THE LISTER HILL NATIONAL CENTER
FOR BIOMEDICAL COMMUNICATIONS

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

TECHNICAL REPORT
LHNCBC-TR-2012-003

Lister Hill National Center for Biomedical Communications Annual Report FY2012

Clement J. McDonald, M.D.
Director

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

LHNCBC
FY2012 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 NLM Personal Health Record (PHR) is a web-based tool for consumers to keep track of their own and their families' health information. The goals of the PHR are to help consumers manage and understand their health care problems, to facilitate federal goals for clinical data interchange using national vocabulary standards, and to determine whether using personal health records can improve adherence to preventive care recommendations and improve consumer health.

Consumers can use the main PHR page to enter key health information, including medical conditions, surgeries, medications, allergies, and immunizations. They can also enter due dates for prescription refills and scheduling appointments. On the main PHR page consumers can also record questions they want to ask their doctor as well as medical contact information, and they can view educational material that pertains to their specific health information via information links to MedlinePlus and other trusted resources. On a separate page of the PHR, consumers can enter data for lab results, radiology reports, and other screening and diagnostic procedures. In addition, they can track measures of wellness including mood, diet, sleep, and exercise, as well as disease-specific parameters such as episodes of asthma or seizure frequency.

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 vocabulary standards in the NLM PHR enables many automated features, such as personalized reminders about preventive care and healthy behaviors based on the specific data the consumer has entered.

In FY2012, LHNCBC researchers and developers continued to improve the capabilities of the PHR. On the technical side, we updated the Ruby on Rails platform to version 3.2. We updated the project infrastructure by switching to a new software version control system and organizing the software into multiple branches to better support a live system. We are well along on an upload feature to allow consumers to incorporate their medical documents (such as radiology or lab reports) in various formats and have developed functionality to track PHR usage statistically. We have also begun work on a feature to allow consumers to upload their health information from a structured Continuity of Care Document (CCD) – which meaningful use regulation requires all health care systems to provide to their patients. We implemented a new user interface that is more visually appealing than the old interface. To accommodate sightless individuals using assistive technology (e.g. text-to-speech), we also created a completely parallel system that replaces graphical interactions with text-based navigation.

On the content side, we developed and implemented several new consumer-oriented panels including a pediatric developmental milestone tracker, Apgar record, and seizure activity log. Several more panels are in development including disease trackers for conditions such as sickle cell disease, migraine headaches, and chronic obstructive pulmonary disease. We also completely restructured the panel classification scheme to make it more intuitive and user-friendly. We reviewed and expanded the list of surgical procedures as well as the word and term synonyms. We began user testing and used user suggestions to redesign certain aspects of the PHR interface. In addition, we revised and expanded the dictionary of help text, which is available throughout the PHR.

This project addresses the long-standing NLM interest in electronic medical records systems and delivery of health care information to consumers and is closely aligned with the NLM strategic plan. It uses the nationally mandated vocabulary standards that NLM has supported, and it provides another consumer entry point to NLM's rich trove of patient-oriented data. We have continued negotiations with Suburban Hospital/Johns Hopkins Medicine for Suburban Hospital to host the PHR. Both parties have signed the Business Associate Agreement, and we are working on finalizing the software license and end user license agreements, both of which have been reviewed by the NIH Office of General Counsel. As part of these agreements, NLM would analyze de-identified data for research purposes. Early research would focus on PHR usability and usage patterns to guide the next round of development and research. Longer term, our goal is to evaluate the effect of using the PHR on consumer health behaviors.

Using Drug Databases to Assess Prescribing Practices and Continuity of Care

Medication history is an important part of patient assessment in emergency care. Studies have shown that a significant proportion of Emergency Department (ED) visits are related to adverse events of drugs. However, gathering such information from the patient is time-consuming, expensive, and, when patients are unconscious, it is infeasible. Studies have shown that patient-provided medication histories are incomplete. Suburban Hospital's ED employed Surescripts, a consortium of major Pharmacy Benefit Managers (PBM), to provide an electronic summary of a patient's full year prescription filling history. LHNCBC created a messaging interface engine, based on the open-source HL7 interface program called *Mirth*, which linked with Surescripts and delivered the prescription records for patients who had checked in for ED care. Before the system went live, Suburban Hospital collected both Surescripts data and patient-provided history for quality assurance. Researchers obtained this information in a de-identified form, then compared the two sources of information. We found that Surescripts information, when available, significantly augmented the manual history, and covered a high proportion (88%) of a patient's current medications.

The concise prescription dispensing history report that we developed (based on the Surescripts data) is now routinely provided for patient care in the Suburban ED. Suburban Hospital ED clinicians have reported that the full year prescription history is also helpful in identifying potential problems of drug compliance, drug seeking behavior, and duplicative prescriptions.

We now have three years of Surescripts records for about 120,000 ED visits and are using this dataset to develop frequency statistics to pick the (most frequent) medications to display when a user types in a part of a medication's name (via an auto-complete algorithm) in other LHNCBC projects. We also have done preliminary analyses of the distribution of drug interaction messages and the frequency at which clinicians will be interrupted by a drug interaction messages under different choices of the severity of the interactions. A very few drugs are responsible for a very large share of such drug interaction alerts. Thesedata also suggest that patients are often taking interacting drugs prescribed by two different practices, such that neither provider would likely get the warning about the interaction because both drugs would not be in the Electronic Medical Record of either provider. To pursue these questions further, we have obtained a database representing one year's worth of prescription information (25 million prescriptions) for a major US metropolitan area.

EMR Database Research and Natural Language System Development

We have used the MIMIC-II de-identified intensive care database under a restricted-use Memorandum of Understanding to conduct retrospective clinical studies on the significance of obesity and metabolic syndrome, interactions between feeding practices and blood transfusions in premature babies with necrotizing enterocolitis, and significance of vitamin levels in ICU mortality and post-discharge survival. *Critical Care* published our study that confirmed the “obesity paradox” (lower mortality odds after an ICU stay for overweight and obese patients compared to normal weight patients).

To facilitate the use of this database by researchers, LHNCBC staff mapped medications from multiple tables to RxNorm ingredients and radiology report headers to LOINC. We continue mapping nursing observations from MIMIC-II local codes to LOINC, RxNorm, and/or SNOMED CT as appropriate. In line with the NLM mission to facilitate access to health information resources, LHNCBC continues as a mirror site to provide access to PhysioNet, a 4.3TB database collection of physiologic waveform tracings gathered from health care institutions world-wide by the MIT researchers who also developed MIMIC-II.

We developed natural language processing (NLP) techniques to extract important clinical variables from the narrative content of x-ray reports, discharge summaries, etc. in the MIMIC-II database. Even the best NLP techniques produce results with errors ranging from 5% to 10%., We have developed statistical methods that combine the large sets of NLP-extracted variables and the small sets of manually coded variables to improve the accuracy and tighten the confidence intervals for clinical studies that use NLP-derived variables. We have made the statistical methods, developed in R programming language, publicly available.

We continued evaluating our clinical NLP and information retrieval methods by participating in the 2012 Text REtrieval Conference (TREC), in which the task was to identify patient cohorts from a large set of de-identified clinical reports for comparative effectiveness studies. The LHNCBC team was the top-scoring team for the second year among 24 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

We continued our collaboration with the American Society for Colposcopy and Cervical Pathology (ASCCP) in the operational use of our Teaching Tool, an image-based system to assess professional knowledge and skills in the field of colposcopy. More than 100 Obstetrics/Gynecology and Family Practice resident programs nationwide (at more than 95 universities and other premier institutions such as the Mayo Clinic) are using our Teaching Tool. To date, resident programs have used the Teaching Tool to administer more than 1,300 individual online exams of two types: Resident Online Exam (ROE) and the Colposcopy Mentorship Program (CMP). Planning is underway to administer a third (and more advanced) Colposcopy Resident Award exam through the Teaching Tool.

The National Cancer Institute (NCI) and their collaborators used another of our imaging programs, the Boundary Marking Tool, at a new site in Senegal (as well as existing sites at the University of Oklahoma Health Sciences Center, Costa Rica, Nigeria, the Netherlands, and Spain) to collect and annotate colposcopy images for biopsy studies and the creation of a worldwide database for cervical cancer research.

We also continued to use our Virtual Microscope (VM) to annotate histology images. We used expert-annotated images (from the University of Oklahoma and from Queens University, Belfast) to test our in-house developed MATLAB modules to analyze the geometry of epithelial segments in the images. The analysis includes finding rectangular regions that span the segments, segmenting the content of these regions, deriving features related to density of nuclei per unit area, and using these features to classify the segments as Normal, or as various grades of abnormality (CIN1, CIN2 and CIN3.) Work is in progress to compare these automatic classifications with “ground truth” from pathologists.

A new approach with potential great utility is to rapidly locate segments of epithelium tissue within large histology images of the uterine cervix. Our goal is to locate epithelium more accurately and one or two orders of magnitude faster than current methods. The method (1) uses compression information stored in the file to roughly separate tissue regions from background, then (2) carries out Graphical Processing Unit (GPU) processing to classify the tissue regions into epithelium and non-epithelium.

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). We also collaborated with academic researchers to develop interactive segmentation capabilities for very large images using 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 (CBIR)

CBIR is an active research area in the imaging research community because systems for image indexing, search, and retrieval all use many tools and techniques developed under this rubric. CBIR finds images (in repositories or the published literature) that are visually similar to a query image, as well as those that are semantically similar, or relevant to a query. For example, one chest x-ray might be visually similar to another, but CBIR uses "semantic similarity" to find a chest x-ray from another person who has the same lung disease.

Several practical systems and tools at LHNCBC rely on CBIR research. Our Open-I system carries approximately 450,000 full-text publications that together include more than 1.2 million figures, including photographs, clinical images, charts, and other illustrations. The system automatically classifies the figures as "regular" images and graphical images, and then sub-classifies"graphical" images as diagrams, statistical figures, flow charts and tables. It sub-classifies "regular" images into: x-ray, ultrasound, CT, MR, etc. We automated these classification steps by extracting more than 15 image features (color, texture, shape), then using those features in a Support Vector Machine (SVM)–based framework. We demonstrated the success of our classification methods in the international ImageCLEFmed competitions in 2011 and 2012, and then incorporated those methods in Open-I.

We apply an additional step of extracting specific regions of interest (ROI) within images, based on the hypothesis that specific parts of an image may contain information that is more relevant to a query or concept, than would the entire image. Journal articles usually highlight these regions by "markers" or pointers – arrows, symbols – that our program must find. We have developed multiple approaches, e.g., edge or region-based segmentation, Markov Random Field (MRF) models to recognize arrow-type pointers, and a neural network-based technique to recognize asterisks. Our methods yield a precision of 85% and recall of 82%. More recent work in combining the detection of pointer boundaries and body (homogeneous pixel intensity area within the boundary) for pointer segmentation has improved the precision significantly over our previous edge-based segmentation – from an initial 25% precision to current 85% precision. Work is ongoing in improving performance for difficult cases in which pointers have exceptionally weak edges, or where the color of the pointer body is similar to that of the background.

The CBIR tasks of pointer identification and ROI extraction are combined with *textual* ROIs from figure captions or the body of the paper. Pairing visual ROIs with the corresponding textual mentions (e.g., "black straight arrow') enables the automatic indexing of the (visual) ROIs and the images of which they are a part. The tagged ROIs can then be used for image retrieval or building a visual ontology. We have developed a dynamic time warping (DTW) method to pair visual and textual ROIs. This pairing algorithm combines visual pointers with textual mentions by grouping recognized pointers by their visual characteristics first, and then searching for the best-matching pointer group with a text mention. Our experiments with ground truth text data shows that it successfully pairs over 96% of recognized true pointers with their textual mentions.

We also used CBIR to develop CervigramFinder, a research tool that automatically indexes and enables the retrieval of uterine cervix images (cervigrams) by shape, color, and texture features. Efficient searching by image features is a significant step toward locating records in large databases of cervigrams and patient data, such as the NCI's Guanacaste and ALTS databases, which contain a total of 100,000 cervigrams.

An LHNCBC project that is developing automated techniques to screen for tuberculosis and other pulmonary diseases is using CBIR to detect image features in chest x-rays. We have developed algorithms to automatically detect ribs, aorta, and other structures, and to segment lung areas. Our ongoing SVM method research aims to extract texture features to classify lungs as normal vs. abnormal.

We are also exploring use of distributed computing and GPUs for computer-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 tracing irregular object boundaries such as the separation between the epithelium and non-epithelial tissue 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

In this project, we investigate models for producing interactive publications that allow readers to actively manipulate images, tracings, and data in the publications – capabilities that we think will be demanded in the future. The project focuses on the standards, formats, and authoring and reading tools necessary for the creation and manipulation of such "interactive publications" that allow users to "interact with" video, audio, bitmapped images, interactive tables and graphs, and clinical DICOM images such as x-rays, CT, MRI, and ultrasound.

We have created tools for viewing and analyzing interactive publications (Panorama) and for authoring them (Forge). These tools are analogous to Adobe's Acrobat Reader and Professional for PDF viewing and editing documents. Panorama was one of 9 semi-finalists out of 70 software tools for life sciences research that were entered in Elsevier's Grand Challenge contest two years ago. After we conducted a formal usability study, we recently enhanceed Panorama by adding bar charts, and the capability to run natively on Mac OSX.

We also extended Panorama to provide Annotation Concepts. A Panorama user may click on text in an interactive publication which is sent to an NLM servlet (RIDeM, developed in-house) 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. Annotation Concepts 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 FY2012 we developed a web-based version (Panorama Lite), using Adobe Flex, that eliminated the burden of having to download the Panorama software before using it. The only requirement to run this new version 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.

Over the past year we have collaborated with two organizations to create interactive publications from their traditional static ones. The first was a publisher (ProQuest) and the second was a government agency (CDC's National Center for Health Statistics (CDC/NCHS)). We created two interactive papers for ProQuest from one of their open-source journals (*Sustainability: Science, Practice and Policy*). ProQuest announced the launch of these papers for public use in a press release.

For CDC/NCHS, we converted one of their key documents – *Health US In Brief* – to interactive form, and hosted it on our Web site. *In Brief* contains summary information on the health of the American people, including mortality and life expectancy, morbidity and risk factors such as cigarette smoking and overweight and obesity, access to and utilization of health care, health insurance coverage, supply of health care resources, and health expenditures. We presented this project to the NLM Board of Regents in September 2012, and received a favorable review. Planning is underway to select other CDC/NCHS documents for conversion to interactive form.

Screening of Chest X-rays for Tuberculosis in Rural Africa

In FY2012, LHNCBC continued its collaborative project with AMPATH (Academic Model Providing Access to Healthcare), an organization supported by USAID that runs the largest AIDS treatment program in the world. This project uses LHNCBC's imaging research and system development to fulfill NIH global health policy objectives. Our objective is to leverage in-house expertise in image processing to screen HIV-positive patients in rural Kenya for evidence of pulmonary tuberculosis (TB) in chest x-rays. 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 in the MOI University

Hospital in Eldoret, Kenya and are being readied for deployment. The team completed design of vehicles to transport the x-ray units – one vehicle is being outfitted as a mobile x-ray truck.

Since the lack of sufficient radiological services in the area suggests the utility of 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 detect texture features characteristic of pulmonary disease,. At present, the algorithms distinguish between 2 cases: abnormal vs. normal. After receiving IRB exemption, we obtained chest x-ray images to use as test and training sets: 400 from Montgomery County's TB Control Program, 850 from a source in India, and 8,200 from Indiana University. We also acquired an open-source Japanese set containing about 250 x-ray images.

Using these x-rays for training and testing, we have developed algorithms for detecting lungs and ribs focusing on region-based features such as log Gabor wavelets that exploit the orientation of anatomical structures. For lung segmentation, we have developed algorithms using region-based level sets and a novel graph-cut segmentation method, yielding an accuracy of about 95%. 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 FY2012, we completed the development of a voice-to-text conversion and recognition tool which is platform agnostic. Independent reviewers are providing feedback on an alpha version of the Web-based "Dialogues in Science" series. We are adding new interviews to the series and updating some of the current interviews.

Computational Photography Project for Pill Identification (C3PI)

Launched in September 2010, *Computational Photography Project for Pill Identification* (C3PI) intends to create an authoritative, comprehensive, public digital image inventory of the nation's commercial prescription solid dose medications. Ultimately our goal is to use these images to enable content-based information retrieval (CBIR) to promote patient drug-safety at the national level. Support for the project in FY2012 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. We are creating a collection of high resolution digital photographs of front and back surfaces of prescription tablets and capsules, 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 FY2012, we generated over 50,000 pictures of 1,500 samples of solid-dose pharmaceuticals from over 150 manufacturers and distributors. In addition, we are acquiring pictures from multiple cameras in a variety of lighting conditions to prepare a data collection for a broad effort or national computer vision/content-based information retrieval challenge (CV/CBIR) for the identification of medications. Staff also coordinated the deployment of a server-based repository and content management system to support distribution and curatorship of the image collection. In July 2012, NLM hosted a workshop of the Structured Product Labeling (SPL) Working Group – a collection of distributors, sponsors, and downstream users of SPL data. NLM staff presented a complete description of our image collection processes, the intended uses of the data, and the information outlets: DailyMed from NLM LO and Pillbox from NLM Specialized Information Services (SIS).

In September 2012, LHNCBC hosted a workshop on the design of a CV/CBIR national challenge. Attendees included members of NIST, the director of the Face Recognition Grand Challenge (FRGC), and a representative of the FRGC academic winner (University of Houston).

The Visible Human Project

The Visible Human Project image datasets 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 datasets 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 3,450 licensees in 64 countries are applying the datasets 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 FY2012, 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 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 FY2012, staff continued collaboration with the NCI'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. 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 FY2012, OHPCC expanded our high resolution electron microscopy research to include processing 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 emerging methods into mature production software for the study of protein structures on the surfaces of HIV and influenza virions. Additionally, LHNCBC staff helped supervise segmentation efforts for data from ion-abrasion electron microscopy to study stem-cell differentiation in murine myocytes. The unpublished results have raised more questions about the mechanisms for adult stem cell development and have suggested new studies in this critical research area.

The 3DI group also 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. We acquired and tested advanced software for managing 3D printing. A new set of models is under development including dosimetry models from CT scans of small animals. This work is conducted in partnership with the NIH National Institute of Allergy and Infectious Diseases (NIAID).

Insight Tool Kit

The Insight Toolkit (ITK) is a public, open-source algorithm library for segmenting and registering high-dimensional biomedical image data. The current official software release – ITKv4.2.1 -- contains over 845,000 lines of open source code, making available a variety of image processing algorithms. ITK can be run on Windows, Macintosh, and Linux platforms, reaching across a broad scientific community that spans over 40 countries and more than 1,500 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 completed the release of ITK version 4 (ITKv4), ending the effort that has been supported by the American Reinvestment and Recovery Act. This program helped fund 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.

Image and Text Indexing for Clinical Decision Support and Education

A picture is “worth a thousand words,” and especially valuable for medical research and clinical practice. Scientists and the lay public can better understand complex biomedical concepts through visual means: e.g., radiographic images, photographs of organs, sketches, graphs or charts. This idea motivated us to develop the OpenI system, which lets users search and retrieve medical citations enriched by relevant illustrations. Users may search by text queries as well as by example images.

In September 2010, the LHNCBC Board of Scientific Counselors reviewed our prototype multimedia search engine, which provided text-based search and image-feature based retrieval of other pre-indexed images visually similar to an image in search results. In FY2012, we presented a more advanced pre-release version to the NLM Board of Regents. Both Boards recommended scaling the system to large collections and continuing research and development of high-accuracy image retrieval methods.

Since the reviews, we have: redesigned the system architecture and refactored the original code; improved the user interface; added functionality (such as filtering images based on their type, filtering journals by clinical specialty, and ranking papers by clinical content); acquired and processed a larger set of scientific publications; and incorporated distributed computing operations (Hadoop™ MapReduce) and enterprise level “bare-metal” virtualization.

The evolution of OpenI has required the development of (1) new ways to represent images as text strings, that could be indexed with traditional search engines such as Lucene (2) improved methods to automatically segment multi-paneled illustrations into single images, and to partition their captions to correspond to single images (3) improved methods to extract pointers (arrows, arrowheads, symbols) within images to identify regions of interest. In other work, we have begun to build a visual ontology and we have also developed methods for segmenting lung and brain tissues and extracting key features such as lung cysts, micronodules, and emphysema.from those images.

OpenI is the first production-quality system of its kind in the biomedical domain, and it gives medical professionals and the public access to visual information from biomedical articles that are highly relevant to their questions, as well as a brief synopsis of the articles. OpenI consistently ranked among the best in international competitions at ImageCLEF, OpenI demonstrated the best retrieval results and placed third in image type classification in a 2012 medical image retrieval evaluation that attracted competitors from academia, industry and clinical settings.

Toward the end of FY2012 we released OpenI to the public. It now provides access to over 1,200,000 figures extracted from more than 450,000 articles in the Open Access subset of PubMed Central. Usage grows by the week and now exceeds 11,000 unique visitors a day.

Turning The Pages (TTP)

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.

The kiosk now presents all eleven books; the Web version offers ten, and the iPad offers six. We are working to make all the books available in all three formats.

This year we added two more books to the iPad version, and we released Andrew Snape's *Anatomy of An Horse (sic)* in all three versions. Snape’s comprehensive and beautifully illustrated book was originally published in London in 1683. We are currently creating the iPad and Web versions of Elizabeth Blackwell's *A Curious Herbal*, previously available only in the kiosk version. We have recreated all the graphics for Blackwell’s book and updated the animation, and plan to launch the iPad and Web versions shortly. We have also added new features to the iPad version such as bookmarks and contextual zoom for curators’ notes.

Longer term, we might develop a reactive 3D implementation system for TTP using several middleware development tools (Unity, Coco's 3D, and Unreal Engine). An advantage to a real-time 3D version is that it will considerably speed up the production of each book and make it possible for other institutions to create their own interactive books with our software. It would also allow for new functionality such as rotating the book 360 degrees, as well as turning multiple pages at once (more intuitive version of our current menu).

To reduce project costs, we have modified our production pipeline to capture the images of new books in-house rather than using commercial scanning companies. We also are investigating newer technologies, such as multi-touch monitors, for the next generation of the TTP kiosks.

Natural Language Processing and Text Mining

Medical Article Records System

NLM’s flagship database, MEDLINE, contains more than 20 million bibliographic records for articles from over 5,500 biomedical journals. To meet the challenge of producing these citations in an affordable way, researchers at LHNCBC develop automated techniques to extract bibliographic data (abstract, author names, affiliations, etc.) from both scanned paper journals and online journals.

While the bulk of citations come to NLM directly from publishers (in XML format), there are still approximately 820 journal titles that arrive in paper form. These papers are processed by the MARS production system, in operation for some years. MARS combines document scanning, optical character recognition (OCR), and rule-based and machine learning algorithms to extract citation data from paper copies of medical journals needed 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.), and pattern matching to identify these entities.

LHNCBC manages and continually improves the MARS system. For example, we are introducing three new features to improve MARS performance: (1) expansion of the MEDLINE character set, (2) capability for the Edit operators to correct errors made by the automated zoning process, and (3) a new user interface design for large screen monitors. The software implementation and integration test for the first two features have been completed, and are being deployed to the production system.

Citations that do come to NLM in electronic form from publishers often contain errors or have missing content. Missing items include: databank accession numbers (e.g., GenBank), NIH grant numbers, grant support categories, Investigator Names, and Commented-on Article information. The capture of Investigator Names can be especially difficult because some articles contain hundreds of such names and capturing the articles commented on by a paper requires operators to open and read other articles related to the one being processed. To automatically extract these fields from online articles, we have developed the Publisher Data Review System (PDRS) whose subsystems are based largely on machine learning algorithms such as Support Vector Machine.

PDRS was put in production in early FY2012, for open-access articles in NLM’s PubMed Central. To extend automated data extraction to *all* online journals on publisher’s sites, including ones with restrictive copyrights, we are developing IMPPOA (*In-Memory Processing for Publisher Online Articles*), a system based on the PDRS platform and its machine learning algorithms, but implemented to process articles in RAM memory rather than downloading the articles to disc, we expect that this approach will eliminate publisher’s concerns about copying articles into an external system disc. IMPOAA: (1) provides data missing from the XML citations sent in directly by publishers, (2) corrects errors in publisher data by extracting data from the articles on their sites and comparing these with the data sent to NLM, and (3) extracts data from articles for which publishers do not send in citations at all.

The systems outlined above rely on underlying research in image processing and lexical analysis which also enables the creation of new initiatives in which these techniques find application, such as the ACORN project.

Automatically Creating OldMedline Records for NLM

There has been a long-standing interest in expanding MEDLINE to include bibliographic records dating back to the late 19th century when Index Medicus was first developed, so we would have a complete electronic record of all citations in Index Medicus. These early citations appear in printed indexes published before 1960, and Library Operations (LO) has collected many of these with considerable manual effort. To automate this process, we have designed the ACORN system – combining scanning, image enhancement, OCR, image analysis, pattern matching, and related techniques to extract unique records from the printed indexes. These tasks are formidable considering the old typefaces and fonts (as well as a mix of different languages) in the indexes – producing a substantial degree of OCR errors. To overcome this problem in one of the indexes (Quarterly Cumulative Index Medicus, or QCIM), we developed a novel pattern matching technique that automatically finds and compares two versions of every citation from the subject and author listings, thereby minimizing the OCR errors encountered in each version. In addition, our system is designed to search MEDLINE as well as sources on the Web to avoid duplicating records that already exist elsewhere, and to further reduce OCR errors. Our system, designed to process indexes appearing both in paper form as well as on microfilm, has three main components: Quality Control, Processing, and Reconcile (for operator verification), which are currently being developed for demonstration in FY2013.

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 uses the MeSH headings from the PubMed related articles which are precomputed by PubMed. 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 and increasing use by NLM indexers to index MEDLINE citations. 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. Our developers also created versions of MTI to assist in indexing the NLM History of Medicine book collection and for general cataloging. Due to its success with certain journals, MTI was designated as the first line indexer for 23 journals totaling 4,205 articles in FY2012. 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 (II) team and Library Operations (LO) are currently working together on developing ways to evaluate the performance of journals in the MTIFL program and to help identify future MTIFL journal candidates.

The II team worked closely with an NLM Associate Fellow whose ongoing project was designed to investigate the feasibility of automating the creation of functional annotations about genes, known as Gene Reference into Function (geneRIF). We have developed a prototype, the Gene Indexing Assistant (GIA), and integrated it into the Data Creation Management System used by the Indexers for testing and evaluation.

The II team is also collaborating with LO to develop a new system designed to aid in indexing Technical Bulletins which are published by LO. The goal is to provide keywords for all past and future Technical Bulletins (1997 forward). The keywords will be automatically assigned by the program and manually reviewed prior to publication.

MetaMap is a critical component of the MTI system and used worldwide in bioinformatics research. Recent work has improved processing speed significantly, added XML (eXtensible Markup Language) output, implemented negation identification, and enabled users to supply their own acronyms/abbreviations list. MetaMap is available on Windows, Macintosh and Linux platforms. Users can 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. In FY2012, users downloaded approximately 1,700 copies of MetaMap, in the Java API and/or the UIMA Wrapper form. Of note, MetaMap is one of the NLM resources integrated into IBM's Watson system for healthcare applications.

Digital Preservation Research

The long-term preservation of documents in electronic form, both born-digital as well as those resulting from scanning paper, is a mandate for NLM as it is for other major libraries and archives. The goal of this LHNCBC project is to investigate and implement techniques for key preservation functions, including: automatically extracting metadata to enable future access to the documents, ingesting the documents and metadata into a storage system, and knowledge discovery from the archived material. To provide a platform for this research, we 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.

We are focusing on two collections. One is a historic medico-legal collection of early 20th century court documents acquired from the FDA. NLM curators are using SPER to preserve the FDA documents, numbering 67,000 in total, and in 2012 they processed over 22,000 of these. These documents and their metadata are in a publicly accessible NLM Web site. The other collection, from NIAID, is a set of conference proceedings of the "US-

Japan Cooperative Medical Science Program on Cholera" (CMSP), an international program conducted over a 50-year period from 1960 to 2011. Our CMSP document preservation activities include: (a) building a full repository for 2,800 research articles on cholera and 8,000 references on CMSP participants such as authors, panelists, attendees and Study Section reviewers, followed by (b) developing a portal for the public to search for research articles, institutes, authors and other participants from this dataset.

We have developed automated metadata extraction (AME) techniques to identify and extract three different types of metadata from the CMSP documents, namely: publication metadata with titles, authors and their affiliated institutions from research articles; investigator metadata with name, role, designation and affiliation of each person from the conference proceedings rosters; and Study Section metadata with names and affiliations of CMSP program reviewers from separate Study Section rosters. The AME processes include (a) layout analysis to recognize different types of information within a document set; (b) evaluation of the effectiveness of models such as Support Vector Machine and Hidden Markov Model for different metadata layouts; and (c) capture of relationships among various entities in the collection from the extracted metadata of different types. In 2012, we added two more conference proceedings, for the years 2010 and 2011, to the CMSP repository. Furthermore, we used the extracted metadata to implement data analysis function for the CMSP document corpus to discover patterns and trends on factors such as important drugs, discoveries, investigators as well as international collaborations under the CMSP program over its 50 year span.

In addition, we 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 the corresponding. We are testing an external tool named LymbaGrid to develop the cholera vocabulary and generate the knowledgebase from the CMSP document corpus.

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.

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 information 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 since July 2009 through the *Evidence-Based Practice* tab in CRIS During the past year, the *Evidence-Based Practice* tab was accessed 28 times a day, on average, by over 890 returning NIH CC users.

In FY2012, we started extending RIDeM services to answer consumer health questions submitted to the NLM customer services that handle about 90,000 requests a year. We are developing a prototype Consumer Health Question Answering (CHQA) system to facilitate answering the requests. The prototype can classify the incoming requests as, for example, questions about health problems or requests to correct MEDLINE citations. Once the request type is recognized, CHQA provides information needed to answer the request. For MEDLINE quality assurance requests, the system automatically finds and retrieves the citation that set off the request, and then extracts information that helps answer the question, for example, the citation publication status. The prototype is capable of understanding simple frequently asked questions (such as, "Is Parkinson's disease hereditary?") and finding sections of the NLM GHR articles that answer these questions.

De-identification Tools

De-identification enables research on clinical narrative reports. We are designing a de-identification system that will remove protected health identifiers from narrative clinical reports according to the provisions of 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.

We completed a version of the software system to be tested at the NIH Clinical Center. This version of the system is designed to de-identify clinical narrative text in the form of Health Level 7 (HL7) version 2. It is capable of utilizing information embedded in various HL7 fields as well as externally provided information such as name lists of the health care providers at NIH.

We designed the Visual Tagging Tool (VTT) - an editor for visualization and markup, which we use to produce gold standards against which to test the CTD system. Although it is designed specifically to facilitate and speed up manual tagging of identifiers that contain protected health information (PHI), we have made it publicly available to the greater NLP community for any kind of lexical tagging and text annotation. (http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/vtt/current/web/index.html)

We also are preparing a large corpus of clinical reports to serve the project as the gold standard based on VTT for de-identification. As of December 2012, the corpus comprises 20,994 clinical reports of 7,571 patients, where every piece of individually identifiable health information has been marked manually. We divided the set of annotated reports into a training set and a test set—the latter of which will be used to evaluate the overall de-identification performance of our system.

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 NIH Clinical Center Laboratory for Clinical Informatics Development 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 instead 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 knowledgebases 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 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). In 2012, we made LITE officially available to users of the OpenInfobutton system, with paper presentations and demonstrations at the American Medical Informatics Association.

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 is nearing completion of a project to enhance the derivational-variants function of the lexical tools. Derivational variants are words related by a word-formation process like suffixation, prefixation or conversion (change of category). Earlier versions of the derivational variant system had 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 2011 release of the Lexical Tools included around 4,500 derivational facts; the 2012 release had nearly 90,000 and the 2013 release will have over 121,000 derivational facts. These new derivational facts now include information on negative derivations and will be used to infer derivational rules.

During the year, our web page had an average of 1,999 unique visitors per month. We had an average of 4,800 downloads per month in 2012. We provided support to 35 internal users, 14 US domestic users and 8 international users.

The 2013 release of the SPECIALIST Lexicon will contain over 469,992 records, representing over 857,502 forms, an increase of over 7,863 records from the 2012 release. Many of the new terms are derived from de-identified clinical records from our own De-identification project and from the MIMIC II database (described in another section).

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 EMR interoperability.

During FY2012, staff investigated the validation of value sets used in clinical quality measures as part of the Meaningful Use incentive program, in collaboration with the Office of the National Coordinator for Health Information Technology (ONC) and the Centers for Medicare & Medicaid Services (CMS), contributing to the creation of the Value Set Authority Center at NLM. We continue to explore the use of Semantic Web technologies for representing the UMLS and semantic predications, with application to literature-based discovery. Finally, we pursued our work on quality assurance in biomedical terminologies by investigating LOINC using description logic tools.

Research activities this year resulted in two journal articles, four papers in conference proceedings, one book chapter 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). We also co-chaired the scientific program committee of the Eighth International Conference on Data Integration in the Life Sciences (DILS 2012).

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 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 that is available to the research community. 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. The Semantic MEDLINE technology was recently adapted for analyzing NIH grant applications, allowing NIH portfolio analysts to track emerging biomedical research trends and identify innovative research opportunities.

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 effort 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, to support formal models of knowledge representation, 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 registered interventional and observational clinical research studies. ClinicalTrials.gov receives over 95 million page views per month and hosts approximately 900,000 unique visitors per month.

Data are submitted to ClinicalTrials.gov through a Web-based Protocol Registration System (PRS) by nearly 12,000 study sponsors including the US Federal government, pharmaceutical and device companies, academic, and international organizations. At the end of FY2012, the site had records for nearly 134,000 research studies, conducted in all 50 states and in over 180 countries. Approximately one-third of the studies are still open to recruitment. For the remaining two-thirds, the recruitment phase is over or the study has been completed. Over 7,400 of the closed studies display summary results tables describing primary and secondary outcomes, adverse events, and characteristics of the participants studied. In FY2012, new registrations were submitted at an average rate of 370 records per week, an increase of 6% from FY2011. The average rate of new results submissions was approximately 70 per week, an increase of 17% from FY2011. The continued growth in the use of ClinicalTrials.gov can be attributed to US laws as well as international recognition of the scientific and ethical importance of registration and results reporting. 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.

NLM established ClinicalTrials.gov in FY2000 –to comply with requirements of 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 expanded significantly to support other international registration policies, such as the International Committee of Medical Journal Editors (ICMJE) policy requiring prospective trial registration as a condition of publication. In FY2012, ClinicalTrials.gov continued to expend significant effort on implementing and educating the public on the most recent law, Section 801 of the Food and Drug Administration Amendments Act of 2007 (FDAAA 801). NLM has been working with the NIH Office of the Director and other NIH Institutes and Centers as well as the Food and Drug Administration (FDA) on a Notice of Proposed Rulemaking (NPRM). When the NPRM is published, it will further elucidate the requirements of FDAAA 801 and request public comment on key implementation issues.

In FY2012, ClinicalTrials.gov staff focused on enhancing educational and outreach materials available to the public, as well as to sponsors and investigators affected by FDAAA 801. A key component of this strategy was the launch of a redesigned Web site with a new graphic design and reorganized written content, which was informed by user feedback, usability studies, and the need to better integrate educational materials about data submission and FDAAA 801. Another key component was targeted education and outreach on the results database and submission requirements. This included hosting two on-site workshops for personnel at academic research institutions and presenting at conferences, participating in webinars, and publishing in trade and peer-reviewed journals. Research projects in FY2012 investigated issues related to the ethical and scientific oversight of clinical research and the quality and timing of results reporting. Additionally, new tools were introduced in the PRS in FY2012 to support more efficient data entry and to expand resources available for uploading content available in other databases, such as NCI’s Clinical Trial Registration Program. ClinicalTrials.gov also continued to provide technical advice and collaborate with other clinical study registries, professional organizations, funders, and regulators in working towards the development of global standards for trial registration and reporting to results databases. For example, a key activity was working with the European Medicines Agency (EMA) on developing a common set of data elements for results submission to both ClinicalTrials.gov and the EMA results database, which is being developed for possible release in 2013.

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 more than 1,900 genetics topics, including more than 800 genetic conditions, about 1,040 genes, 77 gene families, 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. In 2012, a new section on the Encyclopedia of DNA Elements (ENCODE) Project was added to the handbook.

Genetics Home Reference celebrated its ninth anniversary in 2012. In the past year, the project expanded its genetics content for consumers, adding 260 new summaries to the Web site. We intend to continue this rate of production in FY2013, covering additional Mendelian genetic disorders as well as more complex disorders. The team also plans to add more background information to *Help Me Understand Genetics*, including pages on next-generation DNA sequencing and the diagnosis and management of genetic disorders. In FY2012, the site averaged almost 28,000 visitors per day and about 27.6 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.

Staff presented the Web site to several visiting groups, including health and science journalists as part of the Association of Health Care Journalists – NLM Fellowship program. In addition, staff attended several major genetics conferences. At the annual meeting of the American Society of Human Genetics (ASHG), we presented a poster titled, “Genetics Home Reference Ten Years In: Where We Are Now.”

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 60,000 unique visitors each month.

This year, the papers of pioneering surgeon Henry Swan were added to *Profiles in Science*. Staff added 402 transcripts of documents, to make handwritten items searchable and to provide alternatives to PDF format files. Staff also added 10 digital items to the 35 existing *Profiles in Science* collections. Currently, 142,180 image and 27,042 digital items are available on *Profiles in Science*. The Web site now features the archives of 33 prominent scientists and health advocates.

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

In addition to updating the *Profiles in Science* collections during FY2012, LHNCBC staff replaced the backbone of the architecture of the *Profiles in Science* Web site with the open-source Apache Solr enterprise search platform. We also continued to enhance documentation and migrate modules of the *Profiles in Science* digital library data creation software to eliminate dependence on unsupported software while improving reliability and security. We increased the visibility of the digitized items throughout the *Profiles in Science* Web site and search results through the use of preview images. We provided the research community insight into the digital library underlying the *Profiles in Science* Web sites through publication of "The *Profiles in Science* Digital Library: Behind the Scenes" in JCDL '12: Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries.

Evidence Based Medicine - PubMed for Handhelds

Developed and released in FY2003, PubMed for Handhelds facilitates 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.

PubMed for Handhelds is available as an iOS (iPhone/iPad) app and an Android app. As of early December, the iOS app has been downloaded more than 25,000 times. For some search tools (PICO and askMEDLINE) queries from smartphone apps now account for 90% of queries.

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.

During FY2012, we reviewed the codes contained in the National Quality Forum (NQF) eMeasures value sets, a crucial component of the Federal Meaningful Use regulations, to determine which codes were incorrect (did not exist as written or were obsolete), associated with incorrect terminology, or associated with an incorrect definition. We also began reviewing content of the NQF eMeasures for clinical validity and compatibility with clinical workflows.

The CORE Problem List Subset of SNOMED CT

SNOMED CT is a comprehensive, multi-lingual medical terminology for anatomic sites, organisms, chemicals, diagnoses, symptoms, findings, and other such concepts. The problem list is an essential part of the Electronic Health Record (EHR). The adoption of a common standard for the codes in this list prevents duplication of effort and promotes data interoperability. "Meaningful Use" regulations from the Centers for Medicare & Medicaid Services (CMS), require the use of SNOMED CT to code the problem list.

Based on the analysis of the problem list vocabularies and their usage frequencies in seven large-scale US and overseas healthcare institutions, we identified a subset of the most frequently used problem list terms in SNOMED CT and we published it as the CORE (Clinical Observations Recording and Encoding) Problem List Subset of SNOMED CT. The CORE Subset can be a starter set for institutions that do not yet have a problem list vocabulary and this will save significant development effort and reduce variations between institutions. Existing problem list vocabularies can also be mapped to the CORE Subset to facilitate data interoperability.

Since its first publication in 2009, the CORE Subset has received considerable attention from the IHTSDO (International Health Terminology Standards Development Organization), the SNOMED CT user community, EHR software vendors and terminology researchers. It has been installed in various EHR products, and used as a focus for SNOMED CT-related research, mapping projects and quality assurance. The MedlinePlus Connect Project, which facilitates online linkage to patient education information, has mapped the CORE Subset to MedlinePlus health topics. In 2012, a clinical dataset from the Veterans Administration covering over 3 million patients was utilized to enrich the CORE Subset. It required the addition of a moderate number of new concepts (about 300) to cover the most frequently used terms in the new dataset, further lending support to the proposition that a relatively small number of SNOMED CT concepts is sufficient to cover a high percentage of usage in most institutions. We also published a map from 9,000 commonly used ICD-9-CM codes to SNOMED CT to facilitate conversion of legacy clinical data from ICD-9-CM to SNOMED CT codes. The CORE Subset is updated 4 times a year to synchronize with changes in SNOMED CT and the UMLS. The CORE Subset currently contains about 6,000 concepts.

Mapping between SNOMED CT and ICD codes

International Classification of Diseases (ICD) codes are required for public health reporting of population morbidity and mortality statistics. In the US, ICD-9-*CM* (the "Clinical Modification") is also used for reimbursement (soon ICD-10-*CM* will be required for this purpose). Because of this need, many existing EHR systems are still using ICD-based vocabularies to encode clinical data. However, ICD was not designed to capture information that is detailed enough to support clinical care. SNOMED CT is a much better clinical terminology and its use will be required as part of the "Meaningful Use" regulations. To encourage the migration to SNOMED CT, and to enable EHRs to output ICD codes for administrative purposes, various maps between SNOMED CT and the ICD classifications have been developed. We published a SNOMED CT to ICD-10-*CM* rule-based map, covering 15,000 SNOMED CT concepts. This map allows users to encode patient problems in SNOMED CT terms, and then generate the appropriate ICD-10-*CM* codes in real-time for billing or other purposes. To demonstrate the use of the map, we developed the I-MAGIC (Interactive Map-Assisted Generation of ICD Codes) demo tool.

For an international project, in collaboration with the IHTSDO and the World Health Organization (WHO), we developed an analogous rule-based map between SNOMED CT and ICD-10 covering 19,000 SNOMED CT concepts. We adapted the I-MAGIC tool to showcase this map to ICD-10 as well. In a separate project, to help convert legacy ICD-9-*CM* encoded clinical data into SNOMED CT codes, we produced another map from ICD-9-*CM* to SNOMED CT.

We are currently planning to create maps between SNOMED CT and ICD-9-*CM*/ICD-10-*PCS* procedure codes, since SNOMED CT is also designated as the terminology standard for coding clinical procedures in Phase 2 of the "Meaningful Use" incentive program for electronic health records.

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 (CMS) used RxTerms in one of their pilot projects in the post-acute care environment. RxTerms is also used in the NLM PHR, and at least one EHR from a major medical institution in Boston. There is ongoing effort to align data elements between RxTerms and RxNorm, and 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.

RxTerms content and features are now being bundled with the RxNorm standard (the US standard drug reference terminology that is required by meaningful use regulations). During FY2012, we aligned the data model of

RxTerms and RxNorm by creating a new term type in RxNorm to cover the drug-route combination. We also worked with RxNorm on the federal vocabulary standard for identifying prescribed medications and ingredients, especially as needed for identifying drug allergies.

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, we created SOAP-based and RESTful application programming interfaces (APIs), to let users integrate RxNorm functions into their applications. Examples of such functions include mapping drug names to RxNorm, finding the ingredient(s) corresponding to a brand name, and obtaining the list of National Drug Codes (NDCs) for a given drug.

During FY2012, staff aligned RxNorm with the Anatomical Therapeutic Chemical (ATC) classification system developed by the World Health Organization (WHO) Collaborating Centre for Drug Statistics Methodology and widely used in Europe. Many RxNorm drugs are now linked to drug classes in ATC. Similarly, we also linked RxNorm drugs to pharmacologic actions from the Medical Subject Headings (MeSH), and enabled the remapping of obsolete drug codes to the current version of the drug codes.

A major effort this year was the development of RxMix, a graphical interface allowing users to create workflows – complex sequences of API functions – and to execute them on single values or on list of values, in batch. A typical use is the determination of the list of clinical drugs in RxNorm that have a given property in NDF-RT, for example the list of statin drugs, of penicillins. Ability to establish such list is critical for applications, such as allergy checking and clinical decision support.

We have integrated two other drug information sources 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 SOAP and RESTful application programming interfaces (APIs) for RxNorm, RxTerms and NDF-RT, received a combined total of about 50 million queries during FY2012 (a significant increase from 40 million queries last year). Users include clinical and academic institutions, as well as pharmacy management companies, health insurance companies, EHR vendors, and drug information providers. Developers of mobile apps have also started to integrate our APIs into their applications.

LOINC Standards for Identifying Clinical Observations and Orders

Federal Meaningful Use (MU) 2 electronic health records regulations require LOINC codes in lab result messages sent to ordering clinicians (within medical record systems, patient summaries and reports to public health). In FY2012, LHNCBC continued to work with the Regenstrief Institute, major laboratory companies, several NIH institutes, and other organizations to expand the size and breadth of the LOINC database. By the end of FY2012, LOINC had nearly 20,000 users in 148 countries and was translated into eleven languages and dialects. Users can pick any of these languages, search for words in the chosen language, and see the matching LOINC terms in that language plus English. To further expand LOINC globalization, we enabled language-specific web pages in the LOINC web browser (prompts and button labels in native language) – so far, the web browser is completely understandable in two languages besides English (simplified Chinese and Italian).

We worked with Regenstrief and the LOINC Committee to create more than 5,500 new LOINC terms for both laboratory and clinical variables, and the LOINC database now contains nearly 70,000 terms. We released new terms for radiology (RAD), Neuro-QOL, OPTIMAL survey (American Physical Therapy Association), NEMSIS, the CARE long term care hospital (LTCH) survey and core terms for behavior for SAMHSA (Substance Abuse and Mental Health Services Administration). During FY2012, we also edited existing molecular genetics terms to harmonize with Human Genome Organization (HUGO), Human Genome Variation Society (HGVS), and International System for Human Cytogenetic Nomenclature (ISCN) recommended nomenclature.

We worked with four of the eight largest international laboratory instrument vendors to help map or check the mapping of their internal instrument codes to LOINC codes, in order to facilitate electronic reporting of lab results. All eight such vendors now assert they provide LOINC codes for all the test codes their instruments can generate. We also worked with many smaller vendors to find (or create new) LOINC codes to describe the results of their test kits or instruments to fit their needs.

We continued to meet with other NIH organizations that are developing assessment instruments with the goal of closer alignment among NIH standard element development. We are collaborating with other NIH organizations (and Regenstrief Institute) to structure their assessment instruments and registry system values into the LOINC format and incorporate them into the LOINC database - a common framework that includes many kinds of clinical and research variables. We are serving on the Common Data Elements (CDE) Working Group to the trans-NIH BioMedical Informatics Coordinating (BMIC) Committee. We are working with the National Eye Institute (NEI) to restructure its packages of assessment instruments for the National Ophthalmic Disease Genotyping Network (eyeGENE®), and with the NIH/NCATS Office of Rare Disease to revise their CDEs—and we plan to create corresponding LOINC codes. Staff is also working on the NHLBI HUMLO project, with colleagues at NHLBI, NICHD, and NIH/NCATS, to develop CDEs for hemoglobinopathies using standard terminologies such as LOINC and SNOMED CT. We are also working with NINDS on the NINDS CDEs and the Neuro-QOL measures.

Staff enhanced LOINC's support tools and databases, increased database field size to enable support for new LOINC numbers, and deleted 12 obsolete fields.

Newborn Screening Coding and Terminology Guide

LHNCBC has collaborated with the multiple federal, state, and other agencies to standardize all of the variables used in newborn screening (NBS) using national coding standards as required by Meaningful Use Stage 2. Our collaborators include the Health Resources and Services Administration (HRSA), the Centers for Disease Control and Prevention (CDC), the Association of Public Health Laboratories (APHL), the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), and the National Heart, Lung, and Blood Institute (NHLBI). We have created a comprehensive panel of LOINC terms for NBS and continue to create new LOINC terms as new conditions and tests come into play. We also periodically review and update existing codes based on user feedback.

We have mapped all of the NBS conditions to SNOMED CT, and have requested new codes from SNOMED CT as needed. All of these terms are intended for use in an HL7 NBS lab result message. During FY2012, we updated and released a new version of our guidance for the HL7 message implementation and our example HL7 NBS result message. Many state NBS laboratories across the country are adopting our codes and HL7 guidance, including Kentucky, Oregon (which is the regional laboratory for 5 additional states), Colorado, New York, Illinois, and Washington. We actively worked with many of these states during FY2012 and have continued to do so.

Critical congenital heart disease (CCHD) is the latest condition that was added to the NBS Recommended Uniform Screening Panel, and several states are beginning to implement CCHD screening. So we are developing standard codes for reporting CCHD screening results. As part of this CCHD effort, we are working with multiple groups, including Children's National Medical Center, HRSA and other federal agencies, state NBS programs and public health departments, hospitals, and pulse oximetry technology vendors. We represent NLM on the technical advisory team for the HRSA CCHD pilot program grantees, to help them with terminology and interoperability standards, and are on the steering committee for NewSTEPs, APHL's new Newborn Screening Technical assistance and Evaluation Program. 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, and we are working on standardizing NBS genetic testing result reporting.

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 (VIC) 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. Because Argonne National Laboratory ceased AG development and support of its open-source community, our team investigated newer, cloud collaboration tools, including proprietary ones that are standards compliant and that emulate the AccessGrid's pioneering model. Our aim is to identify technologies that might have sufficiently broad appeal to replace the AG platform. So far, the business models and licensing terms have proven more of a problem than the technologies.

The team published a comparison of the major compression/decompression (codecs) available in the OHPCC Collaboratory for High Performance Computing and Communications (Collab) last year, submitted a systematic review of uncompressed video technologies for publication, and drafted a report about the cloud technologies. In addition, the team has initiated review of video applications for mobile devices. The overall plan is to study and test these applications, then depending on the results, launch and test them in real settings.

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 to integrate audio and other enhancements. This work led to several demonstrations of trans-Atlantic uncompressed videoconferencing at data rates of 1.5 gbps between NLM and Masaryk. Substantial work was done with the lead developer of ConferenceXP (CXP) at the University of Washington. The team was able to get the uncompressed as well as the compressed versions of the program to work this year and extensive testing was done with the National Center for Supercomputer Applications (NCSA). Although the technology works, there are latencies due to a variety of possible causes and the team has focused on UltraGrid and iHDTV, with a special interest in UltraGrid. 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 has implemented a 3D version of UltraGrid in the lab, but not over networks. The CXP tests with NCSA have led to some initial testing of their compressed/uncompressed Viper HD technology. The team continues to collaborate with the Rochester Institute of Technology (RIT) and the University of Puerto Rico Medical Campus to test open-source software for compressed HD videoconferencing based on the H.264 video standard and cloud technologies.

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 network capacity made the uncompressed trans-Atlantic videoconferencing demonstrations possible without using dynamic circuit network technology to reserve bandwidth. The team believes DCN technology should be further explored and attempts were made to establish an alliance with Northrup Grumman to test their software. The company has decided to keep development in-house and make it proprietary, however. 3D HD cameras have been acquired and components for 4K video have been purchased to further test the network infrastructure and continue work with Masaryk.

The team installed iHDTV systems at the Medical University of South Carolina (MUSC) to study uncompressed video's use as a diagnostic tool. We selected teledermatology 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 are measuring diagnoses, clinician confidence, decisions to biopsy and physician and patient encounter satisfaction and are comparing telemedicine applications 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. We chose to use iHDTV systems for the study because, at the time, audio was not integrated into UltraGrid and ConferenceXP could only transmit compressed video. The study was delayed substantially by IRB complications, OMB clearance requirements, the death of the PI collaborator, and the subsequent incapacitating illness of her successor. We found a new MUSC PI and data collection has now begun. These problems necessitated an additional site visit to provide refresher training and a run through of the research protocols and methodology. Previous work at MUSC experimenting with video medical interpretation has been published.

Staff continued to work with SIS on a 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. In FY2012, staff delivered posters and presentations on the "virtual computer lab" methodology at the meetings of the American Telemedicine Association, the Medical Library Association, and the Research Centers at Minority Institutions annual symposium.

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, open-source technologies such as Conference-XP, iHDTV, UltraGrid, and the Access Grid as well as cloud technologies and mobile apps.

Last 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 and this year we acquired hardware to build systems capable of transmitting 4K video. We updated the Collab's streaming and web servers and worked on replacing the Google appliance for searching the web server with open-source software. Changes in streaming technology required extensive file conversions of previously recorded video programs. Finally, staff published a retrospective review of the research generated as a result of NLM advance computing initiatives.

Disaster Information Management: Lost Person Finder

NLM's increasing interest in recent years in mitigating the effects of wide area disasters has resulted in information resources and tools from many parts of the library. In our Lost Person Finder (LPF) project, we address the problem of family reunification in the wake of a mass casualty event. LPF systems combine image capture, database and Web technologies, and address both hospital-based and community-wide disaster scenarios.

Web Site and Services

The heart of our system, now operational, is *People Locator* (PL), the main LPF Web site and its MySQL database. Developers extensively customized the open-source Sahana disaster management system to create, for example, a unified site to hold data from multiple disasters, thereby eliminating the need to build multiple Web sites and database instances for different disasters. PL can be searched by hospital counselors, relief workers, or the public. Searching or reporting can be done via computer or through mobile apps using Web services.

During 2012, we made PL more robust. It is now running on enterprise-level, load-balanced dual systems with failover, independent uptime monitoring, and an integrated SOLR indexing engine for fast record retrieval. We also developed a lighter-weight, more-portable version, *PLmobile* for special local use.

Mobile Apps

For hospital-based reporting, the triage process begins with *TriagePic*, a Windows application that hospital staff can use to quickly photograph arriving victims. These pictures, along with general health/triage status and minimal descriptive metadata (e.g., name, age range, gender) are packaged and sent by Web services to PL. In 2012, we moved TriagePic from a laptop-hosted system to tablets, to exploit their touch screens and front- and rear-facing cameras. Developers made the existing Windows 7 version more touch friendly and feature-rich, and deployed most recently on a Samsung 7 Slate. To provide a wide range of attractive platforms for hospital use, we are also developing Android and iPad versions.

In 2012, we further improved ReUnite, a smartphone app originally introduced during the Haiti earthquake for the public and aid workers, to better report and search for those missing during community-wide disasters. Our iOS5/6 version for the iPhone or iPod Touch is available from iTunes as a free download.

Deployments

From the beginning of the project, we have taken part in demonstrations and large-scale multi-institutional drills, as well as stood up events for international disasters, including the Christ Church Earthquake (February 2011), the Japanese Earthquake and Tsunami (March 2011), and the Joplin Tornado (May 2011), among others.

In FY2012 we also: (1) repurposed ReUnite to track patients as they were transferred from one hospital to another and this function was used by a hospital in Indianapolis (March); (2) supported Sahana Software Foundation for a demonstration to the Australian Red Cross (April); (3) served as a standby software expert for Hurricane Sandy and as a backstop-consultants for New York City shelter management software developers (there are commonalities between their software and ours) (October/November); and (4) demonstrated our software at the HALO 2012 counterterrorism conference in San Diego (October).

Face matching research

Our goal here is to enable users to find missing person records through automatic face recognition, a significant extension of our current method of searching by name or other text metadata. Our application faces special challenges: unlike many other systems, our face matching needs to rely on a *single* photo of a person to identify her/his face in other images, and cannot therefore exploit traditional face recognition models that require large training sets. Researchers accomplished substantial work mainly in face localization, a key first step in face matching. Originally operating only on grayscale images, we extended the face localization subsystem to exploit human skin tones that serve as an effective clue for faces in images. With up to 95% accuracy achieved for face localization and detection, our system outscores most existing solutions, including commercial systems like FaceSDK by Luxand, Inc.

In addition, as a way to reduce the search space for face matching, researchers developed a near-duplicate image detector to identify images submitted for the same person. This detector successfully reduced the PL image collection by 33% by discarding such redundant information, and thereby speeding up face matching.

To support research and testing, we need annotated images. For this purpose, we have developed ground-truthing and annotation-judging tools. For instance, our ImageStats tool, originally used in-house for face annotations, was released to handle crowd-sourced annotations of public-domain face images by volunteer students (globally recruited) through the *Google Code-in* contest, through our collaboration with the Sahana Software Foundation.

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.

NLM Media Assets Project

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

- The NLM/APDB Tape Library Archive Project,
- 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, and

- 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 including video teleconferencing and NLM-wide webcasting, and
- 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, and development of interactive tools and applications which utilize video for clinical monitoring and diagnostics),
- The Profiles in Science video modules, and
- 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:

- augmented reality modeling and applications,
- mobile device application development,
- 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, and
- the LHNCBC Research Update Modules.

Interactive Mobile Applications

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.

Planning and development of interactive multimedia for the FY2012 NLM Exhibition "Native Voices: Native People's Concepts of Health and Illness" continued. APDB staff worked with the NLM Director, and staff from the Office of Communications and Public Liaison, Office of Health Information Programs Development, and History of Medicine Division to produce interactive video and videographic content for the exhibition web site and the traveling exhibition program. Based on this outreach plan, APDB produced and developed an interactive iPad app featuring all video materials currently in the physical exhibition. Additional video interviews and segments were incorporated, as well as highly interactive search and retrieval functionality within the app. All video content was encoded in formats for distribution across multiple platforms including the iPad and mobile QR code applications featured throughout the onsite, web, and traveling 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

APDB is now applying digital workflow management and file format standards (originally established for exhibition production support) to convert LHNCBC's large library of historical tape – which contains over four decades of NLM programs – into a viable digital repository accessible for future use. The Branch began testing use of the motion JPEG2000 for long-term archiving. Testing is in progress and results are being monitored. APDB expanded by more than 50 videos the extensive digital video library assembled for the NLM Director's exhibition interview database, which is part of APDB's ongoing effort in digitizing, organizing, and accessibly storing large-scale video libraries.

Biomolecular Visualization

APDB staff continued to collaborate with 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. We applied 3D biological imaging technologies to data using advanced image segmentation methods and computational analysis to obtain an integrated molecular understanding of cellular architecture.

To facilitate a better understanding of the normal and pathological process of the subcellular environment, staff developed novel imaging techniques including linking correlative light microscopy and ion-abrasion scanning electron microscopy as a method of locating, identifying, and studying cellular or sub-cellular structures through the use of 3D reconstructions.

APDB produced 3D models, animations, and medical illustrations for various molecular content areas, including:

- HIV membrane fusion,
- HIV pathogenic distribution from T-cell to astrocyte,
- T-cell to T-cell HIV infection,
- mouse intestinal epithelium,
- dendritic cell/CD4+ T-cell interactions,
- role of GroEL in protein folding, and
- HIV-infected T-cell and neural stem cell junction.

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.

Computing Resources Projects

The Computing Resources (CR) Team has a variety of core projects that build, administer, support, and maintain an integrated and secure infrastructure to facilitate LHNCBC's research and development (R&D) activities. 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 implements the 10-gigabit network, and studies many advanced communication protocols to support LHNCBC collaboration activities and research projects. Another core project implements a network monitoring system that displays network usages in real time. The network management team also participates in the study of Trusted Internet Connection (TIC) consolidation and evaluates 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 a safe IT working environment. One core project studies and implements a unified patch management to improve LHNCBC's overall security measures. Another core project implements the automated security audit system that ensures all system at LHNCBC comply with policies. The security management team also studies and evaluates the network performance impact of web anti-virus software, and coordinates annual penetration testing to ensure network security. The facility management team deploys new IT equipment and servers, including power acquisition, network planning, cabling connection, and space allocation in the central computer room as well as at co-location facilities. Another core project studies, designs, and implements an enterprise console management system that enables LHNCBC to remotely manage large numbers of servers.

The system administration team provides LHNCBC-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 centralized network storage to support Continuity of Operation (COOP) requirements. 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 shared computing resources such as security audit, system buildup, and security certification.

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 FY2012, 49 participants from 16 states and 6 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, research in question answering systems, research on large biomedical data sets. The program emphasizes its focus on diversity through participation in programs for minority students, including 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.