THE LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS

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

TECHNICAL REPORT
LHNCBC-TR-2005-006

The Lister Hill National Center For Biomedical Communications Annual Report FY 2005

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

Annual Report FY 2005

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. The work of each is described below.

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 is engaged in applied research and development in image engineering and communications engineering motivated by NLM's 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 multimedia-rich interactive publications, content-based image indexing and retrieval (CBIR) of biomedical images, document image analysis and understanding (DIAU), image compression, image enhancement, image feature identification and extraction, image segmentation, image retrieval by query by image content, image transmission, optical character recognition (OCR) 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 annually produces the Unified Medical Language System (UMLS) Metathesaurus, 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 Office of the Public Health Service 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. Historian staff developed a system for cataloguing the Office's growing collections of artifacts and documents. In January, the Office mounted its first exhibit which was on the Cadet Nurse Corps Program. The Office of the PHS Historian also initiated a new oral history program. Over the course of the year, office staff have interviewed PHS officers who have worked on a variety of different initiatives and programs, including the closure of PHS hospitals and Hurricane Katrina. 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 of the office 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/.

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 FY 2005, the Center provided training to 47 participants from 12 states and 8 countries. Participants worked on projects in the areas of biomedical knowledge discovery, content-based image retrieval, consumer health informatics, document imaging, image database research, information retrieval research, medical illustration, natural language systems, ontology research, hand-held technology, telemedicine, ubiquitous computing, question-answering systems, machine learning and visualization research.

The Center continues to offer an NIH Clinical Elective in Medical Informatics for third and fourth year medical and dental students. The elective provides an overview of the state-of-the-art medical informatics in a lecture series by Center research staff, and offers an opportunity for independent research under the mentorship of expert NIH researchers. The program maintains its focus on diversity through participation in programs supporting minority students, including the Hispanic Association of Colleges and Universities and the National Association for Equal Opportunity in Higher Education summer internship programs. Established in 2001, the eight-week NLM Rotation Program 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 and the opportunity for students to work closely with established scientists and meet fellows from other NLM funded programs.

Image Processing

The Lister Hill Center performs extensive research and development in the capture, storage, processing, retrieval, transmission, and display of biomedical documents and medical imagery. Areas of active investigation include image compression, image enhancement, image recognition and understanding, image transmission, and user interface design.

WebMIRS

Developed several years ago and still in active use, 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. 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. At the current time there are about 400 users of WebMIRS in 44 countries.

An important new tool, Multimedia Database Tool (MDT), will serve as the next generation WebMIRS system. The MDT will provide 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 specific framework that has been designed has the goal of accommodating new sets of text and images under a very flexible database schema and GUI approach intended to allow new databases to be incorporated with work done only at the level of the database administrator, and not at the software modification level. New features being incorporated for end users of the system include support for multiple levels of system privileges for users and capability for users at authorized levels to make new data entries into database fields. Hence, the system will allow not only data dissemination but distributed data collection as well. The new system is intended to accommodate the existing WebMIRS databases as well as a new text/image database currently being created from a collection of uterine cervix images.

Content-Based Image Retrieval (CBIR)

The goals of this project are to research and implement the latest technological approaches for indexing and retrieving biomedical images by direct use of image data and in association with text related to the biomedical images. Our emphasis is on two-dimensional images, primarily the NHANES spine images, using shape methods on vertebrae in the images, and on NCI cervigrams, using color and texture methods to differentially identify tissue regions and tissue characteristics within these images. We also focus on development of effective CBIR methods to be incorporated into our multimedia database programs (such as the MDT) or into separate, prototype systems for use by the biomedical research and/or clinical communities.

The range of CBIR activities include: the development of integrated CBIR capability for retrieval by shape and partial shape; segmentation and truth data collection for the x-ray images; development of relevance feedback methods for shape retrieval for the spine x-rays; refinement of the Live Wire Segmentation technique; collection of vertebral segmentation data by medical experts; addressing problems of brightness removal and illumination correction in the cervigram images; development of Level Sets Segmentation methods for the spine x-rays; and the development of an integrated shape segmentation system.

A prototype CBIR client/server application was also developed. This application allows remote users to submit shape queries to shapes that have been indexed with the shape space indexing method developed by our Yale collaborators. The system incorporates a Java image server that delivers images using Texas Tech HVSQ compression. It is a set of modules developed in MATLAB and Java and is CEB's first demonstrable prototype for CBIR over the Web.

MARS

Medical Article Records System (MARS) progressed sufficiently to enable NLM to discontinue the manual keyboarding activity entirely. A key element in allowing NLM to eliminate its keyboarding contract is the capability designed in MARS to accommodate foreign language journals (that account for 11% of MEDLINE citations). This requirement introduced new rules to extract vernacular titles (requied in Roman script languages but not in others), and process the second pages of articles (to accommodate abstracts that spill over to a second page). These goals have been achieved by our FLEX software suite that is incorporated in several MARS workstations.

For some years MARS 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.

WebMARS was developed to complement MARS by enabling the extraction of bibliographic citations from online journals. Since a majority of citations now come directly from publishers in XML format, WebMARS functions have been used to develop two other systems that will serve to increase the efficiency of creating citations for MEDLINE so that the expected doubling of the citation rate in a few years can be accommodated through automation, a goal of NLM's Indexing 2015 Initiative. One of the advantages of WebMARS is that all of the bibliographic data contained in the online article may be extracted.

Its first component, Publisher Data Review (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. Currently, this is 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.

An initial version of the PDR system was completed after testing the software to handle databank accession numbers and grant numbers using a training set of several hundred online articles. Currently efforts are under way to discover discrepancies between those detected and actual data in MEDLINE. In collaboration with the Indexing Section, rules are being developed to eliminate these discrepancies, so that PDR can reliably provide operators data missing from the XML citations sent in directly by publishers. PDR will also correct incorrect data sent in by publishers. These features will help reduce the manual effort in creating citations for MEDLINE.
The second component, the WebMARS Assisted Indexing (WAI) system is for the indexers; it will help them search for terms in an article that correspond to biomedical terms in a predefined list. Again, indexers currently have to read through the entire article to confirm the occurrence of these terms, a labor-intensive process. WAI will automatically search through the text and highlight these terms for the indexer to simply confirm and select, thereby reducing manual effort.

AnatQuest: A window into the Visible Human

The goal of this project is to bring the high resolution Visible Human images to the lay public both directly as well as by linking text documents received from Web sources to relevant anatomic objects. This is achieved by two systems: AnatQuest and TILE.

AnatQuest is a Web-mediated system designed to provide widespread access to the Visible Human images for a broad range of users, including the lay public frequently limited to low speed Internet connections. This system is based on a 3-tier architecture in which the first tier consists of Java applets for displaying thumbnails of the cross-section, sagittal and coronal images of the Visible Human Male, from which detailed (full-resolution) views are accessed. The second tier is a set of servlets that process user requests and compress the requested images prior to shipment back to the user. The third tier is the object-oriented database of high resolution VH images and rendered 3D anatomic objects. Low bandwidth connections are accommodated

by a combination of adjustable viewing areas and image compression done on the fly as images are requested. Users may zoom and navigate through the images. Since its release in 2003, it has averaged about 60,000 hits per month, about 5 times the number of hits for the AnatLine system developed in an earlier project intended mainly for the scientific and visualization communities.

Recently, to improve access to VH images through common search engines, e.g., Google, the image labels were displayed below each thumbnail. This strategy has enabled Web crawlers to index the labels, thereby allowing the public direct access to the images through search engines, in excess of 10,000 hits in the first month after release.

TILE (Text to Image Linking Engine) 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. We interpret this goal as adding value to text resources such as PubMED and MedlinePlus by linking to anatomic images. A modular prototype TILE system is being developed to serve as a testbed to investigate the alternatives in 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. For the first (Document Analyzer) function, the MetaMap system, specifically its Java implementation (MMTx), is employed to analyze document text to identify biomedical terms.

Engineering Laboratories and Resources

The R&D relies on laboratories designed, equipped and maintained by the Branch, as well as content resources that support research as follows:

Image Processing Laboratory. The CEB Image Processing Lab is equipped with a variety of high end servers, workstations and storage devices connected by a mix of 100 and 1000 Mb/s Ethernet. The laboratory supports the investigation of image processing techniques for both grayscale and color biomedical imagery at high resolution. In addition to computer and communications resources and image processing equipment to capture, process, transmit and display such high-resolution digital images, the laboratory also archives a variety of image content. Currently, 60,000 images of the uterine cervix from a large National Cancer Institute (Guanacaste) study are being scanned for Web distribution. In addition to these are pap smear and histology images, also from this study. The Image Processing Lab also contains a selection of History of Medicine color images digitized at high resolution from the Library's Arabic and Persian medical manuscript collection.

Document Imaging Laboratory. This laboratory supports DocView, MARS and other research and design projects involving document imaging. Housed in this laboratory are advanced systems to electro-optically capture the digital images of documents, and subsystems to perform image enhancement, segmentation, compression, OCR and storage on high density magnetic and optical disk media. The laboratory also includes high-end Pentium-class workstations running under Windows 2000, all connected by Gigabit Ethernet, for performing document image processing. Both inhouse developed and commercial systems are integrated and configured to serve as laboratory testbeds to support research into automated document delivery, document archiving, and techniques for image enhancement, manipulation, portrait vs. landscape mode detection, skew detection, segmentation, compression for high density storage and high speed transmission, omnifont text recognition, and related areas.

Document Image Analysis Test Facility. Designed, developed and maintained by the Communications Engineering Branch, this off-campus facility houses high-end Pentium workstations and servers that constitute the MARS production system. While routinely used to produce bibliographic citations for MEDLINE, this facility also serves as a laboratory for research into techniques for the automatic zoning, labeling, and reformatting of bibliographic fields from document images, intelligent spell-check by pattern recognition techniques, and other key elements of MARS. These techniques are fundamental to the automated extraction of descriptive metadata for the long term preservation of document images. Besides real time performance data, also collected and archived are large numbers of bitmapped document images, zoned images, labeled zones, and corresponding OCR output data. This collection serves as ground truth data for research in document image analysis and understanding.

Ground truth data for document image analysis

For research in document image analysis and understanding techniques by the computer science and informatics communities, we provide a database named Medical Article Records Groundtruth (MARG). 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 6,300 unique IP visits from 92 countries.

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 2100 licensees in 49 countries. The Visible Human Project® has been featured in well over 850 newspaper articles, news and science magazines, and radio and television programs worldwide.

The Insight Toolkit (ITK), a research and development initiative under the Visible Human Project®, continues to grow and influence the medical community. It is now in its fifth year with a recent official software release ITK 2.2 in FY-05. 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. A consortium of university and commercial groups is executing this work.

The Visible Human Project® has extended awards to extend the software infrastructure into clinical and research applications through the introduction of database management tools, workbenches for tumor volume measurement for possible use in clinical trials, and the sponsorship of web portals for sharing research data and publications. Non-funded researchers from across the world are now testing, developing and contributing to ITK in over 38 countries. ITK users are represented by over 1000 active subscribers to the global mailing list for the project.

NLM has attained its primary goal of creating a strong, usable, public, open-source software infrastructure to support medical imaging research. The National Alliance of Medical Image Computing (NA-MIC), an NIH Roadmap National Center for Biomedical Computing, 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 as the center's driving biological problem. During FY2005, the NA-MIC organization has held 9 separate user-training workshops across the country, as well as an advanced ITK workshop in Lausanne, Switzerland. Other NIH ICs and their sponsored projects are not only taking up ITK, but they are helping to spread its influence and applying the toolkit to real-world problems. In addition to the NIH sponsored projects, a new non-profit organization, the Insight Software Consortium (ISC), was formed in FY2005 to undertake long-term advocacy for ITK and its concerns.

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 a 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 FY2005, the project produced a working proof-of-concept implementation and submitted the results of the research for publication. This project has been developed in conjunction with a visiting faculty fellow and continues to drive toward open-source public distribution.

We have extended and enhanced our pilot project for creating the framework for an archive of volume image data, the National Online Volumetric Archive (NOVA). In FY2005, we added a diverse collection from the Mayo Clinic including multi-modal imaging data from across a wide range of anatomy, physiology, pathology, and even animal models. Also in FY2005, we took some early steps toward management of a cancer imaging collection of digital colonography data that includes the source X-ray CT patient, the radiologist's reports, the results of a virtual colonoscopy scan, images and reports from a endoscopic colonoscopy exam, and pathology reports of any polyps removed during the procedure. These collections are helping to expand the NLM Lister Hill Center mission toward imaging informatics.

Visible Human Educational Collaboration

The National Library of Medicine has endorsed the acquisition and application development through the anatomical and NGI programs which has used raw visible human data in the past decade, to pass from a data to knowledge stage. The consequence of this data acquisition led to the development of educational applications affecting the discipline of gross anatomy, by

contractees awarded to develop knowledge based outcomes that can be potentially applied to basic science curriculum.

In 2005 further contacts to the NLM to give invited sessions to professional anatomy society's have concluded that most educational medical and allied health institutions are very interested and eager to see the continued development outcomes of the Visible Human applications. The timing of creating educational collaboratory's is also important due to the compression of curriculum time in the anatomical sciences, and the merging of some anatomy courses with physiology. The American Association of Anatomists, the American Association of Clinical Anatomists, and the Human Anatomy and Physiology Society strongly endorse the Visible Human's use as a basis of basic science curricula to accomplish these curriculum shifts and changes.

The purpose of this "collaboratory" is to develop academic sharing mechanisms (across advanced broadband networks) for application software directed to basic science curriculum insertion in dental, medical, or allied health school settings, in the discipline of gross and microscopic anatomy. The goal of the program is to make available through multicasting or a similar networked broadcasting arrangement so that academic institutions that have developed software applications based on the Visible Human Project data, can be shared at the other member Institutions of the Collaboratory. This process requires the planned integration of the applications into the existing curriculum schedule at these schools. Therefore, a demonstration of this type of networked software applications, developed by one institution can be linked by networking to other institutions of the Collaboratory. The Collaboratory approach involves image processing, information systems, infrastructure research, multimedia visualization, and training. The process becomes a multidisciplinary effort to take mature digital science program areas and focus the outcomes to creative public use in professional schools Worldwide. It would be appropriate for the Library to fund this knowledge based venture.

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. Their 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 members of the public with comprehensive information about all types of clinical research studies, both interventional and observational. The site has over 23,000 protocol records sponsored by the U.S. Federal government, pharmaceutical industry, academic and international organizations in all 50 States and in over 120 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 7 million page views per month and hosts approximately 20,000 visitors daily. Data are submitted by over 1,440 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 the development and implementation of new standards to promote transparency in clinical research through trial registration. Concurrent with the release of the International Committee of Medical Journal Editors (ICMJE) statement requiring trial registration as a condition for publication in fall 2004, ClinicalTrials.gov expanded its scope to include trials sponsored by members of the international research community. ClinicalTrials.gov also added several new key fields specified by the World Health Organization minimum data set, thereby allowing sponsors to describe their research plans more fully. As a result of these new initiatives, over 9,000 new registrations were received over a five-month period beginning in May 2005.

The Genetics Home Reference (GHR) provides basic information about genetic conditions and the genes or chromosomes responsible for those conditions. Created for affected individuals, their family members, and the public, the site currently includes more than 170 condition summaries and more than 250 gene summaries, which are added at a rate of about 14 new summaries per month. GHR's content expanded to include all 29 conditions recommended by the Health Resources and Services Administration (HRSA) for inclusion in states' newborn screening programs. In addition, GHR added a "spotlight" feature to the web site's home page. The "spotlight" draws attention to important initiatives related to genetic health that may be of interest to the general public.

In the two years since its launch, the site's usage has increased nearly tenfold. GHR continues to receive favorable comments from healthcare providers, patients, family members, educators, and librarians. Recently, online media outlets such as Forbes.com, CNN, and Nature.com have linked to GHR to provide background material for genetics-related stories. A usability survey among the membership of the Genetic Alliance indicated that the site is easily navigable and the content is credible and understandable.

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

Digital Library Research encompasses all aspects of creating, disseminating and preserving digital collections, developing metadata standards, applying emerging technologies and formats, and resolving copyright and legal issues. 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 prime example of these activities is the Profiles in Science project. It uses innovative digital technology

to make available the manuscript collections of prominent biomedical researchers, medical practitioners, and those fostering science and health.

The content of the database was created in collaboration with NLM's History of Medicine Division (HMD), which processes and stores the physical collections, including books, journal volumes, pamphlets, diaries, letters, manuscripts, photographs, audio tapes and other audiovisual resources. The collections of Francis Crick, Albert Szent-Györgyi, and Salvador E. Luria were added this year. An additional 4,500 pages in 600 documents were also added to existing Profiles in Science collections. Presently the database features the collections of sixteen prominent scientists: Christian B. Anfinsen, Oswald T. Avery, Julius Axelrod, Francis Crick, Donald S. Fredrickson, 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. Profiles in Science was recognized by ScientificAmerican.com as one of the top fifty science and technology Web sites and was highlighted in Science Magazine's "Best of the Web" each time a new collection was added.

During this fiscal year, several research projects continued to enhance the effectiveness of Profiles in Science. Enhancements to the underlying Profiles in Science digital library framework included developing methods for detecting and correctly handling changes in master data elements that affect multiple records, adding new database infrastructure to support new functionality, creating additional views of the information in the database, providing new error detection and correction rules, and importing finding aid structure information. New scanning hardware and software were installed, configured, documented, and extensively tested. Protocols for digitizing collections at other institutions were fine tuned and put to the test by NLM and Tennessee State Library and Archives staff. Protocols for digitizing video were also developed and tested on the Szent-Györgyi video clips. Suitable test sets of Profiles in Science master TIFF images were identified and provided to others for automatic metadata extraction experimentation. Development of a new XML-based Web front end and transition to a new XML-based search engine, as well as automated testing and verification tools, continue to be pursued. Several extensive tests were performed to ensure that the transition from the current Web front end to the new XML-based version will be seamless.

Document imaging for the biomedical end-user

The goal of this research area is to apply 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, originally released in January 1998 and subsequently improved over several generations, has 17,099 users in 193 countries, an increase of more than 1,000 new users and 3 countries over last year. In September 2005 alone, there were 82 new users spread over 23 countries registering to use DocView. However, reflecting the declining worldwide use of TIFF for distributing document images (compared with PDF), and the age of the software itself, the use of DocView is expected to decrease.

Another factor leading to a decline in DocView use is the way libraries have used it in tandem with Ariel® software for their interlibrary loan services. Since new versions of the Ariel software issued by the marketer, Infotrieve, are not compatible with DocView (our Web site notifies users of this), the use of our software will drop as libraries change to the new Ariel software. Nevertheless, this changeover is likely to be gradual especially in foreign countries since their purchase of the new Ariel may take longer.

MyDelivery. Seen as a successor to DocView, the goal of the client/server MyDelivery communications system is to enable reliable and secure delivery of very large (gigabyte-sized) files, and large numbers (hundreds) of attachments in a single delivery, especially over unreliable links such as wireless networks. Potential users are medical researchers, clinicians, administrators, librarians and many other health professionals who need to securely exchange electronic medical information residing in a wide variety of file sizes. Examples of large files include document images, digitized color photographs, digitized x-rays, FDA drug applications, and clinical images such as PET scans, MRI scans, CT scans, sonograms, and digital videos. Recent developments have been in six critical areas in the design of MyDelivery: Log Shipping, Network Load Balancing, operation of the client behind proxy servers, operation of the client on all target operating systems, user interface, and managing the client.

DocMorph and MyMorph. The DocMorph system continued to serve both browser-based users (over 12,500 to date: 2000 more than last year) and MyMorph users (over 5,200 users) this year. Of the more than 12,000 registered users, many 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.

By using Simple Object Access Protocol (SOAP) that combines XML with HTTP, MyMorph has been developed as a Web service that significantly improves the DocMorph function used 75 percent of the time, viz., the conversion of files to PDF. MyMorph consists of Windows-based client software and modifications to DocMorph to accommodate SOAP. MyMorph significantly improves user productivity compared to the (conventional) use of DocMorph through a Web browser, particularly for users who need to convert large numbers of files to PDF. This is accomplished by reducing the time required for users to interact with the software. Test results show that MyMorph reduces the user interaction time from hours to seconds for all users regardless of their Internet connection speed. The process of using MyMorph for converting image files to PDF has been integrated into many library document delivery operations worldwide.

MEDLINE Database on Tap (MDoT)

This project, previously known as PubMed on Tap, 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. An objective is to understand how to display data so that users can quickly find the most pertinent information, while limited by the

small screen and restricted bandwidth of handheld computers. Techniques were developed for display and navigation, as well as those for information organization. In addition, a prototype MDoT system was demonstrated.

As our primary method of discovery, we have developed a system that supports MEDLINE search and retrieval from a wireless, Internet-connected PDA. PDA client software for both Palm OS and Pocket PC OS have been developed and are freely available. Also, the MDoT Web site continues to be accessed at the rate of about 5,000 hits every month. This Web site provides information about the project as well as the software, and allows us to solicit feedback from our users and monitor aggregate user behavior.

Two studies were conducted to improve searching and user interaction. The first was a study of LHNCBC's experimental probabilistic search engine. It ranks results by relevance and was originally developed to support ClinicalTrials.gov. This investigation suggested that its search and ranking algorithms might be advantageous to MEDLINE searching at the point-of-care. A second study was conducted to assess the value of the new e–spell function available from Entrez. This study showed that about 10% of MDoT searches would benefit from an automatic spelling corrector. Based on these studies, we designed a new Palm OS version 1.7 of the client with user interface and communications to support both an optional Auto Spell feature and an option to use the Essie search engine in lieu of the PubMed search engine. (This new capability is one of the reasons for changing the name of this project from PubMed on Tap.) Other developments included changing the search icon from a magnifying glass, which apparently is not intuitive, to a big green "Go" button. A new database structure was also developed to record the use of the new options, and the fields and tables were reorganized to facilitate session analysis.

Interactive Publications Research

This research effort was initiated in FY2005 to create a comprehensive, self-contained and platform-independent multimedia document that is an "interactive publication." Following a study of existing open source formats and standards, a prototype 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 images (CT, MRI, ultrasound) following the DICOM standard. 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. A comparison of these in terms of ease of use and development effort was done.

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. On-going work will demonstrate and extend its usability and utility.

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 five NLM search systems to provide results from 15 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 September 2004 in which access to the NLM book, serials, and audiovisual materials

was changed from the former LocatorPlus system to the new NLM Catalog system running under Entrez.

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 2005 MeSH update, and Year End Processing were incorporated during the year as they became available.

A comprehensively redesigned NLM Gateway was released in April 2005. The new interface is cleaner, more straightforward and more understandable than the original, offering direct access to many functions and result sets with a single click of the mouse. The new system runs on Dell/Intel servers using the Linux operating system, making it possible to upgrade capacity at significantly lower cost than with the former Sun/Solaris servers.

As with the original, the new NLM Gateway is a meta-searching system, mapping user queries to appropriate search commands for multiple backend systems and issuing simultaneous searches in 15 databases. Results in terms of hit counts for all 15 databases are shown on one page. With one click, the user can display results from any of the 15 resources. While looking at those results, the user can with one click move to any of the other 14 result sets. Significant infrastructure changes were made to facilitate this direct access to results from specific databases rather than the former categories of results: journal article citations, books and monographs, consumer health, and meeting abstracts. A version of the Gateway with access to 20 NLM databases including five new toxicology-related systems from the NLM Division of Specialized Information Services is in beta testing. It will be released in FY2006.

Digital Preservation Research

This project falls into two broad categories, one concerned with the preservation of documents and the other with video. In each area we focus on some of the key functions of an economical and robust digital preservation system. Two in particular are automated metadata extraction and file migration.

Extracting metadata automatically from the contents of material that need to be preserved, rather than relying on manual entry, is probably the only way large collections can be economically preserved. Techniques are also being developed that automate the migration of files in bulk. This is important for the conversion of files in formats that face obsolescence, largely because they are no longer supported by newer software and modern computers, and will be inaccessible as time passes.

For document preservation, a detailed design was done and a prototype *System for Preservation of Electronic Resources* or 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. The module for metadata extraction from Web pages is internal to SPER while those for TIFF and online journals are implemented via a SOAP interface, so that SPER can access a remote AME system and retrieve extracted metadata in XML format as defined in the METS standard. In addition, a module for the extraction of technical metadata from TIFF file headers is implemented to include many of the items listed in the NISO Z39.87 standard for digital still images.

Research into video preservation focused on identifying an open file format such as Motion JPEG 2000 (MJ2) for archiving digitized video on disk media. Toward this end, a one-

day invitational meeting was organized with about 50 archivists and technologists involved in the long term preservation of video and film. Participants considered the potential of lossless, on-disk video storage in light of the "twilight of tape" as a cost-effective storage medium. Barriers to adopting lossless algorithms were identified, and specific directions to overcome them suggested. Also discussed were current video metadata standards, and recent work in automatic extraction of metadata from video.

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.

Scalable Information Infrastructure (SII) 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, or public health authorities. These significant, network dependent healthcare, health education and research applications have made significant progress as follows:

Wireless Internet Information System for Medical Response in Disasters (WIISARD)
The University of California, San Diego in support of the Scalable Information Infrastructure (SII) initiative is developing an integrated software-hardware system designed to enhance the delivery of medical care at the site of mass casualty disasters. The system, called WIISARD (Wireless Internet Information System for Medical Response in Disasters), is intended for use by a Metropolitan Medical Strike Team (MMST) in response to nuclear, biological or chemical (NBC) events. Improvements to an early prototype system have been implemented, and are ongoing. The system uses wireless networks, GPS, RF tags, and handheld and wearable computers to monitor the position of victims and their vital signs; the position of MMST personnel, and an auxiliary channel for communications with the incident command center; the position of treatment assets; and distribution of NBC weapons plumes. The developing prototype system has been deployed in several regional San Diego area wide drills, and more sophisticated and compact hardware are being developed to overcome some of the problems evident during drills.

Advanced Network Infrastructure for Distributed Learning and Collaborative Research
The Haptic Audio Video Network for Education Technology (HAVnet) project, being conducted at the Stanford University School of Medicine, builds on prior work developing visual and haptic educational applications for anatomy and surgery training. The project includes aspects of self-scaling technology, self-optimizing end-to-end network aware real-time middleware, wireless technology and GIS. These elements are being investigated within a context of two educational test beds, each an extension of prior work developed as part of a previous Next Generation Internet (NGI) contract. One test bed, a Clinical Anatomy suite presents challenges of bandwidth

and latency. The other, a Clinical Skills test bed concentrating on surgical training simulations using haptic devices, presents performance challenges due to low tolerance for network jitter. The project is working to deliver: enhancement and integration of two existing middleware applications, Information Channels and Weather Stations, allowing correlations to be made between network metrics and actual application performance; addition of self-optimizing features to six applications using the core middleware; development of a new application, Anatomy Window, that uses a handheld computer to map a cadaver and present corresponding images derived from the Visible Human data set; development of a Remote Tactile Sensor, capable of capture and transmission of tactile dermatology information over a network; implementation of the anatomy teaching suite over local, national and global networks for use initially in laboratory based teaching, and ultimately in actual field teaching; and implementation of a clinical skills test bed, primarily in early phase and laboratory testing.

Project Sentinel Collaboratory

This project is a partnership between Georgetown University and the Emergency Departments of MedStar Washington Hospital Center and MedStar Georgetown University Hospital. This 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 (I&Ws) of emerging threats to human health.

During FY2005, significant progress was made in the implementation and upgrading of complementary systems for the project's infrastructure. The principles behind Project Sentinel Collaboratory are important for preparing the nation for disaster management, whether those disasters are due to nature, infectious disease, or terrorism. An unexpected opportunity to leverage the infrastructure and project expertise presented itself in the aftermath of Hurricane Katrina. The client server on the VPN (Virtual Private Network) was utilized during this disaster to manage critical data such as registering and tracking evacuees in the metropolitan Washington, DC area. Servers have been added to the Collabortatory in order to support the Hurricane Katrina effort, and the project team is adding Shibboleth to protect/secure personal data and make it accessible to multiple entities requiring access.

National Multi-Protocol Ensemble for Self-Scaling Systems for Health

The Boston Children's Hospital informatics program has developed a secure XML medical record template for individuals, geographic information systems that can collect data on the use of these individual templates, and is negotiating with health care organizations for large scale applications (e.g. state immunization registries).

Advanced Health and Disaster Aid Network (AID-N)

The Johns Hopkins Applied Physics Laboratory is developing a redundant network for the public health facilities of Montgomery County, MD, to be tested in mock public health emergency exercises with the anticipation of permanent installation if tests are successful. The network will include many innovative components including wireless multimedia transmission.

Advanced Network Infrastructure for Health and Disaster Management

The University of Alabama at Birmingham is utilizing emerging secure, high-speed wireless communication in the prehospital emergency medical services (EMS) and public safety

information environment and integrating it with a hospital information system in an academic medical center and an independent EMS dispatching center.

SMART

OHPCC is overseeing the SMART (Scalable Medical Alert and Response Technology,) 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.

Wireless PDA Pubmed Searching

As part of the project, "Information Technology for Low-resource Areas", an organizational computer network was developed for health care organizations in low-resource environments. The network consisted of a central server, wireless access points, a dial-up connection to an Internet Service Provider, and featured the use of open-source software applications (Linux, Apache, PHP, MySQL.) The wireless access points (802.11b, Bluetooth, Infrared) enabled mobile devices such as personal digital assistants (PDAs) and wireless portable computers to disseminate administrative and health care information, and acquire reference resources and decision support tools through the Internet. An open source software electronic medical record software, OpenEMR was initiated. Examples of access to the National Library of Medicine (NLM) and other knowledge sources were shown. It was presented at the American Telemedicine Association Annual Meeting, Denver, Colorado, April 17-20, 2005.

Telemedicine Initiatives

OHPCC participated in talks, demonstrations on state-of-the-art Telemedicine, e-Health projects, and solution 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 2005 Telehealth, e-Health, and healthcare informatics projects and programs designed to address pressing healthcare issues including: advances in biosurveillance; efforts toward a National Health Information Infrastructure; bringing information to the point of care; technologies toward effective disease management and for reducing medical errors; cost saving technologies; disaster and humanitarian assistance projects and others. This program is intended to inform Members of Congress and staff, federal agency officials, healthcare and technology organization representatives, and the public. Honorary Committee Co-chairs include: Senator Kent Conrad (D-N.D.); Senator Mike Crapo (R-ID); Representative Earl Hilliard (D-Alabama); Representative Ernie Fletcher (R-KY); and more than 60 Members of Congress from the House and Senate. 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. This site will archive images for medical education and telemedicine A Medical Image Database at the NLM- http://images.nlm.nih.gov is also online. As individual images are viewed, a search form for MEDLINE/PubMed is automatically filled-out with the title of the image. A scanning digital microscope was acquired for the VM project. Collaboration with the NCI is planned.

The Telemedicine Information Exchange(TIE) is an NLM sponsored web based resource of information about telemedicine and telemedicine related activities maintained by the Telemedicine Research Center, Portland, OR. During FY05 approximately 503 non-NLM bibliographic citations and HPROJ-type records were received at NLM in compliance with the NLM contract. OHPCC staff conducted quarterly teleconferences with the staff of the Telemedicine Research Center, Portland, OR, the parent organization of the TIE.

Videoconferencing and Collaboration

Major upgrades to existing videoconferencing codecs were done in FY-05 and new codecs were added. Demonstrations of steaming and wireless webcasting were done and videoconferencing and webcasting was employed routinely in the OHPCC program. The adverse impacts of network security on the use of videoconferencing and collaboration technology were resolved by adding a new subnet outside the LHC and NLM firewalls that is isolated from all other internal networks. Care has been taken to do regular security updates and virus checks an the machines connected to this subnet.

A distance learning program was conducted in collaboration with SIS, coordinator of NLM’s Adopt-A-School Program. The program involved a series of presentations on medical topics and related health information resources and it was offered in conjunction with the Charles R. Drew University of Medicine and Science. The target audience for the program were students at the King Drew Medical Magnet High School in Los Angeles, one of the few medical magnet high schools for minorities in the United States, that is affiliated with the university. Procedures were developed for recording the presentations and offering them later on demand by webcast using the Collab streaming server. The distance learning program was formally evaluated and found to be effective. A new program is planned for the next fiscal year to assess hands on training.

Work continues in experimenting with new codecs. Digital video compression technologies are being acquired for testing with members of the Internet2 community, since the DV format is being considered as a compression format for the Access Grid. A beta version of Conference XP, an open source, experimental codec from Microsoft Research was installed on the Access Grid computer and tested with the Universities of New Mexico and Puerto Rico.

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 terminology services, modeling and learning methods, medical ontologies, indexing initiative, and semantic knowledge representation.

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 lexicon-building tool; SPECIALIST contains over 295,000 records. Lexical access 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, the lexicon-building tool, has been recently updated and moved to a Linux server. LexBuild is used by the lexicon building team, and 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. The lexical tools have been expanded so that they can be adapted to non-English language vocabularies. MMTx, the JAVA implmentation 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.
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. Development has begun on 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. These developments should result in a more flexible interface with both machines and human users.

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.

Modeling and Learning Methods

Digital information is at the center of scientific endeavor and, if managed carefully, may provide a bridge for scientists and clinicians across disciplines. Today interdisciplinary research is severly hindered by language and terminology barriers such that scientific data and information

generated in a field cannot be easily accessed, correctly interpreted and effectively used by other scientists.

The Modeling and Learning Methods (a.k.a. the muON) project is aimed at developing computational learning methods to enable scientists to utilize crossdisciplinary information effectively. The methods that are being developed in this project are multidisciplinary in nature and are rooted at biomedical informatics, machine learning, computational linguistics, and probabilistic knowledge modeling.

Recent research focused on effective information access by breaking the language barriers between the information and its potential users. Information retrieval, which currently is restricted to searching a corpus of text through a short list of keywords, is being enhanced with ontological, linguistic and probabilistic methods. Current prototypes can access, identify and retrieve biomedical information more effectively than the leading information retrieval engines such as PubMed and Google.

Significant strides have also been made at a more fundamental level. A new probabilistic graphical modeling method called parameter interdependency networks (PIN) has been developed. The method is substantially different from current probabilistic graphical models such as Bayesian networks, which are based on the concept of random variables. The PIN method on the other hand is based on a more fundamental concept called random events. The resulting method enables designers to model the system of interest (e.g., a biological process or a clinical problem) in the desired level of detail and provides a rich set of interactions between the model components. The method has been developed as the underlying modeling theory for the envisioned multidisciplinary information architecture, called multifaceted ontological networks (muON).

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 LHNCBC research team focused on the role of reference ontologies – including the Foundational Model of Anatomy (FMA) and the Chemical Entities of Biological Interest (ChEBI) – for ontology alignment and integration. For example, the FMA was used as a reference ontology for mapping between human and mouse anatomy ontologies. Analogously, a mapping between two important clinical terminologies – SNOMED CT and ICD-9-CM – was derived automatically from knowledge extracted from the UMLS. Such automated mapping techniques are expected to play an important role in the exploitation of electronic patient records (EHRs), and ultimately to contribute to health information technologies.

Work on ontology visualization was also pursued with RxNav, a publicly available, cross-platform application for visualizing drug information represented in RxNorm, one of the standards for use in U.S. Federal Government systems for the electronic exchange of clinical health information. Other research areas of this group include semantic similarity among genes computed from the various model organism databases annotated to the Gene Ontology (GO).

The application of GO-driven similarity to the prediction of functional properties of genes and to the annotation of new genes was investigated. Methods were developed for converting ontologies from frame-based to description logic-based representations, with application to the FMA.

Additionally, an ontological analysis of the UMLS Semantic Network was initiated and is expected to result in a better integration with other ontologies, including upper-level ontologies (e.g., DOLCE, BFO) and ontologies used in other branches of the Federal Government (e.g., Federal Enterprise Architecture Reference Model Ontology (FEA-RMO) and Data Reference Model (DRM)). In the future, the research team will investigate the application of semantic similarity to MeSH descriptors and MEDLINE documents for the purpose of clustering documents and identifying related citations.

Indexing Initiative

The Indexing Initiative project continues to investigate techniques 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 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 match text to Metathesaurus concepts, while the third uses a variant of the PubMed related articles algorithm to find MeSH headings by using the existing indexing terms of articles related to the input text. 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 made available 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. Two recent efforts to improve MTI's accuracy are the extension of MTI input from titles/abstracts to the full text of articles and a word sense disambiguation (WSD) facility to reduce MetaMap ambiguity. The full text effort puts MTI in a position to take advantage of full text as it becomes available.

Journal Descriptor Indexing (JDI)

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. For example, the Journal of Pediatric Surgery is indexed by the JDs Pediatrics and Surgery. 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.

Unified Medical Language System

The mission, scope, and content of the Unified Medical Language System Metathesaurus continued to grow and evolve in FY 2005. During the fiscal year, regular production releases of the UMLS Knowledge Sources were made: 2004AC in October 2004 (with Spanish language names included for 23% of all concepts), 2005AA in January 2005, and 2005AB in June 2005. The 2005AC release is scheduled for November 2005. The number of concepts has increased over this time by 20% to more than 1.2 million. The number of names for concepts has increased by 17% to 5.5 million. There are more than 100 contributing source vocabularies.

The format and content of these 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 installation tool, MetamorphoSys, allows the selection of desired content from the Metathesaurus and writes the desired subset in Rich Release Format (RRF).

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.

MetamorphoSys now includes an integrated Rich Release Format (RRF) Browser that allows users to view their own Metathesaurus subsets in both Raw Data and Concept Report views. This addition 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 Metathesaurus group has begun to promote the Rich Release Format both as a general vocabulary standard and as an input format for vocabulary submission to the UMLS. This process will make the nature and meaning of information from each vocabulary more explicit and allow the use of shared tools for validation, browsing and use, thus increasing interoperability. We are pleased to note that several other vocabulary providers have begun to distribute their vocabularies in RRF.

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, while 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. Recent additions include a mechanism for interpreting hypernymic propositions, and current work addresses arguments of nominalizations, comparative structures, and coordination of predicates. A modification of SemRep, called SemGen, is being developed for identifying and extracting semantic propositions on the causal interaction of genes and diseases from the research literature.

Semantic predications produced by SemRep and SemGen 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. One of these is a tool that cooperates with PubMed to provide an informative and interactive summary of search results. Another will assist domain experts in curating the Genetics Home Reference application.

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 (TTPI)

The TTPI project has two aims: (a) to design efficient methods to reformat the paper volumes in NLM's historic collection to photorealistic "Turning The Pages" (TTP) form, and (b) to extend these virtual books beyond their application as beautiful museum pieces to information systems that augment the material in the originals, including delivery over the Internet.

Originating as a collaboration with the British Library in producing two virtual books, Blackwell's 18th century *A Curious Herbal* and Vesalius' 16th century Anatomy book, in TTP form, we have since made significant progress. The process consists of scanning the pages, enhancing these high quality color images by Adobe Photoshop, creating animated 3D wireframe models of the pages and book cover using Alias Maya (an innovation in our approach), run on a computer by Macromedia Director software, and displayed on a touchscreen monitor in an exhibit kiosk. 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.

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 under consideration: Robert Hooke's *Micrographia*, the first book written about microscopes and in which reportedly the first time the word 'cell' was used.

In 2005 we collaborated with Library of Congress staff to create TTP at their institution. They were invited to a demonstration of our kiosk version of TTP, and a technical discussion of the steps required to create it. In addition to using their own resources and knowledge to accomplish the scanning, image enhancement and 3D modeling, they needed our templates for the final stage (to produce the software providing the interactivity). At our suggestion, they

selected books in the life sciences: one by a Dutch surgeon who spent a decade with pirates in the Caribbean (1678); the other an encyclopedia of flora and fauna in the New World (1635). The first book was shown at the opening of the Kislak Collection, an event at the Library of Congress in April 2005.

"The Library as Place" Flash Web Site

In 2005, APDB developed a Flash Web site containing the entire NLM DVD program, The Library as Place. The original DVD program was based on a two day conference held at NLM which examined the future role of libraries in the digital age. Given the number of conferences and lectures held at the NLM, APDB initially chose to use DVD technology as a means for storing and delivering conference assets including video, audio, slides, Web links, textual documents associated with the conference and also make it highly searchable using visual and keyword search engines on the DVD. NLM distributed over 10,000 copies of the Library as Place DVD that APDB produced. The Flash Web site will serve as a prototype for delivery of future NLM conferences and programs.

Virtual Dialogue Programs

APDB has been providing ongoing support to the NLM Visitors Center in the direction and installation of a series of unique interactive multimedia programs, described as "Virtual Dialogues". These programs profile some of the scientific leaders of our time and are based on extensive recorded interviews with them. The virtual dialogue programs utilize speech recognition technology to allow a user, speaking into a microphone, to interview one of the many experts on a broad range of topics that each scientist has previously addressed in a carefully constructed video interview. Posing questions which may address the scientist's personal life or a particular scientific discussion, the scientist, appearing as video on the computer screen, responds directly to the viewer in an informal, conversational style. Thus far, the Virtual Dialogue series includes interviews with Marshall Niremberg, Joshua Lederberg, Julius Axelrod, Edward Feigenbaum, and Donald Lindberg. An interview with C. Everett Koop is scheduled to be conducted later this year.

Changing the Face of Medicine - Local Legends Program and Web Site

In consultation with the Office of Communications and Public Liaison, (OCPL) and in collaboration with the American Medical Women's Association, (AMWA), APDB has been working on the development, production planning, and review for the NLM/AMWA Local Legends program and Web site. The Local Legends Web site highlights congressionally nominated women physicians from 50 states. The Web site is designed to include video profiles of one representative from each state, as selected by a committee within the AMWA.

Eleven video profiles were completed for the Web site, and feature Jan Carney, M.D. from Vermont, Jessie Ternberg, M.D. from Missouri, Linda Warren, M.D. from Kansas, Joanne Schaefer, M.D. from Nebraska, Janice Gable, M.D. from Virginia, and Mercy Obeime, M.D. from Indiana, Peggy Goodman, M.D. from North Carolina, Mary Virginia Krueger, M.D. from Washington, Marilyn Glassberg, M.D. from Florida, Virginia Caine, M.D. from Indiana, and Sandra Levison, M.D. from Pennsylvania. The completed video segments and full transcriptions were delivered to Digital Outpost for compression according to NLM specifications. Compressed media was then delivered to OCCS for placement on the NLM-based Web site as delivered from an NLM server. APDB participated in all aspects of quality control to assure that the Local

Legends Web site is fully compliant with Section 508 of the Rehabilitation Act by applying Web content accessibility guidelines to the entire site before deployment.

Additional DVDs were prepared in FY2005 including LHNCBC Research Projects, Profiles in Science: Updates, UMLS, AMPA/NLM Annual Meeting, NLM Program Highlights 2005, InformationRx Press Event Highlights, and NLM BOR presentations.