THE LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS

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

TECHNICAL REPORT
LHNCBC-TR-2006-004

The Lister Hill National Center For Biomedical Communications Annual Report FY 2006

Donald W. King, M.D.
Acting Director

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

Lister Hill National Center for Biomedical Communications

Donald W. King, M.D.
Acting Director

The Lister Hill National Center for Biomedical Communications (LHNCBC), established by a joint resolution of the United States Congress in 1968, is a research and development division of the U.S. National Library of Medicine (NLM). Seeking to improve access to high quality biomedical information for individuals around the world, the Center continues its active research and development in support of NLM's mission. The Center conducts and supports research and development in the dissemination of high quality imagery, medical language processing, high-speed access to biomedical information, intelligent database systems development, multimedia visualization, knowledge management, data mining and machine-assisted indexing. An external Board of Scientific Counselors meets biannually to review the Center's research projects and priorities. The most current information about Lister Hill Center research activities can be found at http://lhncbc.nlm.nih.gov/

Lister Hill Center research staff are 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, and academic and industry partners. Staff regularly publish their research results in the medical informatics, computer and information science, and engineering communities. The Center is often visited by researchers from around the world. The Lister Hill Center is organized into five major components: Cognitive Science Branch, Communications Engineering Branch, Computer Science Branch, Audiovisual Program Development Branch, and the Office of High Performance Computing and Communications.

Organization

Cognitive Science Branch
The Cognitive Science Branch (CgSB) of the LHNCBC conducts research and development in information systems informed by research in the mechanisms underlying human cognition. Important research areas encompass the investigation of several techniques, including linguistic, statistical, and knowledge-based methods for improving access to biomedical information. Branch members have developed and continue to augment SPECIALIST, an experimental natural language processing (NLP) system for the biomedical domain. The SPECIALIST NLP tools facilitate natural language processing by helping application developers with lexical variation and text analysis tasks in the biomedical domain. Branch members also actively participate in the Unified Medical Language System (UMLS) project and lead the NLM's Indexing Initiative, whose goal is to develop automated and semi-automated techniques for indexing the biomedical literature. The Branch conducts research in digital libraries and collaborates with NLM's History of Medicine Division on Profiles in Science, a project to digitize collections of prominent biomedical scientists. Several Branch projects address the challenges involved in providing health information to consumers. Branch staff developed and continue to enhance the ClinicalTrials.gov Web site that links patients to medical research and

promotes public awareness of the role of clinical trials; and the Genetics Home Reference Web site, a rich resource for understanding how genetics affects human health. The most current information about the Cognitive Science Branch can be found at http://lhncbc.nlm.nih.gov/cgsb/.

Communications Engineering Branch

The Communications Engineering Branch (CEB) is engaged in applied research and development in digital imaging and communications engineering motivated by the NLM mission-critical tasks such as document delivery, preservation of electronic resources, automated production of MEDLINE records, Internet access to biomedical multimedia databases, reliable information delivery to handheld computers in a clinical setting, and imaging applications in support of medical educational packages employing digitized radiographic, anatomic, and other imagery. In addition to applied research, the Branch also developed and maintains operational systems for production of bibliographic records for NLM's flagship database, MEDLINE.

Research areas include: the design of imaging and database tools for biomedical research (in collaboration with the National Cancer Institute), content-based image indexing and retrieval (CBIR) of biomedical images, the design of multimedia-rich interactive publications, document image analysis and understanding (DIAU), image compression, image enhancement, image feature identification and extraction, image segmentation, image retrieval by *image example and sketch*, image transmission, optical character recognition (OCR), natural language processing to extract outcome statements from MEDLINE, and man-machine interface design applied to automated data entry. CEB also maintains archives of large numbers of digitized spine x-rays, uterine cervix images, and bit-mapped document images that are used for intramural and outside research purposes. Information on these projects appears at http://archive.nlm.nih.gov/

Computer Science Branch

The Computer Science Branch (CSB) applies techniques of computer science and information science to problems in the representation, retrieval and manipulation of biomedical knowledge. CSB projects involve both basic and applied research in such areas as intelligent gateway systems for simultaneous searching in disparate databases, intelligent agent technology, knowledge management, the merging of thesauri and other controlled vocabularies, and machine assisted indexing for information classification and retrieval.

Research issues include knowledge acquisition, knowledge representation, knowledge base structure, knowledge visualization, and the human-machine interface for complex systems. Important components of the research include embedded intelligence systems that combine local reasoning with access to large-scale online databanks.

CSB research staff include the team that has developed the NLM Gateway, the team that has produced the Unified Medical Language System (UMLS) Metathesaurus since 1990, and members closely involved in the Center's training programs. Staff members participate in the meetings of the Internet Engineering Task Force and in other professional specialty activities. The most current information about the Computer Science Branch can be found at http://lhncbc.nlm.nih.gov/csb/.

Audiovisual Program Development Branch

The Audiovisual Program Development Branch (APDB) conducts media development activities with three specific objectives. As its most significant effort, the branch participates in the LHNCBC's research, development, and demonstration projects with high quality video, audio,

imaging, and graphics materials. From initial project concept through project implementation and final evaluation, a variety of forms and formats of visuals are developed, and staff activities include image creation, editing, enhancement, transfer and display. Included in this effort is the production of a series of video modules, reporting the progress of Lister Hill Center Research Projects. These informational and educational video reviews have been released in a variety of media, including Web delivery.

Consultation and materials development are also provided by the branch for the NLM's other information programs. With the mission requirement of the Library expanded to include effective outreach activities to the public, patients, and families, the range and quantity of support that the branch provides to these programs continues to increase. From applications of optical media technologies and teleconferencing, to support for World Wide Web distribution, the requirement for graphics, video, and audio materials has increased in quantity and diversified in format.

The third area of concentration is the engineering of technical improvements applied to media production issues such as image quality and resolution, color fidelity, transportability, storage, retrieval, and visual information compatibility and complexity. In addition to the development by the staff of new techniques and processes, the facilities and hardware infrastructure must reflect state-of-the-art standards in a very rapidly changing field. High definition video is a technology area that has been explored and developed within APDB and represents today’s standard for improved electronic motion imaging quality. Multimedia systems, visualization and networked media are being pursued for the performance, educational, and economic advantages that they offer. Three-dimensional computer graphics, animation techniques, and photorealistic rendering methods have changed the tools and products of the artists in the branch. Digital video and image compression techniques are central to projects requiring storage of large images and rapid visual file transmission.

Included within the branch is the Office of the Public Health Service Historian. The PHS Historian provides information about the history of Federal efforts devoted to public health, preserves and interprets the history of PHS, and promotes historically oriented activities across the U.S. Department of Health and Human Services, in partnership with the History Office of the Food and Drug Administration and the National Institutes of Health Historical Office. Throughout the year, the PHS Historian worked with the Office of the Secretary of HHS providing detailed research on issues relating to past influenza pandemics, specifically the 1918-1919 pandemic. This material was used in speeches given by the Secretary and used to develop a Web exhibit on the history of influenza. The PHS Historian developed a traveling exhibit on the history of the USPHS shown around the country. As part of the ongoing oral history program, staff interviewed several PHS officers, including former Surgeon General Antonia Novello. The PHS Historian worked with various public health offices, non-profits, and organizations, including the Indian Health Service, the Office of the Secretary, and the Office of Public Health and Science (OPHS), to create and mount exhibits. The most current information about the APDB can be found at http://lhncbc.nlm.nih.gov/apdb/.

Office of High Performance Computing and Communications

The Office of High Performance Computing and Communications (OHPCC) serves as the focal point for NLM’s High Performance Computing and Communications (HPCC) activities. OHPCC coordinates NLM’s HPCC planning, research and development activities with Federal, industrial, academic, and commercial organizations while collaborating with Lister Hill Center

research branches and NLM Divisions in the development, operation, evaluation and demonstration of HPCC research programs and projects. In addition, OHPCC plans, coordinates, and administers the interagency HPCC research and development program. Office staff serve as NLM's liaison to scientific organizations at all levels of national, state and international government on planning and implementing research in High Performance Computing and Communications. The major research activities center on the Visible Human Project®, NLM's Next Generation Internet Program, telemedicine, the HPCC Collaboratory, and the 3D informatics research program. The most current information about the Office of High Performance Computing and Communications can be found at http://lhncbc.nlm.nih.gov/ohpcc/.

The remainder of this overview highlights LHNCBC's principal R&D activities and accomplishments.

Biomedical Imaging

The overall goal of this program area is to address fundamental questions that arise in the handling, organization, storage, access and transmission of very large electronic files in general and digitized biomedical images in particular. A special focus is research into these topics as applied to heterogeneous multimedia databases consisting of both images and text. Projects in this area have benefited from collaborators in several universities as well as at agencies such as the National Center for Health Statistics (NCHS) and the National Institute of Arthritis, Musculoskeletal and Skin Diseases (NIAMS). A great deal of effort in the past year has focused on a partnership with the National Cancer Institute (NCI) in their research in cervical cancer caused by the Human Papillomavirus (HPV). Our biomedical imaging work may be broadly divided into Multimedia Database R&D and Content-Based Image Retrieval.

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

- WebMIRS. The Web-based Medical Information Retrieval System (WebMIRS) continues to provide access to images and text from nationwide surveys conducted by the National Center for Health Statistics. At the current time there are 444 users of WebMIRS in 54 countries. This Java application allows remote users to access data from the National Heath and Nutrition Examination Surveys II and III (NHANES II and III), carried out during the years 1976-1980 and 1988-1994, respectively. The NHANES II database accessible through WebMIRS contains records for about 20,000 individuals, with about 2,000 fields per record; the NHANES III database contains records for about 30,000 individuals, with more than 3,000 fields per record. In addition, the 17,000 x-ray images collected in NHANES II may also be accessed with WebMIRS and displayed in low-resolution form. The NHANES II database also contains vertebral boundary data collected by a board-certified radiologist for 550 of the 17,000 x-ray images. This data consists of *x,y* coordinates for approximately 20,000 points on the vertebral boundaries in the cervical and lumbar spine images. Users may do queries for both radiological

and/or health survey data. An example of such a query is: “Find records for all persons having low back pain (health survey data) *and* fused lumbar vertebrae (radiological data)”. The boundary data points are displayable on the WebMIRS image results screen and may be saved to the user’s local disk.

The Digital Atlas of the Cervical and Lumbar Spine remains available for the public on a CD or on the CEB Web site either as a Java applet or a downloaded Java application. The Java application version allows the user to add his/her own grayscale and color images in a special “My Images” section and to annotate and title those images for later use. The Atlas has capabilities to display color images, to add extensive text annotations, and to import/export sets of images and annotations as a package.

In addition, the FTP x-ray archive of 17,000 digitized spinal x-rays continues to be active, with 444 users worldwide. This archive allows access to the x-rays, available both in full 12-bit flat file format and also in TIFF 8-bit format which is easier for many researchers to use.

A suite of newer systems motivated by, but not restricted to, joint research with NCI, are at various stages of development.

- The Multimedia Database Tool (MDT), designed as the next generation WebMIRS system, provides a software framework for the incorporation of new text/image databases in a much more general way than the current WebMIRS, and new features for the database end user that extend current WebMIRS capabilities.

- The Boundary Marking Tool (BMT) provides Web capability to manually mark boundaries on cervicography images, and to manage collected data with a MySQL database. It is in active use by NCI for multiple studies.

- The Virtual Microscope (VM) provides Web capability to view and collect information on histology images from expert observers.

- The Teaching Tool (TT) is a system for training medical personnel in cervix anatomy/pathology. It displays the uterine cervix images and quizzes an observer, and enables an NCI medical expert to tailor exams by specifying images and questions to use on an examination.

Content-Based Image Retrieval

The Content-Based Image Retrieval (CBIR) system provides capabilities for searching for spine vertebrae by shape and/or descriptive text, using a database of several thousand pre-segmented vertebral shapes and text data from the NHANES II database used by WebMIRS. The key characteristics of this system, developed in MATLAB and Java, are that it can operate in networked or standalone modes, uses XML for reporting, and allows the user to select either a more mature or an experimental version of the system.

A significant development is Relevance Feedback (RF), a MATLAB experiment in utilizing feedback from an expert user for CBIR image retrieval. Work is under way to incorporate the capabilities of the CBIR3 and RF systems into SPIRS, our new, Web-based *Spine Pathology and Image Retrieval System*. In addition, development continues on the Pathology

Validation and Collection (PathVa) tool, a Java-based system for segmentation review and editing. This tool will retrieve spine images that have been compressed with methods developed for NLM by Texas Tech, over the Internet, along with the segmented boundary data. Three board-certified radiologists are collaborating with us to review several hundred images and record presence, type, and degree of severity of anterior osteophytes, disk space narrowing, subluxation, and spondylolisthesis. The recorded information is automatically transmitted to a CEB database after validation of the segmented boundaries. Other work includes collaborations with several universities to develop or improve segmentation algorithms, segmentation systems or tools, and shape validation tools.

The Visible Human Project

The Visible Human Project (VHP) image data sets are designed to serve as a common reference for the study of human anatomy, as a set of common public domain data for testing medical imaging algorithms, and as a test bed and model for the construction of image libraries that can be accessed through networks. The Visible Human data sets are available through a free license agreement with the NLM. They are distributed to licensees over the Internet at no cost; and on DAT tape for a duplication fee. The data sets are being applied to a wide range of educational, diagnostic, treatment planning, virtual reality, virtual surgeries, artistic, mathematical, legal and industrial uses by over 2300 licensees in 49 countries. The Visible Human Project has been featured in more than 850 newspaper articles, news and science magazines, and radio and television programs worldwide.

FY2006 saw the continued maintenance of two databases to record information about Visible Human Project use. The first, to log information about the now over 2300 Visible Human Project 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.

During FY2006 an extensive search of the literature was completed, and a Visible Human Current Bibliographies in Medicine (CBM) was produced. This bibliography is an attempt to identify all publications in the scientific and technical literature which discuss the Visible Human Project and its derivative products. This includes citations to journal articles, books and book chapters, conference papers and meeting abstracts, audiovisuals, and liturature discussing the use and applications of the VHP Insight Toolkit (ITK).

The Insight Toolkit (ITK), a research and development initiative under the Visible Human Project, is now in its sixth year with a recent official software release of ITK 2.9. ITK makes available a variety of open source image processing algorithms for computing segmentation and registration of high dimensional medical data on a variety of hardware platforms. Platforms currently supported are PCs running Visual C++, Sun Workstations running the GNU C++ compiler, SGI workstations, Linux based systems and Mac OS-X. Support, development. Maintenance of the software is managed by a community of university and commercial groups, including OHPCC intramural research staff. The ITK continues to have an impact on the medical imaging research community. Researchers are testing, developing, and contributing to ITK in more than 38 countries, with 1200 active subscribers to the global mailing list for the project.

Across NIH, ITK is providing a foundation for new imaging investigations. The National Alliance of Medical Image Computing (NA-MIC), an NIH Roadmap National Center for Biomedical Computing (NCBC), has adopted ITK and its software engineering practices as part

of its engineering infrastructure. NA-MIC is currently using medical imaging techniques to study the physiological sources of schizophrenia and other mental disorders. During FY2006, the NA-MIC organization held 10 user-training workshops across the country and one in Lausanne, Switzerland. In addition, ITK software engineering paractices and tools are having an impact outside the medical imaging community and are influencing some of the world's largest open-source software projects.

3D Informatics

OHPCC's 3D Informatics Program has expanded in-house research efforts around problems encountered in the world of 3-dimensional and higher-dimensional, time-varying imaging. One of our most intense efforts is our project to create *PLAWARe* (Programmable Layered Architecture With Artistic Rendering), a software framework for artistic and non-photorealistic rendering of digital models. This entails the design of a layered, software architecture for implementing medical illustration techniques using computer graphics technologies. In FY-06, the project contributed to a technical exhibition, demonstrating some of its capabilities at the 2006 ACM SIGGRAPH Conference in Boston, MA.

The 3D Informatics Group has continued work on image databases, including ongoing support for the National Online Volumetric Archive (NOVA), an archive of volume image data, as well as our continuing partnership with the NLM Specialized Information Systems Division and the U.S. Veterans Administration to study content-based retrieval methods for medical image databases. In the pharmaceutical identification project, we are assisting in the acquisition of imagery through digital macro-photography of the thousands of prescription pharmaceuticals dispensed routinely by the VA Centralized Mail-Order Pharmacies. Together we are creating a new, updated, visual database of all these products and developing techniques for automatically identifying any product in the inventory from a representative photograph. New OHPCC research has developed computer vision approaches for the automatic segmentation, measurement, and analysis of solid-dose medications.

In FY2006, the 3D Informatics group, along with representatives of NIBIB and two directorates at the National Science Foundation (NSF), sponsored two workshops on Visualization Research Challenges (VRC). The group continues to publish and submit materials for scientific review at national and international venues and continues to organize and participate in regional, national, and international conferences as invited speakers, panelists, journal reviewers, and conference organizing committee members.

DocView: Document imaging for the biomedical end-user

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

- DocView. This windows-based client software was originally released in January 1998 and subsequently improved over several generations. It is widely used by libraries to deliver TIFF documents for interlibrary loan services. It currently has 17,997 users in 195 countries, an increase of 898 new users and 2 countries over last year. In September 2006 alone, there were 65 new users spread over 22 countries registering to use DocView. Although the use of DocView is expected to decrease, the changeover is

likely to be gradual especially in foreign countries since their purchase of the new Ariel software may take longer.

- MyDelivery. The MyDelivery project is seen as a successor to DocView The goal of the project is to develop a new collaborative tool to improve the delivery and exchange of medical and health information, especially in very large files. MyDelivery provides users (biomedical researchers, administrators, librarians, physicians, patients, hospitals, and other health professionals) a fast, easy, and secure method to exchange medical information, regardless of the size of the electronic file in which it resides. The MyDelivery project seeks to overcome three significant obstacles: (1) transmission of large electronic files (i.e. document images, digitized photographs, digitized x-rays, sonograms, CT and MRI scans, and digital video) over the Internet; (2) sending files reliably and securely; and (3) complying with requirements of the Health Insurance Portability and Accountability Act (HIPAA). To solve all three problems, the MyDelivery project focuses on the development of server-based software running on a cluster of Internet-based servers, and the development of client software for use by collaborators. MyDelivery allows two client computers to exchange large files through an intermediary server via a user interface similar to email. In test conditions, the system permits the exchange of files ranging up to several gigabytes in size. Part of the development of MyDelivery has been to create a method of automatically recovering from communication failures due to reduced signal strength. This part of the project has been completed, and successfully tested over unreliable networks. Additional work centers on using public domain software for security certificate generation and use. A technique is being developed that will allow user roaming in an easy manner, without conflicting with existing patents in this area.
- DocMorph and MyMorph. The DocMorph system continued to serve both browser-based users (14,400 to date: 1900 more than last year) and MyMorph users (6500 users) this year. Most of the registered users are biomedical document delivery librarians. DocMorph allows the conversion of more than 50 different file formats to PDF, for instance, to enable multi-platform delivery of documents. Also, by combining OCR with speech synthesis, DocMorph enables the visually impaired to use library information. It has been used by librarians for the blind and physically handicapped to convert documents to synthetic speech recorded onto audio tapes for blind patrons. Most users continue to use it to convert files to PDF to enable multi-platform delivery of documents. DocMorph is available at http://docmorph.nlm.nih.gov/docmorph.

Document Image Analysis and Understanding (DIAU)

Research in DIAU is directed toward developing techniques to implement in production in line with NLM's mission. The projects in this category are MARS and its various spinoffs.

Medical Article Records System

The Medical Article records (MARS) production system has evolved through several generations of increasing capability. Its core engine consists of daemons based on heuristic rule-based algorithms that use geometric and contextual features derived from OCR output to automatically segment scanned pages of journal articles, assign logical labels to these zones, and to reformat

zone contents to adhere to MEDLINE conventions. About a quarter of the total citations in MEDLINE now are created by MARS, the remaining coming in as XML-tagged data directly from publishers.
Changes continue to be made to the MARS production system to accommodate new requirements from indexers. Three MARS software modules (Edit, Reconcile, and Upload) and the validation library have been modified to automatically extract ISRCTN clinical trials and GEO (gene expression omnibus) databank numbers, and Wellcome Trust grant numbers. In addition, changes were made to the Reconcile and Upload modules to list author and corporate author names in the order they appear in the published article.

WebMARS
Efforts continue toward meeting goals of the Indexing 2015 Initiative through the continuing development of two systems relying on WebMARS to assist both operators and indexers. Initial versions of both systems, WebMARS Assisted Indexing (WAI) and Publisher Data Review (PDR) are currently under test.

PDR will provide operators data missing from the XML citations sent in directly by publishers (such as databank accession numbers, NIH grant numbers, funding sources, and PubMed IDs of commented articles) thereby reducing the burden on operators in creating citations for MEDLINE. In addition, incorrect data sent in by the publishers can be corrected by PDR. Correcting the publisher data is currently a labor-intensive process since the operators perform these functions manually by looking through an entire article to find these items, and then keying them in.

WAI will aid indexers in their search for terms in an article that correspond to biomedical terms in a predefined list. WAI will automatically search through the text and highlight these terms for the indexer to simply confirm and select, thereby reducing manual effort. An initial prototype was demonstrated to indexers who provided feedback for improvement. A pilot version of this system is currently being tested in the Indexing Section.

ACORN
This system is intended to extract bibliographic information from 60 volumes of the printed Quarterly Cumulative Index Medicus (QCIM) from 1927 to 1956 to populate the OLDMEDLINE database. The design of the system is rooted in research in document image analysis and pattern matching techniques. With the help of NLM's Preservation and Collection Management Section, the microfilm version of a particular volume (Vol. 59, Jan – June 1956) was scanned and the TIFF images subjected to OCR conversion. Currently, a module is being created to extract journal name abbreviations from the DCMS database to compare against the abbreviations in the microfilm images.

Text to Image Linking Engine
Text to Image Linking Engine (TILE) is designed to transparently link the print library of functional-physiological knowledge with the image library of structural-anatomic knowledge into a single, unified resource for health information, a long term NLM goal. An early prototype of the modular GUI interface to the system now called *Visual PubMed* (TILE-PubMed proxy server) was completed and demonstrated. This system allows a user to search PubMed, and receive citations that are automatically augmented with anatomic images relevant to the article topic.

Research in TILE seeks the best alternatives for the functions needed to accomplish this linkage. These functions are: identifying biomedical terms in a document, identifying the relevant anatomical terms, identifying the images in the image database, and linking the identified terms to the images. Our main research focus is on the second function, the Term Mapper, which associates the biomedical terms in the document to appropriate anatomic concepts through the Metathesaurus concept relation table, and ultimately to images. Since this table typically yields several relationships that can potentially map a biomedical term to multiple anatomical concepts, relevance ranking is then applied.

Medical Article Records Groundtruth

The Medical Article Records Groundtruth (MARG) database is available for research in document image analysis and understanding techniques by the computer science and informatics communities. The data consists of over 1,000 bitmapped images of the first pages of articles from biomedical journals indexed in MEDLINE falling into 9 layout types encountered in MARS production. Included in addition to the page images are the corresponding segmented and labeled zones, OCR-converted and operator-verified data at the zone, line, word and character levels, all in XML format. Also available from this Web site, marg.nlm.nih.gov, is Rover, an analytic tool that may be used to compare the results of a researcher's program with the ground truth data. Rover has been enhanced to allow a visual comparison of researchers' algorithmic results with the ground truth data, as well as some statistical metrics. The MARG server has had over 9,688 unique IP visits from 96 countries.

Information Systems

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

Consumer Health Informatics Research

The Consumer Health Informatics research projects explore the needs, information seeking behavior, and cognitive strategies of health care consumers. The projects' principal goal is to apply medical informatics and information technologies to study ways to develop, organize, integrate, and deliver accessible health information to the members of the public at all levels of health literacy. These projects include the ClinicalTrials.gov and Genetics Home Reference Web sites and the Consumer Health Information Seeking research initiative.

ClinicalTrials.gov provides the public with comprehensive information about all types of clinical research studies, both interventional and observational. The site has over 34,000 protocol records sponsored by the U.S. Federal government, pharmaceutical industry, academic and international organizations in all 50 States and in over 130 countries. Some 47% of the trials listed are open to recruitment, and the remaining 53% are closed to recruitment or completed. ClinicalTrials.gov receives over 11 million page views per month and hosts approximately 29,000 visitors daily. Data are submitted by over 3,370 study sponsors through a Web-based Protocol Registration System (PRS), which allows providers to maintain and validate information about their trials.

ClinicalTrials.gov was actively involved in promoting the standards of transparency in clinical research through trial registration. These standards where communicated to a broad range of US and international stakeholders via presentations and printed materials. As a result of

increasing awareness of the importance of trial registration, over 12,000 new registrations were received over the last year. ClinicalTrials.gov also launched a comprehensive evaluation program, aimed at identifying and meeting user needs of various groups of ClinicalTrials.gov users. ClinicalTrials.gov continues to collaborate with other registries and professional organizations, working towards developing global standards of trial registration.

Genetics Home Reference (GHR) provides basic information about genetic conditions and the genes and chromosomes related to those conditions. Created for the general public, particularly individuals with genetic conditions and their families, the site currently includes summaries for more than 200 genetic conditions, more than 330 genes, and all the human chromosomes. On average, 10 new summaries are added per month. In the past year, GHR's content was expanded to include information about disorders caused by mutations in mitochondrial DNA. This addition required development of new features such as a circular ideogram representing mitochondrial DNA. Companion tutorial materials were also developed for the "Help Me Understand Genetics" Handbook. The new Handbook materials explain the role of mitochondrial DNA in cells and how changes in this DNA affect health. To support GHR's growing content, the site's search algorithm and display of search results were improved to help users find topics of interest. GHR's usage increased more than 60 percent in the past year, and the site is continually recognized as an important health resource.

New in FY2006 is a project that teaches first aid to Hurricane Katrina evacuees (conducted by Southern University and Louisiana State University), a weekly podcast produced in conjunction with the NLM Office of the Director, Library Operations, and OCCS, and a follow-up study of the factors that make it difficult for consumers to understand a medical text.

The consumer health informatics team continues to publish at national and international venues. The team participates in national and international conferences and helped plan NIH's ehealth national conference as well as the Surgeon General's National Meeting on Health Literacy in September. Additionally, members gave 12 invited lectures to universities and institutions around the nation and internationally.

The Consumer Health Information Seeking initiative focuses on understanding and improving access to online health information. One initiative project explores the search and navigation behavior of consumers using health information systems. Another project investigates methods for developing readability assessment metrics to evaluate health-related text intended for consumers of varying health literacy. A third project examines different approaches for using queries in one language (e.g., Spanish) to retrieve relevant documents in another language (e.g., English) to support access to health information for the Spanish-speaking community. A prototype system for providing basic information about clinical trials in Spanish is undergoing usability testing. Finally, the consumer health vocabularies project focuses on mapping words and phrases commonly used by consumers to technical medical terms and concepts.

Digital Library Research

The Digital Library Research project investigates all aspects of creating and disseminating digital collections, including standards, emerging technologies and formats, copyright and legal issues, effects on previously established processes, protection of original materials, and permanent archiving of digital surrogates. Research issues currently in focus are long-term preservation of digital archives, innovative methods for creating and accessing digital library collections, and the development of modular and open information environments. Investigations concerning

interoperability among digital library systems, the role of well-structured metadata, and varying "points of view" on the same underlying data set are also being pursued.

The Profiles in Science® digital library uses innovative digital technology to showcase digital reproductions of items selected from the personal manuscript collections of prominent biomedical researchers, medical practitioners, and those fostering science and health. The content of Profiles in Science is created in collaboration with NLM's History of Medicine Division (HMD), which processes and stores the physical collections. Most collections have been donated to NLM and contain published and unpublished materials, including manuscripts, diaries, laboratory notebooks, correspondence, photographs, journal volumes, poems, drawings, audio tapes and other audiovisual resources. The collections of Edward D. Freis, Virginia Apgar, and Michael Heidelberger were added this year. An additional 825 digital items composed of 6,500 image pages were also added to existing Profiles in Science collections. Presently the Web site features the archives of nineteen prominent scientists: Christian B. Anfinsen, Virginia Apgar, Oswald T. Avery, Julius Axelrod, Francis Crick, Donald S. Fredrickson, Edward D. Freis, Michael Heidelberger, C. Everett Koop, Joshua Lederberg, Salvador E. Luria, Barbara McClintock, Marshall W. Nirenberg, Linus Pauling, Martin Rodbell, Florence R. Sabin, Wilbur A. Sawyer, Fred L. Soper, and Albert Szent-Györgyi. 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.

During this fiscal year, protocols for improving the quality of scanned images were developed and successfully used to disable automatic de-skewing of images and cleanup speckled color/greyscale images. An early digital library, the Regional Medical Programs collection, was fully integrated into Profiles in Science. MeSH terms in use by the project were analyzed; obsolete terms were translated to MeSH 2006 terms, and "discontinued" MeSH terms were noted for future analysis. New queries and statistical reports were developed and implemented to detect unexpected patterns throughout the Profiles in Science data. Development of a new Profiles in Science XML-based Web front end and transition to a new XML-based search engine, particularly development of the Annotations Server utilizing this new architecture, is ongoing.

MEDLINE Database on Tap

MEDLINE Database on Tap (MDoT) seeks to discover and implement systems and techniques to assist mobile clinicians in quickly finding relevant, high quality information addressing clinical questions that arise at the point of care. The primary goal is to present information to users so that they can quickly find the most pertinent parts, despite the limitations placed by the small screen and restricted bandwidth of handheld computers. MDoT explores display and navigation techniques, as well as information organization and content. MDoT also incorporates tools and systems from other LHNCBC projects, such as MetaMap and the Essie search engine. Essie and Google are offered as options, while the primary search engine is PubMed. We also seek to integrate semantic data from the search query and found citations to optimally rank results while maintaining real time response. A testbed system that supports MEDLINE search and retrieval from a wireless, Internet-connected PDA has been developed. Our client software for Palm OS and Pocket PC OS are freely available from the MDoT Web site, http://mdot.nlm.nih.gov/proj/mdot/mdot.php, which experiences between 5000 and 6000 hits every month. The Web site provides information about the project as well as the software, and

allows us to solicit feedback from users and monitor aggregate user behavior. There are over 500 registered users of MDoT, and an unknown number of unregistered ones.

In the past year, MDoT has been evaluated in clinical settings at two institutions, the University of Hawaii and the VA Medical Center in Washington, D.C. In the first, medical residents enrolled in a Medical Informatics elective accompanied medical teams on morning rounds for 4 weeks, using MDoT to seek answers to clinical questions that arose at the point of care. They submitted daily summaries of each scenario and question in 44 rounds of about one hour each, with 187 clinical questions. Using a variety of MDoT options, they found relevant citations for 153 (82%) of the questions. This evaluation was observed by members of the MDoT team, who also conducted a number of workshops throughout the state, providing an overview of the system and hands-on training.

The second evaluation was conducted at the VA Medical Center (VAMC) in Washington, D.C. In this three-part collaboration among NIH/NLM/LHNCBC, the VAMC, and Universidade da Beira Interior, Portugal, a sixth year medical student rounded for 20 days with four different teams, using MDoT to search for answers to questions that arose on rounds. She recorded 144 clinical questions that were asked in context of 78 in-hospital clinical scenarios and 17 topic reviews. Because the VAMC is equipped with WiFi throughout, a significant difference between this study and the Hawaii study, in which residents used PDA/cell phones, is network data rate. One question to address is whether this higher data rate has a notable effect on the ability to find useful information at the point of care. Results showed that answers were found in MEDLINE for 73% of the questions that arose on rounds, an unexpectedly high figure. On average, there were 5.2 queries per question, and less than 4 minutes per question was spent finding relevant citations. Results show that MDoT is fast and effective.

The MDoT evaluation plan calls for a “second opinion” of the selected citations by a senior investigator and expert MEDLINE indexer at LHNCBC. Based on a review of the scenario and clinical question, each selected citation is assigned a score of A (answers the question), B (contains a partial answer or is topically relevant and clearly indicates that the full text might answer the question), or C (does not answer the question).

As part of the MDoT project, outcomes research was conducted toward automatically finding patient outcomes (e.g., the population under study) from MEDLINE citations using knowledge extractors that rely upon NLM Unified Medical Language System and tools. The Extractor system identifies an outcome and determines whether a found outcome pertains to the topic of interest, the type of treatment studied, and the quality of the study.

Interactive Publications Research

The goal of this project is to create a comprehensive, self-contained and platform-independent multimedia document that is an interactive publication (IP), and to evaluate its value for better comprehension and learning. Following a study of existing open source formats and standards, a prototype interactive document was created containing many media objects: text, dynamic tables and graphs, a microscopy video of cell evolution, an animated spine in Flash, digital x-rays, and clinical DICOM images (CT, MRI, ultrasound). Both self-contained (embedded) and folder-type (linked) documents using all these media types were created in four formats: MS Word, Flash, HTML and PDF. The IPs in these formats were compared in terms of ease of use and development effort.

While using such a document, the reader is able to: (a) view any of these objects on the screen; (b) hyperlink from one object to another; (c) interact with the objects in the sense of

exercising control over them (e.g., start and stop video); (d) and importantly, reuse the media content for analysis and presentation. In light of the large sizes of such publications possibly in the range of hundreds of megabytes, research is ongoing toward identifying techniques and protocols for rapid progressive download of the publications, and the development of a Download Manager based on this research.

To demonstrate the value of large tabular ("raw") datasets in an IP, some published articles were acquired from the American Psychiatric Institute for Research and Education, as well as the datasets underlying the tables appearing in the articles. The Institute also sent SAS scripts coding questions to the raw data. The datasets were loaded in SAS as well as CSV forms, and efforts are under way in linking the raw data to the published tables, and in creating hypothetical questions about specific age group and diseases that a reader might have (but which are not directly addressed in the paper). One of the articles is in the process of being converted to an interactive form.

NLM Gateway

The NLM Gateway provides an easy to use one-stop search method that allows users to issue simultaneous searches in a number of NLM information resources from a single interface. The current version interacts with eight NLM search systems that provide results from 23 information resources. Changes to the underlying data structures or to the targeted search systems are carefully tracked and the Gateway modified accordingly. An example is the NLM Gateway release of October 2005 in which access to 100,000 meeting abstracts and health services research records was changed from the former Verity system to the new SE (Search Engine) system developed by LHNCBC staff.

While databases accessed by the NLM Gateway are regularly (sometimes even daily) updated, other resources incorporated into the Gateway itself are also regularly updated. New releases of the Unified Medical Language System (UMLS) Metathesaurus, the UMLS mapping file, the 2006 MeSH update, and Year End Processing were incorporated during the year as they became available.

A new version of the Gateway accessing five additional toxicology-related resources was brought on line in November 2005. Later, changes in searching of the meeting abstracts retrieved search results from standardized XML data, allowing the display of diacritic marks and of additional fields from the records. In May 2006, the Bookshelf, a growing collection of full text biomedical books and other resources, was made accessible through the NLM Gateway. Access to the Household Products Database was also added. This database is a consumer's guide providing information on the potential health effects of chemicals contained in more than 5,000 common household products. In August 2006, access to the Profiles in Science collection was added. Profiles in Science contains archival collections of leaders in biomedical research and public health.

Feedback from users and statistical analysis of user actions helped to inform the planning of new functionality and new display options. Usability testing in the coming year will help in the testing of new ideas for facilitating user input and for creating better displays of system output. The intent is to create user-focused portals that help various categories of users quickly find what they need.

Digital Preservation Research

This project aims to investigate key issues related to the long term preservation of digital material, both digitized documents and video. Our work in document preservation has matured and focuses on two processes: automated metadata extraction and file migration.

For document preservation, a prototype System for Preservation of Electronic Resources (SPER) was developed. SPER is a flexible, modular system that demonstrates key functions such as ingest, automated metadata extraction (AME) and bulk file migration. AME is implemented for the extraction of descriptive metadata from scanned and online journal articles as well as NLM's obsolete Web pages. Bulk file migration is implemented through an existing CEB system, DocMorph. While these functions are developed in-house, for the necessary infrastructure capabilities in SPER we have incorporated into the system, and customized, the latest version (1.4) of MIT's open source DSpace software. The Java client GUI for SPER was enhanced to incorporate batch metadata extraction and ingest for journal article TIFF pages, online journal articles and NLM Web pages (HTML). The GUI was also redesigned to display Web pages and online articles through Java Swing components.

SPER, in an abbreviated form, is being used in the preservation of a new collection at NLM consisting of over 65,000 historical FDA court records. Since the manual identification and entry of descriptive metadata from these records is labor-intensive, our focus is on automated extraction. In collaboration with the curator for this collection, we identified more than a dozen metadata items which could be extracted automatically. Our approach consists of: scanning the paper documents, auto-zoning the TIFF files using OCR output from the scanned documents, feature extraction, optimal feature selection, feature classification using a Support Vector Machine (SVM) classifier, multi-class probability estimation, and statistical parsing using the Stolcke-Earley parsing algorithm.

Infrastructure Research

The Lister Hill Center performs and supports research in developing and advancing infrastructure capabilities such as high-speed networks, nomadic computing, network management, wireless access, and improving the quality of service, security, and data privacy.

Advanced Biomedical Tele-Collaboration Testbed

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

In FY2006, the research team completed the substantial infrastructure required to test the scenarios. Several successful wide area wireless demonstrations of transmitting video and other patient data from ambulances using 3G and mesh cellular technology have been completed.

3D Telepresence for Medical Consultation

This project tests the efficacy of 2D versus 3D representations of video data transmitted in real time in remote clinical consultations. Although the research design could be undertaken without reference to the underlying computer algorithms for acquiring, transmitting, and displaying real time 3D video, the refinement and instantiation of these algorithms and related procedures for camera calibration, head tracking, and display in viable, transparent, user friendly 3D collaboration environments is the ultimate goal.

The research team made substantial progress in implementing the technology infrastructure. A prototype portable camera unit was added to the stationary one and calibrated. The PDA application was completed and all the basic components of the system proposed are in place. The current focus is on optimizing camera and sensor placement, refining calibration and rendering algorithms, and dealing with problems when perspective changes from different points of view, such as occlusion when an intervening object obstruct the view of interest. In addition, the team started experimenting with stereo displays of the 3D data rendered, and progress has been made in collecting data comparing the performance of paramedics in a simulation center working alone, working with a distant 2D video consultation, and with a 3D proxy consultation.

Scalable Information Infrastructure Initiative

NLM's Scalable Information Infrastructure (SII) Initiative is designed to establish testbed applications that demonstrate advanced network capabilities in health care, medical decision-making, public health, health education or biomedical, clinical or health research within the broad research agenda of the NLM. SII projects involve the use of testbed networks linking one or more of the following: hospitals, clinics, practitioners' offices, patients' homes, health professional schools, medical libraries, universities, research centers and laboratories, and public health authorities. Among the applications:

- Wireless Internet Information System for Medical Response in Disasters (WIISARD) at the University of California, San Diego, the Advanced Network Infrastructure for Distributed Learning and Collaborative Research at the Stanford University School of Medicine, and the National Multi-Protocol Ensemble for Self-Scaling Systems for Health at Boston Children's Hospital.
- The Project Sentinel Collaboratory is a partnership involving Georgetown University and the Washington Hospital Center, among others. The project is tasked with building and deploying a data-centric Collaboratory to collect and analyze data from hospitals, clinics, weather services, satellite images of vegetation, mosquito collection, veterinary clinics and other sources in order to develop indicators and warnings of emerging threats to human health. During FY2006, all project infrastructure was completed and the hospital systems were linked. With the infrastructure completed and the advanced data visualization tools in place, the project team will continue to analyze Collaboratory data and explore open source strategies and methodologies in support of flexible access to biomedical data.
- SMART (Scalable Medical Alert and Response Technology) is a system for patient tracking and monitoring from the emergency site that continues through transport, triage,

and transfer from external sites to the health care facility within a health care facility. The system is based on a scalable location-aware monitoring architecture, with remote transmission from medical sensors and display of information on personal digital assistants, detection logic for recognizing events requiring action, and logistic support for optimal response. Patients and providers, as well as critical medical equipment will be located by SMART on demand, and remote alerting from the medical sensors can trigger responses from the nearest available providers. The emergency department at the Brigham and Women's Hospital in Boston will serve as the testbed for initial deployment, refinement, and evaluation of SMART. This project will involve a collaboration of researchers at the Brigham and Women's Hospital, Harvard Medical School, and the Massachusetts Institute of Technology.

- The Tele-Immersive System for Surgical Consultation and Implant is aimed at developing a networked collaborative surgical system for tele-immersive consultation, surgical pre-planning, implant design, post operative evaluation and education. The Personal Augmented Reality Immersive System (PARIS) has been developed, tested, and displayed publicly. The PHANTOM haptic device has been installed on the PARIS system. A Linux PC is used to drive the PARIS system. The PC controls two display devices at the same time. One is the projector on PARIS to display 3D stereo models. The other is an ordinary monitor to display the 2D user interface. The separation of the user interface and the sculpting working space allows much easier and smoother access to different functions of the application. The Physician's Personal VR Display was developed to facilitate consultation from the physician's desk without requiring that the physician go to a specialized facility. This system allows surgeons to do remote pre-operative consultation and post-operative evaluation, as the system enables all participants in a collaborative session to share their viewing angle, transformation matrix, and sculpting tools information over the network.

Wireless PDA Pubmed Searching

Short Message Service (SMS) use in medicine is increasing with applications in monitoring of patients with chronic illnesses, appointment reminders and patient-doctor communications. Txt2MEDLINE is an application that provides access to MEDLINE/PubMed through SMS. A Web version of the application was presented at the 2006 American Telemedicine Association Annual Meeting in San Diego. Several clinical evaluations of the *txt*2MEDLINE application are being done, including at the Prince Georges Health Center and at the University of the Philippines, to evaluate the application's effectiveness and usefulness in clinical practice.

Telemedicine Initiatives

OHPCC participated in talks and demonstrations of state-of-the-art Telemedicine, e-Health projects at the Dirksen and Hart Senate Office Buildings. The congressional Steering Committee on Telehealth and Healthcare Informatics sponsored the talks, demonstration and roundtable discussion as part of its 2006 Telehealth, e-Health, and healthcare informatics projects and programs designed to address pressing healthcare issues. This program is intended to inform Members of Congress and congressional staff, federal agency officials, healthcare and technology organization representatives, and the public. The NLM/OHPCC display featured wireless handheld access to PubMed.

A “Virtual Microscope” website, http://erie.nlm.nih.gov/~slide2go was created to present the progress of the project. Teaching slides from the Department of Pathology medical student collection were digitized and archived work is available online at http://images.nlm.nih.gov/pathlab.

Videoconferencing and Collaboration

Major renovations were undertaken in FY2006, including rear screen and stereo projection and the installation of an Extron switch enabling multiple computing and video sources in the Collaboratory to be directed to various displays. The H.323 videoconferencing system and multipoint conferencing unit (MCU) were upgraded to include a new Tandberg room system with multipoint conferencing, an improved H.264 video codec, and H.239 capabilities for application sharing. Major upgrades were made to the Access Grid (AG) and Conference XP computers and an Access Grid venue server was installed. A dual camera configuration of the AG was developed for 3D stereo video. Several software programs were installed and configured so the team could become familiar with newer collaboration tools. Web and streaming servers were upgraded.

A distance learning program in collaboration with SIS, coordinator of NLM’s Adopt-A-School Program, continued to provide on-site and distance education about varied health science topics and information sources to students at the King Drew Medical Magnet High School affiliated with the Charles R. Drew University of Medicine and Science in Los Angeles. The link from a previous funded NLM funded telemedicine study connecting the school to the University was re-activated to connect the school to Internet2. Programmatically, it eliminated the logistical problems of having to move students from the school to the university, enabled hands on learning experiences in the school’s computer lab, and allowed more classes to participate. The new Tandberg system at the Collab and a new Polycom system at the school with identical capabilities improved the quality of the communication and the recordings made of the sessions considerably. The NIH Office of Science Education participated in the program and conducted several sessions on health science careers.

The webcasts of the bi-monthly Washington Area Computer Assisted Surgery Special Interest Group continued and videoconferencing was added so that there is now two way interaction between those attending the meeting in the Lister Hill auditorium in Bethesda, where the presentations are made, and those in an auditorium at the Allegheny Hospital System in Pittsburgh. Attendees are now able to obtain continuing medical education credits because of this linkage. The team continues to do work with NCBI and the University of Puerto Rico to resolve technical problems in delivering distance education related to NCBI’s information sources. Methods for providing application sharing and image manipulation with low latency have been identified and substantial progress has been made in enabling the instructor at NLM to view each remote student’s desktop.

The Center for Public Service Communication (CPSC) was given a contract to pilot the use of video over IP to provide remote medical interpretation services. CPSC was open to collaborating with the team to do more formal assessment of the technology. As a result, team members have worked with CPSC staff to install and provide training on the technology in public health clinics in Florida and to develop the research methodology.

Language and Knowledge Processing

The Lister Hill Center conducts and supports research in language and knowledge processing to extract usable and meaningful information from biomedical text. This research covers advanced library and terminology services, modeling and learning methods, medical ontologies, indexing initiative, and semantic knowledge representation.

Advanced Library Services

Advanced biomedical information management applications exploit online information to support evidence-based medicine, enable scientific discovery, help translate discoveries into advances in patient care, and provide the basis for individual decision making. Some of the potentially exploitable information available online is in the form of text and is readily accessible only to humans; examples include Medline citations and associated full-text articles, ClinicalTrials.gov, and clinical narratives. Other online information is structured and includes biomedical vocabularies, clinical and molecular biology knowledge bases, and model organism annotation databases. The objective of the Advanced Library Sevices (ALS) project is to normalize and integrate biomedical information, both text-based and structured, into a repository of executable knowledge, a Biomedical Knowledge Repository (BKR), directly accessible to advanced applications including knowledge discovery, multi-document summarization, and question answering.

This project was launched during FY2006 and recently presented to the Board of Scientific Counselors. Two pilot projects were developed as a proof of concept. The gene information resource Entrez Gene was integrated into the Biomedical Knowledge Repository by converting it from XML representation to the Resource Description Framework (RDF). And Semantic Medline, an application that extracts knowledge from selected MEDLINE citations, creates a visual representation (graph) of salient assertions in those documents, and allows users to manipulate the graph interactively.

Work has begun to extract knowledge from the entire collection of documents in MEDLINE, as well as from structured databases, including the UMLS, and to include metainformation to the BKR. Future applications that exploit the repository will focus on a medical subdomain (e.g., cardiovascular diseases) and be based on user input.

Terminology Research and Services

LHNCBC research staff build and maintain the SPECIALIST Lexicon, a large syntactic lexicon of medical and general English that is released annually with the Unified Medical Language System (UMLS) Knowledge Sources. New lexical items are continually added using a lexiconbuilding tool; the SPECIALIST lexicon contains over 330,000 records. The UMLS Lexical tools, including lexical variant generator (LVG), wordind, and norm are distributed with the UMLS as are text processing tools which analyze documents into sections, sentences, and phrases. The SPECIALIST lexicon, lexical tools, and text processing tools are released as open source resources and available under an unrestrictive set of terms and conditions for their use. LexBuild is a lexicon building tool designed to aid the lexicon building team by facilitating entry of lexical information and providing real time quality control. It is updated and revised on an ongoing basis. The SPECIALIST lexicon release tables are annually generated using the LexBuild tool. The SPECIALIST lexicon and tools are UTF-8 compliant and capable of dealing

with non-ascii characters. MMTx, the Java implementation of the MetaMap algorithm is a major application of the SPECIALIST lexical and text tools. A stochastic part-of-speech tagger is being developed for use in MMTx. The tagger will be specifically designed to exploit the SPECIALIST lexicon and will allow tagging of multi-word terms from the lexicon. It will be released as a freely available open source tool. LNHCBC researchers are engaged in porting the Journal Descriptor Indexing (JDI) tool to Java for future release as part of the UMLS lexical tools. The JDI should provide an element of context that can be useful for word sense disambiguation and other natural language processing tasks.

LHNCBC research staff also develop and maintain the UMLS Knowledge Source Server (UMLSKS) that provides Internet access to the UMLS knowledge sources through application programs and a user interface. UMLSKS is updated quarterly to accommodate quarterly UMLS releases. A grid/web services implementation of the UMLSKS backend and an implmentation of the user interface as a portal consisting of user-chosen “portlets” representing different parts and views of the UMLS data have been developed and will soon be released.

The goal of the Terminology Server (TS) project is to provide tools and data to manage diverse medical vocabularies for diverse purposes. Over the past year, the project continued to provide customized data sets using the released versions of the UMLS to several projects such as Clinical Trials and Genetic Home Reference for use in their operational environments. An important function of the TS is to support the customization of terminologies from the UMLS and other sources to satisfy individual project needs. A number of internal tools were developed to handle the data customization needs of the projects identified above, which resulted in periodic releases of data sets containing customized data. One new set of tools and processes added to the TS handles the generation of English-Spanish translation tables for the Spanish version of Clinical Trials. In addition, significant work was done on generating more efficient processing operations of existing vocabulary mapping algorithms, and producing the mapping tables for the current version of the UMLS Metathesaurus data.

The project has continued to develop a set of tools that allow users to do their own data customization and management. In addition, the project will continue to integrate tools with existing applications and provide updates to the application data sets corresponding to the latest releases of UMLS data and other relevant data sets.

A multilanguage search tool for non-English speakers for MEDLINE/PubMed, http://babelmesh.nlm.nih.gov, is continuing. It allows healthcare providers and researchers to search in their native language. Through international collaborations, including WHO Eastern Mediterranean Regional Office in Cairo, more vocabularies were added. With the multilingual search portal, users can now search in Arabic, French, German, Italian, Japanese, Portuguese, Russian, Spanish and English. Comments from international health liaison officers, including DHHS, were very encouraging after a presentation at a Fogarty International Center meeting.

PICO (Patient, Intervention, Comparison, and Outcome) Linguist, http://babelmesh.nlm.nih.gov/pico.php, is an application allowing users to search MEDLINE/PubMed in a more clinical and evidence-based manner. This work is significant because it is the only cross-language search portal on the Internet that allows the input in more than 2 languages. It is also unique because it allows the user to search in character-based languages (non-Latin alphabet), transform it to an English language search and retrieve citations published in any language or language combination. Full-text articles may be linked to the result if published online and available without subscription requirements.

Modeling and Learning Methods

The Modeling and Learning Methods project is aimed at developing computational learning methods to enable scientists to utilize crossdisciplinary information effectively. Crossdisciplinary scientific information, associated always with uncertainty, comes with a multitude of overlapping but unidentical and sometimes conflicting perspectives. In order to cope with such information overload, scientists need assistance from computers and translate their mental models to computational models. Even with the state-of-art computational tools, this translation is often very difficult due to the vagueness of the available information and its questionable reliability, forcing scientists to make artificially restrictive assumptions about the nature of their domain and information.

This project develops an information architecture called multifaceted ontological networks (muON), which is designed to cope with the aforementioned problems. In muON, every perspective of scientific information is captured in a different facet, which may overlap with other facets. Unlike the other ontological approaches, muON can cope with uncertainty via its underlying representation method called parameter interdependency networks (PIN).

PIN, a graphical modeling method being developed as part of this project, is based on probability theory and machine learning. The development and refinement of PIN are being driven by the needs of biomedical studies (e.g., Framingham Heart Study), of which parametric requirements directly determine the design, specifications and representational capabilities of the method.

Representing linguistic information on muON is also being studied in the context of information identification and labeling. The most fundamental information identification and labeling process in computational linguistics is tokenization. Each tokenizer makes a particular set of assumptions, which frequently fail, and the resulting errors are propagated to the subsequent steps of information processing. Experiments with different tokenization methods have strongly supported the necessity of the muON paradigm since it can preserve information in its entirety by representing both agreements and disagreements of different tokenizers concurrently.

Medical Ontology Research

While existing knowledge sources in the biomedical domain may be sufficient for information retrieval purposes, the organization of information in these resources is generally not suitable for reasoning. Automated inferencing requires the principled and consistent organization provided by ontologies. The objective of the Medical Ontology Research project is to develop methods whereby ontologies can be acquired from existing resources and validated against other knowledge sources. Although the UMLS is used as the primary source of medical knowledge, OpenGALEN, the Gene Ontology, and the Foundational Model of Anatomy are being explored as well.

During this fiscal year, the research team focused on relationships in biomedical ontologies. From a formal perspective, we studied dependence relations in the Medical Subject Headings (MeSH), showing how they correlate with statistical relations. While most efforts in biomedical ontology focus on organizing concepts, we analyzed how relationships from the UMLS Metathesaurus relate to relationships in the Semantic Network, paving the way for the development of an ontology of relationships.

Work was pursued on two particular domains: anatomy and molecular biology. New methods were developed for aligning anatomical ontologies, including complex rules to map

groups of concepts. The Foundational Model of Anatomy was converted from its frame-based representation to the description logic language OWL - the Web Ontology Language used in Semantic Web applications. Similarly the gene information resource Entrez Gene was converted to the Resource Description Framework (RDF) in order to integrate it with other resources. Finally, semantic similarity in the Gene Ontology was used to compute similarity between genes and the results were used in several evolutionary biology studies.

The research team continues to work on the creation of an ontology of relationships as it is one critical element of a repository of biomedical knowledge supporting knowledge discovery and reasoning. Future work includes enhancing RxNav, the interface to the drug vocabulary RxNorm, integrating it with other drug information resources. We continue to participate in the progress of the Semantic Web for Health Care and Life Sciences and collaborate with leading ontology centers, including the National Center for Biomedical Ontology.

Indexing Initiative

The Indexing Initiative project investigates language-based and machine learning methods for the automatic selection of subject headings for use in both semi-automated and fully automated indexing environments at NLM. Its major goal is to facilitate the retrieval of biomedical information from textual databases such as MEDLINE. Team members have developed an indexing system, Medical Text Indexer (MTI), based on three fundamental indexing methodologies. The first of these calls on the MetaMap program to map citation text to concepts in the UMLS Metathesaurus. The second approach, the trigram phrase algorithm, uses character trigrams to also map text to Metathesaurus concepts. Finally, the third method uses a variant of the PubMed related articles algorithm to find previously indexed articles that are textually related to the input and then use some of the MeSH headings used to index them. Results from the three methods are restricted to MeSH, if necessary, and combined into a ranked list of recommended indexing terms.

The MTI system is in regular use by NLM indexers to create indexing terms for MEDLINE. MTI recommendations are available to them as an additional resource through the Data Creation and Maintenance System (DCMS). In addition, the indexing terms produced by MTI are being used as keywords to access collections of meeting abstracts via the NLM Gateway. These collections include abstracts in the areas of AIDS/HIV, health sciences research, and space life sciences. The Indexing Initiative staff continues with research efforts designed to improve MTI's accuracy by adding complementary methods of word sense disambiguation (WSD) to the existing facility for reducing MetaMap ambiguity. They have also begun research to extend MTI's recommendations from unqualified MeSH headings to heading/subheading pairs.

Journal Descriptor Indexing

The Journal Descriptor Indexing (JDI) project investigates a novel approach to fully automated indexing based on NLM's practice of maintaining a subject index to journal titles using a set of 122 MeSH terms known as JDs (journal descriptors), that correspond to biomedical specialties. JDI was used as a broad filter to extract from a ten-year MEDLINE text collection of 4.59 million records, those likely to be of genomics interest (39% of the collection), as part of the NLM participation in TREC (Text Retrieval Conference) 2004.

Project staff also developed an algorithm used in a MeSH gene matcher program that contributed to the NLM TREC 2005 (Text Retrieval Conference) effort. This program takes as

input names of genes in the topics for the TREC 2005 ad hoc retrieval task and returns MeSH preferred terms and synonyms from 2004 MeSH, thereby functioning as a query expansion tool for query genes. The program was modified to return additional synonyms created in 2005 MeSH.

Current work involves efforts to use semantic type indexing based on Journal Descriptor Indexing for disambiguation in the MetaMap system, to produce a Java version of the JDI system including the semantic type indexing component to be distributed as an open source tool with the UMLS Natural Language Processing tools, participation in a study on just-in-time answers to clinical questions using MDoT on PDAs, and assistance in creating test data for the full-text collection against which retrieval tasks performed by participants in the Genomics Track of TREC 2006 are to be run.

Project staff continue to collaborate with researchers at Rouen Medical School and at the University and Hospitals of Geneva, who will perform the evaluation that compares automatic assignment of Metaterms by the CISMeF (Catalog and Index of French Language Health Resources on the Internet) system versus automatic assignment of Journal Descriptors by the JDI system, against the human gold standard finalized in May 2006. The group is also participating in a research project to enhance NLM's Medical Text Indexer (MTI) to append MeSH qualifiers automatically to MeSH headings, using automatic indexing rules.

Unified Medical Language System

The mission, scope, and content of the Unified Medical Language System (UMLS) Metathesaurus continued to grow and evolve in FY2006. Most of the UMLS Metathesaurus group efforts have gone into continuing UMLS production operations while undergoing the transition of production operations from LHNCBC to OCCS and LO. A status report of the transition process for all phases following vocabulary inversion was presented to the Board of Regents at its September 2006 meeting. It is anticipated that transition of these phases will be completed in FY07, while transition of vocabulary inversion and the LO transition continue.

The third quarterly release of the UMLS Metathesaurus in calendar 2006 contains more than 1.3 million concepts (an 18% increase over its predecessor) and 6.4 million concept names (a 20% increase). There are more than 100 contributing source vocabularies. The UMLS provides the only way for the US health care community to obtain SNOMED CT, the largest HIPAA standard clinical vocabulary, under the US government license.

The format and content of the underlying biomedical vocabulary files varies widely. Without unifying standards or common tools, it is difficult to understand and use any single vocabulary, and far more difficult to integrate multiple combined vocabularies. The UMLS customization and installation tool, MetamorphoSys, allows the selection of desired content from the Metathesaurus and generates the desired subset in Rich Release Format (RRF) or Original Release Format. MetamorphoSys includes an improved Rich Release Format (RRF) Browser which allows users to view their own subsets in both Raw Data and Concept Report views. This means that any vocabulary in RRF may be reviewed, studied, or compared with views in other applications. This will make it much easier for users to make and then to see, understand, and verify their chosen Metathesaurus subsets in their own applications.

The Rich Release Format contains additional information allowing exact attribution of the sources for all its information. This allows specific mappings between vocabularies, correct inclusion and exclusion of specific sources, and simultaneous representation of a consistent UMLS view along with each source's own view, which may differ.

New development goals for MetamorphoSys include XML output, Section 508 compliance, a mapset browser, multiple instances to compare vocabularies, and a search by code function. In addition, the LHNCBC Metathesaurus group continues to work to refine and promote two standards for vocabulary exchange and UMLS submission: the Rich Release Format (RRF) and the new single vocabulary Terminology Representation and Exchange Format (TREF).

Semantic Knowledge Representation

Innovative applications for providing more effective access to biomedical information depend on reliable representation of the knowledge contained in text. The Semantic Knowledge Representation project develops programs that extract usable semantic information from biomedical text by building on existing NLM resources, including the UMLS knowledge sources and the natural language processing tools provided by the SPECIALIST system. Two programs in particular, MetaMap and SemRep, are being evaluated, enhanced, and applied to a variety of problems in biomedical informatics. MetaMap maps noun phrases in free text to concepts in the UMLS Metathesaurus. SemRep uses the Semantic Network to determine relationships asserted between those concepts.

The MetaMap Technology Transfer program (MMTx) is an exportable, Java-based version of MetaMap that runs under Windows, Mac OS X or Unix/Linux and is provided as a resource to the bioinformatics community. MMTx allows users to exploit the UMLS MetamorphoSys program to exclude or reorder the Metathesaurus vocabularies that MMTx uses. Users can also create MMTx data files independent of the UMLS, and the inclusion of source code with each release allows additional control of processing.

The development of SemRep is based on viable strategies for effective natural language processing and underpins foundational investigations in biomedical information management. At the core of this research is enhancement of linguistic coverage, and SemRep was recently expanded to address pharmacogenomics text. Syntactic and semantic mechanisms were added to accommodate a range of semantic relations, including genetic (gene-disease), genomic (gene-gene), and pharmacogenomic (drug-gene, drug-genome); in addition, relations between genes and population groups and pharmacological relations (drug-disease, drug-pharmacological effect, drug-drug) are now identified.

Semantic predications produced by SemRep serve as the basis for continued work in biomedical information management. Application areas include automatic abstraction summarization and visualization of text from Medline and ClinicalTrials.gov as well as cross-language summarization and question answering. Current project efforts concentrate on exploiting basic research for constructing practical applications. A recent application, Semantic Medline, integrates PubMed searching, advanced natural language processing, automatic summarization, and visualization into a single Web portal. Semantic Medline is intended to help users manage the results of PubMed searches by normalizing core assertions in the citations retrieved. These normalized forms constitute computable knowledge accessible to further manipulation, including condensation by automatic summarization. The normalized and condensed output of Semantic Medline is visualized as an informative graph with links to the original Medline citations. Convenient access is also provided to additional relevant knowledge resources, such as Entrez Gene, the Genetics Home Reference, and Unified Medical Language System Metathesaurus.

Multimedia Visualization

The Lister Hill Center performs extensive research and development in the capture, storage, processing, retrieval, transmission, and display of multimedia biomedical data. Multimedia products include high quality video, audio, imaging, and graphics materials.

Turning The Pages Information Systems

The Turning The Pages Information Systems (TTPI) project brings rare books at the NLM to public view in a compelling way: as photorealistic volumes whose pages may be virtually 'touched and turned.' Visitors to the Library may experience this on kiosks, and those offsite may view the books online. The TTPI project investigates ways to efficiently produce and distribute the TTP books through the Web while maintaining high quality.

Originating as collaboration with the British Library in producing two virtual books, Blackwell's 18th century *A Curious Herbal* and Vesalius' 16th century Anatomy book, we have made significant improvements on the original process. Our process consists of scanning the pages and book cover, enhancing these high quality color images by Adobe Photoshop, and creating animated 3D wireframe models of the pages using Alias Maya run on a computer by Macromedia Director software and displayed on a touchscreen monitor in kiosks. The library patron may 'touch and flip through' each of these books in an intuitive manner that evokes the feel of a 'real' paper volume.

In creating the 3D model using Maya, each pair of page images is texture-mapped to both sides of the wireframe model of a turning page, with a multisource lighting model that provides attractive diffuse lighting, specular highlights and shadows. For each flip, 12 intermediate animation frames are generated and rendered, and then imported into Director.

Three additional books from NLM's historic collection have been added for a current total of five books in TTP form: Paré's surgical treatise, Gesner's *Animalium*, possibly the earliest book in zoology, and Johannes de Ketham's *Fasiculo de Medicina* (1494). A sixth book is being prepared: Robert Hooke's *Micrographia*, the first book written about microscopes and in which reportedly the first time the word 'cell' was used. New technical challenges in converting this book include fold-out pages and the possible inclusion of images of historic and present day microscopes.

Visible Proofs Exhibition Preview DVD

In conjunction with the Office of Communications and Public Liaison (OCPL), and the History of Medicine's Exhibition Program (HMD), APDB produced an event video featuring the new NLM exhibition, "Visible Proofs: Forensic Views of the Body." APDB provided pre-production planning, thematic treatment, and a production schedule to achieve a target delivery date for the overview video in time to be presented at the NLM's Board of Regent's meeting. Interviews were conducted remotely and included Barry Scheck, JD, Innocence Project, New York City, NY; Marciello Fierro, MD, Virginia Medical Examiner, Richmond, VA; Maryland Medical Examiner, Baltimore, MD; Stephen Sherry, Staff Scientist, NCBI, NLM. The DVD has an original, animated opening sequence as well as interstitial animations to support the DNA-based forensic science themes within the exhibition.

In addition, APDB produced a high definition (HD) video detailing the NLM major program accomplishments over the last year. Medical illustrators created 3D animations that brought clarity, understanding, and visual impact to these programs. An APDB producer

compiled the research materials and coordinated with the spokespersons for each of the featured programs. On-camera interviews of the spokespersons were videotaped and, with the research information, incorporated into a production script that highlighted the programs and participants. Dr. Lindberg's comments regarding the programs and the accomplishments were also recorded and incorporated into the program. The video was shown at the May Board of Regents Dinner. Also, APDB provided project management support and HD video recording of several events including the Information Prescription press event held in Naples, Florida, and the NLM Diversity Council-sponsored Artificial Body Parts Symposium.

Training Opportunities at the Lister Hill Center

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

In FY2006, the Center provided training to 46 participants from 13 states and 9 countries. Participants worked on research projects including medical image processing, consumer health informatics, document analysis, grid computing, information retrieval, machine learning, medical illustration, micro-pathology, medical terminology research, natural language processing, medical ontology research, telemedicine, and ubiquitous computing. The program maintains its focus on diversity through participation in programs supporting minority students, including the Hispanic Association of Colleges and Universities (HACU) and the National Association for Equal Opportunity in Higher Education (NAFEO) summer internship programs.

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