Lister Hill National Center for
Biomedical Communications
An Intramural Research Division of the U.S. National Library of Medicine

FY2013 Annual Report

Clement J. McDonald, M.D.
Director

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

Contents

The Lister Hill National Center for Biomedical Communications (LHNCBC), established by a joint resolution of the U.S. Congress in 1968, is an intramural research and development division of the National Library of Medicine (NLM). LHNCBC seeks to improve access to high-quality biomedical information for people around the world. It leads programs aimed at creating and improving biomedical communications systems, methods, technologies, and networks and enhancing the sharing and use of information among health professionals, patients, and the general public. The development of next generation electronic health records (EHRs) to facilitate patient-centric care, clinical research, and public health is an important focus of the LHNCBC as well as an area of emphasis in the NLM Long Range Plan 2006−2016.

The LHNCBC research staff is drawn from many disciplines including medicine, computer science, library and information science, linguistics, engineering, and education. Teams of people from a variety of backgrounds conduct research that often involves collaborating with other NLM divisions, institutes at NIH, organizations within the Department of Health and Human Services, and academic and industry partners.

LHNCBC is organized into five major components: the Cognitive Science Branch (CgSB), the Communications Engineering Branch (CEB), the Computer Science Branch (CSB), the 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. News and information about LHNCBC research activities are available at http://lhncbc.nlm.nih.gov/.

Next Generation Electronic Health Records to Facilitate Patient-centric Care, Clinical Research, and Public Health

These projects target the overall recommendations of the NLM Long Range Plan Goal 3: *Create 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 designed to help consumers track their health information. The goals of the PHR are to:

- Help consumers manage and understand their healthcare problems,
- Facilitate federal goals for clinical data interchange using national vocabulary standards, and
- 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 dates for prescription refills and medical appointments and doctors' contact information and record questions they want to ask their doctor. From the main page, they can find out about their own health issues by consulting MedlinePlus and other trusted resources. 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 frequency of seizures.

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 (for example, Logical Observation Identifiers Names and Codes (LOINC), RxNorm, and Systematized Nomenclature of Medicine−Clinical Terms (SNOMED CT)). By using vocabulary standards, the PHR can automatically remind consumers about the preventive care and healthy behaviors specific to them, based on the data they entered.

In FY2013, LHNCBC researchers and developers continued to improve the PHR. We launched a demonstration site where people can test out the system without actually storing their data.

On the technical side, we updated JQuery from version 1.6 to 1.9 and implemented 1.9s datepicker, a specialized calendar function. We also updated software to Ruby 2.0 and Rails 4.0. We continued working on ways for users to import into the PHR their health data from continuity of care documents (CCD) – which hospitals and doctors will issue to their patients based on regulations to ensure the meaningful use of electronic health records (EHRs).

On the content side, we revised the content and wording of multiple preventive-care reminder rules, including those for lipid screening, bone density, cervical cancer screening, and breast cancer screening. We also added more consumer-oriented terms and lists of synonyms for medical conditions and surgeries, including common orthopedic and cosmetic surgery terms.

We conducted a formal usability study of the PHR and implemented changes based on the results that make the site easier to use. To simplify data entry, we revised the allergy and immunization tables, and we added a section on the main PHR page where users can view previously entered data and enter new data points for screening tests. We redesigned the health-panel organization and display and the health reminders, and we added visual cues to the main PHR page to ensure that new and unread reminders are displayed more prominently. We added tooltips, such as question-mark icons that lead to help messages, throughout the user interface, and we began developing methods for collecting and reporting PHR usage statistics. We are also exploring using the PHR to collect patient-reported data for clinical trials.

This project addresses the long-standing NLM interest in EMR systems and delivery of healthcare 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 will provide another consumer entry point to NLM's rich trove of patient-oriented data.

Using Drug Databases to Assess Prescribing Practices and Continuity of Care

Getting an accurate medication history for Emergency Department (ED) patients is important for their emergency care, especially since a significant proportion of ED visits are related to adverse events from prescription medications. Gathering such information from patients is time-consuming, expensive, and sometimes impossible (such as when a patient is unconscious), and patient-provided medication histories are often incomplete. Since 2009, Suburban Hospital’s ED has used Surescripts, a consortium of major Pharmacy Benefit Managers (PBM), to provide an electronic summary of the prescriptions their patients filled over the past year.

LHNCBC created an electronic messaging interface engine, based on *Mirth* (an open-source Health Level Seven, Inc. (HL7) interface program), which linked with Surescripts and delivered the prescription records for patients who had checked in for ED care at Suburban. Before the system went live, Suburban Hospital collected both Surescripts data and patient-provided histories. LHNCBC researchers obtained this information in a de-identified form, and then compared the two sources of information. We found that Surescripts information, when available, significantly augmented the history obtained in patient interviews by 28 percent (adding 1.1 prescription drugs per patient) and covered a high proportion (88 percent) of a patient’s current medications. These findings were published in the September 2013 issue of the *Annals of Emergency Medicine*.

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. Clinicians there report that the full-year prescription history is also helpful in identifying potential problems of drug compliance, drug-seeking behavior, and duplicative prescriptions. A brief follow-up study by the ED pharmacists showed that patients were indeed taking 85 percent of the medications they had not mentioned during the patient interviews.

We’re also using prescription databases to study adverse drug events due to drug to drug interactions. With a commercial insurance dataset from the Washington, DC, metropolitan area covering 2 million patients over one year, we’re studying the incidence of potentially dangerous drug to drug interactions and whether adverse reactions have actually occurred (for example, has a patient returned to the ED?). We’re evaluating the alerts generated by several commercial and publicly available drug to drug interaction knowledge databases by applying them to this prescription dataset.

EMR Database Research and Natural Language System Development

In FY2013, we continued working with the sixth update of the MIMIC-II electronic medical record (EMR) dataset. These big datasets are helping us to answer clinical research questions. For example, we’re using the de-identified MIMIC-II data under a restricted-use memorandum of understanding to conduct retrospective clinical studies on the:

- Significance of obesity and as a risk for mortality in intensive-care units (ICUs),
- Interactions between feeding practices and blood transfusions in premature babies with necrotizing enterocolitis, and
- Significance of vitamin B12 levels in ICU mortality and post-discharge survival.

Our analysis has shown that in the ICU, overweight and obese patients have a survival advantage over normal-weight patients. Our findings, published in the *Critical Care* journal, were followed by a letter to the editor of the *Journal of the American Medical Association* (*JAMA*). Manuscripts describing the results of the necrotizing enterocolitis and the vitamin B12 studies were submitted for publication.

In line with the NLM mission to facilitate access to health information resources, we continue to serve as a mirror site for PhysioNet, a very large (4.3 TB) and widely-used database of physiologic waveform tracings gathered from healthcare institutions worldwide by the same Massachusetts Institute of Technology researchers who developed MIMIC-II. These waveform data are collected by sensors attached to the patient and are used to monitor different aspects of a patient's clinical status, including heart rate and blood pressure, respiratory status, intracranial pressure, and sleep.

We continue to update the MIMIC-II waveform data and provide access to this collection. We've also continued to develop information retrieval and natural language processing (NLP) techniques for extracting important clinical variables from the rich narrative text in MIMIC-II. For example, in FY2013 we developed and published a method for extracting maternal data from a newborn's clinical notes. This is important because maternal history directly impacts newborn care but is typically only present in the newborn's record as narrative text.

Biomedical Imaging, Multimedia, and 3D Imaging

The objectives of this research area are to:

- Build advanced imaging tools for biomedical research;
- Create image-based tools for medical training and assessment;
- Develop multimedia image/text databases that accentuate database organization, indexing, and retrieval; and
- Develop content-based image-retrieval (CBIR) techniques for automated indexing of medical images by image features.

Imaging Tools for Biomedical Research

The American Society for Copolscopy and Cervical Pathology (ASCCP) has been using one of our image-based systems, the Teaching Tool, to assess the knowledge and skills of colposcopy professionals. More than 100 resident programs in Obstetrics/Gynecology and Family Practice at more than 95 universities and other premier institutions such as the Mayo Clinic have been using the tool. Since we first released the Teaching Tool in May 2010, these programs have administered more than 1,700 individual online exams of two types: the Residents' Assessment of Competency in Colposcopy Exam (RACCE) and the Colposcopy Mentorship Program (CMP) exam. In 2013, we worked with the ASCCP to update much of the content of these exams so that they're now consistent with the new guidelines for managing abnormal screening tests and cervical intraepithelial neoplasia (CIN). Planning is under way for a more advanced colposcopy resident exam that will also to be administered through the Teaching Tool.

Our Boundary Marking Tool, another imaging program, continued to be used by National Cancer Institute (NCI) staff and their collaborators around the world, including in Senegal, Costa Rica, Nigeria, the Netherlands, Spain, and the University of Oklahoma Health Sciences Center. With the tool, researchers have collected and annotated colposcopy images for biopsy studies and created a worldwide database for cervical cancer research.

We also continued research into methods for the computerized analysis and classification of cervical tissue using images collected and annotated by pathologists at the University of Oklahoma with the Communications Engineering Branch (CEB) Histology Image Assistant (CHIA) (formerly called the CEB Virtual Microscope). This work, with collaborators at the Missouri University of Science and Technology, includes applying our algorithm to carry out nuclei segmentation within the epithelial regions of the tissue and automated classification of the epithelium into classes of normal or various grades of abnormal (CIN1, CIN2, and CIN3).

In 2013, we successfully installed and operated the technology we developed to rapidly locate segments of epithelial tissue within large histology images of the uterine cervix. Developed in collaboration with researchers at Texas Tech University, the technique can reliably locate the epithelium one or two orders of magnitude faster than previously reported when the images are stored in formats commonly used for very large images. The method first uses compression information stored in the file to roughly separate tissue regions from background and then uses graphical processing unit (GPU) computation to classify the tissue regions into epithelial and nonepithelial tissue, at a speed 1,500 times faster than previously reported in the literature.

We also collaborated with researchers at Texas Tech University to integrate our advanced algorithms for histology image analysis and tissue classification into a Web-accessible system. Additional collaboration with academic groups included work with researchers at Lehigh University on classifying cervical disease based on comparing images of patients' cervixes with images in a database containing the "ground truth" classification (that is, a classification validated through follow-up) of these images. This way, we can compare the classification performed by the algorithm with the classification determined by human experts.

Content-Based Image Retrieval (CBIR)

Content-Based Image Retrieval (CBIR) is an active research area in the imaging research community since many of its tools and techniques find application in systems for image indexing, search, and retrieval. Goals of this research are to find images in repositories or the published literature that are visually or semantically similar to an image or text query. For example, one chest X-ray might be visually similar to another, but semantic similarity lies in finding another chest X-ray with the same lung disease.

We have developed several practical systems and tools that rely on CBIR research. For instance, our Open-i system allows users to access more than 1.3 million figures from medical journals including photographs, clinical images, charts, and other illustrations. People can sort search results based on different types of images, starting with "regular" and graphical images. Graphical images are further categorized as diagrams, statistical figures, flow charts, and tables. Regular images are further categorized as X-ray, ultrasound, computed tomography (CT), magnetic resonance imaging (MRI), and other modalities. We use more than 15 image modalities to classify the images with a support vector machine (SVM)–based framework. These modalities are image features, such as color, texture, and shape. We successfully used our modality classification system during international ImageCLEFmed competitions, and we incorporated it into Open-i.

To help people find the images they need more quickly, we also extract specific regions of interest (ROI) within images. These regions are often highlighted in publications by "markers" or pointers — such as arrows and symbols; we extract these markers first. We've improved the performance of algorithms we developed for detecting arrows using image-layout analysis.

Authors place arrows on figures to highlight important regions. They can be of different colors, but they're usually in contrast with the background. An automatic arrow-detection algorithm needs to be able to detect these arrows without prior knowledge about their color. Once we have the arrow, we can correlate concepts in the figure caption (text) with relevant regions of interest on the figure, which can aid in improved image-retrieval performance. We're developing additional algorithms to retrieve images that are more relevant to the query, whether the query is submitted as text, photo, or a combination. In order to retrieve relevant images, we first determine how the visual data correspond to the concepts in the query text. Using a method that divides the image into tiles, or image patches, we group each image patch with other similar-appearing patches across all images in the database. We develop a correspondence between representative image patches selected, frequently occurring groups, and key biomedical concepts in the accompanying text. We then apply machine-learning algorithms to extend this labeling to all patches in various groups so that every image patch has a text label associated with it. Because image patches are derived from images, every image is now transformed from a pixel-based representation into an image document where patches are replaced with these group labels.

The advantage of this approach is that not only are we able to map text queries to visual data, we're also able to apply fast, traditional, text-based information-retrieval techniques to image retrieval. We're working toward extending this idea to text-phrase retrieval strategies to find images with relevant local regions of interest. Recent advances in this area have led to a retrieval relevance of 75 percent for image retrieval using text queries such as, "Find lung CT images with ground-glass opacity."

We use spatial layout of pixel intensities within the image to eliminate regions that are not likely to be arrows, then we apply structural information about the arrow shape to identify candidate arrow regions. We used to use Markov Random Field (MRF) models to recognize arrow-type pointers with a precision of 85 percent and recall of 82 percent. Our new algorithmic approach has improved our precision to 94 percent and recall to 87 percent. Furthermore, our algorithm automatically detects whether the arrow is of a lighter or darker color compared with the background, which enables us to successfully apply it to a wide variety of images.

We're also developing a correlation between image ROIs and key biomedical concepts that appear in neighboring text, such as figure captions or other text describing the image content. Image features used to index the entire image may aggregate the details in specific ROIs.

Another example of the role of CBIR in our work is in our development of CervigramFinder, a research tool that automatically indexes and enables the retrieval of uterine cervix images (cervigrams) by shape, color, and texture features. Being able to search efficiently by image features is a significant step toward locating records in large databases of cervigrams and patient data, such as NCI's Guanacaste and ALTS **[Define?]** databases containing a total of 100,000 cervigrams. We've made advances in this area by developing algorithms that use values from several fields from the patient record, such as the woman's age, HPV-infection history, and prior sexual history, along with image features from the colposcopic exams. We're using these data to develop a model for predicting the likelihood that a patient will progress to more severe forms of HPV-based uterine cervical infections, including precancerous cells.

CBIR is also allowing us to improve the use of chest X-rays in an automated approach to detecting tuberculosis and other pulmonary diseases, which could be very useful in resource-poor countries. We've developed algorithms to automatically detect ribs, aorta, and other structures and to segment lung areas. Research continues into extracting texture features to classify lungs as normal vs. abnormal using SVM classifiers.

Other areas of our research include using distributed computing and GPUs for computer-intensive CBIR tasks, especially image segmentation. Through our collaboration with Texas Tech University, for example, we developed a method that uses GPU processing power for interactively following challenging object boundaries, such as the separation between the epithelial and nonepithelial tissue in histology slides of the uterine cervix. To support early detection and improve health care outcomes for people with cervical cancer, we plan to use these segmented epithelial regions to train doctors to detect various stages of pre-cancer.

Interactive Publications

Recognizing the increasing use of multimedia in scientific work, this project investigates and develops models for highly interactive multimedia documents that could transform the next generation of publishing in biomedicine. The project focuses on the standards, formats, and authoring and reading tools necessary for creating and using interactive publications that, in addition to text, contain media objects relevant to biomedical research and clinical practice, such as video, audio, bitmapped images, interactive tables and graphs, and clinical DICOM (Digital Imaging and Communications in Medicine) image formats for X-rays, CT scans, MRIs, and ultrasounds.

We've developed interactive publications containing these data types and tools for viewing, analyzing, and writing them. The tools, Panorama and Forge, are analogous to Adobe's Acrobat Reader and Professional for PDF documents. Panorama, used for viewing and analyzing these publications, was 1 of 9 semifinalists out of 70 entrants in Elsevier's Grand Challenge contest three years ago. We conducted a formal usability study of Panorama in 2013, and then enhanced the software to include bar charts and the ability to run on the Mac OS X operating system.

We extended Panorama to provide Annotation Concepts. Clicking on text in an interactive publication prompts an NLM servlet (RIDeM, developed in-house in 2004) to identify the corresponding Unified Medical Language System, or UMLS, concepts. The servlet returns an XML (eXtensible Markup Language) file to Panorama, which parses it to provide the preferred UMLS term and semantic group, and it provides linkouts for MedlinePlus, eMedline, Family Doctor, and other resources. We are continuing to develop additional features to group concepts by semantic relationships and other factors.

To avoid the need for the large downloads required by our original (desktop) approach, we investigated several Web-based methods for reading interactive publications. In 2012 we developed our first browser-based version (Panorama Lite) using Adobe Flex, thus eliminating the need to download the Panorama software. The only requirement to run it is to have Flash installed. In 2013, we improved and updated Panorama Lite features. Besides offering easy and intuitive usage, this version has better line-chart and graph support and includes tables and subsets similar to the original Panorama. Panorama Lite also features a unique map view that can present data at the county, state, and country levels in a color-coded form so users can visualize geographic patterns relatively easily.

Over the past two years, we've taken this project on the road, collaborating with two organizations to create interactive publications from their traditional, static ones: a publisher (ProQuest) and a government agency (the Centers for Disease Control and Prevention's National Center for Health Statistics, or CDC/NCHS). We created two interactive papers for ProQuest from one of their open-source journals (*Sustainability: Science, Practice and Policy*), and the company announced the launch of these papers for public use in a press release.

For CDC/NCHS, we converted two issues of their key document — the 2011 and 2012 *Health US In Brief* reports — to interactive form and hosted them 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 use of health care, health insurance coverage, supply of healthcare resources, and health expenditures. We also converted another important CDC/NCHS document, the Data Brief No. 115 March 2013, *Death in the United States, 2011*. We expanded Panorama's functionality to support drill-down pie charts for all these interactive papers, and we added usability support for standard deviation charts for *In Brief* using HTML5/javascript.

Screening Chest X-rays for Tuberculosis in Rural Africa

In FY2013, we continued our collaborative project with AMPATH (Academic Model Providing Access to Healthcare), an organization supported by the U.S. Agency for International Development (USAID) that runs the largest AIDS treatment program in the world. Through this project, we're conducting imaging research and developing systems to support NIH efforts to improve global health. Our objective is to use our in-house expertise in image processing to clinically screen HIV-positive patients in rural Kenya for lung disease, with a special focus on tuberculosis (TB) and other lung infections prevalent in patients with HIV. We provided AMPATH with lightweight digital X-ray units that are easy to transport in rural areas. The AMPATH staff will take chest X-rays (CXR) of people and screen them for the presence of disease.

In the past year, one of our X-ray units was mounted in the MOI University Hospital in the town of Eldoret in western Kenya. The images from the X-ray units are in a standardized DICOM radiological-image format. Through advances in technologies for Web-based Access to DICOM Objects (WADO) and the implementation of long-range Wi-Fi in western Kenya, images acquired in the field can be stored in the PACS (Picture Archiving and Communications System, a database system used in hospitals to store medical images) housed at the AMPATH building on the hospital grounds in Eldoret.

Because of the lack of sufficient radiological services in western Kenya, we've been focusing our in-house research effort on developing software that automatically screens the CXR images for disease. Our researchers are developing machine-learning algorithms to automatically segment the lungs; detect and remove ribs, heart, aorta, and other structures from the images; and then detect texture features characteristic of abnormalities, which allows us to discriminate abnormal from normal cases. These machine-learning algorithms, which allow computers to learn so they can do a task without being programmed to do it, require large sets of example X-rays. After receiving an internal review board (IRB) exemption, we explored many options for acquiring chest X-ray (CXR) training sets. We now have about 400 CXRs from Montgomery County's TB Control Program, 850 from a source in India, 2,000 from a hospital in China, 8,200 from Indiana University, and 250 from an open-source set from Japan.

The usable number of CXRs in our collection is less than the total. This is partly because of the marginal quality of many images and the inclusion of lateral views, which we're not considering yet. To create a training set, we manually validate and annotate the images, and in FY2013, we completed this process for the images from Indiana University. To ensure the privacy of the X-rayed subject, we discarded images that included any form of personally identifiable information, such as cardiac pacemakers with inscribed serial numbers, or dental images, which are known to be identifying. Also, we eliminated images with noisy data, such as the stray wires or tubes that are common in hospital-based radiological imaging, since they tend to confuse the classifiers.

Using these X-rays for training and testing, we developed algorithms for detecting relevant anatomy in the chest regions such as lungs, ribs, and the heart. To ensure the acquisition of high-quality images, we are using rib structures to develop other algorithms that detect the planar and positional rotation of the patient because such rotation can lead to incorrect diagnosis. Being able to detect the heart not only allows us to separate it from the lung regions, but we can then also detect abnormalities such as cardiomegaly (enlarged heart) that can be a precursor to congestive heart failure.

In 2013, we significantly improved the lung-segmentation algorithm with a novel atlas-based method that is 96 percent accurate. The algorithm is now better able to detect lung boundaries in CXR images for lungs exhibiting disease pathology, and by increasing the algorithm's computational efficiency we've also enabled it to run much faster (30 seconds versus several minutes). Radiologists from the NIH Clinical Center, Yale University, and the University of Missouri in Columbia helped us by annotating pathology in some of our images. We used these annotated images to train our SVM-based classifier, which uses several features extracted from the X-rays as input, such as histograms of intensity, gradient magnitude and orientation, shape, and curvature. On the basis of these input features, the SVM returns a confidence value, allowing an operator to inspect cases in which the classifier is uncertain. We also compared the performance of the algorithm classifier with that of human experts. We found that they perform similarly (87 percent accurate), but the classifier tends to be more sensitive, yielding nearly twice as many false positives. While not ideal, this oversensitivity does prevent overlooking X-rays that show disease, and it's useful in a resource-constrained setting. We are continuing to research methods for advancing classifier performance by sampling image patches.

Remote Virtual Dialog System (RVDS)

The Remote Virtual Dialog System (RVDS) will make the NLM "Dialogues in Science" series, currently only viewable in the NLM Visitors Center, available anywhere through the Internet. Major support for the project came from stimulus funds made possible through the 2009 American Recovery and Reinvestment Act (ARRA), as well as NLM-appropriated funding. The project involves enhancing programmatic capabilities of the virtual dialogue model. During FY2013, we improved free-text recognition and gathered feedback from independent reviewers on an alpha version of the Web-based version of the series. We also made the Web-based series available in the NLM Visitors Center.

Computational Photography Project for Pill Identification (C3PI)

Launched in September 2010, the Computational Photography Project for Pill Identification (C3PI) is a start on an authoritative, comprehensive, public digital-image inventory of the nation's commercial prescription, solid-dose medications. Working with expert consultants, we're creating a collection of photographs of prescription tablets and capsules, confirming that the images match the description of the medication, and developing and matching the images of the samples to relevant metadata (such as size descriptions, dimensions, color, and the provenance of the sample).

In FY2013, we increased our online collection to more than 70,000 images of nearly 2,000 pill (solid-dose) samples pharmaceuticals from more than 150 manufacturers and distributors. The team generates high-resolution, high-quality pill images from a variety of lighting conditions. The long-term goal of our collaboration is to facilitate the development of computerized tools to help identify medication based on features such as shape and color.

Also in FY2013, as part of a public launch of the C3PI project, we worked with the contractor to create two public online repositories: http://rximage.nlm.nih.gov (which links to RxNav) and http://splimage.nlm.nih.gov. Through these Web sites, we can distribute images of oral, solid-dose medications to the public and to pharmaceutical manufacturers, respectively.

As part of cost management and to reflect the current needs of the C3PI project, we established a new contract, conducted a fair and open competition, and awarded the contract to Medicos Consultants at the end of FY2013. In FY2013, we completed all the research to establish the imaging, database, delivery, and storage protocols for structured product label (SPL) IMAGE files, and entered production mode for C3PI. This was a joint development effort with contributions by both NLM staff and the contractor. In FY2014, we'll acquire data using these protocols.

In October 2013, NLM's Office of Medical Subject Headings (MeSH) hosted a workshop, the DailyMed Jamboree, for a broad group of invited distributors, labelers, federal partners (from FDA, CMS, and DOJ) and other users of NLM pharmaceutical data. About 100 people attended the DailyMed Jamboree in person, and another 500 watched the videocast. We presented a complete description of our image-collection processes, the intended uses of the data, and the information outlets. We also participated in a panel with other government personnel, including SPL representatives from the Food and Drug Administration (FDA).

The Visible Human Project

The Visible Human Project image datasets serve as a common reference for the study of human anatomy, a set of common public domain data for testing medical imaging algorithms, and a test bed and model for constructing image libraries that can be accessed through networks. These datasets are available through a free license agreement with the NLM. We distribute them in their original or in PNG (*Portable Network Graphics*, a raster graphics file format that supports lossless data compression**)** format to licensees over the Internet at no cost and on DVDs 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-surgery uses, in addition to artistic, mathematical, legal, and industrial uses. In FY2013, we continued to maintain two databases for information about how people are using the Visible Human Project: one for information about the license holders and their intended use of the images, and the other for information about the products the licensees provide NLM as part of the Visible Human Dataset License Agreement. More than 1,000 newspaper, magazine, and radio pieces have featured the Visible Human Project since we released the first dataset in 1994.

3D Informatics

We continue to address problems encountered in the world of three-dimensional and higher-dimensional, time-varying imaging through our 3D Informatics Program (3DI). We provide ongoing support for image databases, including the National Online Volumetric Archive (NOVA), which NLM created in 2003 and continues to host. NOVA contains 3D data from numerous medical institutions, including the Mayo Clinic Biomedical Imaging Resource and the Walter Reed Army Medical Center Radiology Department. In FY2013, 3DI added data to NOVA from an ion-abrasion scanning electron microscope from our collaboration with NCI in high-resolution electron microscopy. We continue to serve a broad community with these data and have become leaders in the distribution of medical data to the public.

Throughout FY2013, we continued our collaboration with 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. This work combines high-performance computing with life sciences research, related to the detection and prevention of cancer and infectious diseases. The resulting visuals have provided insights about the character of several immunological cells, cell structures, and the cells' interaction with pathological viruses, including HIV.

We continued our commitment to processing data collected through transmission electron tomography. We successfully published the results of software-development research that uses GPUs for high-performance computing for sub-volume averaging and reconstruction. (The image data collected are equivalent to the serial sections of a CT scan. These sections can be put back together to "reconstruct" the original "volume." Any mathematics performed on part of the "volume" is called "sub-volume.") We're working now to develop emerging methods into mature software that can extend this work on sub-volume reconstruction to single-particle microscopy. If successful, the software will expand our impact to a wider range of molecular targets in systems biology.

This past year, we also helped supervise segmentation efforts for data from ion-abrasion electron microscopy in a study of the disposition of malaria pathogens (that is, where they wind up) in normal human blood cells. We're evaluating the impact of heterogeneous sickle cell hemoglobin on the infiltration, geometry, and spatial relationships of the pathogens in affected erythrocytes. Other work in related areas includes the segmentation and study of vaccinia viruses in normal and mutant human liver cells.

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 is ITKv4.2.2. More than 845,000 lines of openly available source code comprise ITK, and with it, people can access a variety of image-processing algorithms for computing segmentation and for registering high-dimensional medical data. ITK runs on Windows, Macintosh, and Linux platforms, so it can reach across a broad scientific community. It is used by more than 1,500 active subscribers from 40 countries. A consortium of university and commercial groups, including our intramural research staff, provides support, development, and maintenance of the software.

ITK remains an essential part of the software infrastructure of many projects across and beyond 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 applied programming interface (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 Mayo Clinic's Analyze and SCIRun from the University of Utah's Scientific Computing and Imaging Institute, and through developing a new release of VolView, free software for medical volume image viewing and analysis.

Our ongoing efforts in 2013 included expanding and continuing to develop Simple ITK, or sITK. Developed primarily by programmers at NLM, sITK is a simplified layer built on top of ITK. It's intended to facilitate ITK's use in rapid software development and education through the support of scripting languages, primarily Python. The sITK interface conceals the structural and design complexities of ITK, enabling more straightforward, procedure-based programming. Designed to be an interpreted scripting system, sITK supports a typeless, polymorphic data model, thus simplifying the use and expression of ITK in image-analysis education. The current software release of Simple ITK is version 0.7.

3D Printing

In FY2013, the 3DI group participated in the HHS Ignites challenge by proposing, with our colleagues at the National Institute of Allergy and Infectious Diseases (NIAID) and the National Institute of Child Health and Human Development (NICHD), an online repository called the *NIH 3D Print Exchange*. Designed to supply biomedical shape files for education and research to a growing audience of users worldwide, the exchange provides public software built from NLM's Insight and the Visualization Toolkits to automatically generate printable models from X-ray CT data. The project Web site is currently online in beta-testing mode, and a national advisory board is assisting with oversight, testing, and guidance for the project.

The 3DI group also continued to investigate the use of rapid-prototyping technologies in radiology with partners at NIAID. 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. A new set of models is under development, including dosimetry models from CT scans of small animals.

Open-i: Image and Text Indexing for Clinical Decision Support and Education

A picture is worth a thousand words, especially in medical research and clinical practice! Most people can understand complex biomedical concepts more easily if they are presented visually: through radiographic images, photographs of organs, sketches, graphs, or charts. This idea motivates our project, part of the LHNCBC Clinical Information Systems effort, which exploits ongoing research in both natural language processing and content-based image retrieval by processing both text and image features. We developed the Open-i system for finding images and figures in published literature or other sources. Open-i system enables users to search and retrieve medical citations from 450,000 open-access articles in PubMed Central® (PMC). Users may search by text queries as well as by example images. They can filter images by type (e.g. X-ray, graph, etc.), filter journals by clinical specialty, and rank papers by clinical task (e.g., treatment).

Open-i was released to the public in 2012 and is the first production-quality system of its kind in the biomedical domain, Open-i gives medical professionals and the public access to images contained in biomedical articles that are highly relevant to their query, as well as a summary of the articles' take-away messages. The system is available 24/7, and it can handle about 20,000 interactions per day in real time. The quality of the information delivered by Open-i has been evaluated in international medical-image-retrieval "Cross Language Evaluation Forum" competitions (ImageCLEF), in which the system consistently ranks among the best. For example, the system demonstrated the best retrieval results in a 2013 ImageCLEF that attracted participants from academia, industry, and clinical settings.

In 2013 we added a collection of clinical documents to the Open-i open-access biomedical articles. This collection, provided by Indiana University, consists of chest X-ray examinations and accompanying de-identified reports for about 4,000 patients. To prepare the collection for indexing and public release, the Open-i team has reviewed the university's de-identification of the DICOM images and radiology reports and removed images with any possible risk of re-identification (for example, images of teeth).Then, we converted the remaining images to JPEG format for indexing. We also indexed the text reports with NLM's MeSH® (medical subject heading) terms, and converted the reports to Open-i-enriched citations. The resulting collection is indexed and searchable in Open-i either by itself or in combination with the scientific publications.

This past year we also increased the collection of images in Open-i to 1.3 million from PubMed Central articles (from 450,000 in 2012) and 8,000 from the 4,000 Indiana University radiology reports. The demand for Open-i services nearly doubled in 2013 — about 20,000 total visitors, bots, and auxiliary pages a day, compared with 11,000 in 2012) — to accommodate this increase, we upgraded the system architecture, refactored the XML structure, and redesigned the indexes. In addition to online searches, Open-i provides batch-retrieval services for researchers who need access to images. For example, the University of Pittsburgh uses the services to add illustrations to their oral squamous cell carcinoma dataset.

In response to users' requests in 2013, Open-i introduced a search-history list that allows users to see their previous searches. Adjustments to our search algorithm led to an improvement in modality classification from 85 percent to 90 percent. We also modified the auto-complete suggestions in searches, based on information about the presence and frequency of the UMLS terms in the collection of documents indexed in Open-i.

According to data from our internal logs, Open-i serves up to 20,000 distinct users a day; of these, per Google Analytics' estimates, about 5,000 are unique visitors.

The evolution of Open-i continues to be supported by research in these key areas:

- Representing images with text strings;
- Combining global and local representations of image features;
- Improving methods for automatically segmenting multi-paneled illustrations into single images and partitioning their captions to correspond to those single images; and

- Improving methods for extracting pointers (arrows, arrowheads, symbols) within images to identify regions of interest that could be more relevant to a query than the entire image would be.

In other work this past year, we took steps toward building a visual structural framework for organizing information called a visual ontology. We also developed methods for segmenting lung and brain tissues and extracting key features for imaging properties of several pathologies in these tissues (for example, lung cysts, micronodules, and emphysema).

Turning the Pages (TTP)

This goal of this project is to give laypeople access to historically significant and previously inaccessible books in medicine and the life sciences. We build 3D models for books and develop animation techniques that let users touch and turn page images in a realistic way on touch-sensitive monitors in kiosks at NLM or tablets using a high-resolution (“Retina”) iPad app, or click and turn if they’re using an online version. We’ve also built a different 3D model for a “scroll”-type document and applied it to the 1700 BC Edwin Smith medical papyrus, which can be “touched” (or clicked) and “rolled out.” The TTP Web site is very popular, attracting about 300,000 page views a month. The iPad app is also popular: in 2013, 3,270 people downloaded the app for the first time (new users), and more than 16,751 people who already had the app updated it with the newer version of the software, which indicates a steady user base.

The Turning the Pages kiosks at NLM and the NLM Web site now present 11 rare books. In 2013, we added 4 books to the iPad version, bringing that total to 10, with only the Smith papyrus yet to be added. Now included in the TTP Web and iPad versions is Elizabeth Blackwell's *A Curious Herbal* (published between 1777 and 1779), previously available only at the NLM kiosks. We had to recreate all the graphics and update the animation software for that project. We also added new features to the iPad version in 2013, such as bookmarks and contextual zoom for curators’ notes. To reduce project costs, we modified our production pipeline to capture images in-house rather than using commercial scanning companies. Right now, we’re creating the TTP version of a Mongolian Prayer Scroll. We’re also investigating options to create a TTP version of the now-historic genetic charts, created by Nobel Laureate Marshall Nirenberg, that elucidate the process of translating the four-letter code of DNA into the 20-letter alphabet of amino acids, the building blocks of proteins.

For the longer term, we’re studying ways to develop a reactive 3D implementation system for TTP and investigating tools for this purpose, such as Unity, Coco’s 3D, and the Unreal Engine. The advantages of a real-time 3D system are:

- We can produce 3D versions of each book more quickly,
- Other institutions can use our software to create their own interactive books, and
- We can discover new functionalities, such as rotating a book 360° and turning multiple pages at once.

In addition, anticipating the next generation of kiosk design, we’re investigating newer display technologies, such as multi-touch monitors, which enables users to use two or more fingers to perform tasks such as zooming in and out.

Natural Language Processing and Text Mining

Medical Article Records System (MARS)

NLM’s flagship database, MEDLINE®, contains more than 20 million bibliographic records for articles from more than 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 and online journals.

While the bulk of citations now come to NLM directly from publishers (in XML format), nearly 820 journal titles are provided only in paper form. These papers are processed by the Medical Article Records System, or MARS, which we launched in 1996. 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 for MEDLINE. Our algorithms extract the data through a pipeline process: segmenting page images into zones, assigning content labels to the zones (title, author names, abstract, grant numbers, etc.), and pattern matching to identify these entities.

We manage and continually improve the MARS system. For example, we’re introducing three new features: (1) an expanded 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. We have completed the software implementation and the integration test for the first two features, which are now being deployed to the production system used by NLM Library Operations.

Citations that come to NLM in electronic form from publishers often contain errors or have missing content. Missing items include databank accession (identifying) numbers from databases such as GenBank, NIH grant numbers, grant support categories, investigator names, and information about the links between articles and the comments submitted in response to them. The capture of investigator names can be especially difficult because some articles contain hundreds of such names, and capturing the articles that "comment on" another paper (usually an editorial or a review article) requires operators to open and read other articles related to the one being processed. To automatically extract these fields from online articles, we developed the Publisher Data Review System (PDRS), whose subsystems are based largely on machine-learning algorithms such as the support vector machine, or SVM.

PDRS went into production in early FY2012 for open-access articles in NLM's PubMed Central. To extend automated data extraction to *all* online journals on publishers' sites, including the ones with restrictive copyrights, we're developing IMPPOA (In-Memory Processing for Publisher Online Articles). This is a system based on the PDRS platform and its machine-learning algorithms, but implemented to process articles in RAM memory, which is temporary. IMPOAA:

- Provides data missing from the XML citations sent in directly by publishers,
- Corrects errors in publishers' data by extracting data from the articles on their sites and comparing these with the data sent to NLM, and
- Extracts data from articles for which publishers do not send in citations at all.

Because this new system avoids downloading the articles to a disk drive, we expect IMPOAA to eliminate publishers' concerns about copying articles into an external system disk.

The systems outlined above rely on underlying research in image processing and lexical analysis that also enables the creation of new initiatives for applying these techniques, such as the ACORN project (also known as the Automatically Creating OLDMEDLINE Records for NLM project), described below.

Digital Preservation Research

The long-term preservation of documents in electronic form is a critical task for NLM as well as other 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, storing the documents and metadata, and discovering knowledge in the archived material. To provide a platform for this research, we built and deployed a System for Preservation of Electronic Resources (SPER). SPER builds on open-source systems and standards (e.g., DSpace, RDF), and incorporates in-house-developed modules that implement key preservation functions.

Our research focuses on two collections. One is a historic medico-legal collection of 67,000 early twentieth-century court documents from the FDA. These are "notices of judgment" issued by courts against companies that were indicted for misbranding or adulterating foods, drugs, or cosmetics. They offer insights into legal and governmental history dating from the 1906 Food and Drug Act, and illustrate regulatory impacts on public health. Curators in the NLM History of Medicine Division are using SPER to preserve the FDA documents, and in 2013 they processed more than 15,000 of them. These documents and their metadata are in a publicly accessible NLM Web site.

The second collection, from NIAID, is a set of conference proceedings of the US-Japan Cooperative Medical Science Program (CMSP) Cholera and Other Bacterial Enteric Infections Panel, an international program conducted over a 50-year period from 1960 to 2011. For this collection, our activities include: (1) 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 (2) developing a portal where the public can go to search for research articles, institutes, and authors. To support these activities, we developed techniques for automatically extracting three different types of metadata from the CMSP documents:

- 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.

We then used the metadata to implement data-analysis functions for discovering patterns and trends in factors such as important drugs, discoveries, investigators, and international collaborations under the CMSP program over its 50-year span.

In FY2013, we explored building a knowledgebase from the metadata of an archived collection and performing semantic queries against such archives. We hoped to discover important domain-specific information, thus generalizing the data-analysis capability. We selected open-source tools to create a model based on the OWL/RDF structural framework for organizing information for a given collection characterized by its metadata fields and their interrelations, and we developed new tools to generate a knowledgebase from the stored metadata using that framework. We used techniques to semantically query such knowledgebases through a Web browser and then to graphically display the results.

We applied this methodology to the CMSP collection by building a complete CMSP knowledgebase and performing semantic queries to obtain various patterns and trends of interest. We also developed an ontology model for the FDA collection that can be used to generate the corresponding knowledgebase once the collection is fully archived.

Automatically Creating OLDMEDLINE Records for NLM (ACORN)

NLM would like to expand MEDLINE to include all bibliographic records beginning in 1879, when *Index Medicus* was first developed. The earliest citations exist only in printed paper form, and NLM Library Operations (LO) has collected many of these with considerable manual effort. To translate these paper indexes into electronic format, we designed the ACORN system, which combines scanning, image enhancement, optical character recognition (OCR), image analysis, pattern matching, and related techniques to extract electronic versions of the printed indexes. The fact that the printed versions used old typefaces, fonts, and a mix of different languages yields inaccurate OCR results. 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 searches MEDLINE to avoid duplicating records that already exist.

ACORN has three main components: Scanning and Quality Control, Processing, and Reconciliation (that is, when operators verify the extracted information). We completed and delivered the first component to NLM Library Operations on April 2013, and by the end of 2013 the LO staff had scanned and completed quality-control checks of 16 of the 60 QCIM volumes. We are still developing the other two components and will release them for production in 2014.

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. Results from the two basic methods are combined into a ranked list of recommended indexing terms, incorporating aspects of MEDLINE’s indexing policy in the process.

NLM Library Operations (LO) MEDLINE citations indexers regularly (and increasingly) use the MTI system. To facilitate their indexing, MTI provided recommendations for 673,428 articles in FY2013 as an additional resource available 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 specialized versions of MTI to assist in the indexing of 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 (MTIFL) for those journals. As a “first-line” indexer, MTI indexing is also still subject to human manual review. The number of MTIFL journals will grow gradually and should prove to be a time and money saver for NLM.

In FY2013, we added 75 new journals (for a total of 120) to the MTIFL program, which included 9,771 articles. We are collaborating with LO to evaluate how well MTI is performing on indexing the journals that are already part of the MTIFL program by computing standard information-retrieval measures (recall, precision, and fmeasure) and then comparing MTI's indexing recommendations with the final, official MEDLINE indexing. We also work with LO to identify future MTIFL journal candidates.

In FY2013, MTI provided the primary baseline for the international BioASQ challenge for biomedical semantic indexing and question answering (http://www.bioasq.org/). The aim of the challenge is to make biomedical text more accessible to researchers and clinicians. The MTI indexing results provided one of the baselines used in the “large-scale online biomedical semantic indexing” part of the challenge, which is designed to parallel the human indexing currently being done at NLM. The II team provided help and guidance in developing the list of journals used in the challenge, as well as the baseline results. MTI will also provide a baseline for the second year of the BioASQ challenge in 2014.

The II team collaborated with LO to develop a new system to help with indexing LO’s NLM Technical Bulletins. The system automatically supplies keywords for all Technical Bulletins, which are then manually reviewed before publication. In FY2013, our new system processed 1,361 Technical Bulletins, including all archived Technical Bulletins from 1997 to the present. A comment from an LO staffer illustrates how effective the new system has been for them: “The team is using the system to index all Technical Bulletin articles now, streamlining our workflow and making our assigned terms consistent. We are saving a lot of time with each article!”

MetaMap is a system for identifying concepts within text documents. It is a critical component of the MTI system and is also used worldwide in bioinformatics research. In 2013, we significantly improved its processing speed, added XML output, implemented a way to determine whether a statement is positive or negative, e.g., "He has cancer." vs. "He does not have cancer" (known as negation identification), and enabled users to supply their own acronyms-abbreviations list. In 2013, we migrated MetaMap’s legacy lexicon module to a Java-based implementation. MetaMap is available on Windows, Macintosh and Linux platforms. Users can build their own datasets with the MetaMap Data File Builder and run their local version of MetaMap to process documents containing sensitive data via either an embedded Java API (applied programming interface) or a UIMA (unstructured information management architecture) wrapper. In FY2013, users downloaded about 2,000 copies of MetaMap and 900 copies of the Java API and UIMA wrapper.

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 (Open-i, 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 just-in-time access to patient-specific information to be integrated into an existing EMR system. Such patient-specific information includes medications linked to lists of medications for each patient, or formularies, and images of pills, evidence-based search results for patients’ complaints and symptoms, and MedlinePlus information for patient education. For clinical settings without access to the API, a Web-based interface allows information requests to be entered manually.

The InfoBot API integrated with the NIH Clinical Center’s EMR system (CRIS) has been in daily use through the *Evidence-Based Practice* tab in CRIS since July 2009. In 2013, the tab was accessed 609 times a month, on average, by more than 1,380 unique users at the NIH Clinical Center.

Consumer Health Information and Question Answering (CHIQA) system

NLM’s customer services receive about 90,000 requests a year. In 2012, we started to investigate the possibility of automating the process of answering these consumer health questions. In 2013, we developed and evaluated a prototype Consumer Health Information and Question Answering system (CHIQA). The prototype can classify the incoming requests as either questions about health problems or requests to correct MEDLINE citations. Once the request type is recognized, CHIQA generates an answer and submits it to NLM’s reference staff for review. For MEDLINE correction requests, the system automatically finds and retrieves the citation that set off the request, extracts relevant information, and generates an answer. The prototype also understands simple frequently asked questions about causes, treatments, and prognoses of diseases. For these questions, CHIQA finds relevant articles from NLM consumer resources, such as Genetics Home Reference and MedlinePlus, and uses sections of the articles to answer the questions.

De-identification Tools

De-identification allows people to conduct research on clinical narrative reports. We are designing a Clinical Text De-identification (CTD) system that will remove protected health identifiers from narrative clinical reports. The provisions of the Privacy Rule of the Health Insurance and Accountability Act require the removal of 18 individually identifiable health information elements that could be used to identify the individual or 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 de-identifies clinical narrative text in a form of electronic messaging known as Health Level Seven, Inc. (HL7) version 2. It can use information embedded in various HL7 fields as well as externally provided information, such as list of names of the healthcare providers at NIH.

We're using the Visual Tagging Tool (VTT) we designed to produce gold standards against which to test the CTD system. Although we designed the VTT specifically to help the CTD identify protected health information (PHI), the natural language processing (NLP) community is already also using it for other types of lexical tagging and text annotation.

By the end of 2013, we amassed a collection of 21,849 clinical reports of 7,571 patients - in which a human reviewer manually labeled every piece of individually identifiable health information. We can use this collection as the gold standard for testing the CTD system.

Librarian Infobutton Tailoring Environment (LITE)

Infobuttons are links from one information system to another that anticipate users' information needs, take them to appropriate resources, and help them retrieve relevant information. They are mostly found in clinical information systems (such as electronic health records (EHRs) and PHRs) to give clinicians and patients 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 Health Level Seven, Inc. (HL7) - an electronic messaging standards development organization - to develop an international standard to support the communication between clinical systems and knowledge resources. MedlinePlus Connect currently provides an HL7-compliant query capability.

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 the most applicable resources from a large library of known resources. The infobutton manager customizes the links to those resources using appropriate data from the context, and then presents the user with a list of those links. The NIH Clinical Center Laboratory for Informatics Development is working with investigators at the University of Utah and the Department of Veterans Affairs to establish a freely available, HL7-compliant infobutton manager, known as "Open Infobutton" (http://www.openinfobutton.org), that can be a national resource for electronic health record developers and users. With the Open Infobutton, clinicians and patients will be able to obtain the health-related information they need, when and where they need it.

Infobutton managers, including Open Infobutton, require knowledgebases to do their customization work. The knowledgebases are very institution-specific. We developed the Librarian Infobutton Tailoring Environment (LITE), a user-friendly tool that can be used by an institution's medical librarians and provides Open Infobutton with the knowledge it needs to customize its responses to requests from that institution. The system is in beta testing at the University of Utah (http://lite.bmi.utah.edu). In 2013, we transferred the maintenance of LITE to the University of Utah, which will continue to make it open-access and will also develop the software to make it part of an open-source package that can be installed at any institution. A user-evaluation project is now under way in collaboration with Ohio University.

Terminology Research and Services

The Patient Data Management Project (PDM) brings together several activities centered on lexical issues, including developing and maintaining the SPECIALIST lexicon and lexical research. Those lexicon and lexical tools support key NLM applications. A package of lexical-tool applications underlies the MetaMap algorithm we use to find UMLS concepts in biomedical text and to automatically index MEDLINE abstracts. We distribute the lexicon and lexical tools to the medical informatics community as free open-source tools.

In FY2013, we provided support to 35 internal users, 14 U.S. domestic users, and 8 international users. We also enhanced 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). The enhanced derivational-variant system will be part of the UMLS 2014 release. The 2014 release of the SPECIALIST lexicon contains 476,856 records representing more than 875,000 forms, an increase of 6,865 records from the 2013 release. Many of the new records are derived from de-identified clinical records from our own de-identification project and from our work with the MIMIC-II database.

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 FY2013, we investigated the use of terminological resources for mapping between rare-disease information sources, and we developed a framework for aligning pharmacologic classes between MeSH and Anatomical Therapeutic Chemical (ATC) classification. We supported the NLM Value Set Authority Center (VSAC) – a crucial part of the quality measures stemming from the meaningful use regulations – by proposing metrics for assessing the quality of value sets in clinical quality measures. Finally, in collaboration with the Center for Drug Evaluation and Research at the FDA, we developed methods for extracting information about adverse drug events from MEDLINE indexing.

Research activities this year resulted in one journal article, five papers in conference proceedings, seven editorials, abstracts, and posters, 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, the eleventh version of the International Classification of Diseases).

Semantic Knowledge Representation

The Semantic Knowledge Representation (SKR) project conducts basic research, based on the UMLS knowledge sources, in symbolic natural language processing. A core resource is the SemRep program, which extracts semantic predications (relationships — such as interacts with, treats, causes, inhibits, and stimulates — between drug and disease, gene and gene, gene and disease, drug and -drug, etc.) from biomedical text. SemRep was originally developed for biomedical research. We're developing a way to extend its domain to influenza epidemic preparedness, health promotion, and health effects of climate change. In FY2013, we made a downloadable version of SemRep available to the public.

SemRep finds biomedical-related semantic relationships in MEDLINE, and then our Semantic MEDLINE Web application manipulates those relationships. 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 Semantic MEDLINE, which integrates PubMed searching, SemRep predications, automatic summarization, and data visualization. The application helps users manage the results of PubMed searches by creating 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 the 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.

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

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

ClinicalTrials.gov

Established in 2000, ClinicalTrials.gov makes public comprehensive information about registered clinical research studies. It receives more than 95 million page views and hosts about 980,000 unique visitors per month. Nearly 13,000 study sponsors, including the federal government, pharmaceutical and device companies, and academic and international organizations, submit data to ClinicalTrials.gov through a Web-based Protocol Registration System

(PRS). At the end of FY2013, the site had nearly 154,000 research studies, conducted in all 50 states and in more than180 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. More than 10,400 of the closed studies display summary-results tables describing primary and secondary outcomes, adverse events, and characteristics of the participants studied.

In FY2013, new registrations of clinical trials were submitted at an average rate of 400 records per week, an increase of 8 percent from FY2012. The average rate of new results submissions was about 70 per week, consistent with FY2012. We can attribute the continued growth in the use of ClinicalTrials.gov to U.S. laws that require registering and reporting the results of clinical trials, as well as international recognition of the scientific and ethical importance of registration and reporting results. The combined registry-and-results database provides information about ongoing and completed clinical research for patients, healthcare providers, and policy decision makers.

In FY2013, ClinicalTrials.gov staff continued to implement and educate the public about the most recent clinical trial law, Section 801 of the Food and Drug Administration Amendments Act of 2007 (FDAAA 801). We have been working with the NIH Office of the Director, other NIH Institutes and Centers, and the FDA) on a Notice of Proposed Rulemaking (NPRM) that will elucidate the requirements of FDAAA 801 and solicit public comment on key implementation issues.

Following the launch of the redesigned public Web site in FY2012, we started evaluating the public’s experience with the redesign and shifted our focus to evaluating and enhancing the ClinicalTrials.gov data-entry site, the PRS. Sponsors and investigators use the site to submit, update, and maintain information about their studies. On the basis of comments from users to an online survey and other user feedback, we implemented user-interface improvements intended to streamline the data-entry process. We started a usability study of the PRS in FY2013, and we’ll use its findings for further enhancements in FY2014. We continued providing targeted education and outreach on the results database and submission requirements through hosting three on-site training workshops, presenting at conferences, participating in working groups, and publishing in journals. ClinicalTrials.gov research projects in FY2013 included:

- Characterizing prematurely terminated trials registered at ClinicalTrials.gov,
- Illuminating reasons for early termination, and
- Finding out how many and what types of summary results are available from such trials in the ClinicalTrials.gov results database and the published literature.

In April 2013, the LHNCBC Board of Scientific Counselors reviewed the ClinicalTrials.gov program and praised it. The program continued to provide technical advice and collaborate with other clinical study registries, professional organizations, funders, and regulators on working toward developing global standards for trial registration and for reporting to results databases. For example, a key activity continues to be 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 release in FY2014.

Genetics Home Reference (GHR)

The Genetics Home Reference (GHR) Web site offers high-quality information about genetic conditions and the genes and chromosomes related to those conditions. It answers the public's questions about human genetics and the rich technical data from the Human Genome Project and other genomic research. At the end of FY2013, the GHR included user-friendly summaries of more than 2,100 genetics topics, including more than 900 genetic conditions; about 1,250 genes and gene families; all the human chromosomes; and mitochondrial DNA. GHR 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.

GHR celebrated its 10th anniversary in FY2013. In the past year, we expanded the project’s genetics content for consumers, adding 242 new condition, gene, and gene-family summaries and new Help Me Understand Genetics pages about genetic susceptibility and informed consent. In FY2013, the site averaged almost 42,500 visitors per day and about 34.9 million hits per month (an increase of 52 percent and 26 percent, respectively, from FY2012).

In FY2013, we integrated GHR results into NLM’s MedlinePlus Connect. This service enables electronic medical records and other applications that use MedlinePlus Connect to retrieve GHR summaries (along with MedlinePlus content) by using code queries from SNOMED CT (the Systematized Nomenclature of Medicine Clinical Terms). GHR topics, each of which can map to multiple SNOMED CT codes, currently map to more than 2,300 SNOMED CT codes.

GHR continued its formal collaboration with Genetic Alliance, an umbrella organization for condition-specific genetics interest groups, to update existing GHR Web site content and to track new clinical and research developments about particular genetic conditions. We updated about 70 existing GHR topics through this initiative in FY2013. We also performed outreach activities last year to increase public awareness about GHR. We presented the Web site to several groups, including health and science journalists who visited NLM as part of the Association of Health Care Journalists - NLM Fellowship program, clinical and molecular fellows at the National Human Genome Research Institute (NHGRI), and conference attendees at the annual meetings of the National Society of Genetic Counselors and the American Medical Writers Association.

Profiles in Science Digital Library

The *Profiles in Science®* Web site showcases digital reproductions of items selected from:

- The personal manuscript collections of 33 prominent biomedical researchers, doctors, public health practitioners, philanthropists, political leaders, and other people who provided resources, removed barriers, and spearheaded projects to improve the health of the nation and the world, and
- Three thematic collections: the 1964-2000 Reports of the Surgeon General, the history of the Regional Medical Programs, and the Visual Culture and Health Posters.

The site gives researchers, educators, and future scientists all over the world access to unique biomedical information previously accessible only by making in-person visits to the institutions holding the physical manuscript collections. It also serves as a tool for recruiting donations of collections from scientists who wish 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, including to individuals with disabilities. The growing *Profiles in Science* digital library provides ongoing opportunities for future experimentation in digitization, optical character and 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. Collections donated to NLM contain published and unpublished materials, including manuscripts, diaries, laboratory notebooks, correspondence, photographs, poems, drawings, and audiovisual resources. The Web site averages more than 95,000 unique visitors each month, including people seeking an authoritative source of information about current events, such as the February 2013 death of former U.S. Surgeon General C. Everett Koop. When the July 25, 2013, Google home-page doodle featured the late Rosalind Franklin's birthday, *Profiles in Science* received nearly 100,000 unique visitors in a single day.

At the end of FY2013, the 36 publicly available collections contained 27,058 items composed of 142,214 digitized image pages, including transcripts of 10,075 handwritten pages or pages we couldn’t use optical character recognition technology for. We also completed a video biography of Michael E. DeBakey that we’ll be adding soon to the *Profiles in Science* site.

In addition to updating the *Profiles in Science* collections during FY2013, we increased the visibility of the *Profiles in Science* digital items through user-configurable software that allows customers to display or hide large thumbnail images. To enhance the interoperability and availability of the *Profiles in Science* data, we added an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) application programming interface (API). We also improved the reliability and security of the underlying software by enhancing documentation, upgrading open-source software, and replacing modules to eliminate dependence on aging third-party software. We increased compatibility with evolving standards by developing, testing, and documenting alternative digitization protocols for multipage items. We offered the research community insight into the digital library underlying the *Profiles in Science* Web sites by publishing "Improving Software Sustainability: Lessons Learned from Profiles in Science" in the proceeding of the Society for Imaging Science & Technology Archiving 2013 Conference, April 2-5, 2013, and "Why Can't You Just Build It and Leave It Alone?" a guest post for the Library of Congress digital preservation blog, The Signal, on June 5, 2013.

Evidence-Based Medicine: PubMed for Handhelds (PubMed4Hh)

Developed and released in FY2003, PubMed for Handhelds (PubMed4Hh) facilitates evidence-based medical practice with MEDLINE access from almost anywhere via smartphones, wireless tablet devices, netbooks, and portable laptops. PubMed4Hh requires no proprietary software and reformats the screen display as appropriate for the wireless handheld device being used. Clinical filters feature easy access to relevant clinical literature.

Newly developed resources allow people to search MEDLINE through text-messaging. An algorithm to derive "the bottom line" (TBL) of published abstracts allows clinicians to quickly read summaries from almost anywhere. For example, it enables doctors to quickly consult the findings from research to help determine the best course of treatment for a patient. A "consensus abstracts" element provides rapid review of multiple publications with smartphones. A recent review of PubMed4Hh server logs showed that more than 90 percent of queries were clinical in nature.

To evaluate the usefulness of abstracts in clinical decision-making, randomized controlled trials using simulated clinical scenarios were conducted by the Uniformed Services University of the Health Sciences, the Botswana−University of Pennsylvania partnership, and the National Telehealth Center and Philippine General Hospital, Manila. Using simulated clinical cases, these studies demonstrated the usefulness of the app for clinical decision-making.

The PubMed4Hh app is available for iOS (iPhone/iPad) and Android users. Since its release in September 2012, the iOS app has been downloaded more than 138,000 times, by users in the U.S. (44 percent of downloads) and elsewhere (56 percent). Queries from smartphone apps now account for 60 percent of all queries. The total number of searches has tripled since before the introduction of the smartphone app.

Clinical Vocabulary Standards and Associated Tools

Many of our projects in this area continue to promote the development, enhancement**,** and adoption of clinical vocabulary standards. During FY2013, we provided comments and assistance to ONC, Centers for Medicare and Medicaid Services (CMS), and FDA, often on extremely short deadlines, related to health IT clinical vocabulary and messaging standards for quality measures and meaningful-use regulations. We reviewed several hundred quality measures proposed for phase three meaningful-use regulations and gave high-level feedback to CMS related to their terminology, clinical validity, and compatibility with clinical workflows. We also reviewed and provided comments about:

- Electronic messaging in response to draft implementation guides issued by the standards development organization Health Level Seven, Inc. (HL7) and
- Templates for reporting the results of laboratory testing for cancer biomarkers generated by the College of American Pathology (CAP).

The CORE Problem List Subset of SNOMED CT

SNOMED CT is a comprehensive, multilingual medical terminology for anatomic sites, organisms, chemicals, diagnoses, symptoms, findings, and other such concepts. The problem list - a patient's list of active conditions and symptoms - is an essential part of the electronic health record (EHR). Meaningful-use regulations from CMS require the use of SNOMED CT to code the problem list and many other EHR fields.

We analyzed problem list vocabularies and their usage frequencies in seven large-scale U.S. and overseas healthcare institutions, identified a subset of the most frequently used problem list terms in SNOMED CT, and then 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 linkage by medical records systems and other outside sources to NLM's rich consumer sources of medical information, has mapped the CORE Subset to MedlinePlus health topics so that medical records systems could automatically pull this educational material for patients to use. In 2012, the CORE Subset was enriched with a clinical dataset from the U.S. Department of Veterans Affairs covering three million patients.

The CORE Problem List Subset of SNOMED CT currently contains about 6,000 concepts and is published in the UMLS as a specific content view. The CORE Subset is updated four times a year to synchronize with changes in SNOMED CT and the UMLS.

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 ninth version “Clinical Modification”) is also used for reimbursement (soon to be replaced by ICD-10-*CM* in October 2014). 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 for that purpose, 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, we have developed various maps between SNOMED CT and the ICD classifications. We published a SNOMED CT to ICD-10-*CM* rule-based map, covering 35,000 SNOMED CT concepts. This map allows users to encode patient problems in SNOMED CT terms, and then to 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 assisted in the development of an analogous rule-based map from SNOMED CT to ICD-10. We adapted our 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 two more maps to SNOMED CT from 9,000 commonly used ICD-9-*CM* diagnostic codes, and 3,000 ICD-9-*CM* procedure codes.

We are currently investigating the need to create maps between SNOMED and 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, 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. 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’re continuing to align data elements between RxTerms and RxNorm, and we’re reviewing the dose-form information in RxTerms to improve usability. We update RxTerms monthly and bundle it with the full release of RxNorm.

RxNav

When we released RxNav in September 2004, it was a graphical interface browser to the RxNorm database and was primarily designed for displaying relations among drug entities. During FY2013, we expanded RxNav to additional drug information sources: drug classes from the Medical Subject Headings (MeSH) and the Anatomical Therapeutic Chemical (ATC) classification of drugs (developed by the World Health Organization (WHO) and widely used in Europe).

In addition to the RxNav browser, we created SOAP-based and RESTful application programming interfaces (APIs) so system developers and researchers can easily 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.

We released RxMix in FY2013, a graphical interface allowing users to create complex queries to drug information sources (that is, complex sequences of API functions) and to execute them on single values or on a list of values, in batches. For example, users could generate a list of statin drugs, or drugs containing a particular ingredient (such as penicillin), by identifying drugs in RxNorm that have a given property in NDF-RT. This ability enables applications and electronic medical records to create lists that can be used to help automate critical tasks, such as avoiding prescribing medications to which a patient has known allergies and other clinical decision support.

We have integrated two other drug information sources into RxNav: RxTerms, an interface terminology for prescription writing or medication history recording, and NDF-RT (National Drug File - Reference Terminology (NDF-RT) is produced by the U.S. Department of Veterans Affairs), 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 (Simple Object Access Protocol XML-based information exchange protocol) and RESTful (Representational State Transfer system architecture) APIs for RxNorm, RxTerms, and NDF-RT, received a combined total of about 100 million queries during FY2013 – this was double the number in 2012! In FY2013, we presented these APIs at the HHS Health Data Initiative. Users include clinical and academic institutions, 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

Within medical record systems, patient summaries, and reports to public health organizations, Federal Meaningful Use (MU) 2 EHR regulations require lab result messages sent to ordering clinicians to use LOINC (Logical Observation Identifiers, Names, and Codes). In FY2013, we 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 FY2013, LOINC had more than 26,000 users in 157 countries and had been translated into 12 languages, including a new Russian translation. 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, and by the end of FY2013, the Web browser was completely understandable in 12 languages besides English. The Regenstrief Institute and the International Health Terminology Standards Development Organisation (IHTSDO) signed a long-term agreement to begin collaborative work on linking their leading global healthcare terminologies: LOINC and SNOMED Clinical Terms.

We worked with Regenstrief and the LOINC Committee to create more than 1,700 new LOINC terms for both laboratory and clinical variables, and the LOINC database now contains more than 72,000 terms. We released new terms for radiology, toxicology, chemistry, hematology, molecular pathology, antibiotic susceptibility, the 2003 version of U.S. Standard Birth Certificate and Fetal Death Report panels, newborn hearing and critical congenital heart disease (CCHD) screening, the Neonatal Skin Risk Assessment Scale (NSRAS), and other new survey instruments. 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.

To facilitate electronic reporting of lab results, 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. 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 provided technical advice to the CMS Office of Office of Clinical Standards and Quality on creating standardized variables and data elements for all CMS data-collection tools for post-acute-care assessment.

We continued to meet with other NIH organizations that are developing assessment instruments and registry system values with the goal of closer alignment among NIH standard element development efforts. We are collaborating with other NIH organizations (and the 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 serve on the Common Data Elements (CDE) Working Group to the trans-NIH BioMedical Informatics Coordinating (BMIC) Committee. We're also working with colleagues at the:

- National Eye Institute (NEI) to restructure its packages of assessment instruments for the National Ophthalmic Disease Genotyping Network (eyeGENE®),
- National Center for Advancing Translational Sciences (NIH/NCATS) Office of Rare Disease to revise the common data elements (CDEs) for their registry system for rare diseases — and we plan to create corresponding LOINC codes,
- National Heart, Lung, and Blood Institute (NHLBI), NICHD, and NIH/NCATS on the NHLBI Hemoglobinopathies Uniform Medical Language Ontology (HUMLO) project, to develop CDEs for hemoglobinopathies using standard terminologies such as LOINC for the questions and SNOMED CT for the answers, and
- National Institute of Neurological Diseases and Stroke (NINDS) on CDEs and the neurological quality-of-life (Neuro-QOL) measures.

Newborn Screening Coding and Terminology Guide

We've collaborated with the many federal, state, and other agencies to standardize 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 NHLBI. We 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.

During FY2013, we worked closely with many states and vendors on their implementation of newborn screening LOINC and SNOMED CT coding and HL7 messaging standards. We revised many LOINC terms and created several new terms based on requests from NBS programs and laboratories. We created a comprehensive LOINC panel to report screening results for critical congenital heart disease (CCHD), the latest condition added to the HHS Secretary's Recommended Uniform Screening Panel for NBS. We expanded beyond NBS and worked with many states to create a panel of terms for reporting the results of therapeutic diet monitoring for patients with conditions that were diagnosed based on NBS. We also participated as a technical advisor about terminology and interoperability standards for HRSA's CCHD-pilot-program grantees and represent NLM on the Newborn Screening Technical assistance and Evaluation Program (NewSTEPs) Steering Committee and Health Information Technology workgroup.

Communication Infrastructure Research and Tools

We perform and support research to develop and advance infrastructure capabilities such as high-speed networks, nomadic computing, network management, wireless access, security, and privacy.

Videoconferencing and Collaboration

We continue 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 continued monitor the high-definition (HD) open-source work of video conferencing tool (VIC) developers on H.264 compression.

Our team also investigated some new cloud collaboration tools, including proprietary ones that are standards compliant and that emulate a pioneering collaboration model developed by Argonne National Laboratory. Some commercially available cloud technologies lack features required at NLM, are either too complicated or unstable, or have licensing terms that are not cost-effective given the needs of the Library. Currently, there seems to be no compelling rationale to migrate from the H.323 standard videoconferencing appliances we use for NLM program support. The team continued to collaborate with the Rochester Institute of Technology (RIT) and the University of Puerto Rico Medical Campus, however, to test open-source software and commercial cloud technologies for compressed HD videoconferencing based on the H.264 video standard.

We decided to work exclusively with UltraGrid technology for video conferencing because:

- The technology has been enhanced since we started our collaboration with UltraGrid's developers at Masaryk University in Brno, Czech Republic, last year,
- Other uncompressed technologies are receiving less institutional support, and
- The research team at Masaryk is very active and shares our research goals and interests.

The team installed an uncompressed video system at the Medical University of South Carolina (MUSC) for a clinical trial. Given the laboratory focus on UltraGrid, iHDTV technology was moved to the clinical setting. The trial's purpose is to study uncompressed video's use as a diagnostic tool. Investigators selected teledermatology as a research focus because previous research has shown it to be particularly difficult to use standard definition video to perform remote dermatological exams. The study was delayed in 2012 because of the need to obtain IRB and Office of Management and Budget (OMB) clearances, the death of the principal investigator collaborator, and the subsequent incapacitating illness of her successor. A new PI was appointed, and the research is half-way through the data-collection phase.

We continued to work with Specialized Information Services (SIS) on a distance-education outreach program for minority high school students and with the NIH Library to offer National Center for Biotechnology Information (NCBI) database and other bioinformatics training at a distance. In FY2013, we conducted bioinformatics programs with the Charles R. Drew University of Medicine and Science, the University of Maryland, the University of North Carolina at Chapel Hill, the University of Tennessee at Memphis, the University of Puerto Rico Medical Campus, and Virginia Commonwealth University. One program, involving three of the participating institutions, used cloud-computing resources that reduced genomic-analysis overnight computational times to a matter of minutes. The simultaneous use of the technology by so many students in computer labs at different sites stressed the cloud architecture, but proved feasible. Because travel funds were cut, the team also supported several successful virtual site visits with Regional Medical Libraries to review programs funded under NLM's Regional Medical Library Program. The team published a comparison of the uncompressed video technologies and a re-analysis of findings from previous telemedicine research reviews that identified conditions affecting telemedicine-intervention success.

Collaboratory for High Performance Computing and Communication

LHNCBC established the 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 and distance-learning research, both within LHNCBC and outside NLM.

The facility can be configured to such technologies as 3D interactive imaging (with stereoscopic projection), haptics for surgical planning and distance education, and interactive imaging and communications protocols applicable to telemedicine and distance education involving interactive video and application-sharing tools. These protocols enable staff to collaborate with others at a distance and, at the same time, demonstrate much of the internal and external work being done as part of the NLM Visible Human and advanced networking initiatives. The collaboration technologies include tools built around the H.323 and MPEG video compression standards for transmitting video over internet IP, open-source technologies such as Conference-XP, iHDTV, UltraGrid, and the Access Grid, as well as cloud technologies and mobile apps.

In addition to supporting research, the Collab is a vehicle for offering the SIS distance-learning program, biotechnology training, and virtual site visits. We also developed a beta-version search engine, still in testing mode, for the Collab Web server.

Disaster Information Management: Lost Person Finder

NLM's increasing interest in recent years in mitigating the effects of wide-area disasters has led us to develop several information resources and tools. In our Lost Person Finder (LPF) project, we address the problem of how to reunite families separated by mass casualty events. 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 is People Locator® (PL), the main LPF Web site and its MySQL database. We extensively customized the open-source Sahana disaster-management system to create a unified site to hold data from multiple disasters. Missing or found people may be reported to PL by hospital counselors, relief workers, or the public. PL users can report or search for the missing via computer or through mobile apps using Web services. PL is now running on enterprise-level, load-balanced dual systems with failover (to ensure continued availability and operations if one system fails), independent uptime monitoring (to track server availability and performance), and an indexing engine (powered by SOLR software) that can retrieve records quickly.

During 2013, we created a separate demonstration Web site (https://TriageTrak.nlm.nih.gov) so hospitals can report directly or via TriagePic, a reporting app. We designed and implemented new features to improve the chances that search engines will find our site.

In FY 2013, we also:

- Created a prototype interface for visual search (query by face image) that allows the use of optional metadata (such as gender and age) to reduce the search space, thereby increasing search speed;
- Added a feature to allow reporting missing persons with multiple photos, enabling the PL to incorporate certain face-matching project researcher datasets; and
- Added better notification and alerting functions.

In addition to technical improvements, in FY2013 we also secured a formal Authority to Operate (ATO) as required by NIH, and we obtained registered trademarks for People Locator, ReUnite, and TriagePic.

We collaborated with many organizations and agencies interested in disaster mitigation by participating in a FEMA-led Webinar and the Missing Persons Community of Interest (MPCI). We also participated in the Google Summer of Code and Google Code-in international student programs, in cooperation with the Sahana Software

Foundation, by giving university students all over the world project-related opportunities to learn about and contribute to open-source coding and other tasks.

Deployments

People Locator has been deployed in disasters since the Haiti Earthquake in 2010, as well as in demonstrations and large-scale multi-institutional drills with local Bethesda hospitals. In 2013, PL was deployed for the Boston Marathon Bombing (March), Sichuan Earthquake (April), Uttarakhand India Flooding (June), Acapulco Flooding (September), and Super Typhoon Haiyan (November). In the case of Typhoon Haiyan in the Philippines, more than 100,000 missing-person records were posted to PL (many through Google's Person Finder which is interoperable with PL). Of these records, about 20,000 had photos of missing people.

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 and triage status and minimal descriptive metadata (e.g., name, age range, gender) are packaged and sent by Web services to PL.

In 2013 we moved TriagePic from a laptop-hosted system to both iOS and Android platforms, adding AES 128 encryption of patient text data and photos to these apps. In addition, we investigated approaches to design a version of ReUnite that supports multiple screen sizes — from cell phone to tablet, from 4-inch screen to 10-inch — to better support the many new mobile devices as the market rapidly expands.

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 work faces special challenges: unlike many other systems, our face-matching needs to rely on a *single* photo of a person to identify her or his face in other images, so we can't exploit traditional face-recognition models that require large sets of photos to train the system.

In 2013, we made considerable progress in face localization, a key first step in face matching. Traditional face-localization methods fail by either missing a face or finding too many false positives— that is, objects that the algorithm misidentifies as a face. We minimized these errors and made the algorithm more robust by introducing the simple idea that a face will have skin color. A custom skin-color space was developed using Neural Network–based learners. All pixels in the image that are considered "skin-like" are flagged as potential skin zones, and these zones are contrast enhanced. On reprocessing the image we find fewer false-positive and false-negative errors. In comparison tests with state-of-the-art commercial software, our algorithm performs as well as or better than other methods, not only on high-quality images, but also on the lower-quality images typically sent from disaster sites. In addition, with a view to speeding up face localization, which demands a lot of processing power, we have adapted our algorithm for computation using graphical processing units (GPUs) that offer up to 20X speed up over a CPU alone. The method is currently undergoing testing, and is being readied for deployment in PL.

In addition, we improved face-matching methods by studying several state-of-the-art similarity metrics and implementing two that used scale- and rotation-invariant measures. Further, we developed a novel rotation- and scale-invariant line-based color descriptor. These methods not only improve the system's face-matching accuracy, but also make it robust to variation in illumination. The system was deployed on our staging server and is undergoing testing, and we are working to improve search times.

In an effort to make the face matching robust to variations due to illumination, age, and injury, we're developing face-*region*-based matching algorithms. One that we've developed extracts and matches localized eyes and mouth regions. It performs at 94 percent accuracy but is sensitive to the size of the image. Further improvements to this method are under way.

To support research and testing, we need annotated images. For this purpose, we developed machine-learning tools such as ImageStats to gather objective and accurate "ground truth" data. In 2013, in collaboration with the Sahana Software Foundation, student volunteers from around the world, recruited through the Google Code-in contest, improved ImageStats by contributing crowd-sourced software.

Video Production, Retrieval, and Reuse Project

This development area encompasses projects that contribute to the NLM Long Range Plan goal of promoting health literacy and increasing biomedical understanding. The NLM Media Assets Project gives the NLM easy access to audio-video resources for improved biomedical communications.

Movement Disorders Video Database/MDmedia

This ongoing LHNCBC Research Support Project contributes to improving access to high-quality biomedical-imaging information. We're now able to include video in the clinical study of patients with movement disorders and research the role of mobile technologies in the management of Parkinson's disease by patients and their caregivers.

In FY2013, in collaboration with the National Institute of Neurological Disorders and Stroke and the NIH Movement Disorders Clinic, we developed prototype mobile applications (for iPad and Samsung Galaxy), based on the Movement Disorders Video Database (MDVD), to help diagnose and predict the course of movement disorders. On the basis of input from both clinical and patient-advocacy groups, we also developed two prototype apps for research purposes to measure two functional abilities whose impairment can give insight into the severity of movement disorders: finger-tapping speed and accuracy, and rapid "saccades" eye movement.

A new initiative emerged from our collaboration with clinical and patient groups, to explore the value of video and voice data in mobile devices for patients suffering from movement disorders. Phase one of this research effort focused on interviews with patients attending the NIH Clinical Center, soliciting their feedback on how mobile technologies could help them better manage their care between clinician visits. The results of these interviews will be a list of mobile-platform capabilities and associated expected outcomes that patients and clinicians identify as enabling them to manage their care better and improve their overall health. So far, the interviews have suggested that:

- Video and voice records can be valuable.
- Patients could easily use cell phones to share video and voice data with their doctors and have the data included as part of their medical record.
- There is great potential for the use of video and voice journals to monitor and graph symptoms over time.
- Touch-screen technology is more accessible than PC or mouse technology for patients with movement disorders.

Based on these interviews, we will identify viable technologies, then develop and validate these technologies in future studies.

MedlinePlus Interactive Tutorials

In keeping with the LHNCBC mission to develop new technologies and core resources for disseminating biomedical information, we've been engaged in the research and development of augmented-reality modeling and applications, mobile apps, ultra-high-definition imaging research, image-rich Web sites, and audio/video/imaging archives.

In FY2013, with the NLM MedlinePlus team, we developed a design and navigation prototype for new MedlinePlus Interactive Tutorials to enhance their usefulness. To enable broader public distribution, and allow for an equal experience on all operating systems and devices, we created tutorials with HTML5-compliant animation and interactivity.

Native Voice Mobile Exhibition Planning, Development, and Deployment

In FY2013, NLM published two major LHNCBC-produced Native Voices iPad app updates. They feature improved, enhanced interactive designs and 34 new video clips from interviews conducted by the NLM Director with 15 additional tribal representatives, plus high-resolution images of Native objects and artifacts illustrating the connectedness of wellness and Native culture.

The pilot version of the Native Voices Mobile Adaptation traveling exhibit opened at the Spirit Lake Nation in Fort Totten, ND, in October 2013. APDB provided on-site support and location training to the faculty and staff at the Cankdeska Cikana Community College (CCCC).

In collaboration with the NLM Office of the Director, the Office of Health Information Programs Development, the Office of Communications and Public Liaison, Specialized Information Services, and the History of Medicine Division, we developed pilot-testing survey measures to collect feedback from the CCCC installation. The pilot-test results will be incorporated into the future planning, development, and deployment of the traveling exhibit.

Computing Resources Projects

We conduct numerous projects to 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, as well as system administration support for a large number of individual workstations and shared servers.

In FY2013, as part of Federal Green Information Technology (IT) initiatives, we oversaw continuous monitoring of information systems, and we updated hardware and software to comply with Federal Information Security Management Act (FISMA) 2.0 requirements and NIH guidelines. We achieved a 90 percent FISMA compliance rate (10 percent more than NIH requirements and FY2012 baseline). We reduced purchases of personal IT equipment, purchased and deployed network printers as shared staff resources, and cut locally used device purchases by 80 percent. We reviewed all IT purchase requests to verify equipment was needed for research or administration. To avoid interruptions of service, as part of our disaster planning and Continuity of Operations (COOP) risk management efforts, we created a new contact list of Subject Matter Experts (SMEs), continuous monitoring processes, detailed recovery procedures, and centralized data storage. We also upgraded the DMZ (demilitarized zone) perimeter network— a system configuration that provides an additional security buffer — to 10-gigabit capacity with revised Internet Protocol version 6 (IPv6) features (which increase efficiency and network security), and successfully implemented these upgrades without downtime impacts on public service.

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 including multicast video conferencing and IPv6. Another core project implements a network monitoring system that displays network usages in real time. The network management team also participated in the NIH-wide study of Trusted Internet Connection (TIC) consolidation and evaluated the impacts to the NIH and NLM.

The security management team incorporates security operations into firewall administration, patch management, anti-virus management, intrusion monitoring, security and vulnerability scanning, and vulnerability remediation to ensure 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 systems 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 NIH penetration test in FY2013 confirmed the effectiveness of LHNCBC's security practices.

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 Domain Name System (DNS), Network Information Service (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 ensure system availability and performance during emergencies such as 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 postdoctoral fellows and other people beginning their biomedical research careers. Our Medical Informatics Training Program (MITP), ranging from a few months to two years or more, is available for visiting scientists, postdoctoral fellows, and graduate and medical students. Each participant spends between a few months and several years working on LHNCBC research projects under the guidance of 16 LHNCBC research staff mentors. The FAQs include information about the lecture series that is part of this program. Participants also make presentations, write manuscripts, attend and present at professional conferences, and may publish in professional journals.

During FY2013, 39 trainees (including 13 postdoctoral fellows) from 15 states and seven foreign countries received training and conducted research at LHNCBC in a wide range of disciplines including:

- 3D image processing;
- Big data to knowledge (BD2K) based on large biomedical datasets;
- Biomedical ontology and terminology;
- Clinical Health Information Question Answering systems;
- Content-based information retrieval;
- De-identification of medical records;
- Evidence-based medicine systems;
- mHealth-, image-, text-, and document-processing;
- Information retrieval;
- Literature-based discovery;
- Natural language processing;
- Personal health records;
- Pill identification;
- Collaboration tools;
- Semantic Web research; and
- Disaster management.

We emphasize diversity by participating 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. In FY2013, two students from the Washington Internship for Native Students – both from the Apache nation in Arizona – worked on projects with LHNCBC mentors.

The MITP 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. We continue to offer an NIH Clinical Elective Program in Medical Informatics for third- and fourth-year medical and dental students, which offers students the opportunity for independent research under the mentorship of expert NIH researchers. We also host a two-month NLM Rotation Program that 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. We also provide an informatics lecture series and submit project proposals for the NLM Library Associates program.