LISTER HILL NATIONAL CENTER
FOR BIOMEDICAL COMMUNICATIONS

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

TECHNICAL REPORT
LHNCBC-TR-2010-003

The Lister Hill National Center for Biomedical Communications Annual Report FY2010

Clement J. McDonald, M.D.
Director

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

LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS
FY2010 ANNUAL REPORT

Clement J. McDonald, MD
Director

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

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

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

An external Board of Scientific Counselors meets semi-annually to review the Center's research projects and priorities. The most current information about the Lister Hill Center research activities can be found at http://lhncbc.nlm.nih.gov/.

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

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

NLM Personal Health Record

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

The NLM PHR supports the entry and tracking of key measurements and test results, prescriptions, problems, and immunizations, future health appointments. It will produce digital and paper copies of its contents in various formats. Users can get access to MedlinePlus information resources by clicking the icon adjacent to the name of any prescription drug, medical condition, or surgery that they enter into the system. The PHR automatically assigns codes to the medications, observations, and problems as users enter them and utilizes NLM-supported terminologies that HHS recognizes as national standards. By using vocabulary standards and coding, the PHR can "understand" its content and provide numerous benefits such as computer-generated personalized reminders about preventive care or healthy behaviors and automatic calculations based on other values in the PHR (e.g. calculate body mass index based on height and weight entered in the same panel). The PHR provides direct links from patient-entered conditions, drugs, or surgeries to MedlinePlus and other respected consumer information resources. In FY2010, researchers continued to expand and improve the capabilities of the PHR. Developers gave users the ability to set the number of days before their medications expire, to initiate a mail order refill, or to obtain follow-up

testing. The PHR provides easy ways to generate input forms for any kind of health information that the patient wants to track

Researchers reviewed and enhanced the controlled vocabulary for more than 2,000 condition names and synonyms and more than 300 surgery procedure names and synonyms by enriching the synonymy, providing the consumer-friendly name when feasible, and adding SNOMED codes, when available, to these items.

Developers completed a major revision of the decision rules authoring system. In the first version, authors needed to use a syntax that required programming skill. In the latest version, users pick from a series of menus and input fields on a Web form. To build a base rule, the author identifies the PHR table of interest, e.g. medications, conditions, allergies, etc., and then identifies the subject and the predicate of the rule. A rule author can identify a subject of the rule by name, or class. For example, if the users picked medications as their table of interest, they could identify the subject by name, e.g. Simvastatin oral, or by class, e.g. cholesterol lowering drug, and then add criteria such as status = active, or date started was before January 1, 2009. Users could define another rule to determine whether a patient's last LDL cholesterol was greater than 130 mg/dl during the last year. These two rules could be combined into a third rule for reminding PHR record holders to ask the doctor about the use of diet and/or a cholesterol lowering agent if their LDL cholesterol was elevated and their medication list did not contain any active cholesterol lowering agents. As research progresses, the rules will support increasing complexity. Researchers than completely rewrote all of the existing rules, and added a substantial number of new rules using this new system.

Developers added a number of other capabilities to the PHR including:

1. pop-up forms for capturing information from almost any questionnaire or laboratory panel. The only prerequisite is that the information that defines the questionnaire (or lab panel) must be in the PHR's master knowledge database and follow the LOINC form structure. Currently the PHR has access to more than 500 such "questionnaires."
2. a new data table driven rule system that supports reminder rules.
3. new tables for identifying medication content by ingredient and a system to warn PHR users when they enter a prescription that contains the same ingredient as a previously entered one.

The LHNCBC continues to work with NLM and others to ensure that policies concerning the PHR are developed and in place. This young project addresses the longstanding NLM interest of facilitating health care management and is closely aligned with the NLM strategic plan. It will help refine the message and vocabulary standards that NLM has supported and also provide another consumer entry point to a rich trove of patient-oriented data.

Use of Surescripts Prescription Data in Direct Patient Care

An important part of patient assessment in the Emergency Department (ED), medication history can have significant impact on the diagnosis and treatment of a patient's problems. However, manually-acquired medication histories are prone to inaccuracies. An electronic prescription history, which can be obtained through sources like Surescripts, a consortium of major Pharmacy Benefit Managers (PBM) and the largest e-prescribing network in the U.S., has data that could improve these histories. Surescripts handles over two billion prescriptions per year and can provide prescription benefit and history information for an estimated 65 percent of all prescriptions paid for by private insurance in the United States.

Partially-funded by the Bethesda Hospitals Emergency Preparedness Partnership (National Naval Medical Center, NIH Clinical Center, and Suburban Hospital), this project established a connection between Suburban Hospital and Surescripts to allow real-time retrieval of a patient's prescription history, and studied the feasibility and value of using prescription history information from Surescripts in direct patient care. This external source of medication information that can be obtained automatically could be both time- and life-saving in disaster circumstances, when the normal route of obtaining patients' medication history is likely to be disrupted due to over-strained medical staff or special patient circumstances (e.g. unconscious patients).

As a quality assurance process, we compared the history obtained by the ED nurse to Surescripts information for all ED patients for three months. We found that neither source was complete by itself (nearly 20 percent of all potentially current medications were not captured in the manual medication history and 25 percent of the medications were missing from Surescripts data); however, Surescripts does add substantial information to the medication histories collected by the ED personnel.

As part of this project, the developers built a system that connects the Emergency Department (ED) of Suburban Hospital to Surescripts data center. The system 1) receives ADT messages generated by the HL7 engine of Suburban Hospital; 2) filters in five essential key information about the patient (first and last names, gender, date of birth, and zip code); 3) sends them to Surescripts; 4) receives prescription history of the patient if the patient's record is found in Surescripts databases, and; 5) maps the information to the patient's record in the project database. All messages are sent and received in HL7 format. The system was installed at the data center of Suburban Hospital in the first quarter of 2009 and has been in operation since then. Finally, Suburban Hospital conducted a survey about the use of the Surescripts medication history reports and shared their survey results with us. The respondents unanimously welcomed this medication information.

Suburban Hospital changed their HL7 engine in 2010, which affected all HL7 transactions, including the feed that we use to exchange data with Surescripts as well as to collect de-identified research data to study the value of Surescripts' medication history reports. The developers completely redesigned the HL7 interface, the associated relational database schema, and the hardware to represent and capture the new data stream seamlessly.

Surescripts transmits a raw medication history report for each patient that lists every prescription and refill, each on a separate line, that providers find difficult to review and digest. We have developed a pilot report that we believe will be easier to comprehend. In our new approach to reporting, the raw Surescripts data is reprocessed, clustered by medication names, sorted by date, associated with dispensed drug duration information (when available), and at the end, a one-year-prescription history of the patient is plotted on a timeline graph. In the latest iteration of our design, we can also convey pharmacy and prescriber information (when available) and plot each drug dispensation separately so that care providers, with the help of their patients, can distinguish the actual data points from the noise.

EMR Database Research and Development

LHNCBC developed a general purpose longitudinal database structure to investigate secondary use of data collected in electronic medical records and shared this structure with the owners of the MIMIC II database as well as the AMIA-NLP working group. Both groups will use our schema for design and development of their repositories of clinical data.

This year LHNCBC acquired a new update of the MIMIC II EMR dataset. We use this de-identified dataset under a restricted-use Memorandum of Understanding (MOU) to populate our database. The current version of this EMR dataset consists of clinical data for over 32,000 ICU encounters and more than 26,000 patients. It addition to over 11,000,000 laboratory results and over 400,000 clinical notes, the dataset now contains information on a patient's death at any time after the patient's discharge. Altogether, the dataset carries over 200,000,000 discrete observations. The current version of the data contains the Simplified Acute Physiology Score (SAPS) designed to measure the severity of disease and predict mortality for patients admitted to intensive care units. The initial scores were developed by the MIT research team. In the process of replicating the scores, we found additional variables that could be used to compute the scores for almost all patients and reconciled discrepancies with the MIT team. We also determined that some of the extreme values for the variables contributing to SAPS are correct for a given patient and others are device or data entry errors. We are developing methods for quality assurance of data using structured information and clinical notes.

Biomedical Imaging, Multimedia, and 3D Imaging

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

Imaging Tools for Biomedical Research

In FY2010, the LHNCBC worked with the National Cancer Institute (NCI) and the American Society for Colposcopy and Cervical Pathology (ASCCP) to put one of our imaging programs, the Teaching Tool, into operational use for the assessment of professional knowledge and skills in the field of colposcopy. The Teaching

Tool is being used by 52 resident programs in Ob/Gyn and Family Practice, and over 200 individual online exams have been given. In March 2010 the ASCCP acknowledged our work by giving the organization's Award of Merit to two members of the Teaching Tool development team.

NCI deployed another of our imaging programs, the Boundary Marking Tool, at sites including the University of Oklahoma Health Sciences Center and sites in Guanacaste, Costa Rica, and Irun, Nigeria to collect and annotate colposcopy images for the creation of a worldwide database for cervix research. We also began providing public access to the Boundary Marking Tool code under a modified Berkeley open source license, through a link on the CEB Web page. We continued development on both of these tools to improve functionality and usability; part of this effort included carrying out a formal usability study for the Boundary Marking Tool with participants at the biennial meeting of the ASCCP in Las Vegas in March 2010. We also have continued development of the Multimedia Database Tool by adding several thousand graphical annotations of cervix images to the two databases accessible through this tool. Staff are currently working with NCI to add pH measurement data to these databases to support NCI study of the correlation between vaginal microflora and HPV infection. We have begun collaboration with a new group of NCI-sponsored researchers at the University of New Mexico and are working toward expanding the capabilities of the CEB Virtual Microscope tool to support a study of histology diagnoses of the cervix in glass slides versus digital images. If this study indicates positive results for digital microscopy research in this field, we anticipate supporting further NCI histology studies with this tool. Also this year, to benefit our NCI collaborators, we created a new graphing capability to provide a compact visual representation of the spatial distribution of pathology across the surface of the cervix.

In September 2010, along with representatives of NCI and the ASCCP, we presented our work with the Teaching Tool to the NLM Board of Regents.

Content Based Image Retrieval (CBIR)

CBIR is an active research area at the Lister Hill Center with several objectives. One is to develop techniques to introduce automation into our existing cancer research tools. For example, our CervigramFinder automatically indexes and allows retrieval of cervigrams using shape, color and texture features. This system therefore contains the key elements needed to augment the Boundary Marking Tool with an automated assist for the user in marking boundaries of regions of medical significance. Researchers evaluated the CervigramFinder for usability and acceptance at the biannual meeting of the American Society for Colposcopy and Cervical Pathology (ASCCP) in 2010, and developers are using the evaluation results are to improve the tool. CBIR research is supported by the Cervigram Segmentation Tool that combines several image segmentation and shape similarity algorithms. This tool enables researchers to select the optimal algorithms that automatically segment regions for image indexing. We also developed a third tool, MOSES, as a service for the evaluation of segmentations done by multiple experts, to reduce the inter- and intra-observer variability in marking boundaries of significant regions in cervigrams. Together, these three tools form a unified system for cervigram image analysis.

In addition, we developed a hybrid CBIR system for x-rays in the NHANES II collection by linking two geographically separated CBIR systems (SPIRS at the Lister Hill Center, and IRMA at the University of Aachen in Germany) to exploit the characteristics of each. SPIRS/IRMA allows a user to retrieve detailed medical image data in large collections of heterogeneous images of varying modalities, presentations and anatomy. For example, a user may narrow a search to *spinal x-rays* (using IRMA) and then retrieve specific x-rays containing *osteophytes* through SPIRS. In FY2010, we extended the capability of SPIRS with research into improved shape matching methods, resulting in more efficient shape matching as well as a better understanding of the optimal number of shape features needed. We implemented a new feature (distance across shape) and used a Support Vector Machine (SVM) classifier to reduce the *semantic gap* between diagnostic labels used to describe the pathology on the vertebrae and the shape characteristics extracted by feature extraction algorithms.

CBIR has been used to index illustrations in medical journals by using image features, in combination with text processing of figure captions and in-document text mentions. This research is aimed at enriching the user experience of searching for relevant documents by including the contents of medical images, photographs, graphs and other illustrations found in articles. Techniques developed in this work were evaluated in the international ImageCLEF competition in 2010 and found to be successful. Over 15 image features were implemented and used in an SVM-based framework to detect modality (x-ray, ultrasound, CT, MR, etc.) and to compute similarity. Our efforts were ranked second among 12 teams from around the world, many of which were from industrial R&D labs.

We also developed methods to describe images in a *bag-of-words* and a *bag-of-keypoints* representation analogous to those used in text-document retrieval. These were very successful in automatic coarse annotation of images. Research into improved automatic image annotation methods is ongoing.

We have explored the role of CBIR in extracting regions of interest (ROI) in images. One approach to identifying meaningful ROIs, and thereby annotating biomedical-article images, is by first extracting author-placed markups (or "pointers") such as arrows, asterisks, and alphanumeric characters. Novel methods were developed for each type of markup with over 87 percent accuracy in detecting arrows. Further research is ongoing**.**

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

Other avenues explored in this research area are distributed computing and use of GPUs for compute-intensive CBIR tasks, with a particular focus on image segmentation.

Research Toward Next Generation Scientific Publishing

Separate efforts are under way to explore techniques and technologies to create and use new and powerful forms of scientific publications that exploit the increasingly availability of various forms of multimedia. The goal of the first project, Interactive Publications Research (IPR), is to demonstrate a type of highly interactive multimedia-rich document that serves as a model for next-generation publishing in biomedicine. The research focuses on the standards, formats, authoring and reading tools necessary for the creation and use of such interactive publications (IP) containing media objects relevant to the biomedical literature: text, video, audio, bitmapped images, interactive tables and graphs, and clinical DICOM images such as x-rays, CT, MRI, and ultrasound.

The first tool LHNCBC engineers developed, *Panorama*, is for viewing and analyzing video, DICOM clinical images, tables, graphs and animations. Panorama, written in Java, was one of nine semi-finalists out of 70 entrants in Elsevier's Grand Challenge contest a year ago. In FY2010 we developed *Forge*, *a* tool for authors or publishers to create interactive publications. The Forge environment replicates the Panorama look and feel with the additional ability to edit and define the interactivity of each PDF annotation. Panorama is for viewing only and Forge is for both viewing and editing. These tools were recently demonstrated at the 2010 AMIA symposium and described in a publication (Thoma GR, et al. Interactive Publication: the Document as a Research Tool. *Web Semantics: Science, Services and Agents on the WWW*, 2010).

To support the future long term development of these Java tools in an open source environment, developers re-wrote the core code of Panorama and Forge to accommodate Eclipse plug-ins (http://marketplace.eclipse.org/) and thereby extend the functionality of these applications.

Taking advantage of our InfoBot system, we also enhanced the Panorama application to provide Annotation Concepts. This feature provides additional meaning and context to the article text. The text in an IP is sent to an NLM servlet for processing to identify Unified Medical Language system (UMLS) concepts. An XML file is returned to Panorama that is parsed and provides linkouts for Medline Plus, eMedline, Family Doctor, and also provides the preferred UMLS term and semantic group. Further work is ongoing to group concepts by semantic relationships, and other grouping ideas are being explored as well.

Engineers developed a prototype interactive publication and presented it to the National Center for Health Statistics (NCHS). This multimedia document contains Motion Charts using NHANES data from NCHS. Motion Charts is the Google widget version of Gapminder (created in Stockholm in 2005). NCHS is interested in creating IPs of their own but want to delay until they publish their conventional papers in the literature first.

We are also investigating several approaches for online access to interactive publications, since convenient and rapid access to these potentially large documents would be necessary in a practical deployment. In one approach, we modified Panorama to support Java's Web Start technology. Using this technology, standalone Java software applications can be deployed with a single click over the network. Java Web Start ensures that the most current version of the application will be deployed, as well as the correct version of the Java Runtime Environment (JRE). We are in discussion with a publisher to experiment with this and other approaches to provide access to IPs.

In a separate approach to new forms of publishing, LHNCBC is partnering with the Optical Society of America (OSA) to test the use of interactive publications in an operating journal. The goal is to evaluate software and database infrastructure that enables viewing and analysis of curated, supplemental biomedical source data

published in conjunction with peer-reviewed manuscripts, evaluate the educational value of such an infrastructure, and explore the problems of archiving this medium. In order to accomplish these goals, OSA has published four electronic special issues of OSA journals on research topics which lend themselves to interactive publishing. Articles published in these special issues are peer reviewed and fully citable as OSA journal publications indexed in Medline. They are published on the Web in Acrobat format (PDF) with links to source data, videos, and other media objects. The links allow users to quickly and conveniently download these objects and visualize them using interactive viewing software designed to look like an Acrobat plug-in. The viewing software is freely available as a download for all computer platforms. The journal articles and data sets are open access and the source data and associated metadata are searchable and accessible from outside the publication. The project will also serve as an NLM testbed for archiving this new form of publication. The most recent special issue focusing on *Image in Diagnosis and Treatment of Lung Cancer*, published in April 2010, is available at http://www.opticsinfobase.org/oe/virtual_issue.cfm?vid=105.

Screening for Tuberculosis in Rural Africa

The goal of this new project is to develop a biomedical image processing system for clinically screening HIV positive patients in rural Kenya for lung disease with a special focus on tuberculosis, which is rampant. This project leverages the image processing, analysis, and communication expertise at the LHNCBC, and aligns with NIH and NLM policy and strategic planning objectives in global and rural health. LHNCBC initiated this collaborative project with Academic Model Providing Access to Healthcare (AMPATH), the largest AIDS treatment program in the world (more than 100,000 AIDs patients under treatment) and part of a large USAID collaboration.

Chest radiography is important to the detection and proper treatment of tuberculosis and other pulmonary infections which are prevalent among HIV-positive patients. As part of the collaboration, AMPATH is acquiring a lightweight FDA-approved digital x-ray imaging system that can be transported in existing SUVs and/or trucks to remote regions in Kenya at regular intervals in order to screen the high risk individuals and bring them into treatment programs. Initially, LHNCBC will obtain an existing de-identified collection of 1,000 film-based chest x-rays from Kenya that have been photographed using a digital camera. These images will be used to train image processing algorithms and machine classifiers to develop a screening system for relevant pathology.

Once the equipment is deployed, we will routinely collect data about the usage and durability of the equipment, the number of x-rays produced, and the performance of the screening algorithm to determine the value of this light weight approach and its long-term viability of such systems in other parts of the world.

Medical Article Records Groundtruth (MARG)

To exploit the images and labeled data collected routinely from the MARS operation for research, a validated test set for document image analysis was created for the computer science and informatics communities for research into advanced algorithms for data mining. The MARG database consists of images of journal articles, the corresponding OCR data, zones, labels and verified data obtained from the routine operation of the MARS production system.

Remote Virtual Dialog System (RVDS)

The Remote Virtual Dialog System (RVDS) will make the NLM “Dialogues in Science” series, currently only available in the NLM Visitors Center, available anywhere through the Internet. The project involves the enhancement of programmatic capabilities of the virtual dialogue model to make it sustainable and to allow for expanded applications of the model.

Computational Photography Project for Pill Identification

As an exploratory program for content-based information retrieval to promote patient safety at a national level, the LHNCBC has undertaken a new project, *Computational Photography Project for Pill Identification* (C3PI), an authoritative, comprehensive, public digital image inventory of the nation’s commercial prescription solid dose medications. Heretofore, we have conducted most of this work in partnership with the NLM Specialized Information Systems Division and the U.S. Veterans Administration to study content-based retrieval methods for medical image

databases. NLM researchers have developed computer vision approaches for the automatic segmentation, measurement, and analysis of solid-dose medications from these pilot datasets including work on robust color classification tools to help identify prescription drugs. Working with an expert team from Medicos Consultants, we are creating a collection of digital photographs of prescription tablets and capsules, creating high resolution digital photographs of the front and back surfaces of pharmacy samples, confirming that the images match the description of the medication, developing and matching the images of the samples to relevant metadata (including size descriptions, dimensions, color, and the provenance of the sample).

Virtual Microscope (VM) and Virtual Slides

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

The Visible Human Project

The Visible Human Project image data sets are designed to serve as a common reference for the study of human anatomy, as a set of common public domain data for testing medical imaging algorithms, and as a testbed and model for the construction of image libraries that can be accessed through networks. The Visible Human data sets are available through a free license agreement with the NLM. They are distributed in their original format or in .png format to licensees over the Internet at no cost; and on DVD discs for a duplication fee. Almost 3,200 licensees in 61 countries are applying the data sets to a wide range of educational, diagnostic, treatment planning, virtual reality, and virtual surgeries, in addition to artistic, mathematical, legal, and industrial uses. More than 1,000 newspaper or magazine articles or radio programs have featured the Visible Human Project.

In FY2010, we continued to maintain two databases to record information about Visible Human Project use. The first, to log information about the license holders and record statements of their intended use of the images; and the second, to record information about the products the licensees are providing NLM in compliance with the Visible Human Dataset License Agreement.

While designing a database of the parameters and variances defining the normal range of human anatomical structures and the dependencies and covariances between them, an attempt was made to glean the needed statistical data on bone size variation from the existing anatomical literature. Over 1,000 references were scanned. The literature was found to contain a description and images of the variation in shape that could be expected of each bone, but not the mathematical description of it variances.

In FY2011, the LHNCBC will hold a cross-disciplinary workshop made up of key representatives of the disciplines that are relevant to bone size and shape in an effort to find a source for the needed statistical data. These disciplines will include forensic pathology and other forensic sciences, paleontology, radiology, orthopedics, anatomists, prosthesis designers, clothing designers and furniture designers. By inviting the participation of the interested community combined with using Web tools developed under the Interactive Publication Project, developers expect to collect the needed data.

3D Informatics

In FY2010, the 3D Informatics Program (TDI) expanded research efforts concerning problems encountered in the world of three-dimensional and higher-dimensional, time-varying imaging. The LHNCBC provides continuing support for image databases, continues to explore the growing need for image databases, including ongoing support

for the National Online Volumetric Archive (NOVA), an archive a collection of volume image data. This collection contains 3D data from across medicine. The TDI group began migration of these data to the newly installed infrastructure based on the MIDAS software system, the same systems supported as part of an NLM project in Interactive Scientific Publication. Contributors to the collection include the Mayo Clinic Biomedical Imaging Resource and the Walter Reed Army Medical Center Radiology Department. The archive contains such integrated and multimodal data as virtual colonoscopy matched with recorded video from endoscopic interventions, time-varying 3D cardiac motion, and 4D MRI of a human hand. During 2009, the TDI group installed the necessary software and hardware infrastructure, including a Linux server and a MIDAS software system, to support interactive scientific publication at NLM. We continue to serve a broad community with these data, and seek to establish a leadership role through public data distribution.

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

Collaborative work in telemedicine has increased the demand for large display technologies, but the cognitive pathways for understanding how the human visual system processes large-scale digital displays are not well understood. Previous TDI group achievements include the construction of high-end rendering systems for large displays, incorporating multiple GPUs in volume-rendering applications capable of rendering the full VHP Male dataset at real time rates. In 2010, we expanded our investigations in the question of presence and visualization, combining forces with the NCRR-funded Biotechnology Resource Center at the Utah Scientific Computing and Imaging (SCI) Institute and their access to the Magnetic Encephalography (MEG) equipment in the University of Utah, Department of Radiology. We also began a study of the difference between human-scale versus non-human scale perception, seeking changes in cortical processing of reading and eye-tracking from large wall displays versus hand-held book-sized displays. The result of such a study could have a far-reaching impact on how new digital user interfaces are developed in future generations.

In FY2010, the TDI study of the use of rapid prototyping technologies in Radiology generated considerable success. Our goal is to develop witness objects or phantoms, reference models as a public engineering and scientific standard for research in image-based computer-aided diagnosis. We have characterized the x-ray attenuation characteristics of some of the 3D-printing materials available at NIH. and are presently evaluating the use of contrast agents as printing materials to vary the appearance of the 3D models In FY2010, we successfully modified the 3D printing process through the use of contrast agents, primarily sodium iodide, to create 3D models that mimic human tissue when viewed with x-ray CT scanners. The goal is to create complex, anatomically-accurate models to test diagnostics systems and evaluate and compare their performance under known conditions. We were able to create models that correspond to CT scans of the Visible Human Project male dataset and demonstrate the possibilities for modeling soft tissue and metastatic disease. This work is conducted in partnership with the National Institute of Allergy and Infectious Diseases.

TDI also facilitated new collaborations and expanded research in medical imaging throughout 2010. With the appointment of a new postdoctoral fellow, OHPCC and the TDI Group began working on volume rendering of strain images from 3D ultrasound in a collaborative effort with the robotics group at Johns Hopkins University Department of Computer Science. This work was expanded over the summer with the participation of a Medical Informatics Training Program (MITP) graduate research fellow. New collaborations for OHPCC also included new work with the NIH Clinical Center's Center for Infectious Disease Imaging on a long-term study of pulmonary fibrosis, where a substantial cohort of affected patients and a corresponding control group contributed image data to enable research and development for computer aided diagnosis tools in pulmonary medicine. Findings on multimodal correlative medicine were developed and submitted for publication also with the participation of an NIH summer high school intern who is extending the project into an Intel National Science Talent Search project. Finally, OHPCC has published recently developed work on statistical deformation models (SDMs) for morphological image analysis based on Markov Random Fields.

Insight Tool Kit (ITK)

In FY2010, the Insight Toolkit celebrated a decade of service for biomedical image analysis with a symposium at the LHNCBC. The current official software release is ITK 3.20. Over 845,000 lines of openly available source code comprise ITK, making available a variety of image processing algorithms for computing segmentation and registration of high dimensional medical data on a variety of hardware platforms. ITK can be run on Windows, Macintosh, and Linux platforms, reaching across a broad scientific community that spans over 40 countries and more than 1500 active subscribers to the global software list-serve. A consortium of university and commercial groups, including OHPCC intramural research staff, provide support, development, and maintenance of the software.

ITK remains an essential part of the software infrastructure of many projects across and beyond the NIH. The Harvard-led National Alliance of Medical Image Computing (NA-MIC), an NIH Roadmap National Center for Biomedical Computing (NCBC), has adopted ITK and its software engineering practices as part of its engineering infrastructure. ITK also serves as the software foundation for the Image Guided Surgery Toolkit (IGSTK), a research and development program sponsored by the NIH National Institute for Biomedical Imaging and Bioengineering (NIBIB) and executed by Georgetown University's Imaging Science and Information Systems (ISIS) Center. IGSTK is pioneering an open API for integrating robotics, image-guidance, image analysis, and surgical intervention. International software packages that incorporate ITK include *Osirix*, an open-source diagnostic radiological image viewing system available from a research partnership between UCLA and the University of Geneva and the Orfeo Toolbox (OTB) from the Centre Nationale D'Etudes Spatiales, the French National Space Administration. Beyond the support of centers and software projects, the ITK effort has influenced end-user applications through supplementing research platforms such as the Analyze from the Mayo Clinic, SCIRun from the University of Utah's Scientific Computing and Imaging Institute, and the development of a new release of VolView, free software for medical volume image viewing and analysis.

In FY2010, OHPCC and the ITK project ran two competitions to revise, accelerate, simplify, expand, and test the Insight Toolkit. Through the ITK Version 4 (ITK-v4) project, we are attempting to upgrade ITK for emerging computational platforms and meet research needs for the upcoming decade. The development of a major new version of ITK will help to continue our international leadership role in medical imaging research. A companion effort known as the 2010 ITK Algorithms, Adapters, and Data Distribution program (ITK-A2D2-2010), we are revisiting successful programs run by NLM in 2002 and 2004 to test the ITK-v4 major software release, to expand our research community throughout life science research in microscopy and radiology, and collect data as a foundation for imaging research in an open-science world.

We have secured the services of notable software development services of groups including General Electric Global Research, the Mayo Clinic, Harvard University, Kitware, Inc., CoSmo Software, the University of Iowa, the University of Pennsylvania, Ohio State University, Old Dominion University, Carnegie Mellon University, Georgetown University, the University of North Carolina at Chapel Hill, and the University of Utah Scientific Computing and Imaging Institute. The research topics supported by these software development efforts include microscopy, digital histology, tumor micro-environments, zebrafish embryology, deconvolution methods for astronomy and astrophysics, image registration for neurosurgery, tumor volume measurement for lung cancer treatment, and video processing for security applications as well as healthcare. This work is funded through the American Reinvestment and Recovery Act.

Image and Text Indexing for Clinical Decision Support and Education

As part of the Clinical Information Systems effort, the Image and Text Indexing project seeks to automatically identify relevant illustrations in biomedical articles that could provide multimedia assistance to clinical decision making. We developed an experimental search engine, the Image Text Search Engine (ITSE), that augments retrieved bibliographic citations with illustrations (medical images, photographs, charts and other figures), extracted from scientific articles. The retrieval methods implemented in the search engine achieved top performance in the ImageCLEF 2010 image modality classification task.

ITSE retrieves and displays "enriched citations" - structured MEDLINE citations expanded with image-related text and concepts and linked to images and image representations based on image features. In FY2010,

processes leading to generation of enriched citations were organized into a pipeline. Experiments conducted demonstrated that enriched citations improve retrieval of similar patient cases, compared to traditional citations.

Research was also conducted in key areas: ways to organize and display retrieval results using annotated images, improved methods to automatically extract both single- and multi-paneled illustrations from articles, and improved methods to extract pointers (arrows, arrowheads, symbols) within images to identify regions of interest, among others.

The project was presented to the Board of Scientific Counselors in September 2010. The Board commended the effort and also recommended providing the existing image and related text processing methods as services, and scaling the experimental search engine to large scale collections, such as PubMedCentral. To address these recommendations, researchers have started redesigning the system architecture.

Turning The Pages (TTP)

The goal of the TTP project is to provide the lay public a compelling experience of historically significant and normally inaccessible books in medicine and the life sciences. In this project, we build 3D models for books and develop animation techniques to allow users to touch and turn page images in a photorealistic manner on touch-sensitive monitors in kiosks, as well as 'click and turn' in an online version. The online version of TTP is a popular Web site, attracting 1,300,000 page views a month.

Following the release of the kiosk version of the Edwin Smith Papyrus, having first created a 3D model for flat scrolls, in FY2010 we released the online version that may be 'clicked and rolled out'. Commentaries on medical cases and treatments from this 1700 BC work may be viewed in synchrony with their locations in the scroll. Also this year we created the kiosk and Web versions of Hanaoka's Japanese medical manuscript. Hanaoka Seishu's fame rests on his invention of an oral anesthetic that rendered a patient unconscious for long enough to allow him to remove deep tumors.

Ongoing work includes the creation of an iPad version of TTP, with the initial release to include the Japanese manuscript, as well as Brunschwig's Liber de Arte Distillandi. Also in development are the Web and kiosk versions of *Ketab Ajaeb al-makluqat wa Gara eb al- Mawjudat (Marvelous Creatures and Mysterious Species)* compiled by al- Qazwini in the middle 1200s in what is now Iran or Iraq.

Biomedical Image Transmission via Advanced Networks (BITA)

Researchers evaluated performance of NLM-supported networks such as the Interactive Video Outreach and Distance Learning Network for Minority High School Students at Health Science-focused magnet schools, the BHEPP network linking NLM and area hospitals, 802.11a, 802.11b and higher designation wireless network implementations, and networks exhibiting narrow bandwidths, high latency and high jitter. As NLM develops applications for handheld platforms the performance of wireless networks becomes more important. In conjunction with NIH networking staff, we conducted a review of wireless infrastructure at the LHNCBC to address capacity questions raised by the movement of images within and on branch infrastructure by in-house developers of handheld applications (e.g., Turning The Pages on the iPad).

Research staff provided guidance and contacts to the organizers of an NSF-sponsored US-India Networking and e-Science Workshop held in December 2010 in India. The workshop focused on improving collaborative biomedical and science research infrastructure between the two countries using high-performance networks.

We regularly monitor the usable capacity of the existing NLM high-performance connections to the domestic and international research communities via Internet2. General capacity planning discussions internal and external to NLM have taken place focused on preparing for the movement of increasing numbers of sequencing datasets to and from NLM. Staff members continue to investigate methods to measure network performance to and from NLM collaboration sites focusing on the utility of perfSONAR and other tools.

As part of this effort, research staff represented NLM and NIH on an Internet2 Advisory Council, Joint Engineering Team, Large Scale Networking Team, and other forums for high performance/speed networks.

Natural Language Processing and Text Mining

Medical Article Records System (MARS)

Automation is required to accommodate the rapid growth of the MEDLINE database, now exceeding 18,000,000 records. The MARS project aims to develop automated systems to extract bibliographic text from journal articles, in both paper as well as electronic forms. For the approximately 1500 journal titles that arrive at NLM in paper form, a production MARS system combines document scanning, optical character recognition (OCR), and rule-based and machine learning algorithms to yield citation data that NLM's indexers use to complete bibliographic records for MEDLINE. Our algorithms extract this data in a pipeline process: segmenting page images into zones, assigning labels to the zones signifying its contents (title, author names, abstract, etc.), pattern matching to identify these entities, lexicon-based pattern matching to correct OCR errors and reduce words that are incorrectly labeled as errors to increase operator productivity.

In FY2010, LHNCBC staff provided all engineering support for the offsite MARS production facility: installation of upgraded modules, testing, maintenance and operation of all hardware and software for servers, clients and networks, and the necessary system administration. A new capability was introduced this year to accommodate up to three scanned pages in an article because we found that some lengthy abstracts require optical scanning of a third page. This required modifications to several modules, in particular those used by scanning and reconcile (verification) operators.

In addition, to help achieve the goals of NLM's Indexing 2015 Initiative, LHNCBC staff developed the Publisher Data Review (PDR) which Library Operations staff evaluated prior to deployment. The PDR system is designed to provide data missing from the XML citations sent in directly by publishers: such as databank accession numbers, NIH grant numbers, and grant support categories. By providing these missing data, PDR will reduce the manual effort in completing the citations sent in by publishers. PDR also corrects incorrect data sent in by publishers. The automated steps to fill in missing data and to correct wrong data will substantially reduce the load on the operators, eliminating the need to look through an entire article to find this information, and then to key them in. Currently, we are designing machine-learning methods to extract two additional MEDLINE citation fields: Investigator Names and Commented-on Article. If done manually, extracting names of investigators is a labor-intensive effort since articles frequently contain hundreds of such names. Similarly, identifying commented-on articles is a time-consuming process since it requires operators to read other related articles for commented-on information.

We are also developing an automated system to help operators handle cases when publishers do not supply an issue or partially supply certain citation fields to MEDLINE. This system, called WebMARS, is a software tool that operators can use to automatically create missing citations from these problematic issues. Currently, this task requires operators to manually type, copy, and paste data from online articles, a very time-consuming step.

The MARS, PDR and WebMARS systems rely on underlying research in image analysis and lexical analysis (within the Analysis of Images for Data Extraction, or AIDE project), but this research also enables the creation of new initiatives in which these techniques find application. Examples are the ACORN initiative and MARG database described below.

Automatically Creating OldMedline Records for NLM (ACORN)

The ACORN initiative aims to capture bibliographic records from pre-1960 printed indexes (e.g., IM, QCIM, QCICL, etc.) for inclusion in the NLM OldMedline database, thereby creating a complete record of citations to the biomedical literature since Index Medicus appeared in the late 19th century. In FY2010, we continued our investigation of scanning, image enhancement, OCR, image analysis, pattern matching, and related techniques to extract unique records from the printed indexes. Finding that many of the printed indexes are available as microfilm, we decided to scan this medium rather than the paper indexes to take advantage of the lower cost of microfilm scanning. In addition, we investigated Web-based information and existing MEDLINE and OldMedline databases to avoid creating duplicate records and to correct OCR errors in citation information. Researchers designed a prototype consisting of three main components: Quality Control, Processing, and Reconcile. The Quality Control module is completed, and work is proceeding toward implementation of the Processing module which will group OCRed text

into records for operators to verify using the Reconcile component. We expect to deliver the pilot system to Library Operations in FY 2011.

Indexing Initiative (II)

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

The MTI system is in regular, increasing use by NLM indexers to index MEDLINE. MTI recommendations are available to them as an additional resource through the Data Creation and Maintenance System (DCMS). Because of the recent addition of subheading attachment recommendations, indexers now have the option of accepting MTI heading/subheading pairs in addition to unadorned headings. In addition, indexing terms automatically produced by a stricter version of MTI are being used as keywords to enhance retrieval of meetings abstracts via the NLM Gateway. These meetings abstracts span the areas of AIDS/HIV, health sciences research, and space life sciences.

Indexing Initiative development focuses on improving MTI's accuracy and efficiency as well as adding functionality to the system. MTI's accuracy is increasing both incrementally based on indexer suggestions and more broadly due to system wide efforts such as incorporation of enhancements to MetaMap including negation detection and resolution of ambiguity. Recent technical enhancements have increased MetaMap's throughput three- to five-fold. One recent functional addition to MTI is an explanation facility to inform indexers how MTI arrived at specific MeSH recommendations. Another recent development consists of successfully modifying MTI both for use in NLM Cataloging and for indexing the History of Medicine's book collection. Finally, we have shown through recent experiments that MTI recommendations can be used as the initial indexing, subject to the normal review process, for a limited number of journals on which it performs exceptionally well.

Digital Preservation Research (DPR)

This project addresses an important problem for libraries and archives, viz., to retain electronic files for posterity, both documents in multiple formats (e.g., TIFF, PDF, HTML) as well as video and audio resources. Researchers focus on the preservation of digitized documents, and have built and deployed a System for Preservation of Electronic Resources (SPER). SPER builds on open source systems and standards (e.g., DSpace) while incorporating inhouse-developed modules that implement key preservation functions: ingesting, automated metadata extraction and file migration.

NLM curators are using the SPER system to preserve more than 60,000 court documents from a historic medico-legal collection acquired from the FDA. In FY2010, more than 17,000 documents were processed and added to a publicly accessible NLM Web site.

We are also investigating the applicability of SPER to preserve other historical collections of importance such as the NIAID collection of conference proceedings of the "US-Japan Cooperative Medical Science Program on Cholera Research" from 1960 to 2010. Our activities toward this initiative include: (a) recognizing different types of information within a document set through layout analysis, (b) evaluating the effectiveness of models such as Support Vector Machine and Hidden Markov Model for different metadata layouts, and (c) capturing relationships between various entities in the collection from the extracted metadata. SPER is used to perform automated metadata extraction (AME) from this document collection, and is building a related portal of research articles, authors, participating members and institutions. More robust algorithms are being developed to enhance the AME accuracy.

In addition, research has been conducted in recognizing metadata in "Form based" documents, and was applied to prototype the extraction of metadata from the NIH Annual Reports on intramural research projects.

InfoBot

As part of the Clinical Information Systems effort, the InfoBot project enables a clinical institution to automatically augment a patient's electronic medical record (EMR) with pertinent information from NLM and other resources. InfoBot processes free-text clinical notes, extracting problems and interventions, and combining extracted information with other patient-specific information in search queries. Information sources are defined for each specific clinical task in a set of rules. Textual search results are post-processed to extract bottom-line advice and present an overview of the results at-a-glance. The manual option is useful when the API cannot be used by the EMR. An API developed for InfoBot allows integrating patient-specific information (medications linked to formularies and images of pills, evidence-based search results for patient's complaints and symptoms, MedlinePlus information for patient education) in an existing EMR system. For clinical settings that have no means to use the API, a Web-based interface allows information requests to be manually entered.

The InfoBot API integrated with the NIH Clinical Center's EMR system, CRIS, is in daily use through the *Evidence-Based Practice* tab in CRIS since July 2009. In FY2010, we completed the evaluation of this prototype at the Clinical Center using a focus group of floor representatives of interdisciplinary teams, mainly from the nursing staff. In line with the evaluation, we modified the information delivered by InfoBot. For example, evidence search results (for a given patient) were automatically re-ranked according to the votes of other team members. We also added images of pills and a search box that allows a user to modify an automatically constructed search. The most followed links are to information about medications and protocols of clinical trials. These links are followed twice as often as the links to MedlinePlus and MEDLINE publications.

De-identification Tools

De-identification enables research on clinical documents. LHNCBC developers are designing software to de-identify clinical documents that comply with the Privacy Rule of the Health Insurance and Accountability Act of 1996. The provisions of the rule dictate removal of 18 individually identifiable health information elements that could be used to identify the individual, the individual's relatives, employers, or household members.
The project consists of three teams: 1) annotators who mark the corpus and build a gold standard set; 2) system developers who design and implement algorithms, and; 3) evaluators who compare outcomes of the algorithmic system against the gold standard and suggest improvements.

LHNCBC obtained narrative reports for approximately 70,000 individuals (under Institutional Review Board (IRB) exemptions) to develop and test this de-identification system. The record mix includes radiology reports, history and physical exam records, occupational therapy notes, discharge summaries, referrals, consult notes, laboratory data and nursing notes. These records are annotated with the assistance of a software tool that helps the annotators tag all HIPAA identifiers in the reports. In FY2010, we annotated 6,000 documents and reached a set size of 16,000 fully annotated documents for testing and improving de-identification.

The de-identification software system operates through a sequence of pipelined processes: 1) Health Level 7 (HL7) message parsing; 2) part of speech tagging; 3) protected personal health information identification, and 4) redaction. Efforts are directed toward conserving all clinically important information while ensuring none of the individually identifiable health information is included in the results set.

In FY2010, we focused our efforts on improving the system's performance on personal names. In the latest performance test of our system, we used 1,001 clinical records containing 360,879 words, of which 7,294 were personal name words. The system's overall sensitivity for correctly identifying personal name words was 99.9 percent. Researchers plan to further improve performance by incorporating HL7 header information.

Terminology Research and Services

The Patient Data Management Project (PDM) brings together several activities centered on lexical issues, including development and maintenance of the SPECIALIST lexicon as well as lexical research. The lexicon and lexical tools are distributed to the medical informatics community as free open-source tools and also delivered with the UMLS information sources.
Objectives for FY2011 are:

- continued expansion and maintenance of the SPECIALIST lexicon with emphasis on clinical vocabulary

- continued development of the lexical management system
- continued development of the cross-platform version of the SPECIALIST Lexical Tools
- continued development of text processing tools (NLP tools)

Medical Ontology Research (MOR)

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

During FY2010, staff investigated issues including quality assurance in ontologies, the representation of pharmacologic classes in biomedical terminologies, the use of the UMLS for mapping between lay terminologies and professional vocabularies, and normalization techniques for mapping clinical drug names to standard terminologies. Many of these studies leveraged the Semantic Web technologies including RDF - the Resource Description Framework - and triple stores (e.g, Virtuoso), which proved to be useful resources for integrating of biomedical information.

Researchers contributed to the LHNCBC training program by providing mentorship to one undergraduate, five graduate and one post-doctoral students, working with them on issues including data integration for pharmacogenomics studies and the representation of rare diseases in clinical and other terminologies.
In FY2010, our research activities resulted in two journal articles, 15 papers in conference proceedings, and six invited presentations. This research project was favorably reviewed by the Board of the Scientific Counselors of the Lister Hill Center. 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).

Clinical and Translation Science

LHNCBC is undertaking a project to facilitate the discovery of published translational science research cited in MEDLINE/PubMed, which has more than 20,000,000 citations. Using a newly-developed database of translational terms, RxNorm and MeSH, the LHNCBC developed a tool for searching for innovative, novel and promising translational research. The query is initiated using disease processes and/or interventions as search terms. Publications identified as translational in nature are then retrieved with relevant terms highlighted for easy recognition. With interventions, such as drugs in the RxNorm database and disease processes in MeSH that are pertinent clearly identified, the user can quickly find publications to facilitate research, experiment planning and bench-to-bedside applications. An additional benefit from this project is the inclusion of a new term, “translational research,” to the MeSH vocabulary.

Semantic Knowledge Representation (SKR)

The Semantic Knowledge Representation (SKR) project conducts basic research in natural language processing based on the UMLS knowledge sources. A core resource is the SemRep program, which extracts semantic predications from text. SemRep was originally developed for the biomedical research domain and is being extended to influenza epidemic preparedness, public health, and health effects of climate change. The SKR project maintains a database of SemRep predications, which currently holds 25,000,000 predications extracted from 7,200,000 MEDLINE citations (1999 through August, 2010).

SKR efforts support innovative information management applications in biomedicine, as well as basic research. The researchers are using semantic predications to find publications that support critical questions used during the creation of clinical practice guidelines (with support from NHLBI). Semantic processing technology is being adapted to identify selected concepts in clinical narrative. One research project combines semantic predications and the literature-based discovery paradigm for investigating the physiology and pathology of sleep, and another exploits predications and graph theory for automatic summarization of biomedical text. Further, the SKR team is collaborating with academic researchers in using semantic predications to help interpret the results of microarray experiments and to investigate advanced statistical methods for enhanced information retrieval.

Information Resource Delivery for Care Providers and the Public

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

Clinical Research Information Systems

ClinicalTrials.gov provides the public with comprehensive information about interventional and observational clinical research studies. ClinicalTrials.gov receives over 100,000,000 page views per month and hosts approximately 900,000 unique visitors per month. At the end of FY2010, the site had nearly 97,000 protocol records, over 2,400 of which display summary results, conducted in all 50 states and in 174 countries. Approximately one-third of the studies are open to recruitment, and the remaining two-thirds are closed to recruitment or completed. Data are submitted by over 9,500 study sponsors which include the U.S. Federal government, pharmaceutical and device industries, academic, and international organizations, through a Web-based Protocol Registration System, which allows sponsors to maintain and validate information about their studies.

ClinicalTrials.gov was established by the NLM in FY2000 in response to the Food and Drug Administration Modernization Act of 1997 and to support NLM's mission of disseminating biomedical knowledge and advancing public health. Over time, ClinicalTrials.gov has been enhanced to support other registration policies and to implement the requirements of Section 801 of the Food and Drug Administration Amendments Act of 2007 [Public Law 110-85]. In FY2010, the second year after passage of this law, new registrations were submitted at an average rate of 350 records per week. In September 2008, ClinicalTrials.gov launched the "basic results" database, which complements the registry. Registered trials may now include tables of summary results data on primary and secondary outcomes and adverse events, as well as information on the patient populations studied. Since the beginning of its operation, over 3,600 results records have been submitted by 724 study sponsors. The average number of submissions per week has increased, with an average of 50 new results records submitted per week at the end of FY2010. The expanded registration requirements as well as the results database will be further elucidated through rulemaking and NLM is working with the Food and Drug Administration (FDA) on the Notice of Proposed Rulemaking. The registry and results database provide access to information about ongoing and completed clinical research. This information is critical for clinical and policy decision makers.

In FY2010, ClinicalTrials.gov was actively involved in educating the public on the new law, system requirements, and continuing to promote standards of transparency in clinical research through trial registration and results reporting. This information was communicated to a broad range of U.S. and international stakeholders via presentations and peer-reviewed publications. ClinicalTrials.gov continues to collaborate with other registries, professional organizations, and regulators in working towards developing global standards of trial registration and reporting to results databases.

Genetics Home Reference (GHR)

Genetics Home Reference (GHR) is an online resource that offers information about genetic conditions and the genes and chromosomes related to those conditions. This resource provides a bridge between the public's questions about human genetics and the rich technical data that has emerged from the Human Genome Project and other genomic research. Created for the general public, particularly patients and their families, the GHR Web site currently includes user-friendly summaries of almost 600 genetic conditions, more than 800 genes, all the human chromosomes, and mitochondrial DNA. The Web site also includes a handbook called Help Me Understand Genetics, which provides an illustrated introduction to fundamental topics in human genetics including mutations, inheritance, genetic testing, gene therapy, and genomic research.

Genetics Home Reference celebrated its seventh anniversary in 2010. In the past year, the project expanded its genetics content for consumers, adding more than 250 new summaries to the Web site, an increase of about 27 percent from the previous 12 months. Staff intend to continue this rate of production in FY2011, covering additional Mendelian genetic disorders as well as more complex disorders. The team also plans to continue expanding the gene families feature, which currently includes explanations of about 55 families of related genes.

Usage of the GHR Web site continued to increase in FY2010. This year, the site averaged almost 16,000 visitors per day (an increase of about 13 percent over the previous fiscal year) and more than 27,600,000 hits per

month (an increase of 47 percent over the previous fiscal year). GHR continues to be recognized as an important health resource.

This year, GHR staff performed outreach activities to increase public awareness of the Web site. The project continues to support the Information Rx initiative, a free program that enables doctors and nurses to write "prescriptions" directing patients to the GHR Web site for an explanation of genetic disorders and related topics. In other outreach activities, GHR staff presented the Web site to several visiting groups, including students and journalists, and represented the project at several major genetics conferences. Staff members will continue to educate others about this important resource in FY2011.

Profiles in Science Digital Library

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

The content of Profiles in Science is created in collaboration with the History of Medicine Division of NLM, which processes and stores the physical collections. Several collections have been donated to NLM and contain published and unpublished materials, including manuscripts, diaries, laboratory notebooks, correspondence, photographs, poems, drawings and audiovisual resources. This year, the collection of Nobel prize-winning biochemist Daniel Nathans was added to Profiles in Science. One thousand seven hundred thirty-eight transcripts of documents were also added, making handwritten items searchable and providing alternatives to PDF format files. Fifty-five digital items were also added to the thirty-three existing Profiles in Science collections. Currently 26,555 digital items composed of 139,945 image pages are available on Profiles in Science. Presently the Web site features the archives of thirty prominent individuals:

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

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

In addition to updating the Profiles in Science collections during FY2010, LHNCBC staff increased the reliability, security and longevity of the Profiles in Science systems. We installed a server at the NIH Consolidated Co-location Site (NCCS) to ensure availability of the Web site during facilities outages, and servers were synchronized on three different networks. We adjusted Web site update procedures to accommodate multiple servers. Staff also investigated various software for possible future use such as: JHOVE2 for validating the project's digitized files, W3C Markup Validation Service for validating Web pages, Total Validator for validating Web pages as well as Section 508 compliance, Apache PDFBox for examining PDF format files, Sitemap for easier searching by external Web crawlers, and Solr/Lucene for searching. Project staff investigated alternative formats to PDF, and found that DjVu is still the most viable alternative. Staff created a presentation, "Profiles in Science: A Digital

Library," that was delivered at the 2010 International Conference on BioCommunications. We continued to develop procedures and protect the master digital files associated with the Web friendly files seen on Profiles, and to deliver the masters to collaborating institutions. We also continued to automate tasks that were being performed manually, especially those involving quality control and updating.

Evidence Based Medicine - PubMed for Handhelds

PubMed for Handhelds was developed and released in FY2003 to facilitate evidenced-based medical practice with Medline access at the point of care via smartphones, wireless PDA's, netbooks or portable laptops. PubMed for Handhelds (PMHh) requires no proprietary software and reformats the screen display as appropriate for the wireless handheld device being used. In support of evidence-based clinical practice, clinical filters feature easy access to relevant clinical literature. Newly developed resources allow searching Medline through text-messaging. An algorithm to derive “the bottom line” (TBL) of published abstracts was recently added for a clinician’s quick reading at the point of need. New features can create a “consensus” opinion of multiple publications. Recent collaborative projects are ongoing in Botswana, Africa and the Pacific Islands. Philippines. Randomized controlled trials using simulated clinical scenarios are currently underway at the Uniformed Services University and the University of Botswana-University of Pennsylvania to evaluate the usefulness of abstracts.

Clinical Vocabulary Standards

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

The CORE Problem List Subset of SNOMED CT

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

Based on data collected from seven large-scale U.S. and overseas healthcare institutions, a detailed study was done on the nature of the local problem list vocabularies. One significant finding is the low level of overlap between these vocabularies, with an average pairwise overlap of around 20%. However, terms that are shared among institutions were used eight times more frequently than concepts unique to one institution, which lends support to the idea of having a common core of problem list terms across institutions. Since SNOMED CT is a designated standard for problem lists according to the “meaningful use” criteria, a CORE (Clinical Observations Recording and Encoding) Problem List Subset of SNOMED CT, which contained about 6,000 concepts and represented the most frequently used problem list terms, was identified and made available to SNOMED CT users. The CORE Subset can be used as a starter set for institutions that do not yet have a problem list vocabulary based on SNOMED CT. This will save significant development effort and reduce unintentional variations in the choice of terms. Existing problem list vocabularies can also be mapped to the CORE Subset which will facilitate data interoperability. Since

publication, the CORE Subset has received considerable attention within the SNOMED CT user community. The IHTSDO (International Health Terminology Standards Development Organization) used the CORE Subset to focus its quality assurance effort on clinically important concepts. The MedlinePlus Connect Project, which facilitates online linkage to patient education information, has mapped all concepts in the CORE Subset to MedlinePlus health topics. There is ongoing effort to map the CORE Subset to the ICD classifications (ICD-10 and ICD-10-CM) which will promote the adoption of SNOMED CT by allowing re-use of SNOMED CT encoded clinical data. Currently research effort is underway to formally review the coverage and usability of the CORE Subset, in comparison to other existing problem list vocabularies.

RxTerms

Originally created for the Personal Health Record Project, RxTerms serves as an efficient drug interface terminology to facilitate electronic capture of prescription information and linkage to standard identifiers in RxNorm, the U.S. national drug terminology standard. To overcome the problem of excessively large pick-lists and long drug names in RxNorm, RxTerms segments the information contained in a RxNorm clinical drug into two portions: ingredient with route (e.g. Amoxicillin (Oral-pill)), and strength with dose form (e.g. 500 mg Tabs). This has been shown to improve data entry efficiency. RxTerms includes other usability-enhancing features like the inclusion of common synonyms and abbreviations, “tall-man lettering” to distinguish look-alike drug names and improved liquid dose concentration information. RxTerms is updated every month with the full monthly release of RxNorm.

RxNav

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

The RxNorm API was further developed and we designed a RESTful version of the API, compatible with the Representational State Transfer (REST) architecture. Two other drug information sources were integrated with RxNav: RxTerms, an interface terminology for prescription writing or medication history recording; and NDF-RT, a resource that links drugs to their pharmacologic classes and properties, including indications, contra-indications and drug-drug interactions.

The production version of RxNav has received about 10,000,000 queries in FY2010. Users include clinical and academic institutions, as well as pharmacy management companies, health insurance companies, EHR vendors, and drug information providers. In the future, the integration with NDF-RT will be refined. Specialized applications relying on the API will be developed, e.g., for mapping large amounts of terms and codes to RxNorm, and for crosswalk purposes between drug vocabularies through RxNorm.

Collaboration with Centers for Medicare and Medicaid Services

In collaboration with the Centers for Medicare and Medicaid Services (CMS), the LHNCBC continues to provide help and guidance regarding standard terminologies for medications and clinical problems. To assist CMS in the implementation of their Continuity Assessment Record Evaluation (CARE) data entry form, we shared with them RxTerms and the PHR Problem List Terminology. We have received some preliminary data in anonymized form captured in the CMS systems and can use that to study the adequacy of the two interface terminologies.

Standards for Identifying Clinical Observations and Orders

In FY2010, LHNCBC expanded the LOINC database. The Regenstrief Institute and the LOINC committees are committed to expanding LOINC’s globalization and usability. Last spring, LOINC completed the development of a web-based LOINC browser based on the Lucene Search engine, http://search.loinc.org. LOINC currently supports nine languages including both Simplified Chinese and Korean. The most recent additions are Greek and Italian.

We are working with the major laboratory companies and the American Clinical Laboratory Association (ACLA) to clarify the content of many of the most frequently ordered *test panels* and mechanisms to represent them in LOINC. This has the dual advantage that the major commercial laboratories help us define the standard approach and then often change their internal systems to conform to it. Laboratory instrument vendors are now linking LOINC codes to their outputs and making the linkage available in the package insert or on their web sites. Other work that is under development is the list of the top 2,000 laboratory test result terms, with guidance about which terms to use in what situation and an improved mapping tool.

LOINC is collaborating with PhenX (https://www.phenxtoolkit.org/) and PROMIS to incorporate their survey instruments fully within the LOINC database. Working with NHGRI, LHNCBC has enhanced the robustness and usability of the PhenX Toolkit, which allows researchers to review and select high priority measures and recommended protocols for inclusion in genome-wide association studies (GWAS) and other broad-based genomics studies. LHNCBC and NHGRI are also working together to map PhenX measures to controlled clinical vocabularies, such as LOINC and SNOMED-CT.

Newborn Screening Coding and Terminology Guide

Newborn screening is an important part of public health, but use of test results is complicated by wide variations among states in the ways tests are conducted, results recorded, and by paper-based communications. The current situation can delay rapid attention to a child's health problems, and it creates frustration and extra work for parents, health care providers, and public health authorities.

Combining standard coding, terminology and electronic messaging methods for newborn screening can improve the quality of health care for children. Moreover, public health agencies will be better equipped to observe and compare nationwide trends from newborn screening test results, which will also support efforts of the biomedical research community at NIH and elsewhere to improve newborn screening methods and evaluation. This large project had many partners including the HHS Office of the National Coordinator (ONC) for Health Information Technology, the Health Resources and Services Administration (HRSA), the Centers for Disease Control and Prevention (CDC), the American College of Medical Genetics (ACMG), the Regenstrief Institute and the Federal Advisory Committee on Heritable Disorders in Newborns and Children.

NLM collaborated with HRSA to create the HRSA/NLM guidance for sending electronic NBS results using HL7 messages and LOINC and SNOMED CT codes. The guidance is available on the NLM Newborn Screening Coding and Terminology Guide Web site, at http://newbornscreeningcodes.nlm.nih.gov. The goal of the guidance is to provide a standard framework for reporting newborn screening results in an electronic message whose contents can be accurately interpreted by recipient electronic health information systems for use in care, follow-up and analysis. Adoption of this framework will enable the meaningful use and comparison of data from different laboratories.

Staff worked with HRSA, CDC, the Association of Public Health Laboratories (APHL), and the National Newborn Screening and Genetics Resource Center (NNSGRC), to hold a workshop about Newborn Screening Laboratory Results and Health Information Exchange in November 2010. The goal of the workshop was to showcase implementation and adoption of the HRSA/NLM guidance for sending newborn screening results electronically using standardized universal HL7 messages and LOINC and SNOMED CT codes, distill best practices, gain feedback from the states about needed additions or edits to the guidance, and gather information for other related ongoing and planned initiatives, including developing codes for the confirmatory and diagnostic testing that occurs after newborn screening for some infants, triggered by the newborn screening results. The workshop brought together representatives from 30 states, several federal agencies, the three vendors developing and operating newborn screening laboratory information systems (PerkinElmer, Natus Medical/Neometrics and OzSystems), and members of other organizations that actively represent the newborn screening community.

Several states are using the HRSA/NLM guidance to develop standard reports of newborn screening results. We have reviewed prototype messages from all 3 of the major vendors for NBS labs that utilize the NLM/HRSA specification for reporting NBS results, and worked closely with at least 5 states (Pennsylvania, Kentucky, and New York) that are in the process of implementing the guidance. During the last year, we have been collaborating with HRSA and the Public Health Informatics Institute (PHII) about data elements and message standards related to their work on an implementation guide for sending electronic newborn screening lab orders using HL7 ver 2.x.

Communication Infrastructure Research and Tools

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

Videoconferencing and Collaboration

Researchers continued to investigate, review, and develop collaboration tools, research their application, and use the tools to support ongoing programs at the NLM. In our ongoing work with uncompressed high definition video over IP, we determined strengths and weaknesses of each of the three technologies (iHDTV, UltraGrid, and Conference XP) and continue to overcome problems encountered in the delivery of uncompressed video due to differing platforms and user machine memory capabilities.

iHDTV is the only system sufficiently robust to use in a clinical trial of uncompressed videoconferencing for telemedicine. Staff members are monitoring the progress of UltrGrid's developers and are collaborating with the developers of the uncompressed ConferenceXP program to solve the uncompressed video transmission problem. We are also monitoring the HD open source work of VLC developers that applies H.264 compression, since VLC is used by the AccessGrid, an open source collaboration tool widely deployed in universities and research centers and used in the OHPCC Collaboratory for research work and to support NLM programs. Staff have developed a manuscript comparing the main video technologies used in the Collab, including those doing standard definition video and compressed HD and some uncompressed technologies.

Staff also completed a study comparing patient, provider, and interpreter ratings of clinical encounter quality when interpretation services were provided in-person and by video and phone. In-person was rated highest and phone lowest. The study compared conventional standard definition video to phone and in person interpretation. A follow up study using lower quality video (less than full screen) and cell phone technology assessing video interpretation in pharmacy settings was started. Extensive tests were done with VSee, a low bandwidth video program.

Results of two co-location studies of videoconferencing were combined to assess learning outcomes and collaborative behaviors when students were co-located or dispersed. There were no differences in performance because all the medical students scored well on the exam, but ratings of interactivity were highest for the dispersed videoconference. This finding was attributed to the ways videoconferencing channeled communication and its realism and synchronicity.

Staff continued to work with SIS on distance education outreach program for minority high school students and with the NIH Library to offer NCBI database and other bioinformatics training at a distance. In FY2010, staff conducted outreach programs with the University of North Carolina at Chapel Hill, the University of Tennessee at Memphis, the University of Maryland at Baltimore, the Virginia Commonwealth University, and the Rochester Institute of Technology.

Staff conducted a retrospective review of NLM OHPCC Telemedicine, Next Generation Internet, and Scalable Information Infrastructure Project publications identifying computing, communication, and health science application themes in this diverse corpus of work.

OHPCC Collaboratory for High Performance Computing and Communication (Collab)

The Collab was established as a resource for researching, testing, and demonstrating imaging, collaboration, communications and networking technologies related to NLM's Next Generation Network initiatives. Staff use this infrastructure to test new technologies of interest to NLM and to conduct ongoing imaging, collaboration and distance learning research both within OHPCC and outside NLM. When appropriate, it is leveraged to support other activities and programs of the NLM. The facility can be configured to support a range of technologies, including 3D interactive imaging (with stereoscopic projection), the use of haptics for surgical planning and distance education, and interactive imaging and communications protocols applicable to telemedicine and distance education involving a range of interactive video and applications sharing tools. The latter enables staff to collaborate with others at a distance and, at the same time, demonstrates much of the internal and external work being done as part of NLM's Visible Human and advanced networking initiatives. The collaboration technologies include a complement of tools

built around the H.323 and MPEG standards for transmitting video over IP and open source technologies such as the Access Grid. Staff upgraded the H.323 technology this year and acquired 3D display and DVD playback technology that are being integrated into the remote control technology for displaying output from collaboration tools.

DocView Project: Tools for Using and Exchanging Library Information

The goal of this project is to conduct R&D on advanced tools allowing libraries and users to access biomedical information. In FY2010, research focused on the completion of MyDelivery development, including an Applications Programming Interface (API), and release of its source code. MyDelivery is a novel Internet communications system designed to deliver very large Gigabyte-sized files and large numbers of files, especially over potentially unreliable networks such as wireless used by an increasingly mobile population. It features a unique memory-based server architecture that buffers user data briefly in memory, and avoids storage of user data on server hard disk, where it could be compromised. Health science applications often require the use and exchange of information contained in very large files (e.g., digitized x-ray images, sonographic images, digital video files, MRI, CT scans, PET scans, and scanned document images). Targeted for use in clinical, research, administration, and library environments, the MyDelivery system will be capable of reliably communicating biomedical information contained in files of virtually any size over networks of all types, including potentially unreliable ones. Developers released source code and extensive system documentation for the MyDelivery client, server, and API in June 2010.

As part of the DocView project, research and system engineering continues to maintain and improve the operation of DocMorph, a Web-based server providing users remote image and information processing capabilities via the Internet. This system now accepts more than fifty file formats, including black and white images, grayscale and color images, text and word processing files, to produce four outputs: PDF files, TIFF files, text, and language translation. DocMorph averages 1,500 conversions daily, and 1,400 unique users monthly.

DocMorph's 24,000 registered users include several hundred libraries that use DocMorph as part of their interlibrary loan services. While DocMorph is generally accessed via a Web browser, the MyMorph client software allows users to perform large scale conversion of thousands of files at a time. MyMorph has more than 16,000 registered users, many of whom are document delivery librarians in small libraries around the country, using MyMorph as an important component of their daily document delivery operation.

Computing Resources Projects

The Computing Resources (CR) Team accomplished a number of core projects to build, administer, support, and maintain an integrated and secure IT infrastructure that facilitates the research activities of the LHNCBC and thereby augments the overall effectiveness of research staff members. The integrated secure infrastructure encompasses network management, security management, facility management, storage and backup management, and system administration support for a large number of individual workstations and shared servers.

The network management team plans, implements, tests, deploys, and operates high-speed network connectivity locally as well as over Internet and Internet2. The core projects include studying and planning the implementation of central network management for effectively responding to network alerts and malfunctions; 10 Giga BPS network to support research projects that require high-speed communication capacity; and an enterprise device management system to update large number of network devices uniformly.

The security management team incorporates security operations into firewall administration, patch management, anti-virus management, intrusion monitoring, security vulnerability scanning and remediation, and penetration testing to ensure a safe working environment from an overall security perspective. The core projects include studying and planning the implementation of a security auditing process, asset management, and configuration management for the consistency and integrity of LHNCBC security profiles.

The facility management team facilitates the deployment of products and servers, including power acquisition, network planning, cabling connection, and space allocation in B1 computer room as well as co-location in Sterling, Virginia. The core projects include studying and planning the implementation of redundant LHNCBC infrastructure in the B1 computer room; new network-wiring schemes to the offices in corridor 28 and 30 at B1 level; coordinating and facilitating the NCCS's move to a new data center, and Intelligent Platform Management Interface (IPMI) for effective monitoring on the large number of devices in the B1 computer rooms.

The system administration team provides center-wide IT services, such as DNS, NIS, centralized storage and backup, printing, and remote access to ensure an efficient operation across the Center. The core projects include studying and planning the implementation of Domain Name System Security Extensions (DNSSEC) that is required by OMB, an enterprise data mirroring system that utilizes different media at multiple locations for data safety and integrity; unified communication to enhance research collaboration; evaluation of Windows 7, Windows 2008 and Red Hat Linux 5 platforms for LHNCBC desktop and server deployments. Additionally, the system administration team and other members support Continuity of Operation (COOP) and Federal Information Security Management Act (FISMA) compliance, and provide operation assistance and troubleshooting functions for shared computing resources.

Disaster Information Management

Lost Person Finder (LPF)

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

The hospital-based LPF system includes means to photograph victims at the triage station of a hospital, and to capture these pictures and descriptive metadata (name, age range, and identifying features) in TriagePic, an application developed for the triage staff to stage the patients to appropriate treatment areas in the hospital. This data enters a MySQL database which can be searched via a Web site built by customizing the open-source Sahana disaster management system. The LPF system features a "Notification Wall" that displays images of victims on computers as well as large auditorium screens for family or staff. In 2009 and 2010, we participated in large-scale multi-institutional drills (*Collaborative Multi-Agency Exercise* or CMAX) and demonstrated TriagePic usage, search capability, and the Notification Wall displays at the Navy and Suburban hospitals in Bethesda.

The community-wide system was developed rapidly in January 2010 in response to the earthquake disaster in Haiti. Building on components of the existing LPF system, developers created the Haiti Earthquake People Locator (HEPL) Web site whose main element was a searchable "Interactive Notification Wall" displaying pictures and metadata of missing as well as found people. This data could be sent to HEPL via computer, cell phone, and an iPhone application, ReUnite, developed in-house and made freely available through Apple's iTunes (ReUnite, NLM's first iPhone app, may also be used with iPod Touch and iPad, and is available for any future disaster).

The engineering effort to develop HEPL involved collaborations throughout the LHNCBC and other NLM divisions. Since it was necessary for HEPL to be understood in English, French and Kreyol, language translations were drawn from various NLM divisions as well as from DC Crisis Camps. Initially, HEPL acquired missing person data by scraping CNN iReport records, and later our system was made interoperable with the Google People Finder site. Eventually, more than 50,000 records (photos and metadata) were made searchable through HEPL. HEPL also offered links to the Google site, metasearch engines, and Haiti-specific content created by NLM's Specialized Information Systems division.

In an effort to provide a system capable of being used to respond to any disaster anywhere, developers are designing a prototype unified system, NLM Person Locator (PL), to hold data from multiple disasters, thereby eliminating the need to build multiple Web site/database instances.

Since personally identifiable information (PII) such as pictures and names are collected by LPF, staff have initiated a Certification and Accreditation Process and developed a System Security Plan. An extension of the original provisional OPM data collection permission has been granted approval through 2013.

Video Production, Retrieval, and Reuse Project

This development area encompasses four projects. The NLM media assets project and the NLM support project contribute to the NLM-wide audio-video support of the NLM Long Range Plan goal of promoting health literacy and increasing understanding. The LHNCBC research support project and the core resources project contribute to ongoing LHNCBC information development projects, working to improve access to high quality biomedical imaging information.

The still image, graphics, and video support staff provide ongoing capability to all of the NLM. This work includes production, post-production, and authoring services for the development of Internet video, kiosk interactive multimedia, and DVDs. This area of focus includes support to maintain the audio, video, and multimedia capability in the NLM board room, auditorium, and other conference areas.

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

The focus on video compression codecs for small screen delivery, navigation, and search capabilities is an ongoing area of research related to the work of the exhibition as well as many other areas of NLM's information programs. Extensive development work continued toward the planning and demonstration of interactive multimedia for the FY2011 NLM Exhibition "Native Concepts of Health and Illness." As an extension of this work, staff developed prototype iPhone and iPad applications. Based on these prototypes, the NLM plans to integrate iPhone applications into the upcoming exhibition.

Training Opportunities

Working towards the future of biomedical informatics research and development, the LHNCBC provides training and mentorship for individuals at various stages in their careers. The LHNCBC Informatics Training Program (ITP), ranging from a few months to more than a year, is available for visiting scientists and students. Each fellow is matched with a mentor from the research staff and participates actively on Center research projects.

In FY2010, the LHNCBC provided training to 51 participants from 15 states and seven countries. Participants worked on research projects including 3-D informatics, clinical information systems, content-based information retrieval, de-identification of medical records; image, text and document processing, information retrieval research, interactive publication research, medical ontology research, medical terminology research, mobile computing, natural language processing, personal health record and telemedicine projects.

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

The ITP sponsors a Clinical Informatics Postdoctoral Fellowship Program to attract young physicians to NIH to pursue research in informatics. This program is run jointly by the LHNCBC and the Clinical Center to bring postdoctoral fellows to labs throughout NIH, though funding is from the LHNCBC. We continue to offer an NIH Clinical Elective in Medical Informatics for third and fourth year medical and dental students. The elective offers students the opportunity for independent research under the mentorship of expert NIH researchers. We also host the eight-week NLM Rotation Program which provides trainees from NLM-funded Medical Informatics programs with an opportunity to learn about NLM programs and current LHNCBC research. The rotation includes a series of lectures covering research being conducted at NLM and the opportunity for trainees to work closely with established scientists and fellows from other NLM-funded programs.