THE LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS

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

TECHNICAL REPORT
LHNCBC-TR-2004-006

The Lister Hill National Center For Biomedical Communications Annual Report FY 2004

Alexa T. McCray, Ph.D.
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 2004

Introduction

The Lister Hill National Center for Biomedical Communications, 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. An organization chart with the names of Branch and Office Chiefs is on the inside back cover of this report.

Organization

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 multiple databases, intelligent agent technology, knowledge management, the merging of thesauri and controlled vocabularies, data mining, and machine-assisted indexing for information classification and retrieval. Research issues include knowledge representation, knowledge base structure, knowledge acquisition, 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 NLM's Gateway, the team that annually produces the Unified Medical Language System Metathesaurus, and the staff who coordinate the Center's training programs. The most current information about the Computer Science Branch can be found at http://lhncbc.nlm.nih.gov/csb/.

Cognitive Science Branch

The Cognitive Science Branch (CgSB) conducts research and development in computer and information technologies. Important research areas encompass the investigation of a variety of techniques, including linguistic, statistical, and knowledge-based methods for improving access to biomedical information. Branch members actively participate in the Unified Medical Language System project and collaborate with other NLM research staff in the Indexing Initiative project, the goal of which is to develop automated and semi-automated techniques for indexing the biomedical literature. The branch also conducts research in digital libraries and collaborates with NLM's History of Medicine Division on Profiles in Science, a project to digitize the archival collections of prominent biomedical scientists. Several branch projects address the challenges involved in providing health information to consumers. ClinicalTrials.gov is an important resource for the public and, additionally, serves as a testbed for conducting consumer health informatics research, and the Genetics Home Reference provides complex information about genes and diseases to the public in easily understood language. The most current information about the Cognitive Science Branch may 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 image engineering and communications engineering motivated by NLM's mission-critical tasks such as document delivery, archiving, automated production of MEDLINE records, Internet access to biomedical multimedia databases, 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 content-based image indexing and retrieval of biomedical images, document image analysis and understanding, image compression, image enhancement, image feature identification and extraction, image segmentation, image retrieval by image content, image transmission and video conferencing over networks implemented via asynchronous transfer mode and satellite technologies, optical character recognition, and man-machine interface design applied to automated data entry. CEB also maintains archives of large numbers of digitized spine x-rays and bit-mapped document images that are used for intramural and outside research purposes. The most current information about the Communications Engineering Branch can be found at http://lhncbc.nlm.nih.gov/ceb/.

Audiovisual Program Development Branch

The Audiovisual Program Development Branch (APDB) conducts media development activities with several specific objectives. As its most significant effort, the branch participates in the Center'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. Consultation and materials development are also provided by the branch for NLM's other information programs. From applications of optical media technologies and teleconferencing to support for web distribution, the requirement for graphics, video, and audio materials continues to increase in quantity and diversification of format. In addition to the development of new techniques and processes, the

facilities and hardware infrastructure must reflect state-of-the-art standards in a rapidly changing field. Included within APDB is the Office of the Public Health Service Historian. The office preserves and disseminates information about the history of Federal efforts devoted to public health. The most current information about the Audiovisual Program Development Branch 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 2004 the Center provided training to 53 participants from 17 states and 8 countries. Participants worked on projects in the areas of biomedical knowledge discovery, content-based image retrieval, consumer health systems research, document imaging, image database research, information retrieval research, medical illustration, natural language systems, ontology research, hand-held technology, web services research, user interface research, telemedicine, ubiquitous computing, Unified Medical Language System research, and visualization research. The Center continues to offer a successful NIH Clinical Elective in Medical Informatics for third and fourth year medical students. The elective provides an overview of the state-of-the-art of medical informatics in a lecture series by nationally and internationally known speakers, and offers an opportunity for independent research under the mentorship of expert NIH research staff. 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 NLM Rotation Program continues to grow. The eight-week rotation program for trainees from NLM funded Medical Informatics programs provides these individuals 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.

Additional information about Lister Hill Center training opportunities is available at the Center's web site under "Training Opportunities". Interested individuals will find descriptions of each of the training programs, including specific application procedures.

Language and Knowledge Processing

Developing SPECIALIST, an experimental natural language processing system for the biomedical domain, is the focus of the Center's natural language processing work. The SPECIALIST system includes several modules based on the major components of natural language: lexicon, morphology, syntax, and semantics. The lexicon and morphological component are concerned with the structure of words and the rules of word formation. The syntactic component addresses the constituent structure of phrases and sentences, while the semantic component seeks to extract biomedical content from text. All components of the SPECIALIST system rely heavily on the domain knowledge in the Unified Medical Language System Knowledge Sources.

Terminology Research and Services
Lister Hill Center research staff build and maintain the SPECIALIST Lexicon, a large syntactic lexicon of medical and general English terminology released annually with the UMLS Knowledge Sources. New lexical items are continually added to the Lexicon using a lexicon building tool, LexBuild, developed and maintained by the lexical systems research team. LexBuild allows researchers to enter items directly into a central database via a web browser. A new version of LexBuild featuring internal checks to prevent common data entry mistakes and logical inconsistencies was deployed in FY 2004. The FY 2005 SPECIALIST Lexicon release tables will be generated entirely using the new LexBuild tool. The SPECIALIST Lexicon increased by over 32% to 242,000 lexical items in the FY 2004 release. Lexical access tools, including a lexical variant generator (LVG), wordind, and norm, are also distributed as open source resources with each UMLS release. During this past year the group also developed several tools to manage diverse vocabularies for a range of language and information processing purposes. The team recently achieved a significant milestone in providing customized UMLS data to several projects, including ClinicalTrials.gov, Profiles in Science, and the Genetics Home Reference. A number of internal tools were also developed to handle data customization. These incorporate UMLS updates and provide client applications with periodic releases of customized data and the latest terminology enhancements.

Semantic Knowledge Representation
Innovative methods 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 used to address a variety of problems in biomedical language and information processing. MetaMap maps noun phrases in free text to

concepts in the UMLS Metathesaurus. 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. Users are able to create MMTx data files independently of the UMLS. MetaMap Technology Transfer source code is included in the MMTx release, and an error reporting and tracking system ensures that problems reported by users are effectively addressed.

SemRep is a tool that uses the Semantic Network to determine the relationship asserted between concepts developed in MetaMap. SemRep serves as the basis for ongoing research initiatives in biomedical information management, such as projects for extracting medical and molecular biology information from text, processing clinical data in patient records, and research in knowledge summarization and visualization. Recent enhancements to SemRep's linguistic coverage include the addition of a mechanism for interpreting hypernymic propositions. Current work addresses arguments of nominalizations, comparative structures, and coordination of predicates. Semantic predications produced by SemRep serve as the basis for continued work in automatic abstraction summarization of biomedical text, including MEDLINE citations and an online encyclopedia. SemGen, a modification of SemRep, is being developed for identifying and extracting semantic propositions on the causal interaction of genes and diseases from MEDLINE citations. Project staff are also developing methods for automatically suggesting appropriate images as illustrations for anatomically oriented text.

Indexing Initiative

The Indexing Initiative investigates concept-based indexing methods for the automatic selection of subject headings in both semi-automated and fully automated indexing environments at the NLM. The goal of the Indexing Initiative is to obtain retrieval performance equal to or better than performance of systems using manually assigned index terms. A prototype indexing system for testing indexing methods, the Medical Text Indexer (MTI), is being tested by NLM indexers. MTI recommendations are available to all indexers as an additional resource available through NLM's Data Creation and Maintenance System. In addition, results of the MTI system are being used as keywords for AIDS/HIV, health sciences research, and space life sciences collections of meeting abstracts that are not manually indexed.

On-going improvements to MTI continue to be made. Short-term, incremental changes arise from requests made by indexing staff or by a desire to incorporate more of NLM indexing policy into the system. Longer term goals include a word sense disambiguation effort to improve MTI's accuracy. The team has also begun to investigate the use of the full text of articles in addition to their work with MEDLINE titles and abstracts. Additional work investigates an 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) corresponding to biomedical specialties. The JD system associates JDs with words in titles and abstracts in a three-year training set of 1,378,597 MEDLINE records. Each record "inherits" the JDs from the journal in the record. A word in the training set can then be described by a list of JDs ranked according to the number of co-occurrences between the word and the JDs. Text as input to the system can be indexed based on averaging the word-JD co-occurrences for the words in the text that are also in the training set, ranking the JDs in decreasing order of these averages. The journal descriptor approach 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 NLM's participation in the Text Retrieval Conference (TREC 2004).

Unified Medical Language System

Unified Medical Language System research regularly develops and distributes multi-purpose, electronic knowledge sources and associated lexical programs. The Metathesaurus, Semantic Network, and SPECIALIST Lexicon are used by system developers to enhance patient data, create digital libraries, retrieve web and bibliographic data, apply natural language processing, and improve decision support. The Metathesaurus represents multiple biomedical vocabularies organized as concepts in a common format providing a rich terminology resource in which terms and vocabularies are linked by meaning. The Semantic Network allows users to investigate relationships among semantic types and relations and retrieve a list of Metathesaurus concepts assigned to a particular semantic type. Finally, the data in the SPECIALIST Lexicon provides users with the syntactic and morphologic information about each of its lexical items.

The Metathesaurus continues to grow in size, scope, and mission. As of FY 2004, there are more than one million concepts with 5 million names from 117 source vocabularies in 15 languages. The Metathesaurus is now released in a new "Rich Release" format that 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. Following the July 2003 announcement by the Secretary, HHS of a government license for nationwide use of SNOMED CT, this widely used standard vocabulary for US clinical medicine has been added to the Metathesaurus. The Metathesaurus installation and configuration program called MetamorphoSys has been enhanced to offer easy extraction of pre-computed subsets, for example all HIPAA (Health Insurance Portability and Accountability Act) vocabularies, or selected natural language processing names. This feature will assist users in many areas including regulatory compliance in electronic medical records. The Metathesaurus team has successfully met several new challenges including meeting increasing demand for frequent updates; developing methodologies for mappings between vocabularies; and the development of tools to meet the changing needs of an expanding community, especially of clinical users.

A significant change in the method of delivery of the UMLS Knowledge Sources to users has occurred along with the increase in size of the Metathesaurus to 18 Gigabytes. Approximately one third of all users now access the UMLS through the UMLS Knowledge Source Server, one third request the files on DVD-ROM, and one third download the full Knowledge Sources online. The Metathesaurus group has developed a multi-platform Java program that allows users to decompress, customize, and install the Knowledge Sources on local machines, and has added browsers for users who create local subsets.

Modeling and Learning Methods

The Modeling and Learning Methods project seeks to develop new modeling methods that enable researchers to rapidly construct effective computational models from large datasets. The objectives of the project are to develop machine learning methods that automate the process of constructing probabilistic models for identifying relevant information among large datasets and corpora, mapping identified information to networks of ontologies, accessing queried

information accurately, and answering user queries through mining the data located in heterogeneous information sources. Interest in probabilistic models ranges over a wide spectrum of biomedical fields, including computational biology; biomedical, clinical, and healthcare informatics; and epidemiology. The objectives of the project will be evaluated with a set of suitable metrics such as receiver operating characteristic that measure the performance of prospective models in terms of sensitivity and specificity in reaching their target functions. Depending on the domain of the models and the problems of interest, domain subjects and/or experts might be needed to determine the gold standards or the target functions for the performance evaluations of the models if such gold standards or target functions are not readily available. FY 2004 research focused on identifying information represented in textual data (e.g., MEDLINE abstracts) using UMLS tools, the SPECIALIST parser, and MetaMap. New computational methods in modeling textual and numerical data are being developed. Staff participated in TREC 2004 and competed in the Physiological Data Modeling Contest at the 21st International Conference on Machine Learning.

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 foundational issues and explored the ontological properties of resources such as SNOMED CT and the Foundational Model of Anatomy. Non-lexical approaches to identifying dependence relations in ontologies were studied, with application to acquiring associative relations in the Gene Ontology. A generic framework was also developed for computing semantic similarity from a taxonomy and the frequency of its nodes in a corpus, applicable to the Gene Ontology, MeSH, and WordNet. Finally, the team pursued work on visualization by developing RxNav, an application for navigating drug information in the RxNorm model. In the future, the research team will investigate a semantic similarity approach to comparing lists of MeSH descriptors assigned to MEDLINE documents and to identifying functionally related gene products annotated with the Gene Ontology.

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.

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. VHP data sets are available through a free license agreement with the NLM. Data sets are distributed to licensees over the Internet at no cost and on DAT tape for a duplication fee. Worldwide use of the data sets continues to grow as they are applied to a wide range of educational, diagnostic, treatment planning, virtual reality, virtual surgeries, artistic, mathematical, and industrial uses by over 2000 licensees in 48 countries. The Visible Human Project has been featured in well over 850 newspaper articles, news and science magazines, and radio and television programs worldwide.

FY 2004 saw the continued maintenance of two databases to record information about Visible Human Project use. The first database logs information about VHP license holders and records their plans for using the images. The second database records information about the products that licensees are developing. The Insight Toolkit (ITK), a research and development initiative under the Visible Human Project, completed two official software releases in FY 2004. 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. Additional ITK awards have been made 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 are now testing, developing and contributing to ITK in over 30 countries. Research institutions, including the Mayo Clinic, the Imperial College of London, Georgetown University, the University of Utah, Kitware, Harvard University, Cognitica, the University of Pennsylvania, and U.S. National Library of Medicine staff participated in demonstrations and technology exhibits at the November 2003 annual Radiological Society of North America conference in Chicago. Tutorials on how to use ITK were presented at the IEEE Vis2003 conference in Seattle, the SPIE Medical Imaging Conference in San Diego, and the MICCAI 2003 conference in Montreal. At the end of FY 2004, the NIH Roadmap Initiative for Bioinformatics and Computational Biology awarded a 5-year cooperative agreement to the National Alliance of Medical Image Computing. This $20 million national center for biomedical computing has adopted ITK and its software engineering practices as part of its engineering core.

3D Informatics

During FY 2004 the 3D Informatics Program has continued to mature and develop its in-house research efforts around problems encountered in the world of 3-dimensional and higher-dimensional, time-varying imaging. Research is continuing in the areas of image-based implicit rendering, research and systems trials for ITK, and haptic latency analysis for surgical simulation. The team has extended and enhanced its pilot project for creating the framework for an archive of volume image data, the National Online Volumetric Archive (NOVA). This project includes the physical implementation of the pilot archive for volume image data, as well as a tutorial for data submission, meta-data structure management tools using XML, and web

page structure. The meta-data structure management were refined, published and presented at the 2004 SPIE Medical Imaging conference. Research is continuing in an effort to create a software framework for artistic and non-photorealistic rendering of digital models entitled, Programmable Layered Architecture With Artistic Rendering (PLAWARe). The framework will consist of a layered software architecture for implementing medical illustration techniques using computer graphics technologies. PLAWARe adopted the infrastructure from the ITK software engineering methodologies in FY 2004, including the construction of DART dashboards and the use of CMake and generic programming principles. Additional work includes: research of implicit surface and its application to surface generation for efficient rendering of anatomic objects, research of finite element modeling and simulation system for human colon straightening and its application in virtual colonography, and research on geometric mapping using the index-check-and-correct method for human colon polyp detection in helical CT datasets.

In September 2004, the HPCC office together with representatives of the National Institute of Biomedical Imaging and Bioengineering and two directorates at the National Science Foundation, sponsored a workshop on visualization research challenges. The 28-member panel drew national and international participants from industry and academia to begin a discussion on the current grand challenges in visualization and imaging research.

AnatQuest

While the Visible Human images have been used by biomedical scientists and developers worldwide, the goal of this in-house project is to provide widespread access to the Visible Human images for a broad range of users, including the lay public. In line with this goal, a web-mediated system, AnatQuest (available at anatquest.nlm.nih.gov), was developed. 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.

Current work is proceeding in two directions. The first is to increase the number and type of rendered images (beyond the current 300 surface-rendered structures) to make the collection more useful for the public. This would require registering all of the cryosection slice images, segmenting and labeling anatomic structures on each slice, and using these to create surface- and volume-rendered images. The second direction taken in this project addresses a long term NLM goal, that is, 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. This may add value to text resources such as PubMED and MedlinePlus by linking to anatomic images. For this purpose, project staff are developing a modular prototype system (Text to Image Linking Engine, or TILE) to serve as a testbed to investigate the alternatives in the functions needed to accomplish this linkage. These functions involve 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.

WebMIRS

The Web-based Medical Information Retrieval System, a Java application, allows remote users to access data from two surveys conducted by the National Center for Health Statistics. These are 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. Through the WebMIRS graphical user interface, a user may construct a query for the NHANES II or NHANES III data. WebMIRS allows the user to save the returned data to the local disk drive, where it may be analyzed with appropriate statistical tools such as the commercially available SAS and SUDAAN software. The WebMIRS NHANES II database also contains vertebral boundary data that was collected by a board-certified radiologist for 550 of the 17,000 x-ray images in WebMIRS. 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 this type of 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.

WebMIRS enhancements include collaborative work with Texas Tech University to develop an advanced compression capability custom tailored to the image characteristics of the x-ray images, to allow delivery of the WebMIRS images in compressed form rather than in the low-resolution form as at present. Software written in Java has been developed for the decompression at four different levels. Work is now underway to improve the performance efficiency of the decompression, before the code is incorporated into the WebMIRS system. Significant progress was made toward the development of the next generation WebMIRS system, the Multimedia Database Tool. This system will provide a software framework for the incorporation of new text/image databases in a much more general way than the current WebMIRS and provide 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 flexible database schema and GUI approach that is 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.

Online X-ray Archive

The complete set of 17,000 NHANES II x-ray images in the full-resolution form in which they were digitized was made publicly available in FY 2000. These images are available by FTP and have been accessed by researchers from both within the U.S. and also from international sites. Staff created the ImViewJ software, a downloadable Java application, which allows users to view images at their full spatial resolutions (e.g., 1463x1755 for the cervical spine images, 2048x2487 for the lumbar spine images). Coordinate data collected under the supervision of a radiologist at Georgetown University are also available for 550 images. This coordinate data defines landmark points for each vertebra in a manner commonly used in the field of vertebral morphometry, and serves as reference data to aid in creating and evaluating the performance of

image processing algorithms for segmentation of the vertebrae. This coordinate data is publicly available on the FTP site along with TIFF 8-bit versions of the corresponding x-ray images. Users may access this coordinate data either through the FTP archive or through the WebMIRS system. The number of TIFF 8-bit images publicly available was increased to 1,000 in FY 2004.

Content-Based Image Retrieval

The overall goal of the content-based image retrieval (CBIR) project is to develop methods for effective extraction of biomedical information from biomedical digital images, with the current concentration being on the NHANES II spine x-rays. The focus is both on indexing the image data and search of those data. For example, for the 17,000 NHANES II images, the only indexing data originally available was the collateral alphanumeric data collected in the questionnaires and examinations; no indexing information derived directly from the images was originally available, and the high cost of employing radiological experts to compile such data by physical viewing and interpreting each image makes it unlikely that such information will ever be acquired by purely manual means. These circumstances could be reversed if reliable, biomedically-validated software could produce image interpretations automatically. Even in the more likely case that only semi-automated methods should prove feasible, the reduction in labor costs could be sufficient to allow the creation of databases of significant biomedical information where this is not currently economically feasible. This is the motivation for research into computer-assisted image indexing. Computer-assisted image searching is a potential enabler of enhanced information extraction from a database that has already been indexed.

During the current year new and substantially extended CBIR capability was developed with the implementation of the latest version, CBIR3. Highlights of the system, developed in MATLAB and Java, are that it can operate in networked or stand-alone modes, uses XML for reporting, and allows the user to select either a more mature or an experimental version of the system. CBIR3 differs from its predecessors in that all data (text, images, and segmentations) are now stored on a centralized MySQL database. Each user is allocated a unique login that grants them certain rights and privileges. The system supports access to multiple data sources that can be selected by the user. CBIR3 also provides a validation sub-mode for expert review, validation, and pathology indication for indexed images. CBIR3 currently allows vertebral shape segmentation using the Modified Active Contour Segmentation and LiveWire segmentation techniques. In addition it has a well defined interface allowing the addition of more techniques. It is now possible to segment images in a database-controlled sequential mode that remembers the user's state when he/she stopped working. The last image and vertebra segmented are saved and automatically brought up the next time the same user segments. Another new feature of CBIR3 is that it allows text searching on the complete NHANES II dataset through the familiar WebMIRS interface. WebMIRS (standalone) has been linked with CBIR3 for allowing hybrid text and image searches. For image queries, CBIR3 supports query by sketch and query by image example. Query shape can be generated by sketch, choosing it from the existing shapes on the database, or by supplying an image and segmenting it to obtain a shape. The query shape can subsequently be edited by moving points, adding points, and removing points.

Digital Archive of Uterine Cervix Images

Work continued in FY 2004 towards the creation of an archive database of the 60,000 – 100,000 digital images of the uterine cervix collected by the National Cancer Institute. This work

included analysis of color models, standards and technology for the digital capture of color information from 35 mm slides with high color fidelity, and similar issues related to retaining the color across digital output devices such as monitors and printers. MATLAB programs were created to enable the comparison of images digitized at different scan densities or at different compression levels. A Nikon 4000 slide scanner was acquired, and 200 uterine cervix slides were scanned to generate evaluation data. For each of these slides, a medical expert labeled regions of interest by marking rectangular boundaries and entering labels such as "acetowhite", "mosaicism", "punctation", and "vasculature" with a MATLAB tool that was developed for that purpose. A compression study was conducted to allow the comparison of uncompressed uterine cervix images with those compressed using the Hybrid Multiscale Vector Quantization method developed by Texas Tech University. Multiple medical experts participated in the study, which used 50 test images compressed at eight different compression levels. A similar study was conducted to determine a suitable spatial resolution to use for digitizing the 35 mm slide collection.

Engineering Laboratories

The Image Processing Laboratory 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. The equipment includes a Sun Enterprise 4500 server with dual 400 MHz CPUs, and 1.5GB memory, and a SunFire 280R server with dual 1.2 GHz CPUs, 3 GB memory, and two internal 73 GB SCSI disks. Additional computers in the lab include two Sun Ultra 10 workstations, each with a 440 MHz CPU, 512 MB memory, and an external 36 GB SCSI disk; and two Sun Ultra 10s, each with a 300 MHz CPU and 512 MB memory. All of these machines run the Solaris 9 operating system. Large-scale magnetic storage is provided by a Network Appliance FAS960 which is a network-attached storage device connected by redundant Gb/s Ethernet connections and provides 24TB of RAID storage. For the ultra-high-resolution display of x-ray images, two E-systems Megascan monitors provide image display at a spatial resolution of 2048x2560 pixels. The laboratory also contains specialized equipment and software for device calibration and color profile creation. This includes a USB-interfaced MonacoOPTIX colorimeter, capable of color measurement from emissive sources, for CRT and LCD monitor color calibration, and used with MonacoOPTIX software; and a USB-interfaced GretagMacbeth Eye-One spectrophotometer, which measures color in the 380-730 nm range, with resolution of 10 nm, from both emissive and reflective sources, used with MonacoProof software, for the creation of standard color profiles which characterize the color I/O of devices such as scanners, monitors, and printers using the International Color Consortium standard.The Document Imaging 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. The laboratory also contains rack-mounted, networked processors running all recent versions of Windows-based operating systems to support the DocView, DocMorph and MyMorph projects. This provides an easily-configurable test platform for simulating a variety of potential user environments, including those with firewalls, for testing, modifying and improving software developed in these projects.

The Document Image Analysis Test Facility is an off-campus facility that 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 spellcheck by pattern recognition techniques, and other key elements of MARS. In addition, 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

Multimedia Research and Development

Multimedia research and development efforts concentrate on the engineering of technical improvements applied to issues such as image quality and resolution, color fidelity, transportability, storage, and visual communication. In addition to developing new methods and processes, Lister Hill Center facilities and hardware infrastructure reflect state-of-the-art standards in the rapidly changing field of multimedia research and development. High definition video, for example, represents today's standard for improved electronic, motion imaging quality.. Multimedia systems, scientific visualization and networked media are being pursued for their performance, educational, and economic advantages. 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. CD-ROM, DVD and DVD-ROM technology for capturing media assets including video, audio, web information, and computer text slides continue to be explored. Web-links within these assets are used for updating program content and providing links to additional information tools. A template allowing the simultaneous viewing of multiple interactive windows, including speaker video, slides, and an interactive index was developed to improve access to program content on CD-ROM, DVD and DVD-ROM technology. By selecting any one slide from the index, two other windows immediately synchronize to that point in the presentation. Using the new template technology, project staff developed a symposium DVD-ROM "The Library As Place: A Symposium on Building and Renovating Health Sciences Libraries in the Digital Age" and a Conference DVD "From Double Helix to Human Sequence - And Beyond" featuring over 10 hours of video, web access, video, and additional information on each disc.

Together with NLM's Office of Communications and Public Liaison and the HMD Exhibition Program, project staff have worked with MacNeil/Lehrer Productions to launch the *Changing the*

Face of Medicine: Profiles in Achievement web-enhanced DVD in FY 2004. The highly interactive DVD features 12 physician profiles, a mentoring program profile, and 200 web links as an information resource tool for users. The interactive DVD was awarded a 2004 Web DVD Excellence Award by the DVD Association of America. As an element of the *Changing the Face of Medicine* exhibit, the NLM is working on the planning and production phases of video and web programs featuring the *Local Legends* program, a collaborative project between the NLM and the American Medical Women's Association (AMWA). 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. The first video interview with the Washington, D.C. local legend, Janelle Goetcheus, M.D., was conducted at the Columbia Road Health Services clinic, and additional video content was produced in the clinic and on the streets of Washington. These materials were edited into an overview video featuring the NLM/AMWA program and presented at the annual AMWA meeting in San Diego, CA in February 2004. The overview video of Dr. Goetcheus won a 2004 Telly Award. Thirty-three on site video interviews with nominees were conducted at the annual meeting to select state representatives of the Local Legends program. All aspects of the Local Legends web site design have been completed and approved by the NLM Local Legends development team and the AMWA. Future work will include the development of additional Local Legends video profiles.

Additional projects illustrate a variety of technological advancements. *Project 20*, a 15-minute videotape chronicling the last 20 years of the NLM, highlighted the growth of MEDLINE, the development of Grateful Med, Internet Grateful Med, Free MEDLINE, UMLS, the creation of NCBI, and other significant events in the History of NLM. A prototype DVD-ROM based on the NLM Dream Anatomy exhibition was completed in FY 2004. The DVD features a video overview, a gallery and timeline, and a virtual tour of the exhibit. The narrated program also features high definition video of the exhibition, video graphics, and an original musical score. Web links to NLM's Dream Anatomy exhibition web site and a fully functional search tool are also available when the DVD is viewed on a computer. Additional DVDs were prepared in FY 2004 including; *LHNCBC Research Project Video DVD, 2004 Collen Award; Dr. McDonald's Life and Career*, and *NLM Board of Regents; Saving Lives and Saving Money, Newt Gingrich*.

Information Systems

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

Profiles in Science

The Profiles in Science web site uses innovative digital technology to make available the manuscript collections of prominent biomedical researchers, medical practitioners, and those fostering science and health. Database content is created in collaboration with the History of Medicine Division, which processes and stores the physical collections. Most collections have been donated to the NLM and contain published and unpublished materials, including books, journal volumes, pamphlets, diaries, letters, manuscripts, photographs, audio tapes and other audiovisual resources. The Visual Culture and Health Posters, as well as the collections of C. Everett Koop and Wilbur A. Sawyer were added in FY 2004, bringing the total number of

archives for prominent biomedical researchers, medical practitioners, and those fostering science and health to 13: Christian B. Anfinsen, Oswald T. Avery, Julius Axelrod, Donald S. Fredrickson, C. Everett Koop, Joshua Lederberg, Barbara McClintock, Marshall W. Nirenberg, Linus Pauling, Martin Rodbell, Florence R. Sabin, Wilbur A. Sawyer and Fred L. Soper. The Reports of the Surgeon General (1964 - 2000), the history of the Regional Medical Programs (1964 - 1976), and Visual Culture and Health Posters are also available on Profiles in Science.

In FY 2004, project staff continued to enhance the effectiveness of Profiles in Science. The web site was upgraded to more powerful hardware with up-to-date applications and operating system software. Enhancements to the underlying digital library framework included a new database infrastructure, the creation of additional ways to view information, and faster methods for extracting records in ASCII format. New error detection and correction rules and methods for automatically updating data were also added. Protocols for digitizing collections at other institutions were developed and tested in collaboration with the Wellcome Library staff, United Kingdom. Development began in FY 2004 on a *Historical Events and Prominent Scientists Timeline* to highlight the major historical events (e.g., polictical, medical, scientific, and social) that occurred at the time of the major achievements of the scientists represented in the collection. Changes to the current Metadata Entry and Editing Program were made in preparation for moving the program to a web interface. Detailed analysis of workflow in obtaining copyright permissions identified changes needed in the database and user interface for tracking permissions. Finally, the development of an XML-based web interface and transition to an XML-based search engine, as well as automated testing and verifcation tools, continue to be pursued.

MARS

Document image analysis and understanding research combined with database design, graphical user interface design for workstations, image processing, string pattern matching, lexical analysis, speech recognition and related areas underlie the development of MARS (Medical Article Records System), a system to automate the production of MEDLINE citation records from biomedical journals. 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. For some years, its production version has been used to extract bibliographic data to populate MEDLINE. Two other techniques to obtain such data have been manual keyboarding and XML-tagged data directly from publishers. To meet the NLM's goal of discontinuing the keyboarding contract and thereby realizing savings, MARS design faced the challenge of having to process journals currently handled manually. These journals include ones with page background in color or gray shades which greatly compromise OCR accuracy. Experiments were conducted with grayscale scanners comparing different approaches to eliminating these atypical backgrounds, and the best approach was found to be by using a library developed with functions in the FineReader OCR toolkit. This library was embedded in the inhouse-developed scan software, and preliminary results from a test set of 101 articles showed that low-confidence characters occur at about the same rate with these grayscale scanners as with the monochrome scanners in production, i.e., effectively eliminating the

deleterious effects of gray and color backgrounds. Following the completion of these tests, the grayscale scanners have been placed in production.

The scanning software has also been modified to improve quality control. Images from poorly scanned documents cause OCR errors and compromise downstream processes. Conventional QC relies on the operator viewing the images and deciding on their quality. This is highly subjective and is not always reliable. To make this step more robust, a commercial library from ScanSoft has been incorporated in the Scan module to detect low-confidence characters and calculate those as a percentage of the total number of characters on the page. This percentage figure provides the operator a quantitative measure of image quality. Another key element in allowing NLM to eliminate its keyboarding contract is the requirement for MARS to accommodate foreign language journals, which account for 11% of MEDLINE citations. This requirement introduces new rules to extract vernacular titles (in Roman script languages but not in others), and process the second pages of articles (formerly only one page needed to be processed). These have been achieved by the FLEX software suite that incorporates new code in several MARS workstations. Starting with journals in French, German, Italian and Spanish, MARS enhanced by FLEX now processes five Western European languages using Roman script and three using Cyrillic script as well.

WebMARS is a system to extract bibliographic data from online journals. A prototype system has been developed to combine downloading and classification of journal articles followed by zoning, labeling and reformatting algorithms to identify and extract the data. The NLM Board of Regents was recently given a talk covering the history of automated bibliographic data extraction from 1996 when NLM's keyboarding contract ran into difficulties, through the evolution and increasing automation in the MARS system, and focusing on the design and functions of WebMARS. A key point in the presentation was a comparison of the relative labor required in producing citations with keyboarding, MARS, XML citations from publishers and WebMARS (which promises to result in the least amount of labor.) WebMARS is undergoing testing with over 60 journal titles. Tests comparing WebMARS output against existing MEDLINE citations for past issues have been useful in refining the labeling and reformatting algorithms.

An additional prototype has been developed to handle meeting abstracts. Testing with four volumes was successful and the prototype is ready for demonstration. "Meeting Abstracts" refers to the proceedings of important conferences in HIV, AIDS and other topics of current importance. The contents of these proceedings are not simply 'abstracts' as conventionally understood, but include most other bibliographic information: title, author names, affiliations, etc., in addition to abstracts, but are not full papers and, most important for automation, do not follow the familiar layouts of typical biomedical journal articles. The unconventional layout of Meeting Abstracts requires a modification of the existing zoning, layout and reformatting rules. For instance, since author names are arranged differently from a typical journal (all names in a single line, and separated by semicolons), the existing reformatting rules in MARS required changes to accommodate this format.

Ground Truth Data for Document Image Analysis

In August 2003, the Medical Article Records Groundtruth database was released 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 nine layout types encountered in MARS production. Included are the corresponding segmented and labeled zones all in XML format (e.g., OCR-converted and operator-verified data at the zone, line, word and character levels). Also available from this web site 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.

DocView

DocView facilitates the delivery of library documents directly to the patron via the Internet in multiple ways, but it is most commonly used by library patrons to receive scanned journal articles from libraries that use Ariel software for interlibrary loan services. While Ariel, developed by Research Libraries Group, and now a product of Infotrieve, is used by libraries and document suppliers routinely to send documents via Internet to similar organizations, there are few options for end users to directly receive them. DocView helps fill this void by allowing end users to receive documents sent by Ariel via a modified form of File Transmission Protocol (FTP). DocView also enables users to retain the received documents in electronic form, view the images, organize them into "folders" and "file cabinets", electronically bookmark selected pages, manipulate the images (zoom, pan, scroll), copy and paste images, and print them if desired. In addition, DocView serves as a TIFF viewer for compressed images received through the Internet by other means, such as web browsers. Users may receive document images either via Ariel FTP or Multipurpose Internet Mail Extensions protocols. With DocView, users may also forward documents to colleagues for collaborative work. 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. The MyMorph web service consists of Windows-based client software and modifications to DocMorph for accommodating the Simple Object Access Protocol. In-house testing has shown that 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.

Document Preservation

Project staff have begun the design and implementation of a flexible, modular software framework that may be used as a prototype for investigating techniques to preserve NLM's digital resources in a cost-effective manner. A prototype system called SPER (System for the Preservation of Electronic Resources) has been developed. The system allows ingest, metadata extraction and file migration, and the identification of minimum required technical metadata for document files. Developing SPER required careful attention to proposed standards and models for digital preservation and preservation metadata schemas, including the NISO X39.87 proposed standard for digital still images. SPER relies on open source, platform-independent components, as well as current open resources and tools which already provide some functionality required by SPER. A JavaServer Faces-based GUI was chosen to provide the web interface for SPER users or operators. The SPER prototype was implemented in FY 2004, with a first phase model designed to convert TIFF images to PDF documents and/or JPEG2000 images. The *Profiles in Science* collection and MARS document images will be used as test sets.

Additional research and development efforts on metadata extraction and prototype design strategies of SPER will address issues with metadata elements, strengthen tools that automatically learn journal-specific rules using both geometric and contextual features, and strengthen systems that automatically learn the 2D layout models of document page images using Bayesian learning algorithms. In-house tools (e.g., DocMorph, MyMorph) are being studied as potential tools for electronic preservation. Modifications of DocMorph and MyMorph to produce PDF/A files from image-based files are being explored. This work may lead to a system, accessible from any point on the Internet, that allows users to mass-migrate image-based file collections to a standard archival format. Additional research is being conducted to identify key issues related to the preservation of video.

Turning The Pages Information Systems

Turning the Pages Information Systems research seeks to design more efficient methods to translate paper volumes from the NLM's historic collection to photo realistic electronic form, extend the virtual books into information systems, and to increase the accessibility of historical documents for the public. After the initial development of the Turning the Pages (TTP) format, research began to transform the initial TTP design into a usable information system (TTP+). Research focused on a "discovery" and a "storyline" model as directions for TTP+. The TTP+ version of Blackwell's *Herbal* uses the "discovery" model, retaining the photorealism of the original TTP while allowing a patron to "travel" to live sites on the Internet. For example, from highlighted text on the St. John's Wort page, users can go to various search engines (e.g., PubMed, ClinicalTrials.gov, USDA) and obtain citations or general information on St. John's Wort. The TTP+ version of Vesalius' *Anatomy in Photorealistic* uses the "storyline" model and contains images from other sources (e.g., rendered Visible Human images, pictures of Italian cities, etc.). Images are interlinked to present the consumer with several multimedia "stories," including *Man of Padua* and *Modes of Portraying Anatomy*.

Two methods have been investigated in order to combine all existing virtual books for kiosk display. A monolithic approach bundling all software into one file was pursued. Memory limits imposed by the Windows OS rendered this method unscalable. On the other hand, a modular approach where the code for each book is accessed on selection by the user, provided a scalable method more suitable for the addition of future books. In addition, while developed under the Windows OS environment, the TTP code has also been successfully tested for operation on a Mac computer running OS X. Future goals are to continue developing efficient, high quality methods for producing and distributing TTP books as more historical books are selected.

NLM Gateway

The NLM offers a number of Internet-based information resources, each with its own user interface. The NLM Gateway provides an easy to use, "one-stop" search method that allows users to issue simultaneous searches in fifteen NLM information resources using five retrieval methods from a single interface. The NLM Gateway continued to grow and evolve in FY 2004 with several additions and enhancements. NLM Gateway access was added for the MedlinePlus Health Tutorials, MedlinePlus Current Health News, Online Mendelian Inheritance in Man, Hazardous Substances Data Bank, TOXLINE Special, and the Genetics Home Reference. NLM's book, serials, and audiovisual materials were migrated from LocatorPlus to the new NLM Catalog under "Entrez", substantially increasing the searching capabilities in the

collection. The NLM Gateway language table was updated with the latest Machine Readable Cataloging language codes. Targeting PubMed, enhancements include the addition of a LinkOut feature for PubMed citations and direct links on the Document Ordering page for PubMed Central articles. A spell checker that automatically searches both British and American spellings of words was also incorporated. Author name truncation for searching was added to the Meeting Abstracts Collection and the Health Services Research Projects database, and approximately 15,000 abstracts were added to the Meeting Abstracts collection.

A comprehensively redesigned NLM Gateway Version 2.0 entered early testing in FY 2004. The new user interface will provide clear, easy to understand, and a cleaner navigation to different areas of the composite result set. At the same time, the new interface will continue to execute simultaneous searches in 15 information resources. The targeted release for the new user interface is early FY 2005.

Consumer Health Informatics Research

Exploring consumer information needs, information seeking behavior, and cognitive strategies, consumer health informatics research projects utilize informatics methods and information technologies to study methods to develop, organize, integrate, and deliver accessible health information to consumers with all levels of health literacy.

ClinicalTrials.gov provides comprehensive, up-to-date information about federally and privately supported clinical trials throughout the United States and many other parts of the world. The system grew out of 1997 legislation requiring the U.S. Department of Health and Human Services, through the National Institutes of Health, to establish a registry for both federally and privately funded trials "of experimental interventions for serious or life-threatening diseases and conditions," thereby broadening the public's access to information on potential interventions for a wide range of diseases. Launched in February 2000, ClinicalTrials.gov provides patients, families and members of the public easy access to information about the location of clinical trials, their design and purpose, criteria for participation and, in many cases, further information about the disease and intervention under study. There are also links to individuals responsible for recruiting participants to each study. Because clinical trials bridge biomedical research conducted in laboratories and applied clinical research in humans, information in this area is often difficult for non-specialists to read. ClinicalTrials.gov is designed to help members of the public make sense of the information provided. The site includes general resources to help people understand what clinical trials are, including a glossary of common terms used to describe clinical trials, and a list of frequently asked questions about human research. In addition, each study is presented in a standard format that helps readers quickly identify important elements of a study, such as its purpose, criteria for participation, locations of the trial sites, and contact information. Furthermore, to provide additional context, study records also point users to relevant health topics at the NLM's consumer health web site, MEDLINEplus, which contains easy-to-read information to help patients research their health questions. Some study records also contain links to published literature, either for background information or study results.

A web-based Protocol Registration System (PRS) allows providers to maintain and validate information about their trials. New views of protocol summaries are supported by geographical location, date added (e.g., last seven days and last 30 days), and by patient recruiting status. A

Spanish-language prototype system using Spanish-English cross-language information retrieval technology was developed and is undergoing extensive testing. ClinicalTrials.gov was the recipient of Harvard University's prestigious 2004 Innovations in American Government Award in recognition of its significant achievements. HHS Secretary Tommy G. Thompson noted that ClinicalTrials.gov is a good example of how government can improve access to vital health care information for all Americans.

The Genetics Home Reference is an integrated web-based information system designed for consumers and others to learn about specific genetic conditions and the genes or chromosomes associated with those conditions. The research results made possible by the Human Genome Project are increasingly being made available in scientific databases on the Internet, but because of the often highly technical nature of these databases, they are not readily accessible to the lay public. The goal is to provide a bridge between the clinical questions of the public and the richness of the data emanating from the Human Genome Project. The Genetics Home Reference web site provides basic information in a question and answer format on the nature of genes and how they give rise to various conditions and diseases. The site currently includes more than 100 condition summaries and more than 160 gene summaries, over half of which were added during FY 2004. Additional FY 2004 improvements include a new feature that provides information about chromosomes and chromosomal disorders. Several new topics (e.g., pharmacogenomics, multifactorial disorders, and imprinting) were also added to "Help Me Understand Genetics", the site's genetics handbook. Genetics Home Reference achieved significant site navigation improvements in FY 2004 with a redesigned home page, as well as newly designed browse, search, and help features. Targeted links were also added throughout the site. The site was integrated with MedlinePlus, Gateway, PubMed Linkout, and the "What's New" series in order to help consumers locate the Genetics Home Reference web site.

Further Consumer Health Informatics Research focuses on understanding and improving access to online health information. Technologies are being developed that provide measures of text difficulty that help determine the suitability of health-related documents for consumers at different literacy levels. New approaches for providing timely access to consumer health information in order to accommodate the diverse needs of people in the U.S. and abroad are being pursued through cross-language information retrieval research. Finally, the Consumer Health Vocabularies project focuses on mapping words and phrases commonly used by consumers to technical medical terms and concepts.

Research Infrastructure and Support

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.

Communication and Collaborative Technologies
Lister Hill Center staff is actively involved in research to develop technologies that will facilitate easy access to biomedical information through devices such as Personal Digital Assistants (PDA's), wireless portable computers, mobile phones, and other emerging devices.

PubMed on Tap is a research and development project to develop accessible biomedical information at the point of care through handheld devices used by clinicians and other mobile health care providers. User interface, content selection, content organization, and system performance are necessary for effective access to information. Initial research is focused on the design of a user interface for search and retrieval of MEDLINE bibliographic citations through PubMed. Initial content selection is involved with categorizing citations returned in response to a query, creating multi-document summaries for clusters of highly related documents, and single-document descriptions containing features specific only to a given document in the cluster. System performance research is focused toward discovering design factors that ensure the speed and reliability of the hardware and software required for accurate and timely retrieval of data. Areas of investigation include choice of parsers, efficient use of a database to store recent queries and citations, and load testing. A prototype system, developed for PDAs running the Palm operating system, was built and tested in FY 2003. The software uses the PDA's wireless communication interface and HTTP protocol to communicate with a servlet residing on a proxy server. The proxy server communicates with PubMed through the Entrez programming utilities (e.g., Esearch, Efetch and Elink). The proxy server stores queries, results, and citations to provide a quick response to recurring queries and fast delivery of frequently requested citations. The proxy server also monitors performance measures and accumulates aggregate statistics to help in developing clustering and ranking tools. The client program is responsible for the user interface and for storing user-specific information, such as preferred search strategies or recurring queries. FY 2004 upgrades, implemented as a result of user feedback, have significantly improved PubMed on Tap usability.

PubMed for Handhelds also explores hand-held technology for use in the clinical setting. During FY 2004 several new features were introduced. PICO (Patient/Problem, Intervention, Comparison, and Outcome) is a method used for developing well-formulated clinical queries. This format can also be used for structuring literature searches and may be helpful to those interested in evidence-based medicine. In support of users of newer handheld devices that feature WAP browsers (mobile phones, hybrid PDA-phones) the system has been reformatted. Current services offered are clinical queries, systematic reviews, PICO searching without filters, journal abstracts browser, and access to ClinicalTrials.gov.

Additional projects targeting the use of handheld devices as a portal to information dissemination continued to expand in FY 2004. The Biomedical Informatics and Pathology departments at the Uniformed Services University collaborate with Center staff to provide wireless (e.g., infrared, Bluetooth, 802.11b) PDA access to PubMed, MEDLINE, and other NLM databases during small, medical student group discussions. PDAs will allow students to electronically submit reports and case summaries, which is expected to enhance their interactions with teachers.

The ASKLEPiOS project (Access to Services and Knowledge, multiLingually, Everywhere, Portably, in Open Source) seeks to explore the integration of portable wireless hand-held devices together with non-mobile computer servers and telephones. The integration framework is built with open source tools and includes internet-based telephony, videoconferencing, wireless data services, speech recognition/synthesis services, and a robotic chat service. The framework provides the needed "middleware" layer upon which applications relevant to the mission of the NLM can be built. Portable personalized devices with visual and speech-based interfaces may

prove helpful in delivering health care to an increasingly multicultural and multilingual society. Through collaboration with external groups, the project focuses on technologies such as information servers, speech synthesis and recognition software, handheld personal computing devices, wireless networking, and the public-switched telephone network.

The Collaboratory for High Performance Computing and Communication investigates innovative means for assisting health science institutions in their use of online distance learning technologies. The Collaboratory also explores advanced computer and network technologies for distance interactivity, including wireless technology and virtual reality research. Major upgrades to existing videoconferencing codecs were accomplished and new codecs were added in FY 2004. Several significant demonstrations were performed using videoconferencing technology, both at NLM and off site at national meetings. Demonstrations of streaming and wireless webcasting were done and videoconferencing and webcasting were employed routinely in program activities. One significant upgrade was the purchase of a Click-2-Meet videoconferencing server that allows end points to tunnel through firewalls. The new software required a significant upgrade in computer hardware. Hardware upgrades were also needed for webcasting and it appears that dual processing machines are increasingly required. Experiments continued using the conventional h.323 videoconferencing technology with Charles R. Drew University of Medicine and Science and its affiliated medical magnet high school for minorities, the King-Drew Medical Magnet High School. A pilot videoconference featuring NLM librarians was completed. The h.323 videoconferencing technology was also employed in a virtual site visit of the NLM funded medical informatics program at the University of Missouri. As a result of the phase out of Access Grid version 1.1, another major upgrade was undertaken with the Collaboratory Access Grid node. A commercially developed software application was purchased in order to use the commercial software for standard applications, while also experimenting with open source beta versions. In addition to utilizing the new software, the audio for the current node was upgraded with a state of the art echo cancellation system. Additional microphones to accommodate group interaction were also purchased. The Access Grid node was used in NLM's tutorials on advanced networking at the 2003 Annual Meeting of the Radiological Society of North America, co-sponsored by NLM and Internet2.

The EtherMed database of web accessible health professions educational materials continued to be expanded through collaborations with colleagues at the University of Utah, UCLA, and the University of Oklahoma. A major FY 2004 upgrade allows outside individuals to nominate web sites and enter information for later review. After initial testing, this improvement is expected to simplify the task of identifying sites to be included in the database. Another major review of EtherMed was completed using an NLM developed set of search queries. Additions to the database are being held until the research is complete. A research study in collaboration with the University of Alabama, Birmingham is expected to start in the near future.

Scalable Information Infrastructure

The purpose of the Scalable Information Infrastructure (SII) initiative is to encourage the development of health related applications of scalable, network aware, wireless, geographic information systems, and identification technologies in a networked environment. The initiative focuses on situations that require, or will greatly benefit from the application of these technologies in health care, medical decision-making, public health, large-scale health

emergencies, health education, and biomedical, clinical and health services research. Projects must use test-bed networks linking one or more of the following: hospitals, clinics, health practitioners' offices, patients' homes, health professional schools, medical libraries, universities, medical research centers, laboratories, or public health authorities.

FY 2004 began the first year of a three year effort for eleven Scalable Information Infrastructure (SII) research contract awards. Several SII projects have already made notable progress; an early prototype system using wireless networks, GPS, RF tags, and handheld and wearable computers was developed by the University of California, San Diego; an auditorium-scale presentation of 3D anatomy and collaborative surgery with haptics was conducted with Stanford University and collaborators in Australia; a monitoring system was implemented for the Project Sentinel Collaboratory information security program at Georgetown University; a secure XML medical record template for individuals was developed at the Children's Hospital in Boston, and; significant progress was made in viewing and manipulating 4D datasets through the "4D Visible Mouse Project" at the Pittsburgh Supercomputing Center, Carnegie Mellon University.

Telemedicine

The Telemedicine Information Exchange, sponsored by the NLM, is a web-based resource of telemedicine and telemedicine related activities maintained by the Telemedicine Research Center in Portland, OR. During FY 2004, approximately 727 non-NLM bibliographic citations and other records were delivered to the NLM. The University of Pennsylvania Dental School completed its NLM-sponsored project during this past year. Given the declining manpower in dentistry, limited training facilities, and the increasing cost of dental education, the project considered the feasibility of providing a distributed program of dental instruction.

The virtual microscope project has been initiated by in-house staff. The project team has developed a web-based system that allows users to view an image in an interactive manner, simulating the experience of examining a slide under a microscope. Potential applications of the tool include medical education, quality control and diagnostic proficiency surveys, and telemedicine. Staff continue to participate in the monthly meetings of the multi-agency Joint Telemedicine Working Group. Participating in this group, Lister Hill Center staff made a formal presentation to Congress and the Administration on state-of-the-art Telemedicine and e-Health projects and solutions.