these images. In the past year, this work has included research, design, and development of segmentation, feature extraction, and tissue classification algorithms for these images.

We also collaborated with academic researchers in projects to develop interactive segmentation capabilities for very large images using Graphical Processing Units (GPUs). Developers successfully installed this capability in an in-house system equipped with two GPU processors, and used it for the segmentation of Gigabyte-sized histology images. Additional collaboration with academic groups included work toward developing high-fidelity image compression techniques for mobile platforms and work in biomedical case-based (text and image) information retrieval.

Content Based Image Retrieval

As a significant part of imaging research at LHNCBC, Content Based Image Retrieval (CBIR) is an active area with several objectives related to the development of tools and systems. One objective is to improve the state of the art in techniques to find visually similar images. A second goal is to extend image matching from simply finding images that are visually similar to those that are also meaningful in a particular context. This research helps introduce automation into our existing cancer research tools. For example, the CervigramFinder automatically indexes and allows retrieval of cervigrams using shape, color and texture features. This system therefore contains the key elements needed to augment the Boundary Marking Tool with an automated assist for the user in marking boundaries of regions of medical significance. Evaluated in 2010 for usability and acceptance at the biannual meeting of the American Society for Colposcopy and Cervical Pathology (ASCCP), during FY2011 developers used the evaluation recommendations to improve the Boundary Marking Tool.

Investigators have used CBIR to index illustrations in medical journals by using image features in combination with text processing of figure captions and in-document text mentions. This research is aimed at enriching the user experience of searching for relevant documents by including the contents of medical images, photographs, graphs and other illustrations found in articles. Biomedical journal articles contain a variety of image types that can be broadly classified into two categories: regular images and graphical images. Graphical images can be further classified into four classes: diagrams, statistical figures, flow charts, and tables. Over 15 image features were implemented and used in a Support Vector Machine (SVM)-based framework to detect modality (x-ray, ultrasound, CT, MR, etc.) and to compute image similarity. Techniques developed in this work were evaluated in the international ImageCLEF competition since 2009 and were again found to be successful. In 2011 our efforts were

ranked highly among 17 teams from around the world, many of which were from industrial research and development labs. We also developed methods to describe images in a *bag-of-words* and a *bag-of-keypoints* representation, analogous to those used in text-document retrieval. These were very successful in automatic coarse annotation of images. Researchers continue to improve automatic image modality detection and annotation methods and are incorporating these findings into our OpenI search system.

We have explored the role of CBIR in extracting regions of interest (ROI) in images. One approach to identifying meaningful ROIs, and thereby annotating biomedical-article images, is by first extracting individual figure panels from multi-panel images in biomedical articles. From each image panel author-placed markups (or "pointers") such as arrows, asterisks, and alphanumeric characters are automatically detected. We developed novel methods that applied a combination of Markov Random Fields, Hidden Markov Models, Active Shape Models, and Particle Swarm Optimization techniques for each type of markup with over 90% accuracy in detecting arrows and alphanumeric characters. We also developed a neural network-based optical character recognizer that is specialized to recognize single characters in images, a challenging task for off-the-shelf commercial OCR packages. Further research is ongoing**.**

Investigators are also using CBIR in a new project for screening digital chest x-rays for pathology, such as tuberculosis and other pulmonary diseases, that are prevalent in third world countries. As an initial step in this project, we have developed image content analysis methods to automatically detect lungs and ribs in the x-ray images. Further research into image feature extraction and machine learning methods for detecting and classifying images is ongoing.

Other avenues explored in this research area are distributed computing and use of GPUs for compute-intensive CBIR tasks, with a particular focus on image segmentation. Through our collaboration with Texas Tech University, we developed a method that uses GPU processing power for interactively following challenging object boundaries such as the separation between cancer and non-cancer cells in histology slides of the uterine cervix. We can then use the segmented regions to train classifiers to detect various stages of pre-cancer.

Interactive Publications

The goal of this project is to conduct research into models for highly interactive multimedia documents that could transform the next generation of publishing in biomedicine. The project focuses on the standards, formats, authoring and reading tools necessary for the creation and use of such *interactive publications* that, in addition to text, contain media objects relevant to biomedical research and clinical practice: video, audio, bitmapped images, interactive tables and graphs, and clinical DICOM images such as x-rays, CT, MRI, and ultrasound.

In this project, LHNCBC has created interactive publications containing these data types, developed tools for viewing and analyzing interactive publications (Panorama) and for authoring such documents (Forge). These tools are analogous to Adobe's Acrobat Reader and Professional for PDF documents. Panorama, written in Java, was one of 9 semi-finalists out of 70 entrants in Elsevier's Grand Challenge contest a year ago. Recent enhancements to Panorama include bar charts, and the capability to run natively on Mac OSX, following a formal usability study of the tool.

We also enhanced Panorama to provide Annotation Concepts. A Panorama user may click on text in an interactive publication which is sent to an NLM servlet that identifies the corresponding UMLS concepts. The servlet returns an XML file to Panorama which parses it to provide the preferred UMLS term and semantic group. This also provides linkouts for Medline Plus, eMedline, Family Doctor, and other resources. Further work is ongoing to group concepts by semantic relationships, and investigators are exploring other grouping ideas.

In order to avoid the need for large downloads which would be required by our current approach for some public, we investigated several web-based approaches. In one approach, we modified Panorama to exploit Java's Web Start technology. This technology allows standalone Java software applications to be deployed over the network with a single click. In a second approach, we developed a Web browser version of Panorama (Panorama Lite) using Adobe Flex, thus eliminating the need to download the Panorama software. The only requirement to run it is to have Flash installed. Besides offering easy and intuitive usage, this client version has better line chart and graph support, and includes tables and subsets similar to the original Panorama. A feature unique to Panorama Lite is a Map View that can present data for example, at the county, state, and/or country level, in a color-coded form to readily visualize geographic patterns. LHNCBC has recently collaborated with a publisher to create two interactive

publications from their full text articles. These are hosted on our Web site and available for use through either a browser (Panorama Lite) or with Web Start.

Considering the long-term development of Panorama and Forge in an open source environment, we have taken steps to support third-party developers to create (for example) new viewers for electrocardiograms. For this purpose, we have ported the core code of these tools to support Eclipse plugins, enabling any open source developer to develop new functionality without needing to modify our code.

In a separate approach to new forms of publishing, LHNCBC partnered with the Optical Society of America (OSA) to test the use of interactive publications in an operating journal. The goal was to evaluate software and database infrastructure that enables viewing and analysis of curated, supplemental biomedical source data published in conjunction with peer-reviewed manuscripts, evaluate the educational value of such an infrastructure, and explore the problems of archiving this medium. In order to accomplish these goals, OSA published four electronic special issues of OSA journals on research topics which lend themselves to interactive publishing.

Articles published in these special issues are peer reviewed and fully citable as OSA journal publications indexed in Medline. They are published on the Web in Acrobat format (PDF) with links to source data, videos, and other media objects. The links allow users to quickly and conveniently download these objects and visualize them using interactive viewing software designed to look like an Acrobat plug-in. The viewing software is freely available as a download for all computer platforms. The journal articles and data sets are open access and the source data and associated metadata are searchable and accessible independently from the publication.

Initial feedback from authors, editors, and readers was positive, though some readers faced a learning curve which included software reader program installation and navigation problems. We made mid-course corrections to solve most of these problems. The reader software remained radiology based and therefore posed a problem to readers who were not familiar with that interface. Eighty percent of readers reported an enhanced experience and 50% reported increased learning and understanding.

Screening of Chest X-rays for Tuberculosis in Rural Africa

The LHNCBC has begun a collaborative project with AMPATH (Academic Model Providing Access to Healthcare), an organization that implements the largest AIDS treatment program in the world, and supported by USAID. This new project exploits the convergence of imaging research and system development at LHNCBC and NIH policy objectives in global health. Our objective is to leverage in-house expertise in image processing to clinically screen HIV-positive patients in rural Kenya for lung disease with a special focus on tuberculosis (TB). Since chest radiography is important to the detection of TB and other pulmonary infections prevalent in HIV-positive patients, we have provided AMPATH with lightweight digital x-ray units readily transportable in rural areas. Their staff will take chest x-rays (CXR) of the population and screen them for the presence of disease. These x-ray units are already on site and are being readied for deployment by vehicle.

Since the lack of sufficient radiological services in the area suggests the need for automation to perform the screening, our in-house research effort focuses on developing software to automatically screen for disease in the CXR images. Our researchers are developing algorithms to automatically segment the lungs, detect and remove ribs, heart, aorta and other structures and then to detect texture features characteristic of abnormalities, leading to a 2-class discrimination between abnormal vs. normal case. These machine learning algorithms require sufficiently large training sets (i.e., example x-rays), and to acquire these many options were explored and IRB exemption received. We reached agreement with Montgomery County's TB Control Program which provided about 200 usable de-identified x-rays. We also received a small initial set of x-rays from the lightweight units in Kenya which are also suitable for training the algorithms.

Using these x-rays for training and testing, we are developing an algorithm for detecting lungs and ribs focusing on region-based features such as log Gabor wavelets that exploit the orientation of anatomical structures. A robust identification of lung shape plays a role in detecting TB in CXR since many abnormalities (e.g., pleural effusion) exhibit deformation in lung shape. After extracting the lung fields, the algorithm measures various geometric features that discriminate between normal and effused cases. Ongoing work is in identifying the most successful geometric features.

In parallel, we are developing a binary SVM classifier that uses several features extracted from the x-rays as input, such as histograms of intensity, gradient magnitude and orientation, shape and curvature. Based on these input features, the SVM returns a confidence value, allowing an operator to inspect cases in which the classifier is

uncertain. This initial classifier, showing an accuracy of about 80%, serves as our starting point for ongoing optimization.

Remote Virtual Dialog System (RVDS)

The Remote Virtual Dialog System (RVDS) will make the NLM "Dialogues in Science" series, currently only available in the NLM Visitors Center, available anywhere through the Internet. Support for the project is coming from stimulus funds made possible through the 2009 American Recovery and Reinvestment Act. The project involves the enhancement of programmatic capabilities of the virtual dialogue model to make it sustainable and to allow for expanded applications of the model. During FY2011, we developed a voice-to-text conversion and recognition tool which would be platform agnostic and introduce as little time delay as possible. Additions to the "Dialogues in Science" series and updates of some current series members are in the early stages of development.

Computational Photography Project for Pill Identification

Launched in September 2010, *Computational Photography Project for Pill Identification* (C3PI) is an authoritative, comprehensive, public digital image inventory of the nation's commercial prescription solid dose medications. This effort is directed toward content-based information retrieval (CBIR) to promote patient drug-safety at the national level. Support for the project in FY2011 came from stimulus funds made possible through the 2009 American Recovery and Reinvestment Act.

Initially working in partnership with the NLM Specialized Information Services Division and the US Veterans Administration to study content-based retrieval methods for medical image databases, researchers developed computer vision approaches for the automatic segmentation, measurement, and analysis of solid-dose medications from these pilot datasets including work on robust color classification tools to help identify prescription drugs. Now, working with an expert team from Medicos Consultants, we are creating a collection of digital photographs of prescription tablets and capsules, creating high resolution digital photographs of the front and back surfaces of pharmacy samples, confirming that the images match the description of the medication, developing and matching the images of the samples to relevant metadata (including size descriptions, dimensions, color, and the provenance of the sample).

In FY2011, the contracted team generated over 25,000 images of 1,056 samples of solid-dose pharmaceuticals. The team generates high-resolution, high-quality reference images of each sample. In addition, the contractor is acquiring images from alternate cameras in a variety of lighting conditions to prepare a data collection for a broad effort or national CBIR challenge for the identification of medications. LHNCBC staff are currently preparing a server-based repository and content management system to support distribution and curatorship of the collection.

Virtual Microscope (VM) and Virtual Slides

LHNCBC has created an archive of virtual slides from the teaching set of glass slides from the Department of Pathology of the Uniformed Services University and other collaborating institutions. Researchers digitize, segment, and process each slide 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. The system provides annotations and automatic linking to Medline/PubMed. A related collection of images allows users to search images and automatically link to Medline citations. In collaboration with the Massachusetts General Hospital, researchers are exploring the feasibility of a Virtual Slide mobile application and its use in training. Other collaborations will study its use in telemedicine and teleconsultation. The recent proliferation of iPhone and iPad devices required a modification of software to enable non-Flash capable devices to display virtual slides. Virtual slides are now accessible using both Flash capable and non-Flash capable computers and handhelds.

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 testbed 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 in their original format or in .png format to licensees over the Internet at no cost; and on DVD discs for a duplication fee. Almost 3,300 licensees in 61 countries are applying the data sets to a wide range of educational, diagnostic, treatment planning, virtual reality, and virtual surgeries, in addition to artistic, mathematical, legal, and industrial uses. More than 1,000 newspaper or magazine articles or radio programs have featured the Visible Human Project.

In FY2011, staff continued to maintain 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.

3D Informatics

The 3D Informatics Program (3DI) continued its research mission to address problems encountered in the world of three-dimensional and higher-dimensional, time-varying imaging. LHNCBC provides continuing support for image databases and continues to explore the growing need for image databases, including ongoing support for the National Online Volumetric Archive (NOVA), an archive a collection 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. The archive contains such integrated and multimodal data as virtual colonoscopy matched with recorded video from endoscopic interventions, time-varying 3D cardiac motion, and 4D MRI of a human hand. In anticipation of new sources of data from research partners contributing to the Insight Toolkit, the 3DI group is updating MIDAS software system and adding additional disk space. We are cultivating sources among research teams in confocal microscopy, and we are seeking collections derived from Visible Human Data including segmentations, annotations, and processed information. We continue to serve a broad community with these data, and seek to establish a leadership role through public data distribution.

Throughout FY2011, staff continued collaboration with the National Cancer Institute's Laboratory for Cell Biology and with teams within LHNCBC to visualize and analyze complex 3D volume data generated through dual beam (ion-abrasion electron microscopy) and cryo-electron tomography. This investigation centers on analysis of the spatial architecture of cell-cell contacts and distribution of HIV virions at immunological synapses formed between mature dendritic cells and T cells. The work combines high performance computing with life sciences research, accelerating and empowering investigators in the detection and prevention of cancer and infectious diseases. The resulting visuals have enhanced the understanding and discoveries in the character of several immunological cells, cell structures and their interaction with pathological viruses including HIV.

In FY2011, OHPCC has grown its commitment in high resolution electron microscopy research, expanding our work to include the processing of data collected through transmission electron tomography. We are currently attempting to adapt research software that uses graphics processing units (GPUs) for high performance computing for sub-volume averaging and reconstruction. We are working to develop the research implementation into mature production software for the study of protein structures on the surfaces of HIV and influenza virions.

In FY2011, the 3DI group continued to investigate the use of rapid prototyping technologies in Radiology. We analyzed the x-ray attenuation characteristics of the 3D-printing materials available at NIH. and are presently evaluating the use of contrast agents as printing materials to vary the appearance of the 3D models In April 2011, we published our early work where we modified the 3D printing process through the use of contrast agents, primarily sodium iodide, to create 3D models that mimic human tissue when viewed with x-ray CT scanners. The goal is to create complex, anatomically-accurate models to test diagnostics systems and evaluate and compare their performance under known conditions. We were able to create models that correspond to CT scans of the Visible Human Project male dataset and demonstrate the possibilities for modeling soft tissue and metastatic disease. Our ongoing effort now has begun to focus on resolution and contrast measurement of our methods to ensure the precision and accuracy of our radiological models. This work is conducted in partnership with the National Institute of Allergy and Infectious Diseases.

Insight Tool Kit

The Insight Toolkit (ITK) is a public, open-source algorithm library for the segmentation and registration of high-dimensional biomedical image data. The current official software release is ITK 3.20. Over 845,000 lines of openly available source code comprise ITK, making available a variety of image processing algorithms for computing segmentation and registration of high dimensional medical data on a variety of hardware platforms. ITK can be run on Windows, Macintosh, and Linux platforms, reaching across a broad scientific community that spans over 40 countries and more than 1500 active subscribers to the global software list-serve. A consortium of university and commercial groups, including OHPCC intramural research staff, provide support, development, and maintenance of the software.

ITK remains an essential part of the software infrastructure of many projects across and beyond the NIH. The Harvard led 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. 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. International software packages that incorporate ITK include *Osirix*, an open-source diagnostic radiological image viewing system available from a research partnership between UCLA and the University of Geneva and the Orfeo Toolbox (OTB) from the Centre Nationale D'Etudes Spatiales, the French National Space Administration. Beyond the support of centers and software projects, the ITK effort has influenced end-user applications through supplementing 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.

This year, LHNCBC and the ITK Project coordinated the efforts of groups including General Electric Global Research, the Mayo Clinic, Harvard University, Kitware, Inc., CoSmo Software, the University of Iowa, the University of Pennsylvania, Ohio State University, Old Dominion University, Carnegie Mellon University, Georgetown University, the University of North Carolina at Chapel Hill, and the University of Utah Scientific Computing and Imaging Institute. The research topics supported by these software development efforts include microscopy, digital histology, tumor micro-environments, zebrafish embryology, deconvolution methods for astronomy and astrophysics, image registration for neurosurgery, tumor volume measurement for lung cancer treatment, and video processing for security applications as well as healthcare. A beta release of ITK-version4.0 was released in September, 2011 with a projected release of the stable version in December. Work is expected to continue until June 2012 with a planned release of ITK-v4.2. This work is funded through the American Reinvestment and Recovery Act.

Image and Text Indexing for Clinical Decision Support and Education

This work seeks to exploit ongoing research in both natural language processing and content-based image retrieval by combining processing by text as well as by image features to index the open access journal literature. In this contribution to our Clinical Information Systems effort, techniques are developed to automatically identify relevant figures in biomedical articles (illustrations, clinical images, graphs, diagrams, etc.) that could provide multimedia assistance to clinical decision making. Following the recommendations of the September 2010 Board of Scientific Counselors that evaluated our initial prototype multimedia search engine, we focused on scaling up this system to large collections. In addition to redesigning the system architecture and refactoring the existing code, we improved the user interface, adding functionality (such as filtering images based on their type, filtering journals by clinical specialty, and ranking papers by clinical task, for example, treatment), and processing a larger set of scientific publications acquired from PubMed Central.

Our new experimental multimedia search engine, OpenI, retrieves and displays "enriched citations" - structured MEDLINE citations augmented by image-related text and concepts, and linked to images and image representations based on image features. The document and image processing system that generates these enriched citations, as well as the search engines, are hosted on a multi-computer cluster with a shared file system for distributing computing operations (Hadoop™ MapReduce) and enterprise-level bare-metal virtualization. Due to the demands of high-performance distributed computing, we implement virtualization that enables us to control the number of logical processors, as well as which logical processor runs on a specific physical core. This ensures

predictable performance and scalability as well as tight resource control. Additionally, the cluster design implements the fault-tolerant features needed to ensure high availability of cluster resources. The system is capable of processing up to 65 concurrent information requests per second. The collection currently contains about 20,000 open access articles from PubMed Central, and 250,000 images. OpenI was demonstrated at various forums including the 2011 AMIA symposium, the University of Delaware, UC San Diego and elsewhere.

Staff also conducted research in key areas: ways to represent images with strings, these strings then indexed using traditional search engines such as Lucene; improved methods to automatically segment multi-paneled illustrations into single images, and to partition their captions to correspond to single images; improved methods to extract pointers (arrows, arrowheads, symbols) within images to identify regions of interest, among others. Steps have been taken toward building a visual ontology. We have also developed methods for segmenting lung and brain tissues and extracting key features for imaging properties of several pathologies in these tissues (for example, lung cysts, micronodules, and emphysema.)

Turning The Pages

The goal of the Turning The Pages (TTP) project is to provide the lay public a compelling experience of historically significant and normally inaccessible books in medicine and the life sciences. In this project, we build 3D models for books and develop animation techniques to allow users to touch and turn page images in a photorealistic manner on touch-sensitive monitors in kiosks at NLM, as well as 'click and turn' in an online version. We have also built a 3D 'scroll' model for the 1700 BC Edwin Smith medical papyrus which is 'touched (or clicked) and rolled out'. The online version of TTP is a popular Web site, attracting more than a million page views a month.

In FY2011 we developed an iPad application (app) containing two of these virtual books: Hanaoka Seishu's *Surgical Casebook*, and Hieronymus Brunschwig's *Liber de arte Distilland*. This app has been downloaded more than 3300 times. We are currently adding Robert Hooke's *Micrographia* to the iPad, and developing an iPhone version.

This year we released the kiosk, Web, and iPad versions for *Ketab Ajaeb al-makluqat wa Gara eb al-Mawjudat (Marvelous Creatures and Mysterious Species)* compiled by al-Qazwini in the middle 1200s in what is now Iran or Iraq, and made in Mughal India.

We also launched a complete redesign of the TTP Web site, built around an open source content management system (Wordpress), while conducting research toward a realtime 3D version of TTP. Work in early 2012 will include adding two books to the kiosk and Web versions: Andrew Snape's *The Anatomy of a Horse*, and a Mongolian prayer scroll.

Natural Language Processing and Text Mining

Medical Article Records System

In this project the goal is to introduce automation in many aspects of creating MEDLINE, especially in light of its rapid growth. The MEDLINE database now exceeds 20 million records. The Medical Article Records System (MARS) project, in operation for some years, aims to develop automated systems to extract bibliographic text from journal articles, in both paper as well as electronic forms. For the approximately 1000 journal titles that arrive at NLM in paper form, a production MARS system combines document scanning, optical character recognition (OCR), and rule-based and machine learning algorithms to yield citation data that NLM's indexers use to complete bibliographic records for MEDLINE. Our algorithms extract this data in a pipeline process: segmenting page images into zones, assigning labels to the zones signifying its contents (title, author names, abstract, etc.), pattern matching to identify these entities, lexicon-based pattern matching to correct OCR errors and reduce words that are incorrectly labeled as errors to increase operator productivity.

In FY 2011, LHNCBC staff provided all engineering support for the offsite MARS 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. Developers introduced three additional features to improve MARS performance to support: (1) expansion of the MEDLINE character set, (2) rezoning capability for

Edit users, and (3) user interface for larger monitors used for the Reconcile (operator-verification) stage. These improvements required modifications to several subsystems, in particular the Edit and Reconcile modules.

Developers created and released a new system, Publisher Data Review (PDRS), in June 2011. This system is designed to provide data missing from the XML citations received from publishers: such as databank accession numbers, NIH grant numbers, grant support categories, Investigator Names, and Commented-on Article information. By providing these missing data, PDRS reduces the manual effort in completing the citations sent in by publishers, as well as correct their errors. The automated steps to fill in missing data and to correct wrong data substantially reduces the load on the operators, eliminating the need to look through an entire article to find this information, and then to key them in.

Investigator Names and Commented-on Article are the two most recent fields extracted by PDR. If done manually, extracting names of investigators is a particularly labor-intensive effort since articles frequently contain hundreds of such names. Similarly, identifying commented-on articles is a time-consuming process since it requires operators to open and read other related articles for commented-on information. We have designed and implemented machine-learning methods to extract these two fields.

To assist indexers to automatically retrieve "check tags", indexing terms that are pre-defined (e.g., 'human', 'female', etc.), we developed the CheckTagger system. This prototype Web-based system locates such terms in the article being indexed, and presents them to the indexer for selection, eliminating the need to read through the article to identify the ter. The prototype was demonstrated to the Indexing section staff in October.

Staff are developing another system, WebMARS, to address cases where NLM is missing a journal issue or when citation data from publishers is incomplete. WebMARS is a software tool that operators can use to automatically create missing citations from these problematic issues. This eliminates the current manual labor on part of the operators to type, copy, and paste data from online articles, a very time-consuming step.

The MARS, PDR and WebMARS systems rely on underlying research in image analysis and lexical but this research also enables the creation of new initiatives in which these techniques find application, such as the ACORN project.

Automatically Creating OldMedline Records for NLM

The Automatically Creating OldMedline Records for NLM (ACORN) initiative 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. This year we continued our investigation of scanning, image enhancement, OCR, image analysis, pattern matching, and related techniques to extract unique records from the printed indexes. Finding that many of the printed indexes are available as microfilm, we decided to scan this medium rather than the paper indexes to take advantage of the lower cost of microfilm scanning. In addition, we investigated Web-based information and existing MEDLINE and OldMedline databases to avoid creating duplicate records and to correct OCR errors in citation information. Researchers designed a prototype ACORN system consisting of three main components: Quality Control, Processing, and Reconcile. We have completed the Quality Control module and intend to deliver a pilot ACORN system in FY 2012.

Indexing Initiative

The Indexing Initiative (II) 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.

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). Because of the recent addition of subheading attachment recommendations, indexers now have the option of accepting MTI heading/subheading pairs in addition to unadorned headings. Versions of MTI have also been created to assist in indexing NLM's History of Medicine book collection and for use in Cataloging. In addition, indexing terms automatically produced by a stricter version of MTI are being used as keywords to enhance retrieval of meetings abstracts via the NLM Gateway. These meetings abstracts span the areas of AIDS/HIV, health sciences research, and space life sciences.

MTI's overall performance has dramatically improved this year due primarily to increased collaboration with Library Operations (LO) that has provided several rules based on their in-depth knowledge of the indexing process. Discussions with LO have also produced improvements due to a change in technical focus, emphasizing the precision of recommendations rather than recalling as many topics as possible.

Due to its success with certain journals, MTI was designated as the first line indexer for 23 journals totaling 3,211 articles. As a first line indexer (MTIFL), MTI indexing is still subject to the normal manual review process. The number of MTIFL journals will grow gradually and should prove to be a time and money saver for NLM.

The Indexing Initiative team also worked closely with two NLM Associate Fellows, whose projects were designed to explore automatic assistance to different aspects of the indexing process. One project evaluated the feasibility of automatically indexing comment articles for MEDLINE (30,000 each year). Results of the study showed that approximately 70% of terms assigned by indexers to comment articles matched terms assigned to the article being commented on. We added automation of the comment indexing to the DCMS system in October 2011. Automatically assigning the commented on article indexing to the comment articles will save approximately $280,000 per year in contract indexing costs, while maintaining high quality indexing for these articles. The second project investigates the feasibility of automating the creation of functional annotations about genes, known as Gene Reference Into Function (geneRIF). This project is ongoing.

MetaMap is a critical component of the MTI system and a leading tool around the world in bioinformatics research. Recent work has provided significant speed improvement, XML (eXtensible Markup Language) output, negation identification, and user supplied acronyms/abbreviations list. MetaMap is also now available on Windows, Macintosh and Linux platforms. Users can now build their own data sets with the MetaMap Data File Builder and access their local version of MetaMap via either an embedded Java API (Application Programming Interface) or UIMA (Unstructured Information Management Architecture) wrapper. We had approximately 700 downloads in 2011 for MetaMap, 180 for the Java API and 50 for the UIMA Wrapper. Of note, MetaMap is one of the NLM resources integrated in IBM's Watson system for healthcare applications.

Digital Preservation Research

The Digital Preservation Research (DPR) project addresses an important mandate for libraries and archives: to retain electronic files for posterity as well as to retrieve information from preserved documents through semantic search. To preserve digitized documents, researchers have built and deployed a *System for Preservation of Electronic Resources* (SPER). SPER builds on open source systems and standards (e.g., DSpace, RDF) while incorporating inhouse-developed modules that implement key preservation functions: ingesting, automated metadata extraction and knowledge discovery.

NLM curators are using SPER to preserve more than 60,000 court documents from a historic medico-legal collection acquired from the FDA. In FY2011, NLM processed more than 20,000 documents and added them to a publicly accessible NLM Web site. In addition, SPER is being used to preserve another important collection, from NIAID, comprising conference proceedings of the "US-Japan Cooperative Medical Science Program on Cholera" (CMSP), a program conducted over a 50-year period from 1960 to 2010. Our activities toward this initiative include building a full repository for this collection with more than 10,000 documents, 2,500 research articles and names and affiliations of 6,000 investigators dealing with cholera. We extracted metadata from the document contents using automated metadata extraction (AME) techniques, and then built a portal for research articles, authors, investigators and institutions. The AME processes include: (a) layout analysis to recognize different types of information within a document set; (b) evaluating the effectiveness of models such as Support Vector Machine and Hidden Markov Model for different metadata layouts; and (c) capturing relationships among various entities in the collection from the extracted metadata.

Investigators are conducting research toward knowledge discovery from information preserved in this repository by (a) developing a domain-specific vocabulary, (b) generating RDF graphs or triples from the preserved information using this vocabulary and natural language processing techniques, and (c) building a knowledgebase accessible over the Web. The CMSP repository is being used as the prototype for this research task.

RIDeM/InfoBot

As part of the Clinical Information Systems effort, the RIDeM (Repository for Informed Decision Making) project seeks to automatically find and extract the best current knowledge in scientific publications. The knowledge is provided to several applications (OpenI – a multimodal literature retrieval engine, Interactive Publications, and InfoBot) through RESTful Web services. Developers expanded the services this year to extract salient information from patients' case descriptions.

The related InfoBot project enables a clinical institution to automatically augment a patient's electronic medical record (EMR) with pertinent information from NLM and other resources. The RIDeM API developed for InfoBot allows integrating patient-specific information (e.g., medications linked to formularies and images of pills, evidence-based search results for patient's complaints and symptoms, or MedlinePlus information for patient education) into an existing EMR system. For clinical settings that have no means to use the API, a Web-based interface allows information requests to be manually entered.

The InfoBot API integrated with the NIH Clinical Center's EMR system, CRIS, is in daily use through the *Evidence-Based Practice* tab in CRIS since July 2009. Since then, we have found that the most followed links are to information about medications and protocols of clinical trials. These links are followed twice as often as links to definitions of terms, MedlinePlus and MEDLINE publications. Links to entry-pages of search systems (such as CINAHL) are used rarely. Investigators modified the API to accommodate changes to CRIS, without interruption of services.

De-identification Tools

De-identification enables research on clinical narrative reports. We are designing a software system to de-identify clinical reports that comply with the Privacy Rule of the Health Insurance and Accountability Act. The provisions of the rule dictate removal of 18 individually identifiable health information elements that could be used to identify the individual, the individual's relatives, employers, or household members.

In a recent study, we tested four prominent state-of-the-art systems designed to recognize personal names in free text along with our Clinical Text De-identification (CTD) system. The CTD, with a hierarchical mean sensitivity greater than 99.9%, was better able to detect names in clinical reports than the methods of the other four systems.

We also developed a method to measure the risk of privacy breach by estimating the probability of identifying patients through the undetected personal names that remained uncensored in the text. In this study, CTD was the only system that fully protected the patient privacy without using hospital staff rosters and patient master index.

This study examines reports from one of the largest number of patient reported studies in the literature, using established as well as newly devised methods and metrics.

Librarian Infobutton Tailoring Environment (LITE)

Infobuttons (http://www.infobuttons.org) are context-aware links from one information system to another that anticipate users' information needs, take them to appropriate resources, and assist with retrieval of relevant information. To date, infobuttons are mostly found in clinical information systems (such as EHRs and PHRs) to provide clinicians and patients with access to literature and other resources that are relevant to the clinical data they are viewing. The Laboratory for Informatics Development (LID) has worked with HL7 to develop an international standard to support the communication between clinical systems and knowledge resources. MedlinePlus Connect currently provides an HL7-compliant query capability.

In order to increase the usefulness of infobuttons, they are typically linked not to a specific resource, but to an "infobutton manager" that uses contextual information (such as the age and gender of the patient, the role of the

user, and the clinical data being reviewed) to select from a large library of known resources those that seem most applicable to the situation. The infobutton manager customizes the links to those resources, using appropriate data from the context, and presents the user with the list of custom-selected, customized links. LID is working with investigators at the University of Utah and the Veterans Administration to establish a freely available, HL7-compliant infobutton manager, known as "Open Infobutton" (http://www.openinfobutton.org) to be a national resource for EHR developers and users, providing all clinical systems users with the capability of integrating knowledge at the point of care.

Infobutton managers require knowledge bases to enable them to perform their customization work; Open Infobutton is no exception. The knowledge in these knowledge bases is very institution-specific, including the applications that might call the infobutton manager, the types of questions users might have, and the resources available for resolving those questions at the particular institution (local documents, site licenses, etc.). The Librarian Infobutton Tailoring Environment (LITE), is a user-friendly tool for that can be used by an institution's medical librarians (or someone acting in that role) to provide Open Infobutton with the necessary knowledge for it to customize its responses to requests from that institution. The system is currently in alpha testing now in an installation at the University of Utah (http://lite.bmi.utah.edu).

Terminology Research and Services

The Patient Data Management Project (PDM) brings together several activities centered on lexical issues, including development and maintenance of the SPECIALIST lexicon as well as lexical research. The lexicon and lexical tools are distributed to the medical informatics community as free open-source tools and also delivered with the UMLS information sources.

The Lexical Systems Group recently began a project to enhance the derivational-variants function of the lexical tools. The derivational-variants function uses a set of derivational facts and rules to generate or identify derivational variants of input terms. Derivational variants are words related by a word-formation process like suffixation, prefixation or conversion (change of category). The current derivational variant system has only suffix rules and facts. These rules and facts are hand entered and curated. In order to add suffixation and conversion functionality to the system, the PDM team has developed a method to automatically extract candidate pairs of words that may be derivationally related, which helps automate the creation of rules and facts for suffixation and conversion.

The SPECIALIST Lexicon and Lexical tools are open source and freely downloadable. During the year, our web page had an average of 3,229 unique visitors per month. We had an average of 1,700 downloads per month in 2011. We count 14 internal users, 27 known academic users and 22 known international organizations among our users.

The 2012 release of the SPECIALIST Lexicon will contain over 462,000 records, representing over 830,000 forms, an increase of over 13,000 records from the 2011 release. Many of the new terms are derived from de-identified clinical records from our own De-identification project and from the MIMIC database. We plan to further extend the lexicon by adding consumer-level medical vocabulary. The Consumer Health Vocabulary recently added to the UMLS Metathesaurus will supply some consumer terms. We also plan to obtain a frequency list of consumer terms quoted in clinical records through a collaboration with the University of Utah.

Medical Ontology Research

The Medical Ontology Research (MOR) project focuses on basic research on biomedical terminologies and ontologies and their applications to natural language processing, clinical decision support, translational medicine, data integration and interoperability.

During FY2011, staff investigated issues including quality assurance in ontologies, the representation of pharmacologic classes in biomedical terminologies, and approximate matching techniques for mapping clinical drug names to standard terminologies. Many of these studies leveraged the Semantic Web technologies including RDF - the Resource Description Framework - and triple stores (e.g, Virtuoso), which proved to be useful resources for integrating of biomedical information.

Researchers contributed to the LHNCBC training program by providing mentorship to four graduate and two post-doctoral students, working with them of issues including data integration for pharmacogenomics studies, quality assurance in the UMLS, and web services composition.

Research activities this year resulted in two journal articles, five papers in conference proceedings, two book chapters and five invited presentations. We continue to collaborate with leading ontology and terminology centers, including the National Center for Biomedical Ontology, the International Health Terminology Standards Development Organization (SNOMED CT) and the World Health Organization (ICD 11).

Semantic Knowledge Representation

The Semantic Knowledge Representation (SKR) project conducts basic research in symbolic natural language processing based on the UMLS knowledge sources. A core resource is the SemRep program, which extracts semantic predications from text. SemRep was originally developed for biomedical research. Researchers are developing a general methodology for extending its domain, currently to influenza epidemic preparedness, health promotion, and health effects of climate change.

The SKR project maintains a database of 60 million SemRep predications extracted from all MEDLINE citations. This database supports the Semantic MEDLINE Web application, which integrates PubMed searching, SemRep predications, automatic summarization, and data visualization. The application helps users manage the results of PubMed searches by outputting an informative graph with links to the original MEDLINE citations and by providing convenient access to additional relevant knowledge resources, such as Entrez Gene, the Genetics Home Reference, and UMLS Metathesaurus.

SKR efforts support innovative information management applications in biomedicine, as well as basic research. The project team is using semantic predications to find publications that support critical questions used during the creation of clinical practice guidelines (with support from NHLBI). Investigators are devoting significant research to developing and applying the literature-based discovery paradigm using semantic predications. One such project is investigating the physiology of sleep and associated pathologies, such as declining sleep quality in aging, restless legs syndrome, and obstructive sleep apnea; another exploits predications and graph theory for automatic summarization of biomedical text. Further, the SKR team is collaborating with academic researchers in using semantic predications to help interpret the results of microarray experiments, to investigate advanced statistical methods for enhanced information management, and to address the information needs of clinicians at point-of-care.

Information Resource Delivery for Researchers, Care Providers, and the Public

The LHNCBC performs extensive research in developing advanced computer technologies to facilitate the access, storage, and retrieval of biomedical and consumer health information.

ClinicalTrials.gov

ClinicalTrials.gov provides the public with comprehensive information about interventional and observational clinical research studies. ClinicalTrials.gov receives over 50 million page views per month and hosts approximately 800,000 unique visitors per month. At the end of FY2011, the site had nearly 115,000 protocol records, nearly 4,700 of which display summary results, conducted in all 50 states and in over 175 countries. Approximately one-third of the studies are or will be open to recruitment, and the remaining two-thirds are closed to recruitment or completed. Data are submitted by over 9,300 study sponsors which include the U.S. Federal government, pharmaceutical and device industries, academic, and international organizations, through a Web-based Protocol Registration System, which allows sponsors to maintain and validate information about their studies.

ClinicalTrials.gov was established by the NLM in FY2000 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. Since that time, ClinicalTrials.gov has undergone enhancements to support other registration policies and to implement the requirements under Section 801 of the Food and Drug Administration Amendments Act of 2007 [Public Law 110-85]. In FY2011, new registrations were submitted at an average rate of 350 records per week. In September 2008, ClinicalTrials.gov launched the "basic results" database, which complements the registry. Registered trials may now include tables of summary results data on primary and

secondary outcomes and adverse events, as well as information on the characteristics of the participants studied. Since the beginning of its operation, over 6,200 results records have been submitted by over 700 study sponsors. The average number of submissions per week has increased, with an average of 60 new results records submitted per week at the end of FY2011. The requirements for the expanded registry as well as the results database will be further elucidated through rulemaking. NLM is working with other institutes and centers and the Office of the Director at the NIH and the Food and Drug Administration (FDA) on a Notice of Proposed Rulemaking. The combined registry and results database provides access to critical information about ongoing and completed clinical research for patients, healthcare providers, and policy decision makers.

In FY2011, ClinicalTrials.gov was actively involved in educating the public and data providers on the new law, developing new features to improve the usability of the system, and promoting standards of transparency in clinical research through trial registration and results reporting. This information was communicated to a broad range of U.S. and international stakeholders through presentations, workshops, a series of webinars, and peer-reviewed publications. ClinicalTrials.gov continues to collaborate with other registries, professional organizations, and regulators in working towards developing global standards of trial registration and reporting to results databases.

Genetics Home Reference (GHR)

Genetics Home Reference (GHR) is an online resource that offers 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 patients and their families, the GHR Web site currently includes user-friendly summaries of 670 genetic conditions, more than 900 genes, all the human chromosomes, and mitochondrial DNA. The Web site also includes a handbook called *Help Me Understand Genetics*, which provides an illustrated introduction to fundamental topics in human genetics including mutations, inheritance, genetic testing, gene therapy, and genomic research.

Genetics Home Reference celebrated its eighth anniversary in 2011. In the past year, the project expanded its genetics content for consumers. Specifically, GHR staff added more than 200 new summaries to the Web site in FY2011. We intend to continue this rate of production in FY2012, covering additional Mendelian genetic disorders as well as more complex disorders. The team also plans to continue expanding the gene families feature, which currently includes explanations of about 60 families of related genes. This year, the site averaged more than 21,700 visitors per day and about 33 million hits per month. GHR continues to be recognized as an important health resource.

This year, GHR staff performed outreach activities to increase public awareness of the Web site. The project 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 visiting journalists and students at a local university medical school. Staff members attended and represented the project at several major genetics conferences, and will continue to educate others about this useful resource in FY2012.

Profiles in Science Digital Library

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 unique biomedical information previously accessible only to patrons able to make an in person visit to the institutions holding the physical manuscript collections. *Profiles in Science* 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. It 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 *Profiles in Science* collections are consistently popular. The Web site averages over 77,000 unique visitors each month. This year, the collections of pioneering surgeons Charles R. Drew and Clarence Dennis were added to *Profiles in Science.* Staff added 1,152 transcripts of documents, making handwritten items searchable and providing alternatives to PDF format files. Staff also added 31digital items to the 33 existing Profiles in Science collections. Currently 141,583 image pages that constitute 26,868 digital items are available on *Profiles in Science*. The Web site now features the archives of thirty-two prominent scientists and health advocates:

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

Evidence Based Medicine - PubMed for Handhelds

PubMed for Handhelds was developed and released in FY2003 to facilitate evidenced-based medical practice with Medline access at the point of care via smartphones, wireless tablet devices, netbooks or portable laptops. PubMed for Handhelds (PubMedHh) 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 abstracts allows a clinician to quickly read summaries at the point of need. A “consensus abstracts” element provides rapid review of multiple publications with smartphones at the point-of-care. This corresponds well with a recent review of PubMedHh server logs that showed that more than 90% of queries were clinical in nature. Randomized controlled trials using simulated clinical scenarios concluded recently at the Uniformed Services University, University of Botswana-University of Pennsylvania and the National Telehealth Center and Philippine General Hospital, Manila to evaluate the usefulness of abstracts in clinical decision making. We also developed and submitted an iOS (iPhone, iPad devices) app for PubMed for Handhelds.

Clinical Vocabulary Standards and Associated Tools

Multiple projects in this area continue to promote the development, enhancement**,** and adoption of clinical vocabulary standards. The CORE Problem List Subset of SNOMED CT is published in the UMLS as a specific content view. RxTerms facilitates the use of RxNorm as an interface for medication orders. Inter-terminology mapping promotes the use of standard terminologies by creating maps to administrative terminologies, which allows re-use of encoded clinical data. The Newborn Screening Guide combines terminology and electronic messaging systems to facilitate care and research related to newborn screening. Another effort focuses on the development of a consumer-friendly medical problem and procedure terminology. LHNCBC continues to play an important role in the UMLS project in research related to the various UMLS knowledge sources and providing support in UMLS production and user support. The inter-terminology maps are also available through the UMLS.

The CORE Problem List Subset of SNOMED CT

The problem list is considered to be an essential part of the Electronic Health Record (EHR) by various sanctioning bodies and medical information standards organizations, including the Institute of Medicine, Joint Commission, American Society for Testing and Materials and Health Level Seven. An encoded problem list is also one of the core objectives of the "meaningful use" regulation of EHR published by the Department of Health and Human Services (HHS). Problem lists have value beyond clinical documentation. Common uses include the generation of billing codes and clinical decision support. To drive many of these functions, an encoded problem list (as opposed to data entered as free-text) is often required. However, most institutions use their own problem list vocabularies. This lack of a common standard leads to duplication of effort and impedes data interoperability.

Based on data collected from seven large-scale US and overseas healthcare institutions, a detailed study was done on the nature of the local problem list vocabularies. One significant finding is the low level of overlap between these vocabularies, with an average pairwise overlap of around 20%. However, terms that are shared among institutions were used eight times more frequently than concepts unique to one institution, which lends support to the idea of having a common core of problem list terms across institutions. Since SNOMED CT is a designated standard for problem lists according to the "meaningful use" criteria, a CORE (Clinical Observations Recording and Encoding) Problem List Subset of SNOMED CT, which contained about 6,000 concepts and represented the most frequently used problem list terms, was identified and made available to SNOMED CT users. The CORE Subset can be used as a starter set for institutions that do not yet have a problem list vocabulary based on SNOMED CT. This will save significant development effort and reduce unintentional variations in the choice of terms. Existing problem list vocabularies can also be mapped to the CORE Subset which will facilitate data interoperability. Since publication, the CORE Subset has received considerable attention within the SNOMED CT user community. The IHTSDO (International Health Terminology Standards Development Organization) used the CORE Subset to focus its quality assurance effort on clinically important concepts. The MedlinePlus Connect Project, which facilitates online linkage to patient education information, has mapped all concepts in the CORE Subset to MedlinePlus health topics. There is ongoing effort to map the CORE Subset to the ICD classifications (ICD-10 and ICD-10-CM) which will promote the adoption of SNOMED CT by allowing re-use of SNOMED CT encoded clinical data. We have an ongoing collaboration with the Mayo Clinic, Intermountain Healthcare, and Vanderbilt University to evaluate the CORE subset. Investigators completed a study of the use of post-coordination to expand the coverage of the CORE problem list. We released a new update of the Subset in August, based on the newest release of SNOME CT and remapping of terms previously unmapped. The remapping process added 137 new concepts. The Veterans Administration has provided us with a new dataset which will add additional concepts to the CORE Subset.

RxTerms

RxTerms is a free, user-friendly and efficient drug interface terminology that links directly to RxNorm, the national terminology standard for clinical drugs. The Centers for Medicare and Medicaid Services has used RxTerms in one of their pilot projects in the post-acute care environment. It is also used in the NLM PHR. RxTerms is available for download from the NLM Web site. There is ongoing effort to align data elements between RxTerms and RxNorm. Investigators are currently reviewing the dose form information in RxTerms to improve usability. RxTerms is updated every month with the full monthly release of RxNorm.

RxNav

Released in September 2004, RxNav was first developed as an interface to the RxNorm database and was primarily designed for displaying relations among drug entities. In addition to the browser, SOAP-based and RESTful application programming interfaces (APIs) were created, enabling users to integrate RxNorm in their applications. Examples of use include mapping drug names to RxNorm, finding the ingredient(s) corresponding to a brand name, and obtaining the list of NDCs for a given drug.

During FY2011, staff released SOAP and RESTful APIs for two other drug information sources also integrated with RxNav: RxTerms, an interface terminology for prescription writing or medication history recording; and NDF-RT, a resource that links drugs to their pharmacologic classes and properties, including indications, contra-indications and drug-drug interactions.

Usage of RxNav and the APIs has increased significantly, from 10 million queries last year to over 25 million queries in FY2011. Users include clinical and academic institutions, as well as pharmacy management companies, health insurance companies, EHR vendors, and drug information providers. In the future, an application facilitating the use of APIs will be developed, e.g., for mapping large amounts of terms and codes to RxNorm, for querying pharmacological classes (in NDF-RT) from codes in RxNorm, and for crosswalk purposes between drug vocabularies through RxNorm.

Electronic Reporting of Units of Measure Standards

During FY2011, LHNCBC developed a Table of Example UCUM codes for Electronic Messaging, enumerating the Unified Code for Units of Measure (UCUM) units codes. UCUM codes are unambiguous and computable units of measure that include tables for converting one unit to another of the same dimension, e.g., pounds to kilograms. UCUM codes have been adopted as *the* unit of measure by many standards bodies. The example units table includes the 815 units of measure including the vast majority of units of measure used in clinical medical codes. This table will assist adoptees' use of UCUM in electronic health records, public health, and research projects. The content for this table was derived from Regenstrief Institute and Intermountain Healthcare.

We delivered this table to the HHS Office of the National Coordinator for Health Information Technology (ONC) Standards and Interoperability Framework Committee for their repository, and the Laboratory Results Interface Initiative for reference and inclusion in an HL7 Implementation Guide. Many Organizations including HL7 and IEEE support UCUM, and groups in the US and Canada are already using the UCUM table, released in September 2011.

LOINC Standards for Identifying Clinical Observations and Orders

In FY2011, LHNCBC continued to work with the Regenstrief Institute (RI), major laboratory companies, several NIH institutes, and other organizations to develop the size and breadth of the LOINC database. By the end of FY2011, LOINC had over 13,000 users in more than 140 countries and supported nine languages. We worked with RI and the LOINC Committee to create more than 6,000 new LOINC terms for both laboratory and clinical variables. These new terms included those from the PROMIS and PhenX survey instruments and were created as a result of collaborations with NIAMS and NHGRI, respectively. We continued to work with the major laboratory companies and the American Clinical Laboratory Association (ACLA) to clarify the content of many of the most frequently ordered test panels.

We released the Top 2000+ list of lab tests that represent over 98% of the test volume carried by three large organizations that cover both inpatients and outpatients, as well as a companion Mapper's Guide that includes information on the development and organization of the Top 2000+ lab tests as well as advice and guidance about which codes to choose in specific situations (http://loinc.org/downloads/usage/obs). The goal of the Top 2000+ list and Mapper's Guide is to help hospitals, providers, laboratories, researchers, and others with the task of mapping their local codes to LOINC. We updated both the LOINC web search tool (http://search.loinc.org) and the Regenstrief LOINC Mapping Assistant (RELMA®) so that the search results can be constrained to the top 2000+ tests.

During FY2011 we also reviewed, revised and distributed version 1.2 of the Common Lab Orders Value Set (http://loinc.org/downloads/usage/orders), which contains 332 laboratory tests that comprise over 95% of the lab test order volume in the U.S. We worked with RI to develop and publish the first version of The Table of Example UCUM Codes for Electronic Messaging (http://loinc.org/usage/units), which is based on content from Intermountain Healthcare and is a guide for mapping common lab result units to standard UCUM units. We also helped develop the HL7 Version 2 Implementation Guide: Clinical Genomics; Fully LOINC-qualified Cytogenetics Model, Release 1.

Newborn Screening Coding and Terminology Guide

In collaboration with the Health Resources and Services Administration (HRSA), the Centers for Disease Control and Prevention (CDC), and the NIH *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), LHNCBC has created and maintained LOINC codes for all variables used in newborn

screening intended for delivery in an HL7 message implementation. Staff also invested special effort to develop a consensus approach to reporting findings of newborn screening for hemoglobinopathies, severe combined immunodeficiency (SCID), and lysosomal storage disorders (LSDs). These specifications are being adopted by many state newborn screening laboratories across the country. Kentucky, Oregon (which is the regional laboratory for 5 additional states), Colorado, New York, and Illinois are furthest along in the implementation work.

LHNCBC is in the process of developing codes for newborn screening of Critical Congenital Heart Disease in addition to the Recommended Uniform Screening Panel of the Secretary's Advisory Committee on Heritable Disorders in Newborns and Children. LHNCBC is also working with our partners to standardize data collection and coding for short- and long-term follow up, beginning with the laboratory tests for confirmation and diagnosis of conditions targeted by newborn screening.

Communication Infrastructure Research and Tools

LHNCBC 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.

Videoconferencing and Collaboration

LHNCBC continues to investigate, review, and develop collaboration tools, research their application, and use the tools to support ongoing programs at the NLM. In our work with uncompressed high definition video over Internet Protocol (IP), we determined strengths and weaknesses of each of the three technologies (iHDTV, UltraGrid, and Conference XP) and we continue to overcome problems encountered in the delivery of uncompressed video due to differing platforms. We are monitoring the High Definition (HD) open source work of Video Conferencing Tool developers regarding H.264 compression. VIC is used by the AccessGrid, an open source collaboration tool widely deployed in universities and research centers and used in the OHPCC Collaboratory for research work and to support NLM programs. Finally, we are reviewing newer, cloud collaboration tools. We published a comparison of the major compression/decompression (codecs) available in the OHPCC Collaboratory for High Performance Computing and Communications (Collab), and we are in the process of writing systematic reviews of the uncompressed and cloud technologies the research team is studying. The overall research effort involves studying and testing collaboration technologies technically in our laboratory first and then deciding which warrant further applied research in clinical or educational settings and which might be put into use supporting NLM programs.

Until recently, iHDTV was the only uncompressed video system sufficiently robust to use in a clinical trial but OHPCC staff have worked with the developers of UltraGrid at Masaryk University in Brno, Czech Republic and it has improved considerably. Most notable has been the integration and synchronization of audio along with the video. We also provided video cards to the lead developer of ConferenceXP (CXP) at the University of Washington so he could directly test the programs he wrote to do uncompressed video for that system. The research team was only able to get the compressed video to work for CXP. The team concentrated on UltraGrid and ConferenceXP technologies because development of iHDTV has ceased. Consequently, dual collaborations are planned with both Masaryk and the University of Washington, with a special interest in Masaryk, since the OHPCC and Masaryk research groups share an interest in 3D HD videoconferencing, 4K video, varied forms of HD compression, and the use of dynamic circuit networks (DCN) to ensure quality of service. The team will continue to collaborate with the Rochester Institute of Technology (RIT) to test open source software for compressed HD videoconferencing based on the H.264 video standard. RIT, Manchester University and others are refining the software and incorporating it into the AccessGrid, a technology used in the Collab for distance learning.

The installation of a 10 Gigabits per second (Gbps) network in the Collab has greatly facilitated collaboration with other institutions and our ability to test uncompressed video. Prior to installation, the technologies could only be tested back to back because they consumed bandwidth exceeding our network's capacity. The teams at NLM and Masaryk were able to conduct videoconferences using uncompressed HD video with synchronized audio over their research networks without the use of dynamic circuit networks (DCN) and to demonstrate this capability to the LHNCBCBoard of Scientific Counselors. Eventually, DCN will be tested because traffic over the research network backbone is unpredictable.

Collaborators installed iHDTV systems at the Medical University of South Carolina (MUSC) to study uncompressed video's use as a diagnostic tool for teledermatology. We selected teledermatology was chosen as a research domain because previous research has shown it to be particularly difficult to use standard definition video to do remote dermatological exams. We measured diagnoses, clinician confidence, decisions to biopsy and physician and patient encounter satisfaction and compared under the following conditions: 1) when patients are examined in-person, 2) when patients are examined using uncompressed high definition video, 3) when patients are examined using compressed high definition video using a standard employed by all major commercial videoconferencing manufacturers, and 4) when patient data (history and photos) are used to assess patients by typical store and forward methods. iHDTV systems were chosen for the study because, at the time, audio was not integrated into UltraGrid and ConferenceXP could only transmit compressed video.

We completed follow-up research with MUSC on video medical interpretation. The follow-up study used lower quality video (less than full screen) and cell phone technology to assess video interpretation in a pharmacy setting at an outpatient clinic. We also did extensive tests with VSee, a low bandwidth video program. Results were mixed; however, the effort proved the feasibility of and requirements for doing video medical interpretation over cellular networks and we will publish the results in FY2012.

Staff continued to work with SIS on distance education outreach program for minority high school students and with the NIH Library to offer NCBI database and other bioinformatics training at a distance. In FY2011, staff conducted bioinformatics programs with the University of North Carolina at Chapel Hill, the University of Tennessee at Memphis, and Virginia Commonwealth University.

OHPCC Collaboratory for High Performance Computing and Communication

LHNCBC established the OHPCC Collaboratory for High Performance Computing and Communication (Collab) as a resource for researching, testing, and demonstrating imaging, collaboration, communications and networking technologies related to NLM's Next Generation Network initiatives. Staff use this infrastructure to test new technologies of interest to NLM and to conduct ongoing imaging, collaboration and distance learning research both within LHNCBC and outside 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 enables staff to collaborate with others at a distance and, at the same time, demonstrates much of the internal and external work being done as part of the NLM 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. Staff upgraded the H.323 technology last year to include compressed HDvideo and 3D display and DVD playback technology. This year we acquired a 3D camcorder for the purpose of using it and/or dual non-3D HD camcorders to transmit 3D HD video in future videoconferencing research.

Computing Resources Projects

The Computing Resources (CR) Team has a variety of core projects that builds, administers, supports, and maintains an integrated and secure infrastructure to facilitate the research and development (R&D) activities at LHNCBC and thereby augments the overall effectiveness of research projects. The integrated secure infrastructure contains network, security, and facility management, and system administration support for a large number of individual workstations and shared servers.

The network management includes the planning, implementation, testing, deployment and operation of high-speed networks over Internet and Internet-2. One core project implemented the 10-gigabit network, and studied many advanced communication protocols to support LHNCBC collaboration activities and research projects. The network management team also participated in the study of Trusted Internet Connection (TIC) consolidation and evaluated the impacts to the NIH and NLM.

The security management team incorporates security operations into firewall administration, patch management, anti-virus management, intrusion monitoring, security and vulnerability scanning, and vulnerability remediation to ensure an IT working environment that is safe from overall security perspectives. One core project studied and implemented a unified patch management to improve LHNCBC's overall security measures. Another

core project implemented automated security audit system that ensures all system at LHNCBC comply with policies. The security management team also studied and evaluated the network performance impact of web anti-virus software to the NLM, and delivered secure ID to all servers.

The facility management team facilitates products' and servers' deployments, including power acquisition, network planning, cabling connection, and space allocation in the central computer room as well as at co-location facilities. One core project studied, designed, and implemented an enterprise console management system that enabled LHNCBC to remotely manage large numbers of servers.

The system administration team provides center-wide IT services such as DNS, NIS, data backup, printing, and remote access to ensure an efficient business operation. Core projects include Federal Information Security Management Act (FISMA) compliance facilitation and support, and Continuity of Operation (COOP) process establishment. Other projects include a centralized ticketing system for better customer support, and an enterprise secure remote access system to meet emergency requirements like pandemic flu. Additionally, the system administration team supports the shared computing resources such as security audit, system buildup, and security certification.

Disaster Information Management

Lost Person Finder

The Lost Person Finder (LPF) project, seeking to develop systems for family reunification in the aftermath of a mass casualty event, was initiated as part of the Bethesda Hospitals Emergency Preparedness Partnership (BHEPP). The systems developed in this project combine image capture, database, and Web technologies, and address both hospital-based and community-wide disaster scenarios.

The hospital-based LPF system includes means to photograph victims at the hospital's triage station, and to capture these pictures, general health status, and descriptive metadata (name, age range, gender, identifying features) using TriagePic, a Windows application for triage staff. This data enters a MySQL database which can be searched via a Web site or via web services built by extensively customizing the open-source Sahana disaster management system. The LPF system also features a "Notification Wall" that displays images of victims on both computers and large auditorium screens for family or staff. In 2009 and 2010 we participated in large-scale multi-institutional drills (*Collaborative Multi-Agency Exercise* or CMAX) and demonstrated TriagePic usage, search capability, and the Notification Wall displays at the Navy and Suburban hospitals in Bethesda. Because the LPF collects personally identifiable information (PII) such as pictures and names, we developed and received approval for a Certification and Accreditation Process and a System Security Plan.

One of the lessons learned in 2010 was the desirability of a unified Web site and database, capable of being used to respond to any disaster anywhere. We developed a prototype unified site, NLM Person Locator (PL) to hold data from multiple disasters, thereby eliminating the need to build multiple Web site and database instances. Staff further developed and tested the unified PL Web site during a multi-agency disaster drill (October 2011, Capital Shield 2012) at Suburban Hospital. We also deployed it for disasters worldwide, including the Christ Church Earthquake (February), the Japanese Earthquake and Tsunami (March), the Joplin Tornado (May) and the Eastern Turkey Earthquake (October).

For Capital Shield 2012, we ported our TriagePic application to a touchscreen tablet platform running Windows 7, and used by hospital staff successfully. The tablet has both forward (3 Megapixels) and rear facing cameras and a touchscreen interface allowing easy image and metadata capture by stylus or touch. Nurses and patient registrars at Suburban Hospital learned to use the tablet's camera and touchscreen interface easily, literally just before the drill started.

We continue to customize and enhance the Sahana open-source software, evolving suitable modules to address LPF's specific needs and findings from the drills and real disasters. Developers improved responsiveness through the development of SOAP-based web service modules, enabling bidirectional communications of data to and from the Web site. We improved search capability through use of Solr/Lucene technology and SQL query optimization for low latency searching. Further, in an international exercise at a workshop in Spain, we tested automatic data interchange and mirroring with the Google Person Finder system in May using PFIF (Person Finder Interchange Format). We refined the automated mirroring process and used it in the Eastern Turkey Earthquake.

Ongoing research in this project includes:

- Developing techniques in image and face matching to de-duplicate records, detect and localize faces, and subsequently match faces using image feature-based methods.
- Experimenting with new communications protocols to collect victim data prior to arrival at a hospital, e.g., from ambulances.
- Refining Web services between TriagePic (or our iPhone app, ReUnite) and PL, to facilitate control and correction of reports sent to the site.
- Investigating the incorporation of our systems into a hospital's normal workflow, with a focus on the optimum mix of privacy, security and openness.
- Extending ReUnite to other platforms: iPad as well as the Android platform.

Video Production, Retrieval, and Reuse Project

This development area encompasses four projects that contribute to the NLM Long Range Plan goal of promoting health literacy and increasing biomedical understanding.

The NLM Media Assets Project provides the NLM with easy access to audio-video resources for improved biomedical communications. This includes:

• The Hypervideo Personal Digital Library/ Digital Video Library (a computer aided search, retrieval and viewing database).
• The NLM/History of Medicine Exhibits Audiovisual Assets Management.
• Archival management of the Visible Human Project film and digital image dataset.

The NLM Support Project provides NLM with the audio/video support and development needed to promote and augment NLM's operation. This includes:

- Support for the maintenance and operation of the NLM state-of-the-art auditorium, board room and conference rooms.
- Ongoing production, post-production, and authoring services for the development of Internet video, interactive multimedia for large-screen and tablet devices and displays, and Blu-Ray DVD production.

The LHNCBC Research Support Project contributes to improving access to high quality biomedical imaging information. This project includes:

• The APDB/NCI collaboration on 3D visualization of molecular structures and functions in the discovery of disease and treatment.
• The Movement Disorders Database (a digital archive of movement disorders patients going through diagnostic routines).
• The Profiles in Science video modules.
• The Visible Human imaging and visualization research.

The LHNCBC Core Resources Project provides research into developing new technologies for disseminating biomedical information. This project includes:

• The LHNCBC Research Update Modules.
• Ultra High Definition Imaging Research.
• Ongoing design and development of image-rich web sites in support of biocommunications.
• Audio/video/imaging archiving and asset management.

A number of LHNCBC projects require videographics, interactive multimedia development, imaging, animation, or video production as part of the overall project objectives. A major effort in this area is improvement of rendering times for videographics and 3D visuals and animations for DVD and other interactive multimedia productions.

Extensive work continued toward the planning and development of interactive multimedia for the FY2011 NLM Exhibition "Native Voices: Native People's Concepts of Health and Illness." APDB staff worked with the Director, NLM and CgSB staff to review the extensive video interview database to establish major content areas based on thematic indexing of all interview transcripts. Based on this work, APDB contributed to the successful

production of interactive video kiosk programs within the exhibition which opened this year. We used the highest quality video format standards throughout, from production to encoding and compression and display. For this project we encoded, video in formats for distribution across multiple platforms including iPad and mobile smart phone applications, which are featured throughout the exhibition. Our focus on video compression codecs for small screen delivery, navigation, and search capabilities is an ongoing area of research related to the work of the exhibition as well as many other areas of NLM's information programs.

Digital Video Archive

The extensive digital video library assembled for the NLM Director's exhibition interview database has added to APDB's ongoing effort in the digitization, organization and accessible storage of large-scale video libraries. Digital workflow management and file format standards established for exhibition production support are now being applied to a major project within APDB to convert the large, historical tape library within LHNCBC, containing over four decades of NLM programs into a viable digital repository accessible for future use.

Biomolecular Visualization

APDB staff continued to collaborate with the National Cancer Institute's Laboratory for Cell Biology and with OHPCC to visualize and analyze complex 3D volume data generated through dual beam (ion-abrasion electron microscopy) and cryo-electron tomography. In this work we focus on the analysis of the spatial architecture of cell-cell contacts and distribution of HIV virions at immunological synapses formed between mature dendritic cells and T cells.

APDB staff are working with NCI scientists on a novel imaging technique developed by the NCI: Ion-Abrasion Scanning Electron Microscopy, which slices and images stacks of data at the microscopic level. One recent exploration yielded a remarkable 3D volumetric model revealing the transfer of HIV from T-cell along filopodia to astrocyte. The final illustrative image was a finalist in the National Science Foundation's Visualization Challenge.

The APDB-produced medical illustrations and animations have illuminated the character of several immunological cells, cell structures and their interaction with pathological viruses including HIV. The resulting visuals have enhanced the understanding and discoveries in the character of several immunological cells, cell structures and their interaction with pathological viruses including HIV.

Training and Education at LHNCBC

LHNCBC is a major contributor to the training of future scientists and provides training for individuals at many stages in their careers. Our Informatics Training Program (ITP), ranging from a few months to two years or more, is available for visiting scientists and students. Each fellow is matched with a mentor from the research staff and participates actively in LHNCBC research projects.

During FY2011, 61 participants from 15 states and 7 countries received training and conducted research in a wide range of disciplines: 3-D image processing, biomedical ontology research, biomedical terminology research, content-based information retrieval, de-identification of medical records, evidence-based medicine systems, image, text and document processing research, information retrieval research, literature-based discovery research, natural language processing research, personal health record research, pill identification research, research into collaboration tools, semantic web research and systems for disaster management.

The program maintains its focus on diversity through participation in programs for minority students and emphasizes the Hispanic Association of Colleges and Universities and the National Association for Equal Opportunity in Higher Education summer internship programs.

The ITP also sponsors a Clinical Informatics Postdoctoral Fellowship Program, funded by LHNCBC, to attract young physicians to NIH to pursue research in informatics. This program is run jointly with the Clinical Center to bring postdoctoral fellows to labs throughout NIH. LHNCBC 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. We also host a two-month NLM Rotation Program which provides trainees from NLM-funded Medical Informatics programs an opportunity to

learn about NLM programs and current LHNCBC research. The rotation includes a series of lectures showcasing research conducted at NLM and provides an opportunity for trainees to work closely with established scientists and fellows from other NLM-funded programs.