THE LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS

A research division of the National Library of Medicine

Lister Hill National Center For Biomedical Communications Annual Report FY 2003

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
FY 2003 ANNUAL REPORT

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 National Library of Medicine. 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 the 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 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.

Lister Hill Center research activities fall into several broad categories. Our training program brings many talented individuals to the Center to learn from and collaborate with our research staff. Language and knowledge processing research involves basic research in medical language processing and medical knowledge representation. Image processing research involves the development of algorithms and methods to effectively process biomedical images of all types. We develop and continue to support a number of information systems, all of which are informed by our basic research activities.

The Lister Hill Center is organized into five major components. The work of each is described below, and 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. Staff members participate in the meetings of the Internet Engineering Task Force. CSB staff also coordinate the many training activities of the Center. 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 involve 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 collections of prominent biomedical scientists. Several Branch projects address the challenges involved in providing health information to consumers. ClinicalTrials.gov, developed by the Branch, is an excellent testbed for conducting consumer health informatics research, as is the newly released Genetics Home Reference, which provides information about genes and diseases to the public. The most current information about the Cognitive Science Branch can be found at http://lhncbc.nlm.nih.gov/cgsb/.

Communications Engineering Branch

The Communications Engineering Branch (CEB) is engaged in applied research and development in 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 the NLM's other information programs. 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. 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 rapidly changing field. Included within the Branch 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. It coordinates NLM's HPCC planning, research and development activities with Federal, industrial, academic, and commercial organizations, and it collaborates with Lister Hill Center research branches and NLM Divisions in the development, operation, evaluation and demonstration of HPCC research programs and projects. In addition, it 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 HPCC. The major research activities of the office center on the Visible Human Project, NLM's Next Generation Internet Program, including 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 Medical Informatics Training Program, ranging from a few months to more than a year, is available for visiting scientists and students. Fellowship programs may be as short as eight weeks or as long as one year, with the possibility of being renewed for a second year. 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 open to all interested members of the NLM and NIH community.

In FY 2003 we provided training to 48 participants from 18 states and 6 countries. Participants worked on projects in the areas of biomedical knowledge discovery, consumer health systems, history of medicine, image database research, information retrieval research, just-in-time

information systems, knowledge based research, natural language processing, ontology research, palm technology, semantic Web research, text mining, distance education, and visualization. We continue 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 preceptorship of expert NIH research staff. We maintain our focus on diversity through our 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 existing NLM Medical Informatics Trainees provides fellows an opportunity to learn about National Library of Medicine 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 website under "Training Opportunities". Interested individuals will find descriptions of each of the training programs including specific application procedures..

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. Natural Language Processing Research investigates how to build and improve systems that understand the meaning of human language in order to mediate between the questions of users and the information systems they seek to use. The successful integration of Lister Hill Center developed research techniques with other information retrieval strategies has the potential of contributing to the resolution of some of the most difficult problems underlying biomedical information management.

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: the 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 linguistic and domain knowledge in the Unified Medical Language System knowledge sources.

Lexical Systems
The Lexical Systems group builds and maintains 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 developed and maintained by the lexical systems research team. LexBuild allows researchers to enter items directly into a central database via a Web browser. As items are

created, the Java-based system eliminates time consuming uploading and unnecessary errors. New items are flagged for review when they are ready for release. The SPECIALIST Lexicon increased by over 12% to 183,000 lexical items in the FY 2003 release.

The SPECIALIST lexicon records the spelling variation inherent in English orthography, and this, together with a set of spelling suggestion rules and techniques, has been incorporated into the lexical access tools, which are freely available to the research community. Several additional modules developed by the Lexical Systems group have been completed recently and are now available independently as tools for a variety of natural language processing projects. These modules include a tokenizer, a lexical look-up utility, and a noun phrase extractor. Lexical access tools, including LVG, wordind, and norm, are also distributed with the UMLS.

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 informatics. 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. MMTx allows users to exploit the UMLS MetamorphoSys program to exclude or reorder specific Metathesaurus vocabularies. Users also 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.

The development of SemRep, a tool that uses the Semantic Network to determine the relationship asserted between concepts developed in MetaMap, underlies the increased understanding of viable strategies for effective natural language processing. 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. A modification of SemRep, called SemGen, 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.

Word-sense ambiguity in language constitutes a major impediment to accurate management of biomedical text through automatic strategies. The semantic knowledge representation project recently implemented a general framework for research in word-sense disambiguation. The framework depends on UMLS Metathesaurus concepts provided by MetaMap. The implementation is written in Java and includes modules that accommodate multiple disambiguation methods, as well as an "arbitrator" for managing output from these methods.

Project resources are being applied to a variety of research initiatives aimed at identifying specific biomedical information in MEDLINE citations, including semantic predications asserting a treatment relationship between drugs and diseases. Several projects focus on molecular biology. One project seeks to identify genes, gene products, and gene functions in abstracts and compares this information to that found in the Gene Ontology. Another project supports comparison of protein function by identifying protein-protein interactions in text. A third project uses semantic information to support text-based knowledge discovery systems in molecular biology.

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 National Library of Medicine. 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 is based on three core indexing methodologies. The first methodology calls on the MetaMap program to map citation text to concepts in the UMLS Metathesaurus. The second methodology, the trigram phrase algorithm, uses character trigrams to match text to the Metathesaurus. The third methododology uses a variant of the PubMed related articles algorithm to find MeSH headings by using existing indexing terms on articles similar to the input text. Results from the three methods are restricted to MeSH and combined into a ranked list of recommended indexing terms. Substantial progress has been made in applying the MTI system to both semi-automated and fully automated indexing environments at the NLM. MTI recommendations are now available to all indexers. In addition, results of the MTI system have recently been assigned as keywords for AIDS/HIV, health sciences research, and space life sciences collections of meetings abstracts that will not be manually indexed. These collections are accessible via the NLM Gateway. As part of the research underlying the Indexing Initiative, the Journal Descriptor (JD) project investigates fully automated indexing based on the NLM's practice of maintaining a subject index to journal titles using a set of 127 MeSH terms corresponding to biomedical specialties. The system associates journal descriptors (JDs) with words in journal titles and abstracts in a two-year training set of approximately 910,000 MEDLINE records. Each record "inherits" the JDs from the journal title in the record, and then each word in the training set can be described by a list of JDs ranked according to the number of co-occurrences between the word and the JDs.

Unified Medical Language System

Unified Medical Language System research regularly develops and distributes multi-purpose, electronic knowledge sources and associated lexical programs. Products such as 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 2003, there are more than 900,000 concepts with 2.5 million names from 102 source vocabularies in 15 languages. The scope of the Metathesaurus is also growing to include the current and candidate Department of Health and Human Services (DHHS) standard vocabularies under the Health Insurance Portability and Accountability Act (HIPPA) in a common format and with increasing interconnections. The Metathesaurus' mission has grown to include the distribution of HIPPA vocabularies for clinical use in the United States.

In July 2003, DHHS Secretary Tommy Thompson announced the government-wide license for the SNOMED Clinical Terms (SNOMED-CT), a key clinical vocabulary. Under the government-wide license, SNOMED-CT will be distributed in the Metathesaurus and will be freely available for all U.S. health care systems. SNOMED-CT contains 344,000 concepts with 913,000 names and 1.3 million relationships. The resulting increased visibility and national expectations for the Metathesaurus have added new demands for quality, currency, and the incorporation of additional vocabularies and mappings. In order to uphold NLM quality standards, all new and updated vocabularies are required to be demonstrably complete, contain full attribution, and be correct in the Metathesaurus. The new "Source Transparency" requirement has led to a major redesign of the Metathesaurus editing, production, and release management systems. A new Rich Data Format (MR+) for transparent releases has been defined and is targeted for a first release with SNOMED-CT early in calendar 2004.

Work on the NLM's new clinical drug vocabulary RxNorm and on the Gene Ontology will continue until the 2003AC release of the UMLS Metathesaurus. Significant efforts in FY 2003 have made possible the creation and editing of RxNorm within the Metathesaurus editing system.

Collaborating with researchers from the University of Amsterdam, Lister Hill Center staff have completed the development of an interactive editing and collaboration interface for the International Classification of Primary Care (ICPC) medical vocabulary. The ICPC contains concepts in 20 different languages including Hebrew, Japanese, Russian and Greek with their character sets represented in Unicode. A platform independent Web-based system using the open source tools Apache/PHP/MySQL has been developed. The Web-based ICPC system allows multilingual display and editing for clients that have no Unicode capability by means of a Java applet and server-side Unicode manipulation.

The UMLS Knowledge Sources are made available over the Internet through the Knowledge Source Server (KSS). The KSS incorporates several features that allow fast and direct access to UMLS data. For example, users can request information about a particular concept in the Metathesaurus, including definitions, semantic types, and synonyms as well as other concepts that are related to the input term. The Knowledge Source Server also accommodates navigation in the Semantic Network, allowing users to investigate relationships among semantic types and relations or to retrieve a list of Metathesaurus concepts assigned to a particular semantic type. Finally, data in the SPECIALIST lexicon provides users with syntactic and morphologic information about each lexical item. The most recent release of the Knowledge Source Server

incorporates several features designed to enhance performance by allowing faster access to UMLS data, providing flexibility through a rich API set, and facilitating scalability in handling ever-increasing user loads and constituent vocabularies. The architecture includes a Web server implemented as a collection of Java servlets that provide quick and easy access to UMLS data.

The KSS server software connects through the Internet to a backend Remote Method Invocation (RMI) server, which processes all requests for data by first accessing a relational database to obtain relevant information and then forwarding the data through the Internet to the requestor. Open source software from Apache was used for the development of all aspects of the system.
In addition to enhancements to the user interface, XML has been incorporated into the design of the Knowledge Source Server to provide flexibility in delivering data to users. There is an object model for Metathesaurus data that allows users to access XML documents produced by the Knowledge Source Server and to manipulate the data in an object-oriented fashion within their programs. The object model provides a mechanism for representing concepts and related data consistently among developers.

The Terminology Server provides tools to manage diverse medical vocabularies for various purposes. Throughout FY 2003, the project achieved a significant milestone in providing customized vocabulary data sets to ClinicalTrials.gov, Profiles in Science, and Genetics Home Reference. An important function of the Terminology Server is to allow individual users to customize terminologies from the UMLS and other sources. Filters are being developed that will help users select subsets of medical terms. The first filter will identify UMLS term variants suitable for natural language processing. Another task of the Terminology Server is to develop models for handling UMLS data retrieval, data maintenance and periodic data updates. In addition, research is focused on developing tools to create and edit "local," non-UMLS terminologies. The project will continue to integrate tools with existing applications and provide updates to the application data sets corresponding to the latest releases of the UMLS.

Medical Ontology Research
Medical Ontology research focuses on the development of a medical ontology that will enable various knowledge processing applications to communicate with one another. Creating a usable ontology requires the definition, organization, visualization, and utilization of semantic spaces created from biomedical knowledge processing applications. Although the UMLS is used as the primary source of medical knowledge, OpenGALEN, Gene Ontology, and WordNet are being explored as well. During the past year, research focused on two subdomains of biomedicine: ontology and terminology. In one project, the representation of anatomical concepts in the Foundational Model of Anatomy and GALEN were compared. In other projects, the research team studied terminological differences across clinical records and biomedical literature. Redundancy in hierarchical relations in the UMLS was also investigated. Finally, the team developed methods for navigating between phenotype and genotype information as well as visual methods for exploring the UMLS semantic groups.

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) 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 1800 licensees in 47 countries. FY 2003 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.

FY 2003 has seen the development of certain anatomical applications that are focused on allied health, undergraduate medical and dental education, and continuing medical education. The concept has been to develop applications that bring visible human data into the classroom and laboratory teaching environments. A second focus has been the final period of performance of the Anatomical Methods contracts designed to address mitigation of method-based artifact generation, and to increase the resolution for detection, discrimination and identification of minute neurovascular structures. These studies have been conducted at University of Colorado and Brigham and Women's Hospital.

With research support from the NLM, the University of Colorado Health Science Center, Center for Human Simulation has developed a first release web site version of a head and neck atlas titled "Functional Anatomy of the Visible Human: Version 1.0 The Head and Neck." The atlas is designed in educational modules covering the topics of mastication, deglutition, phonation, facial expression, extra-ocular motion, and hearing. QuickTime movies have been produced using live human subjects portraying the function of the regional anatomy described from a surface anatomy perspective. Tools include basic anatomic structure identification, a model builder, orthogonal plane browser, and links to the PubMed Web site for automatic key word searches of the literature.

The Visible Human Dissector was added to the atlas website in FY 2003. The Dissector provides access to 3D renderings of Visible Human anatomy as a virtual cadaver. The virtual cadaver includes identified anatomical structures and the cross-sections from which they were derived. The Dissector can be used as a free-form reference or navigated with user-created lesson plans. This unique application provides the ability to approach the human body from any combination of traditional views, including cross-sectional, regional, systemic, clinical, surface and surgical anatomy perspectives. Additional content in the basic anatomical structures involving small dimension neuromuscular connections, clinical cases, and surgical approach discussions to certain relevant procedures was also added to the atlas web site. Certain muscles that were

segmented in the Mastication module were deformed using spline technology to mimic normal musculoskeletal system function.

Two groups are investigating advanced anatomical methods using the Visible Human data with research support from the NLM. Brigham and Women's Hospital is investigating the problem of soft tissue expansion due to the use of frozen tissue required for the cryosectioning process used by the University of Colorado to create the original Visible Human datasets. The problem appears to be solved through the development of a completely different tissue preparation method. First the teeth are de-mineralized in order to achieve improved sectioning. In the original datasets these small objects became brittle and broke off. The hardware used was modified to allow for MRI registration with fiducial screws manufactured from an MRI compatible aluminum alloy. Artery filling (red) and vein filling (blue) was demonstrated to sub-millimeter level. Preliminary results indicate that a new, complete data set at a resolution of 0.1 mm (100 microns) in each of the three dimensions with all artifact problems successfully eliminated can be created. The investigators have already demonstrated an increased voxel resolution in the head from the Visible Human female's 0.33 mm to a resolution of 0.15 mm. Each new transection was cut at a section thickness of 142 microns on an ultracryomicrotome. This allows the collection of a slice of a complete transection, in contrast to the milling method used in the original Visible Human technique. Single intact transections have been histologically stained to differentiate neurovascular structures from adjacent connective tissues.

The Colorado investigators are examining techniques to improve their original milling based method. Tissue differentiation through multi-spectral imaging to enhance automated segmentation is being attempted. Ultraviolet illumination and visible wavelength fluorescence appears to be very promising. Spectral imagery recordings were taken following excitation in the ultraviolet region of the spectrum. What are interpreted as distal peripheral nerves in muscle tissue were seen for the first time in 100 micron sections identified by their intrinsic fluorescence. Recordings were made of the reflectance patterns as a narrow aperture ultraviolet scanner captured the absorbance characteristics of the anatomic structures. Spectral profiles of the basic tissue types were obtained. Surface freezing between slices has been successfully accomplished and automated with manual intervention every hour. Continuous cutting has been achieved for a time of 36 hours. This supports the concept of continuous cutting 24 hours per day from head to foot of an entire human. This process will reduce banding present in the Visible Human Male data and stabilize tissues with a continuous freeze.

The Insight Toolkit (ITK), a research and development initiative under the Visible Human Project, began official software releases of the software during FY 2003. 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 consortium includes General Electric Global Research, Kitware, Insightful, the University of North Carolina Chapel Hill, the University of Pennsylvania, the University of Utah, Harvard University, the University of Pittsburgh, and Columbia University.

FY 2003 saw explosive growth of the ITK research community. Accompanying the official ITK 1.0 software release, NLM made additional awards to exercise and integrate the software infrastructure into clinical and research applications. Research institutions including the Mayo Clinic, the Carnegie Mellon University Robotics Institute, Georgetown University Medical School, Imperial College London and Guys Hospital, have joined the development team. Non-funded researchers from across the world are now testing, developing and contributing to ITK in over 30 countries. At the end of FY 2003, ITK v1.4 was released, including the components developed as part of the funded research supported by NLM. By the end of this fiscal year, we will have attained our primary goal of creating a strong, usable, public, open-source software infrastructure to support medical imaging research.

3D Informatics

During FY 2003 the 3D Informatics Program has continued to mature and develop its in-house research efforts around problems encountered in the world of 3-dimensional (x,y,z and x,y,t) 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. We have extended and enhanced our pilot project for creating the framework for an archive of volume image data, the National Online Volumetric Archive. 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.

Research is continuing on template guided interventions, including joint work with the USUHS Orthopedics Department and the Bethesda National Naval Medical Center Radiology Department on total hip resurfacing arthroplasty and the planning and fabrication of surgical templates with NNMC. A surgical planning workstation and custom-built drill template was devised from CT scans to accurately place guide pins in the femur head during total hip resurfaces. This work is related to our previous work in Patient Specific Surgical Instrumentation for spine surgery. Efforts are also continuing in the area of data driven modeling with implicit surfaces with colleagues at UMBC (smooth surface generation from binary volumes) and UNC-C (reconstructing implicit surfaces from contours from arbitrary ultrasound slices). Finally, we have begun explorations in artistic and non-photorealistic rendering of digital models. A project in laser scanning of physical artifacts was undertaken in FY 2003 as well as the software design of a layered architecture for implementing medical illustration techniques using computer graphics technologies.

AnatQuest

AnatQuest provides widespread access to the Visible Human images for a broad range of users, including the lay public, which is frequently limited to low speed Internet connections. It build on earlier projects which focused on developing an object-oriented database for the images, establishing an FTP server for access to the high resolution version of the images, and developing tools for processing the images. The AnatLine system developed earlier allows access through anatomic terms to high resolution cross-sectional images and segment masks (useful for rendering anatomic objects). The tools developed to use AnatLine are VHParser and VHDisplay. The first is for unpacking the data files into their individual components (cross-section images, byte masks, coordinate and label tables, etc.), and VHDisplay is for displaying both cross-sectional and rendered images.

The new system, AnatQuest, is a Web system 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. 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. In addition to its main purpose, AnatQuest serves as an access point for AnatLine as well as for about 300 surface-rendered objects, the majority of which were created at the Lister Hill Center and the rest acquired from outside sources (e.g., VoxelMan). Also through AnatQuest, the public can access the FTP server for bulk transfer of high resolution image files.

An initial prototype of a system linking MEDLINEplus to the anatomic image database has recently been developed. This system relies on a proxy server developed to intercept user requests to MEDLINEplus. The proxy server first retrieves the MEDLINEplus page that satisfies the user query. In parallel, the proxy server sends the user query to the AnatQuest image server which uses the UMLS Knowledge Source Server and a term mapper module to map the query terms (mostly disease terms) to the corresponding anatomical structures. The term mapper module addresses the problem of identifying appropriate anatomical terms corresponding to the biomedical terms in the document. These biomedical terms are likely to be disease terms rather than explicitly anatomical ones. The module uses the location-of concept relationship in the UMLS Metathesaurus to map a biomedical term to a related anatomical term. For example, the term “pneumonia” (a disease) could be mapped to “lung”, the underlying organ for the disease. The links to these images are then inserted by the proxy server as hotlinks in the image section of the MEDLINEplus page, which is then returned to the user.

In addition to the Web-mediated version of AnatQuest, a kiosk version was developed for the Dream Anatomy exhibit at NLM as a Java application suitable for onsite patrons using a touchscreen monitor. To eliminate dependency on the network for the retrieval of VH images, and to speed up the kiosk operation, the images were also stored in the local machine. The effort to redesign the GUI in the AnatQuest web-mediated system for the kiosk application has been to tailor the displayed icons, buttons and other graphical elements to allow convenient human interaction via a touchscreen.

WebMIRS

The Web-based Medical Information Retrieval System (WebMIRS) allows 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. WebMIRS allows a user to control a graphical user interface to construct a query for the NHANES II or NHANES III data. A sample query might be equivalent to the statements: “Find records for all individuals who reported chronic back pain. Return their age, sex, race, age when the pain began, and longest duration of

pain. Also, return the record data required for statistical analysis and display their x-ray images." 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. These data consist 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 done this year 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 decompression at four different levels. Project staff have begun the implementation of new design architecture to provide a software framework for the incorporation of new text/image databases in a much more general way than the current WebMIRS, and to provide new features for the database end user that extend current WebMIRS capabilities.

Digital Atlas of the Cervical and Lumbar Spine
This is a dataset of cervical spine and lumbar spine images with interpretations validated by a consensus of medical experts, along with software to display and manipulate the images. The images in the Atlas were chosen from the 17,000 images collected in the NHANES II survey. For the cervical spine images, the Atlas contains numerical interpretations or "grades" for anterior osteophytes and disc space narrowing, on a scale from 0-3, with 0 being "normal" and 3 being "most abnormal"; and also interpretations for subluxation, on a 0-1 scale, with 0 being "normal" and 1 being "abnormal". Similarly, for the lumbar spine images, the Atlas contains interpretations for anterior osteophytes and disc space narrowing, on a scale from 0-3. The Atlas user may display single or multiple images in order to view, for example, all grades from normal to most abnormal of anterior osteophytes in the cervical spine. Image processing capability is provided to assist in contrast enhancement for viewing of detail. The Atlas may be accessed either as a Java applet, or downloaded as a Java application, from the project website. In addition, we provide a version of the Java application on CD. The Java application version allows the user to add his/her own images (either grayscale or color) in a special "My Images" section, and to annotate and title those images for later use. This year the Atlas was enhanced by the addition of capabilities to display color images, add extensive text annotations, and import/export sets of images and annotations as a package.

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. For viewing the x-rays, we have created the ImViewJ software, a Java application that may be downloaded from our Web site and which allows the viewing of the images at their full spatial

resolutions (1463x1755 for the cervical spine images, 2048x2487 for the lumbar spine images). For 550 images we also have coordinate data collected under the supervision of a radiologist at Georgetown University. 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.

Content-Based Image Retrieval

The Content-Based Image Retrieval (CBIR) project develops methods for effective extraction of biomedical information from digital images of the spine. CBIR focuses on the computer-assisted indexing of image data, as well as the ability to search image data. Computer-assisted image searching is a potential enabler of enhanced information extraction from a database that has already been indexed. The most popular form of this type of search is query by example or a variant, query by sketch. In query by example, the user inputs an image from a set of choices provided by the system or by providing a new image, and queries the database with respect to one or more characteristics of the example image (e.g., shape, histogram, or texture). In query by sketch, the input image is replaced by a sketch of the image made using drawing tools provided by the system. In either case, the system analyzes the input into component features and searches the database for images with similar features. Results are usually returned as a similarity ranking.

Developing a computer-assisted indexing system poses many challenges. For example, the only indexing data available for the NHANES II images is the collateral (alphanumeric) data collected in questionnaires and examinations. There is no indexing information available that has been derived directly from the images. The prohibitive cost of employing radiological experts to compile and interpret indexing data means that it is unlikely that NHANES II indexing information will ever be acquired manually. However, indexing data might be acquired if reliable, biomedically validated software could automatically produce image interpretations. Even the development of semi-automated methods could sufficiently reduce labor costs to allow the creation of databases of significant biomedical information. CBIR research seeks to develop computer-assisted image indexing to acquire data, and at the same time, reduce the overall costs of indexing.

An initial prototype Content-Based Image Retrieval system (CBIR1) was implemented in FY 2001 for the retrieval of images based on simple vertebral shape models. The program allows users to specify a search for up to nine control points and the geometric configuration of these points to define an approximate vertebral shape. The prototype database contains 100 cervical and lumbar images with the ability to rotate and scale every vertebra in each image to identify the best match to the input shape. Alternatively, the user may specify an example vertebra and the program will search for the best shape match to the example. In FY 2003, A second CBIR prototype (CBIR2) was implemented. CBIR2 is significantly enhanced, including an *indexing function* with the capability to perform active contour segmentation, create detailed representations of vertebrae boundaries, and to convert boundaries into multiple shape representations (e.g., global shape descriptors, invariant moments, polygon turn functions, and Fourier descriptors). In addition, a retrieval function supporting the retrieval of shapes by any of

the shape representations was implemented. CBIR2 also includes NHANES text data and supports query by sketch, image example, or text, in addition to hybrid text and image-based queries. The MySQL database system was incorporated into the retrieval function for the storage and retrieval of text data. Current CBIR work is directed towards the completion of the segmentation functions for indexing, analysis of effectiveness of the various shape methods implemented for spine x-rays with significant osteoarthritis features, implementation of spatial data trees for feature vector organization, and the creation of a database of segmented vertebrae of significant size and accuracy to serve as test-bed data for ongoing CBIR work.

Engineering Laboratories

The Document Imaging Laboratory supports 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, optical character recognition 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 100 Mb/s Ethernet, for performing document image processing. Both in-house developed and commercial systems are integrated and configured to serve as laboratory test-beds to support a variety of research. The Image Processing Laboratory is equipped with a variety of high end servers, workstations and storage devices connected by 100 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, the laboratory also archives a variety of image content. Most of the machines housed in the laboratory are equipped with multiple networking ports (e.g., FDDI, ATM, Ethernet, fast Ethernet) which allow, in addition to standard networking capabilities on the local Ethernet, the capability of alternate physical communications channels. ATM switches connect the Ethernet and FDDI networks to other local area networks throughout the Lister Hill Center, the Internet, experimental ATM, Abilene, and the infrastructure for the Next Generation Internet and Internet-2 initiatives. The Document Image Analysis Test Facility is an off-campus facility containing high-end Pentium workstations and servers for the MARS production system. While routinely used to produce bibliographic citations for MEDLINE, this facility also serves as a laboratory for research into techniques that are fundamental to the automated extraction of descriptive metadata for the long term preservation of document images. Techniques include automatic zoning, labeling, and reformatting of bibliographic fields from document images, as well as intelligent spell-check by pattern recognition and other key elements of MARS. Besides real time performance data, the Document Image Analysis Test Facility also collects and archives 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, is being used as the future for improved electronic image quality. Multimedia systems, scientific visualization and networked media are being pursued for their performance, educational, and economic advantages. Project staff conduct research in three dimensional computer graphics, innovative animation techniques, and photorealistic rendering. Research into digital video and image compression techniques is being used in projects requiring the storage of large images and rapid data transmission. Project staff completed an evaluation of one of the leading digital media asset management systems designed for the video environment in FY 2003. The system includes a video database fully integrated with a logging and digitizing system. Based on the evaluation results, the system vendor's engineers have been working with Center staff to redesign the software.

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 (e.g., PubMed). 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, project staff developed two prototypes; a Board of Scientific Counselors meeting as a CD-ROM; the 2002 Leiter Lecture ("Genomics, Medicine and Society" with Dr. Francis Collins) as a CD-ROM and as a DVD.

Three prototype DVDs representing the Once and Future Web Exhibition were developed in FY 2003. The initial prototype DVD demonstrated overall design concepts, the implementation of Web-enhanced DVD technology, a navigable text viewer and virtual object manipulation within the program. Source media for the DVD included still imagery, three-dimensional video graphics and high definition video. Feedback on the content, organization and design was incorporated into the second prototype. A new segment of the program focusing on the NLM's Telemedicine Program, including video content, has been added. MPEG encoding, DVD authoring and Web DVD programming were ongoing throughout the development process. The third DVD, Prototype version 1.1, was demonstrated in FY 2003. Additional feedback on interface design, navigation and content will be integrated into the final design and production of the DVD. Additional enhanced video graphic animations have been completed and included in the "Show Me How it Works" section of the program. The identification and integration of Web sites to enhance fixed DVD content was finalized and incorporated into the program's companion Web portal. Investigation of Web-enhanced DVD programming for delivery on multiple platforms has been a key element in the overall development of all prototypes.

In consultation with the Office of Communications and Public Liaison, and the HMD Exhibition Program, project staff have been working with MacNeil/Lehrer Productions on planning the developmental phases of a Web-enhanced DVD for the NLM exhibition, Changing the Face of Medicine: Celebrating Americas Women Physicians. The DVD is to serve as the prototype for the subsequent Local Legends DVD, a collaborative project between the NLM and the American Medical Women's Association.Video interviews of twelve physicians identified for inclusion in the first DVD have been conducted in twelve cities, and they include Dr. Tenley Albright, Dr. Julie Louise Gerberding, Dr. Donna Christian-Christensen, Dr. Nancy Snyderman, and Col.

Rhonda Cornum. These interviews are part of an up-close and personal interactive video profile of each doctor. A youth mentoring program in California was video recorded and will be included in the prototype. All of the video was recorded on high definition video to assure video quality for production and archive purposes. Overall interface and navigational designs were developed, and include transitional segments featuring young adults.

Information Systems

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

Digital Library Research
Digital library research investigates all aspects of creating and disseminating digital collections including standards development, investigation into emerging technologies and formats, discussion of copyright and legal issues, effects on previously established processes, the protection of original materials, and the permanent archival of digital surrogates. Research issues currently include the long-term preservation of digital archives, innovative methods for creating and accessing digital library collections, and the development of modular and open information environments. Investigations concerning interoperability among digital library systems, the role of well-structured metadata, and varying "points of view" on the same underlying data set are also being pursued.

The Profiles in Science website 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 collections of Donald S. Fredrickson, Fred L. Soper and Florence R. Sabin were added in FY 2003, bringing the total number of archives for prominent biomedical researchers, medical practioners, and those fostering science and health to eleven: Christian B. Anfinsen, Oswald T. Avery, Julius Axelrod, Donald S. Fredrickson, Joshua Lederberg, Barbara McClintock, Marshall W. Nirenberg, Linus Pauling, Martin Rodbell, Florence R. Sabin, and Fred L. Soper. The Reports of the Surgeon General (1964 - 2000) and the history of the Regional Medical Programs (1964 - 1976) are also available. on *Profiles in Science*.

In FY 2003, project staff continued to enhance the effectiveness of Profiles in Science.
A new user interface was developed to standardize consumer navigation, a link to the Reports of the Surgeon General was added and three new categories were established, Biomedical Research, Health & Medicine, and Fostering Science & Health. Enhancements to the underlying Profiles in Science digital library infrastructure include improvements to the back-end database and the development of new methods for viewing database data, detecting and correcting errors, and automatically updating data. Finally, the development of an XML-based front end and transition to a new XML-based search engine continue to be pursued.

The Lister Hill Center collaborates with the History Office of the Food and Drug Administration and the National Institutes of Health Historical Office on preservation efforts for the Public Health Service (PHS). An exhibit on the history of smallpox, on display at the NLM in 2002, was developed into an online exhibit for the NLM Web site in FY 2003. Staff worked to develop a database of resources on the history of African-Americans in medicine and also conducted research on the history of the Public Health Service's involvement in National Negro Health Week. Staff continue to answer historical queries about the history of the PHS and actively work to preserve documents and artifacts related to PHS history.

MARS

Document image analysis and understanding research combined with database design, GUI design for workstations, image processing, string pattern matching, lexical analysis, speech recognition and related areas underlie our development of MARS (Medical Article Records System), a system to automate the production of MEDLINE records from biomedical journals. From bitmapped images of the first page of the articles, this system is designed to automatically extract the article title, author names, institutional affiliations and the abstract. Research investigations center on the identification of rules for algorithms for page segmentation, zone labeling, OCR error correction, affiliation ranking and other essential functions. Manual input is limited to entering fields other than the ones automatically extracted, as well as verifying the text before the records are made available to indexers. In FY 2003, research focused on improving the operation of MARS, developing a system to extract bibliographic data from online journals, and enhancing the processing of publisher-supplied citations in XML format. The work in this project also contributes to the automatic extraction of metadata from electronic resources that need to be archived for preservation purposes.

Anticipating an increasing availability of online (Web-based) journals in the future, we conducted research toward the design of a system that automatically extracts MEDLINE citation data from such journals. Areas of investigation included such techniques as breadth first search algorithm and constraint satisfaction methodology, fuzzy rule-based methods, format conversion methods (to convert PDF to HTML) and Web-based GUI design tradeoffs. Based on this research, modules were created to download issues and articles from the Web, zone and label the relevant text, implement MEDLINE conventions in reformatting the title and author fields, and selecting the correct affiliation from the many that usually appear. Specific developments included a journal-specific learning algorithm to automatically search for and download journal issues and articles, and to classify them as HTML or PDF documents; fuzzy rule-based algorithms to automatically zone and label the bibliographic data; automatically convert PDF documents to HTML for processing; and a stress-test study to investigate the adequacy of the existing cluster-based failover system in MARS to include future online journal processing.

The modules outlined above were integrated to create a prototype system called WebMARS, which was tested to automatically extract data from online journals, with and without publisher-supplied XML data. The latter function is of interest because the publishers who send NLM their XML-coded citations usually leave out Grant Numbers, Databank Accession Numbers and other fields requiring manual entry at NLM. This function requires matching the XML journal title with the title in the Web page, looping through each issue to match articles by author names and

article titles, and computing a confidence value. Tests with 3,000 XML citations and 300 journal issues on the Web are planned toward identifying a threshold for this confidence value, beyond which the online article and the XML citation can be reliably linked for further processing. Performing this function in addition to creating citations automatically from Web journals, WebMARS augments the citations sent in by the publishers, and thereby reduces the labor required in production. The prototype system was tested under realistic conditions to estimate the labor required to create complete citations as a comparison to the labor required in the alternative approaches: keyboarding, MARS, and publisher-supplied XML data, and found to be significantly more efficient than these.

Ground truth data released to the public

In August 2003, a database named Medical Article Records Groundtruth (MARG) 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 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 website is Rover, an analytic tool that may be used to compare the results of a researcher's program with the ground truth data. A paper describing MARG was presented and published at the Symposium on Document Image Understanding Technologies (SDIUT 2003) held in April. In the first two weeks after it was announced at an international conference in document analysis and recognition (ICDAR 2003, Edinburgh), MARG was accessed by more than 1,750 registered users.

DocView

The DocView project enables users to store, view, manipulate, copy, paste, email and print bitmapped images delivered through the internet. The program also serves as a TIFF viewer for compressed images. Users may receive document images via Ariel FTP or Multipurpose Internet Mail Extensions (MIME) protocols. Library patrons, for example, often use DocView to receive scanned journal articles from libraries that use Ariel software for interlibrary loan services. The number of DocView registered users increased 21% in FY 2003 (14,500 users in 181 countries). DocMorph provides additional functionality for DocView by providing the technology for users to convert files into various formats. The system enables users to convert more than 50 different file formats to PDF. Also, by combining OCR with speech synthesis, DocMorph assists the visually impaired in using library information. DocMorph continues to be used by librarians for the blind and physically handicapped to convert documents to synthetic speech that is recorded onto audio tapes. The number of DocMorph registered users increased 40% in FY 2003 (8,000 registered users). Research on DocMorph usage and the availability of new technology have pointed out new opportunities for improvements and innovations. The availability of Simple Object Access Protocol (SOAP) that combines XML with HTTP has allowed us to create a web service that significantly improves the DocMorph function used 75 percent of the time, viz., the conversion of files to PDF. This web service (MyMorph) consists of a Windows-based client software and modifications to DocMorph for accommodating SOAP. Inhouse 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. 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. This new capability is finding frequent use by document delivery librarians, and also by organizations that have used it for mass file migration.

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 electronic form, extend the virtual books into information systems, and to increase the accessibility of historical documents for the public. In 2001–2002, the NLM and the British Library collaborated in the production of two virtual books, Blackwell's *Herbal* and Vesalius's *Anatomy in Photorealistic* to create the "Turning the Pages (TTP)" format. The pages of the two books were scanned into the computer as high quality color images. The images were manually processed by Adobe Photoshop, animated by Macromedia Director, and displayed on a touch screen monitor. Consumers were able to "touch and flip through" each book on a touch screen monitor. After the initial development of the 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, etc.) 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*.

Paré's *Surgical Treatise* and Gesner's *Animalium* were selected for TTP conversion in FY 2003. Pages were digitized and enhanced to remove artifacts, edge effects and lighting non-uniformity. Compared to the initial processing of the first two books, project staff have improved the development procedures for Pare's and Gesner's books. A 3D wireframe model was developed in Maya, a modeling and animation system. The 3D wireframe model texture-maps each pair of page images to both sides of the wireframe of a turning page. A multisource lighting model provides diffuse lighting, specular highlights and shadows. For each flip of the page, twelve intermediate animation frames are generated, rendered and then imported into Director. Another improvement exploits the characteristics of the wireframe model. The node attributes within the model can be adjusted, allowing different rates and styles of curvature to be expressed during page flipping. For example, there is a choice of three flip behaviors depending on where the finger is placed on the page to start flipping (e.g., the upper right corner of the page will flip over if a finger is placed at the top right of a page). Both new books have been completed in TTP form, with Paré's book extended to the TTP+ format using the "storyline" model with explanatory visuals.

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 simultaneously searching nine document collections using five retrieval methods from a single interface. Several enhancements of NLM Gateway occurred in FY 2003. One of the most significant enhancements is the result of a collaborative effort with the Indexing Initiative. In order to improve retrieval results, all of the Gateway's meeting abstracts have been automatically indexed by the Indexing Initiative systems. Other enhancements include the addition of search filters that will allow user-specified views of the NLM information from several data collections with an effect similar to the earlier searching of AIDSLINE, TOXLINE and SPACELINE. User selectable search subsets for AIDS, Bioethics, History of Medicine, and Space Life Sciences in PubMed have been added to the NLM Gateway. The subsets are available through a pull-down menu on the Limits page or with use of a new field qualifier in advanced searches. Phrase detection has been added so that users no longer have to put search phrases in double quotes. Modifications to accommodate a new PubMed applications program interface (API), a new version of the Voyager integrated library system underlying LOCATORplus, and new XML output from NLM's Document Creation and Maintenance System have been incorporated. An API for the NLM Gateway has also been completed and is being evaluated.

PubMed on Tap

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 Personal Digital Assistants (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.

Consumer Health Informatics Research

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

In the spring of 2003 we launched the Genetics Home Reference, an integrated Web-based information system designed for consumers and others to learn about specific genetic conditions and the genes that are 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. Our 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 provides basic information, in a question and answer format, on the nature of genes and how they give rise to various conditions and diseases. For each condition, the site provides information about the specific genes linked to it, how common the condition is, and what its symptoms and available treatments are. For each gene, the site provides information about the normal function of the gene, its chromosome location, the conditions linked to the gene, and whether any gene therapy is available. Each description includes a glossary as well as alternative names for the gene or condition being described. In addition, each condition or gene description links directly to pertinent information available on a variety of other resources, including MEDLINEplus, ClinicalTrials.gov, PubMed, Gene Tests, Gene Reviews, LocusLink, and Online Mendelian Inheritance in Man.

As a further guide to users of the site, we have developed a resource called, "Help me understand genetics", which explains, together with diagrams and other visuals, some basic concepts in genetics. This resource offers, for example, easy to understand explanations of DNA, genes, proteins, chromosomes, and how genes control the growth and division of cells. In addition, the resource has sections on the nature of genetic disorders, genetic consultation and testing, gene therapy, and genomic research.

The system architecture includes three primary modules. The first module is for content collection and work flow management; the second is a "publisher" module that retrieves data from the content manager and relevant external sources, and the third is the public web site that presents the gene and condition descriptions, interlinking these with related resources, such as MEDLINEplus, LocusLink, and ClinicalTrials.gov. An important aspect of the content manager is that it tracks the status of each description, including whether it has yet undergone expert review. No description is released to the public until it has been reviewed by one or more external experts who are, in most cases, board-certified medical geneticists or molecular biologists. The system makes extensive use of the Unified Medical Language System and its constituent vocabularies. The UMLS, in most cases, provides definitions and synonyms for the glossary, and MeSH terms are used to facilitate linking to other NLM resources. The Gene Ontology is used for browsing genes by their function, by the biological processes in which they are involved, or by their cellular structure. The Genetics Home Reference currently focuses on single gene or polygenic conditions that are also topics on MEDLINEplus, the National Library of Medicine's primary consumer health site. As knowledge of genetics expands, the interrelationships between genes and diseases will continue to unfold, and the site will continue to reflect these developments.

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. ClinicalTrials.gov was launched in February 2000 and 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 website, 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.

The number of daily visitors to the site increased by over 50% from 8,000 daily visitors in 2002 to 12,000 daily visitors in 2003. The site increased the number of protocol records by over 25% from 6,600 protocol records in 2002 to 8,300 records in 2003.

Research Infrastructure and Support

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

Next Generation Networking
The National Library of Medicine completed its program to define Next Generation Internet (NGI) capabilities that will allow the NGI to be used routinely in health care, public health and health education, as well as biomedical, clinical and health services research. Collaborative capabilities include quality of service, security and medical data privacy, nomadic computing, network management, and infrastructure technology. Principal investigators were invited to the NLM for a reverse site visit as a conclusion to the NGI projects. Each of the fifteen sponsored projects were given 45 minutes to present an overview of their work and the overarching lessons learned. A video of each talk and its associated PowerPoint presentation will be posted on the NGI website. The NGI networks are being used for multimedia applications involving voice and video. The Abilene network supports full Internet Protocol multicast. In this mode, NLM can receive and transmit multicast voice and video sessions.

NLM's Internet 2 connection to the MAX GigaPOP (Mid Atlantic Exchange Gigabit Point of Presence) was increased from 155 megabits per second (OC-12) to gigabit speed (gig-E) in FY 2003. A full, native multicast is broadcast via the Center's Internet2 connection allowing the Center to implement a multimedia node on the Internet 2 Access Grid. Multimedia nodes are a collection of nodes that transmit and receive a variety of audio and video media that may be used for teleconferences and meetings. The high bandwidth and Quality of Service (QoS) characteristics of Internet2 permit the Access Grid to pass high quality audiovisual signals between nodes.

NLM continues to collaborate on the Multilateral Initiative on Malaria in Africa. Working with Infinite Global Infrastructures personnel, Lister Hill Center staff investigated approaches to overcoming the drawback of limited bandwidth in the satellite links to Africa. For example, a review was conducted of Redwing Satellite Solutions, the space segment and Internet access provider located near London. A presentation was given on the NLM's communications work and the performance measurement task at an NLM-sponsored event in Kenya.

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. Eleven Scalable Information Infrastructure research contract awards were made at the close of FY 2003.

Applications of smart card technology continue to be explored at the Lister Hill Center. Smart cards are credit card sized plastic cards with an embedded circuit chip. Cards may be used for security authentication and for data storage. Recent applications involve the use of biometrics, the storage of biomedical information (e.g., thumbprint, iris scan), in order to increase smart card security. The Lister Hill Center continues to co-sponsor the Western Governors' Association Health Passport Project, one of the largest health-oriented smart card pilot programs in the United States. The Health Passport Project stores data from multiple Federal, State and local agencies on cards used by clients receiving health benefits such as well-child care, checkups, immunizations and food benefits. Phase II of the Health Passport Project is under way in the San Diego area. Phase II will incorporate biometric authentication on the smart card, digital certificates, and trusted third party systems to facilitate the safe, encrypted transfer of private medical and demographic information over the Internet.

Utilizing its technical expertise, Lister Hill Center staff provide technical consultation and representation in a variety of environments. Project staff provided technical consultation and coordination for the Lister Hill Center's participation at the Radiological Society of North America's conference in Chicago in FY 2003. Primary responsibilities included the implementation of a Gigabit Ethernet connection from the Internet2 backbone to the McCormick Place Convention Center. Engineering staff provided technical advice and cost options for the telecommunications link that will be developed between the NLM and the site for the NLM's

backup databases. Staff also represented the NLM and NIH at the Joint Engineering Team (JET), the Internet2 Applications Strategy Council and Health Sciences Advisory Group.

The NLM continues to sponsor the Telemedicine Information Exchange (TIE), a web-based resource of telemedicine and telemedicine related activities maintained by the Telemedicine Research Center in Portland, OR. During FY 2003, approximately 526 non-NLM bibliographic citations and other records were received by the TIE. Staff continue to participate in the monthly meetings of the multi-agency Joint Telemedicine Working Group (JTWG). 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.

The Collaboratory for High Performance Computing and Communications
The Collaboratory for High Performance Computing and Communications 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 done in FY 2003 and new codecs were added. Several significant demonstrations were performed using the 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 conversion of the MPEG2 high bandwidth, high quality videoconferencing codecs from Litton to those of StarValley, since the Litton codecs were no longer supported. Several upgrades were made to the Wavelet videoconferencing codec as it continued to be refined. Finally, and perhaps most importantly, an Access Grid node was installed allowing NLM to experiment with multicast videoconferencing and participate in this form grid technology in collaboration with others in the Internet2 research community. The new MPEG2 codec and the Collaboratory's traditional h.323 codec were used in demonstrations of differences in Internet2 and commodity Internet capabilities in a week long tutorial at the Radiological Society of North America's annual meeting in Chicago. The MPEG2 codec also was employed in a demonstration of collaboration and virtual reality distance learning technology between NLM and Stanford University for the Library's Board of Regents. NLM used its Access Grid technology in multiple demonstrations with Project TOUCH, a collaboration between the Universities of New Mexico and Hawaii. East Carolina University, the University of Arkansas, and the University of Utah also participated. Wavelet technology was demonstrated at the annual meeting of the American Society of Clinical Pathology/College of American Pathologist annual meeting in Washington. Videoconferencing technology used by NGI contractor George Mason University was demonstrated at the CENDI meeting held at NLM and the MACAW workstation using the technology in the collaboratory was employed for the demonstration.

Experiments were started testing the use of conventional h.323 videoconferencing technology with NLM's Adopt-A-School Partner, Wilson High School, in Washington, DC. Some preliminary tests were also done with the Drew Medical Magnet School in Los Angles. Additional tests are planned with the aim of doing a pilot distance learning program for minority students interested in health sciences. 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. Finally, an assessment was made of alternative display technology to accommodate both the Access Grid's multi-screen displays and stereo images.

All videoconferencing and collaboration efforts have been encumbered significantly this year due to firewall policies. The firewall continues to plague current efforts to use h.323 and other videoconferencing codecs. Technologies identified for penetrating firewalls require at least one end point to be outside a firewall. Operators of the Internet2 Commons videoconferencing service recommend placing collaboration tools outside of firewalls and every NGI project funded by NLM using collaboration tools identified contending with firewalls as a critical problem during the reverse site visit.

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. Another major review of the database was conducted. A major upgrade was made to the hardware, web server, SQL server and ColdFusion server software needed to run EtherMed. Collaboration with the Heal Education Assets Library (HEAL) program funded by NLM and NSF continued, focusing on ways the HEAL program could regularly harvest EtherMed records for inclusion in their database. Several improvements to EtherMed's search methods were identified. Search terms are now highlighted in retrieval and a contract has been made to prioritize search results based on number of search term hits. A decision was made to delay the research study with the University of Alabama at Birmingham and to implement HEAL harvesting technology until these improvements were made.

System Security and Advanced Network Planning

System Security and Advanced Network Planning research focuses on computer security, the NLM network, the Next Generation Internet (Internet2 and NGI), and the upgrading of Lister Hill Center systems. A gigabit/second capacity firewall system, installed in FY 2003, has helped reduce security problems. Work on the LHC Network has continued to improve its performance and reliability. Two core routers for the LHC network are redundantly attached both to the OCCS network and to the edge routers throughout the Lister Hill Center. In addition, gigabit/second links have been provided to some desktop workstations and some servers. FY 2003 security improvements have made Lister Hill Center systems less vulnerable to external security attacks. However, the increasing prevalence of worms and viruses necessitates constant vigilance in order to keep systems up to date.