THE LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS

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

TECHNICAL REPORT
LHNCBC-TR-2007-002

The Lister Hill National Center For Biomedical Communications Annual Report FY 2007

Clement J. McDonald, M.D.
Director

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

LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS
Clement J. McDonald, M.D.
Director

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

The 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, other organizations within the Department of Health and Human Services, and academic and industry partners. Staff regularly publish their research results in the medical informatics, computer and information science, and engineering communities. The Center is often visited by researchers from around the world.

The Lister Hill Center is organized into five major components: Cognitive Science Branch (CgSB), Communications Engineering Branch (CEB), Computer Science Branch (CSB), Audiovisual Program Development Branch (APDB) (which currently includes the Office of the Public Health Historian), and the Office of High Performance Computing and Communications (OHPCC).

An external Board of Scientific Counselors meets biannually to review the Center's research projects and priorities. The most current information about the Lister Hill Center research activities can be found at http://lhncbc.nlm.nih.gov/. The Center's principal research activities and accomplishments are described in the remainder of this chapter.

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

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

NLM Personal Health Record (PHR)
The early version of this new project aims to help individuals who are caring for their elderly parent(s) and/or young children. This initial version of the NLM PHR supports entry and tracking of key measurements and test results, prescriptions, problems and immunizations. It also will produce digital and paper copies of its contents in various formats. The current version gives access to MedlinePlus information resources about prescriptions and (soon) ailments through one click on the name of recorded medication or ailment. The PHR has logic that can remind the care giver about preventive care interventions that are due (e.g. it is time for the annual flu shot, or ask your doctor about use of cholesterol lowering drugs in patients with high cholesterol). The

current system attaches codes to the medications, observations, and problems it carries. These codes come from terminologies supported by NLM and designated as national standards by HHS. The automatic inclusion of these codes within the NLM PHR will facilitate automatic downloading of clinical information to the PHR in future versions of the system.

De-identification Tools

De-identification can unlock the research potential of long term clinical records, and no well-supported and freely available de-identification tools exist. Taking advantage of experience with de-identified procedures within the NCI Shared Pathology Informatics Network (SPIN) grant in my previous position, and existing Lister Hill Center tools that recognize sensitive information such as dates, person names and locations, text, numbers, and speech, LHNCBC initiated an effort to develop an open source text de-identification tool. This project is in its early stages and has been using a scrubbed 70-patient database to test the initial tools. We are negotiating with a number of sources for narrative clinical data for test purposes. The investigators have completed their IRB training and will be taking HIPAA (Health Insurance Portability and Accountability Act) training as well.

Clinical Data Entry Tools

The initial goal of this project is to develop a tool that can generate data entry forms dynamically based on specifications stored in a database. Currently the development platform chosen is Ruby on Rails, an open-source web application framework. One potential area of use is in the data capture function of personal health records. Several terminology resources from the UMLS (e.g. RxNORM, ICD9-CM) are used in some data entry fields that require a set of controlled terms. Further development will involve work with very large databases of de-identified patient data. Additional reusable software tools will be created, including those involving biostatistical analysis with the “R” package.

Collaboration with Centers for Medicare and Medicaid Services (CMS)

LHNCBC has assisted CMS in the development of many aspects of Medicare’s Post Acute Care data collection project that will be demonstrated in the spring of 2008. One of Medicare’s goals in this project is to standardize and meld different data collection forms from four different post-acute care settings. LHNCBC proposed and demonstrated a Web approach to auto-completion of entered text, much of which has been adopted by Medicare for this demonstration project. LHNCBC has proposed and delivered the full content of many of the look-up tables that Medicare will use in this project (including a revised subset of RxNorm) and shaped their data conceptualization to fit a LOINC/HL7 model. A memorandum of Understanding (MOU) to formalize this collaboration is now in final review stages within CMS.

Concept Recognition in Narrative Clinical Reports

The LHNCBC Natural Language Processing group has embarked on an effort to examine all of the words and phrases in a narrative clinical report, identifying the clinical concepts and converting them into UMLS Concept Unique Identifiers (CUIs). Tools already exist for converting strings found in published literature into concepts. Clinical text presents the additional challenge of dealing with telegraphic phrasings (the absence of grammatically complete sentences) and distinguishing positive from negative statements about a finding, disease or symptom. The goal here is not full understanding but the ability to comb through and find important clinical concepts in narrative records. Such tools would have widespread use in quality assurance, clinical research, and decision support, with the right level of sensitivity and specificity. The NLM has existing tools that can facilitate this effort.

Biomedical Imaging and Multimedia

The overall goal of this major research area is to address fundamental questions that arise in the handling, organization, storage, access and transmission of very large electronic files in general and digitized biomedical images in particular. A special focus is research into these topics as applied to heterogeneous multimedia databases consisting of both images and text. Projects in this area have benefited from collaborators in several universities as well as at agencies such as the National Center for Health Statistics (NCHS) and the National Institute of Arthritis, Musculoskeletal and Skin Diseases (NIAMS), and a continuing partnership with the National Cancer Institute (NCI) in their research in cervical cancer caused by the Human Papillomavirus (HPV). These projects address the recommendations of the NLM Long Range Plan Goal 1: Seamless, Uninterrupted Access to Expanding Collections of Biomedical Data, Medical Knowledge, and Health Information.

Interactive Publications Research

This effort is a major priority under NLM LRP Recommendation 1.4 to "Evaluate interactive publications as a possible means to enhance learning, comprehension, and sharing of research results." LHNCBC investigators have developed prototype interactive publications (IP) using Adobe tools-based procedures created by the IP Research team. One is an article by CEB authors (published in the SPIE Proceedings) augmented by dynamic tables and graphs, a microscopy video of cell evolution, an animated spine in Flash, digital x-rays, and clinical DICOM images (CT, MRI, ultrasound). In a separate development, IPs were created from two published articles and the raw results for the underlying studies acquired from the American Psychiatric Institute for Research and Education (APIRE), and the raw data were linked to every (tabular) result presented in the articles. Starting with SAS scripts provided by the APIRE authors, we generated executable files and SAS scripts to reproduce every table and statistical test. In discussions of the interactive capabilities with the authors, we identified a set of core data analysis capabilities needed by readers who do not intend to export data for rigorous statistical analysis. The descriptive exploratory analysis tools (means, confidence intervals, linear and non-linear regression analysis) are being developed for ITAG+, a data visualization tool being created in-house.

Steps were taken to recruit a publisher to host IPs for dissemination. Our IP prototypes were demonstrated to the publisher, including the functions of converting tables to graphs, zooming into graphs, creating subsets of the tabular data, zooming into images and changing contrast in DICOM images.

In light of the large sizes of such publications - up to hundreds of megabytes, we are studying techniques and protocols to download such publication rapidly and progressively. We are also developing a Download Manager on the basis of these efforts.

Multimedia Database R&D

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

WebMIRS (Web-based Medical Information Retrieval System)

Developed some years ago and still in active use, WebMIRS continues to provide access to images and text from nationwide surveys conducted by the National Center for Health Statistics. At present, there are 492 users of WebMIRS in 56 countries. This Java application allows remote

users to access data from the National Heath and Nutrition Examination Surveys II and III (NHANES II and III). The NHANES II database 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, 550 of the 17,000 x-ray images collected in NHANES II contain vertebral boundary data collected by a board-certified radiologist. Users may do queries for both radiological and/or health survey data.

Digital Atlas of the Cervical and Lumbar Spine
The Digital Atlas remains available for the public from the CEB Web site either as a Java applet or a downloaded Java application or as a CD version of the Java application. The Java application version allows the user to add images (either grayscale or color) in a special "My Images" section, and to annotate and title those images for later use.

In addition, the FTP x-ray archive of 17,000 digitized spinal x-rays continues to be very active, with over 500 users worldwide. This archive allows access to the x-rays, available in full 12-bit flat file format and also in TIFF 8-bit format which is easier for many researchers to use.
A suite of newer systems motivated by, but not restricted to, our joint research with National Cancer Institute (NCI), are at various stages of development.

Multimedia Database Tool (MDT)
The MDT extends current WebMIRS capabilities. This new system is intended to accommodate the existing WebMIRS databases and the new text/image database currently being created from the collection of uterine cervix images from NCI. New features allow for both data dissemination and distributed data collection. A working implementation of the MDT is regularly used for demos by NLM and NCI. It currently operates on a database of 60,000 JPEG cervigrams and associated clinical text data from the Guanacaste Project. The latest version includes capability to query and display images segmented with the Boundary Marking Tool.

Boundary Marking Tool (BMT)
This tool provides Web capability to manually mark boundaries on cervicography images, and to manage collected data with a MySQL database. The BMT has supported several studies of the uterine cervix which were carried out by NCI. Dr. Mark Schiffman of NCI presented results from the 20-observer, 919- patient, BMT study done earlier this year, to the 23rd International Papillomavirus Conference and Clinical Workshop in Prague, Czech Republic. The BMT was also used in a pilot study to determine appropriate methods for selecting biopsy sites in colposcopic images. This collaborative work using the BMT was used in a paper, "Colposcopy at a crossroads", which was published in the August 2006 edition of the American Journal of Obstetrics and Gynecology. BMT results were also used in the July 2007 paper, "Visual appearance of the uterine cervix: correlation with human papillomavirus detection and type", published in the same journal. In addition to this cervix-related work, the BMT has been used to mark and label dermatological lesions for the NCI Viral Epidemiology Branch, headed by Dr. Jim Goedert for a study which was completed in 2007 to assess the effects of nicotine patch treatment on the regression of Kaposi sarcoma lesions, and the study results have been summarized in the paper, "Treatment of classic Kaposi sarcoma with a nicotine dermal patch: a phase II clinical trial."

Virtual Microscope (VM)
The VM, currently at an intermediate level of development, will provide Web capability to view and collect information on histology images from expert observers. There are both a simple demo system of basic histology image and data collection capability and a fully-functional system, currently being used to support a multiple-observer study of lung histology images, in

collaboration with the NCI Genetic Epidemiology Branch, the NCI Cell and Cancer Biology Branch, and their medical collaborators in Italy.

Teaching Tool (TT)
This system is for training medical personnel in cervix anatomy/pathology. It displays uterine cervix images and quizzes an observer in the categories of medical knowledge, pattern recognition, and patient management, and enables a medical expert to tailor exams by specifying images and questions to use on an examination. A prototype system is available for experimentation by NCI and American Society for Cervical Pathology and Colposcopy (ASCCP) experts, which includes capability to administer and score the ASCCP "Resident's Online Exam."

Visual Triage Study (VTS) software
The study goal is to determine whether non-oncologists can answer screening questions as reliably as oncologists. Two phases of the Visual Triage Study have now been completed, the latest in Peru. The VTS software is now supporting a third, follow-on phase of this study, which has observers in Costa Rica, Ghana, Peru, and Thailand.

For our work with NCI, these systems are interrelated through the data that is used. The MDT distributes images and text data from the NCI Guanacaste and ALTS projects; the BMT allows the collection of additional graphical and text data that is added to the MDT database for distribution; similarly, for the VM; the TT uses data collected by the BMT to create the content for the examinations it supports, and the VTS software uses images and text data from Guanacaste and ALTS.

Content-Based Image Retrieval (CBIR)
This project investigates approaches for query and retrieval of biomedical images by direct use of image data, possibly in association with text related to the biomedical images. The emphasis is primarily on the NHANES spine images, using shape methods on vertebrae in the images, and on NCI cervigrams, using color and texture methods to differentially identify tissue regions and tissue characteristics within these images. One goal is to develop effective CBIR methods that may be incorporated into our multimedia database programs (such as the MDT) or into separate, prototype systems for use and evaluation by the biomedical research and/or clinical communities. A third goal is to develop geospatially distributed computing approaches to multiscale CBIR. As envisioned, the results from this effort would allow export and Web-based reuse of open methods that may be particular to local pathologies in various image collections to interact with others that are designed for gross image analysis and classification.

Currently there are two CBIR systems and an interface for a geospatially distributed computing approach to multiscale CBIR. Our earlier CBIR system, CBIR3, has been entirely redeveloped to have a Web-based focus. It has been renamed SPIRS (Spine Pathology Image Retrieval System) and it supports two shape retrieval methods as well as retrieval by descriptive text, using a database of several thousand pre-segmented vertebral shapes and text data from the NHANES II database. The new design is multi-platform and allows the system to run in any Web browser, as long as the host platform has the Java Runtime Environment (JRE) installed.

Using a similar philosophy, the Web-based CBIR system has been developed for NCI uterine cervix images. This system supports retrieval from a database of over 900 images that have been expert-segmented. (Don't know what it means to segment radiology image- ideally should have something more explanatory To create a query, the user specifies weighted combinations of color, texture, size, and locations of marked region of interest. This system is designed as a CBIR prototype suitable for medical experts to use, experiment with, and provide feedback to the

engineering community. In addition, we have created an internal engineering version of the Web-based system that allows more fine-grained, technical query specification over a broader range of image features and query methods.

SPIRS also exports its algorithms through a query interface to its core algorithms (dubbed cSPIRS) using structured XML queries. A test environment has been setup with the IRMA (Image Retrieval for Medical Applications) project at Aachen University in Germany. The IRMA project aims at classification of images using overall appearance, but is insensitive to local pathologies in these individual image groups. In contrast, SPIRS supports queries of local pathology in digitized x-ray images of the spine.

In data collection, we have segmented an additional 600 images, bringing the total number of vertebral outlines to 13,000. We are currently using 7,000 shapes in our SPIRS system. We have obtained a 3-way expert data collection of 500 images using our Path-Va system. In addition, we have over 1,500 shapes that have expert-marked boundaries.

The Visible Human Project

The Visible Human Project image data sets are designed to serve as a common reference for the study of human anatomy, as a set of common public domain data for testing medical imaging algorithms, and as a test bed and model for the construction of image libraries that can be accessed through networks. The Visible Human data sets are available through a free license agreement with the NLM. They are distributed to licensees over the Internet at no cost; and on DAT tape for a duplication fee. The data sets are being applied to a wide range of educational, diagnostic, treatment planning, virtual reality, and virtual surgeries, in addition to artistic, mathematical, legal, and industrial uses by over 2,450 licensees in 49 countries. The Visible Human Project has been featured in more than 900 newspaper articles, news and science magazines, and radio and television programs worldwide.

FY2007 saw the continued maintenance of two databases to record information about Visible Human Project use. The first, to log information about the license holders and record statements of their intended use of the images; and the second, to record information about the products the licensees are providing NLM in compliance with the Visible Human Dataset License Agreement.

A new edition of the NLM Current Bibliographies in Medicine, Visible Human Project was made available on the Visible Human Project Web site. This publication contains over 900 citations and covers the period from January 1987 through March 2007. The Visible Human Project bibliography is an attempt to identify all publications in the scientific and technical literature which discuss the Visible Human Project and its derivative products.

In FY2007, a planning workshop, “VHP: Scope and Scale for the Future,” assembled an expert panel of radiologists, anatomists, pathologists, computer scientists, and engineers from across the country to advise NLM on future directions for the Visible Human Project. Topics such as human variation, community data annotation, algorithm validation, and multiscale anatomy emerged as leading areas of interest.

3D Informatics

The 3D Informatics Program has expanded research efforts concerning problems encountered in the world of 3-dimensional and higher-dimensional, time-varying imaging. Among its many projects, the 3D Informatics (TDI) Group has continued work on image databases, including ongoing support for the National Online Volumetric Archive (NOVA), an archive of volume image data. This collection contains 3D data from across medicine. Contributors to the collection

include the Mayo Clinic Biomedical Imaging Resource and the Walter Reed Army Medical Center Radiology Department. Integrated and multimodal data such as virtual colonoscopy matched with recorded video from endoscopic interventions, time-varying 3D cardiac motion, and 4D MRI of a human hand appear in the archive.

The 3D Informatics group continues its partnership with the NLM Specialized Information Systems Division and the U.S. Veterans Administration to study content-based retrieval methods for medical image databases. In the pharmaceutical identification project, we are assisting in the acquisition of imagery through digital macro-photography of the thousands of prescription pharmaceuticals dispensed routinely by the VA Centralized Mail-Order Pharmacies. Together we are creating a new, updated, visual database of all these products and developing techniques for automatically identifying any product in the inventory from a representative photograph. New OHPCC research has developed computer vision approaches for the automatic segmentation, measurement, and analysis of solid-dose medications. In particular, recent focus has been on robust color classification tools to help identify prescription drugs.

3D Telepresence for Medical Consultation

This project tests the efficacy of 2D versus 3D representations of video data transmitted in real time in remote clinical consultations. The technology infrastructure is being developed at the University of North Carolina and its efficacy is being researched there with help from colleagues at other institutions. The research team continues to make substantial progress in implementing the technology infrastructure. A prototype portable camera unit was added to the stationary one and calibrated. The PDA application was completed and all the basic components of the system proposed are in place. The current focus continues on optimizing camera and sensor placement, refining calibration and rendering algorithms, and dealing with problems when perspective changes from different points of view, such as occlusion when an intervening object obstructs the view of interest.

Advanced Network Infrastructure for Distributed Learning and Collaborative Research

This project builds on previous work with HAVnet (Haptic Audio Visual Network for Educational Technology) and is collaboration between Stanford University and the University of Wisconsin at La Cross. The project's focus is on developing visual and haptic applications for anatomy and surgical training and includes aspects of self scaling technology, self-optimizing end-to-end, network aware, real time middleware, wireless technology, and GIS. The technology is being developed and refined in the context of teaching anatomy and surgical skills and addresses issues concerning network bandwidth and latency and the integration of 3D visualization, haptic, and real time online collaboration tools.

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

Work on the remote stereo viewer and haptic probe was completed and preliminary trials suggest videoconferencing is essential for dermatologists to see and communicate with patients while

using the haptic device. Research was conducted on sense of touch and ability to detect thickness and resistance of membranes as part of the effort and to compare this feedback to haptic feedback generated by computer. The SPRING surgical simulator engine and its Remote Tactile Sensor component were made open source. The engine allows building of software modules providing haptic feedback for simulated surgical tools. Work continued on the iAnatomy collaboration with the Northern Ontario School of Medicine involving the use of the stereo viewer for anatomy teaching and distance learning and a similar collaboration with the University of California - Davis is being explored.

Insight Tool Kit (ITK)

The Insight Toolkit, a research and development initiative under the Visible Human Project, is now in its sixth year with a recent official software release of ITK 3.4. ITK makes available a variety of open source image processing algorithms for computing segmentation and registration of high dimensional medical data on a variety of hardware platforms. Platforms currently supported are PCs running Visual C++, Sun Workstations running the GNU C++ compiler, SGI workstations, Linux based systems and Mac OS-X. Support, development, and maintenance of the software are managed by a community of university and commercial groups, including OHPCC intramural research staff. The ITK continues to have an impact on the medical imaging research community. Researchers are testing, developing, and contributing to ITK in more than 40 countries, with more than 1500 active subscribers to the global mailing list for the project.

Across NIH, ITK is providing a foundation for new imaging investigations. The National Alliance of Medical Image Computing (NA-MIC), an NIH Roadmap National Center for Biomedical Computing (NCBC), has adopted ITK and its software engineering practices as part of its engineering infrastructure. NA-MIC is currently using medical imaging techniques to study the physiological sources of schizophrenia and other mental disorders. Staff members participate as science officers and lead science officer for the NIH-Roadmap for the NA-MIC consortium.

ITK also serves as the software foundation for the Image Guided Surgery Toolkit (IGSTK), a research and development program sponsored by the NIH National Institute for Biomedical Imaging and Bioengineering (NIBIB) and executed by Georgetown University's Imaging Science and Information Systems (ISIS) Center. IGSTK is pioneering an open API for integrating robotics, image-guidance, image analysis, and surgical intervention. The external advisory board for IGSTK includes members of the Lister Hill staff.

From 2002 to 2007, approximately 20 purchase orders were awarded for reference data sets and enhanced algorithms to support the further development of ITK. This effort supported the integration of ITK into research platforms such as the Analyze from the Mayo Clinic, SCIRun from the University of Utah's Scientific Computing and Imaging Institute, and the development of a new release of VolView, free software for medical volume image viewing and analysis. Among the data acquisitions for NLM, the Mayo Clinic Biomedial Imaging Resource has provided over 100 datasets collected across dozens of animals and clinical cases representing a wide cross section of anatomy, pathology, modality, and pre- and post-operative clinical conditions.

Image and Text Indexing for Clinical Decision Support

The title of a publication is not always sufficient in determining the Evidence-Based Practice (EBP) relevance of a publication. Given that medical illustrations often convey essential information in compact form, this project seeks to automatically identify illustrations from the articles that could help clinicians evaluate the potential usefulness of a publication in a clinical

situation. We explored feasibility of automatic image annotation by utility for EBP, and if such images can be reliably extracted from the original articles.

Our study showed that images presented in clinical journals can be successfully annotated by their usefulness in finding evidence to assist a clinical decision. The feasibility of automatic image classification with respect to its utility in finding clinical decision support demonstrated in this study provides several venues for further exploration. We plan to study the influence of augmenting bibliographic references retrieved from a database search with images; new ways of organizing and presenting retrieval results using annotated images; and further improvement in the automatic single and multi-panel image extraction, annotation, and complementary text extraction.

Turning The Pages Information Systems

Continuing to bring the magnificent rare books at the NLM to public view, a sixth book has been added to the TTP collection: Robert Hooke's Micrographia, the first book written about microscopes and in which reportedly the word "cell" was first used. New technical challenges in converting this book included the handling of fold-out pages and the inclusion of images of historic and present day microscopes. Library visitors may touch and turn the pages of these books on onsite kiosks, and online users may use the Web version of TTP (TTP Online).

A 3D wireframe model of a scroll document has recently been completed. The objective is to create a 'touch and scroll' function for an ancient Egyptian medical document, the Edwin Smith papyrus. It is the world's earliest known medical document, written in hieratic around the 17th century BCE, but thought to be based on material from a thousand years earlier. It is a textbook on trauma surgery, and describes anatomical observations and the examination, diagnosis, treatment, and prognosis of numerous injuries in exquisite detail. The Edwin Smith papyrus shows that the heart, vessels, liver, spleen, kidneys, and bladder were recognized, and that the blood-vessels were known to be connected to the heart. Other vessels are described, some carrying air, some mucus, while two to the right ear are said to carry the breath of life, and two to the left ear the breath of death. The physiological functions of organs and vessels remained a complete mystery to the ancient Egyptians.

TTP Online in French (Tournez Les Pages) was launched on Bastille Day, July 14, 2007, thanks to translations done by the director of the Bibliothèque Interuniversitaire de Médecine et d'Odontologie (BIUM) in Paris. Tournez Les Pages offers explanations of the text, curators' notes and instructions for a francophone audience spread over many countries around the globe. On TTP Online's main introductory page, a user may click on an icon of the French tricolor to go to the equivalent page in French. Translations for five books are planned.

Video Retrieval and Reuse Project

In response to the NLM Director's request for an automated system for storing and retrieving an historic collection of videos, APDB developed the Personal Digital Library (PDL) application which operates on a personal computer (Mac or PC). Videos are stored on an external 160 GB drive connected to the PC. The system software was an outgrowth of an application originally designed for the Movement Disorders Video Database Project, developed by APDB, which dynamically linked patient metadata to patient videos. Upon launching the PDL application, a multi-windowed interface reveals all the titles in the video library. There are currently 122 titles, listed alphabetically, occupying 143 GB of disc space. Titles can also be viewed by category. The entire library can be searched for any word contained in the audio track of any video or in the video's metadata. The results of a search immediately show all instances across the video library where that search term appears. The user can then play any of those search results within the

multi-windowed interface or select the full screen option. QuickTime videos are encoded using the H.264 codec also known as MPEG 4 AVC. The system allows the user to bookmark video, create QuickTime video clips, and create, manage and merge multiple video libraries. The PDL application is fairly compact, occupying 23 MB and when operating the executable uses less than 3 MB of system memory. The program is written with open source code using the Eclipse Platform and Java development tools. The PDL was installed on the NLM Director's laptop in December 2006. The PDL was demonstrated to the LHNCBC Board of Scientific Counselors in May 2007.

In addition, APDB completed production of the Visible Proofs exhibition DVD program, based on the NLM History of Medicine's Exhibition scripts. The final program contains highly interactive rich media content, including additional video and animation materials to enhance the content of the exhibition. New production elements include high-definition (HD) video of the entire exhibition structure and case materials, and seventeen animated segments, based on the exhibition booklet, "The New Forensic Science", which feature scientific theories and tools currently at the forefront of forensic science. Additional video interviews were conducted with Mike Sappol, PhD, Exhibition Curator, Clyde Snow, PhD, (Norman, OK), Sir Alec Jeffries, (London, England) and Brian Andresen, (Livermore, CA.) and Kirk Bloodsworth from the Washington, DC-based Justice Project. The graphical user interface design represents the exhibition themes and incorporates the existing color palette of the exhibition. Also, APDB provided project management support and HD video recording of several events including the production of the Collen Award video, the conversion of the Barbara McClintock Profiles in Science module to HD DVD, and DVD version of the Paul Ortega interview.

Automated Concept Extraction from Documents

Research in this area is directed toward developing techniques and algorithms to extract bibliographic data from biomedical journal articles, both digitized and Web documents, to build MEDLINE citations. The projects in this category are MARS and its various spin-offs and the Indexing Initiative.

Medical Article Records System (MARS)

The MARS production system currently generates bibliographic data for 500 articles per day, the remaining citations coming in as XML-tagged data directly from publishers. MARS has evolved through several generations of increasing capability. Its core engine consists of daemons based on heuristic rule-based algorithms that use geometric and contextual features derived from OCR output to automatically segment scanned pages of journal articles, assign logical labels to these zones, and to reformat zone contents to adhere to MEDLINE conventions.

Center researchers continue to make changes to the MARS production system to accommodate new requirements from the NLM Indexing section. We modified three MARS software modules (Edit, Reconcile, and Upload) to support Unicode. Two conversion libraries (ASCII-to-Unicode and Unicode-to-ASCII) were created to read and write zone information from and to the MARS database. The ASCII data in zones is converted to Unicode for the operators conducting the Reconciling and Editing stages so they can see the diacritics as they appear in the article. The operators' output in Unicode is then converted back to ASCII for storage in the MARS database prior to eventual upload to DCMS, the database used by indexers to complete the indexing before citations are included in MEDLINE.

WebMARS

The goals of the Indexing 2015 Initiative are being addressed by CEB's development of two systems relying on WebMARS to assist both operators and indexers. The initial versions of both systems, WebMARS Assisted Indexing (WAI) and Publisher Data Review (PDR) are currently under test. PDR is designed to provide operators data missing from the XML citations sent in directly by publishers (such as databank accession numbers, NIH grant numbers, funding sources, check tags and PubMed IDs of commented articles) thereby reducing the burden on operators. In addition, incorrect data sent in by the publishers can be corrected by PDR. Correcting or augmenting the publisher data is a labor-intensive process since the operators currently perform these functions manually by looking through an entire article to find these items, and then keying them in.

The second system, WAI, is for the indexers; it will help them search for terms in an article that correspond to biomedical terms in a predefined list. Again, indexers currently have to read through the entire article to confirm the occurrence of these terms, a labor-intensive process. WAI will automatically sift through the text and highlight these terms for the indexer to simply confirm and select, thereby reducing manual effort. In all, the systems consist of seven software modules.

ACORN

The ACORN (Automatically Creating OldMedline Records for NLM) system is intended to extract bibliographic information from 60 volumes of the printed Quarterly Cumulative Index Medicus (QCIM) from 1927 to 1956 to populate the OLDMEDLINE database. The design of the system is rooted in research in document image analysis and pattern matching techniques. A Web-based Reconcile module is being designed to allow operators to access, verify, and create QCIM citations through the Internet.

Validated Test Set for Document Image Analysis

By the end of May 2007, the Medical Article Records Groundtruth (MARG) database had 12,224 unique IP visits from 96 countries. That is an increase of 2,536 visits since September, 2006. MARG provides TIFF images of biomedical journal articles and corresponding operator-verified OCR data, page segmentation and labeling results. This data set, obtained from the normal operation of the MARS production system, is used by researchers in the computer science and informatics communities conducting document image analysis to validate their own zoning and labeling algorithms.

Indexing Initiative

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

The MTI system is in regular, increasing use by NLM indexers to index MEDLINE. MTI recommendations are available to them as an additional resource through the Data Creation and

Maintenance System (DCMS). This year MTI recommendations are being augmented by the attachment of subheadings to some of the MeSH headings it recommends. Indexers will now have the option of accepting MTI heading/subheading pairs in addition to unadorned headings. In addition, indexing terms automatically produced by stricter version of MTI are being used as keywords to access collections of meeting abstracts via the NLM Gateway. These collections include abstracts in the areas of AIDS/HIV, health sciences research, and space life sciences.

Automatic Extraction of Outcomes from Published Documents

Originally part of the MDoT project, research was conducted toward automatically finding patient outcomes (e.g., the population under study) from MEDLINE citations using knowledge extractors that rely upon NLM Unified Medical Language System and tools. Our Extractor system identifies an outcome and determines whether a found outcome pertains to the topic of interest, the type of treatment studied, and the quality of the study. We evaluated the ability of the Extractor both to find outcomes in general, and to find high quality outcomes that answer specific clinical questions. Possible application areas might include clinical trials design, EMR, and a patient-oriented service. Developed to provide access to the repository, a server accepts requests containing information about a patient (at present, current problems, age and medications) and searches MEDLINE via any of three search engines (Essie, PubMed, or the RIDeM database). The extracted information is sent to the client. The repository will be evaluated in a planned pilot study of supporting Evidence Based Nursing Practice at the NIH Clinical Center.

Digital Preservation Research

In line with the NLM mandate to preserve the medical literature, the goal of this project is to investigate key issues related to the long term preservation of digital material, both for digitized documents and video. Our work in document preservation is more mature, and has resulted in a prototype System for Preservation of Electronic Resources (SPER). SPER is a flexible, modular system that demonstrates key functions such as ingest, automated metadata extraction (AME) and bulk file migration. AME is implemented for the extraction of descriptive metadata from scanned and online journal articles as well as obsolete NLM Web pages. Bulk file migration is implemented through an existing system, DocMorph. While these functions are developed in-house, for the necessary infrastructure capabilities in SPER we have incorporated into the system, and customized, the latest version (1.4) of MIT's open source DSpace software. Our Java client GUI has the capability to do batch metadata extraction and ingest for journal article TIFF pages, online journal articles and NLM Web pages (HTML).

SPER, in an abbreviated form, is being used in the preservation of a new collection at NLM consisting of over 65,000 historical Food and Drug Administration court records (Notices of Judgment) from the early 20th century. Since the manual identification and entry of descriptive metadata from these records is labor-intensive, our focus is on their automated extraction. In collaboration with the curator for this collection, we identified more than a dozen metadata items which could be extracted automatically. After scanning the paper documents and performing OCR, the system then auto-zones the TIFF files, performs feature extraction and optimal feature selection, classifies text lines in the documents by a Support Vector Machine (SVM) classifier, and extracts the specific metadata by text pattern matching. Currently, the system extracts all metadata for a notice of judgment in less than 10 seconds on an ordinary Pentium machine. Research was also conducted into video preservation. This effort centered on identifying an open file format such as Motion JPEG 2000 (MJ2) for archiving digitized video on disk media. This effort was guided by the findings of a one-day invitational meeting (at NLM) in 2005 with about 50 archivists and technologists involved in the long term preservation of video and film.

Information Resource Delivery for Care Providers and the Public

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

Clinical Research Information Systems

ClinicalTrials.gov provides the public with comprehensive information about all types of clinical research studies, both interventional and observational. The site has over 43,000 protocol records sponsored by the U.S. Federal government, pharmaceutical industry, academic and international organizations from all 50 States and in over 140 countries. Some 44% of the trials listed are open to recruitment, and the remaining 56% are closed to recruitment or completed. ClinicalTrials.gov receives over 19 million page views per month and hosts approximately 29,000 visitors daily. Data are submitted by over 3,400 study sponsors through a Web-based Protocol Registration System, which allows providers to maintain and validate information about their trials.

ClinicalTrials.gov was actively involved in promoting the standards of transparency in clinical research through trial registration. These standards were communicated to a broad range of U.S. and international stakeholders via presentations and peer-reviewed publications. As a result of increasing awareness of the importance of trial registration, more than 9,000 new registrations were received over the last year. The results of an online evaluation aimed at identifying the needs of various groups of ClinicalTrials.gov users were used to inform the design of a new user interface.

ClinicalTrials.gov continues to collaborate with other registries and professional organizations, working towards developing global standards of trial registration. In anticipation of a new mandate to develop a clinical trial results database, ClinicalTrials.gov convened two expert meetings to discuss scientific issues related to submitting and displaying published and unpublished clinical trial results, including the design of a feasibility study.

Genetics Home reference

Genetics Home Reference (GHR) provides basic information about genetic conditions and the genes and chromosomes related to those conditions. This online resource provides a bridge between the public's questions about human genetics and the rich technical data emerging from the Human Genome Project. Created for the general public, particularly individuals with genetic conditions and their families, the site currently includes summaries for more than 230 genetic conditions, more than 395 genes, all the human chromosomes, and information about disorders caused by mutations in mitochondrial DNA. On average, nine new summaries are added per month. Additionally, new tutorial materials have been developed for the “Help Me Understand Genetics” Handbook. The latest Handbook topics include direct-to-consumer genetic testing, the International HapMap Project, and an introduction to genome-wide association studies. GHR’s usage, as measured by the number of hits per day, increased more than 60% in the past year, and the site is continually recognized as an important health resource.

Profiles in Science Digital Library

The Digital Library Research project investigates all aspects of creating and disseminating digital collections, including standards, emerging technologies and formats, copyright and legal issues, effects on previously established processes, protection of original materials, and permanent archiving of digital surrogates. Research focuses on long-term preservation of digital archives and new techniques for creating and accessing collections. The investigators are also exploring ways to achieve interoperability among digital library systems, using well-structured metadata, and varying “points of view” on the same data sets.

Profiles in Science showcases digital reproductions of items selected from the personal manuscript collections of prominent biomedical researchers, medical practitioners, and those fostering science and health. The content of Profiles in Science is created in collaboration with the History of Medicine Division of NLM, which processes and stores the physical collections. Most collections have been donated to NLM and contain published and unpublished materials, including manuscripts, diaries, laboratory notebooks, correspondence, photographs, journal volumes, poems, drawings, audio tapes and other audiovisual resources. The collections of Harold Varmus, Rosalind Franklin, Mary Lasker, and Sol Spiegelman were added this year. An additional 147 digital items composed of 599 image pages were also added to the 22 existing Profiles in Science collections. Presently the Web site features the archives of 23 prominent individuals:

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

The 1964–2000 Reports of the Surgeon General, the history of the Regional Medical Programs, and Visual Culture and Health Posters are also available on Profiles in Science. The Society of American Archivists awarded Profiles in Science the Philip M. Hamer and Elizabeth Hamer Kegan award "in recognition of successful efforts to increase public awareness through the use of archival or manuscript materials." New additions to Profiles in Science were highlighted in Science Magazine's "NetWatch" and JAMA's "Health Agencies Update."

Developers made several optimizations to software algorithms underlying the next generation Profiles in Science Web server, resulting in significant performance improvements affecting the ingest of data, copying of images, and display of Web pages. They also developed methods to identify handwritten items and target them for creation of transcripts to improve searching.

User Focused Portals

NLM Gateway

The NLM Gateway is an ongoing production system that communicates with multiple NLM information resources. Since these resources are frequently updated, improved, and otherwise modified, the Gateway must change with them. The Gateway's XML parser was replaced to fix a particular issue and upgrade to the latest technology. All TOXLINE data was consolidated on the SIS web portal, combining toxicology citations from PubMed and TOXLINE Special. Gateway search translations were changed so the user would continue to retrieve only the TOXLINE special citations. Similarly, the DART (Developmental and Reproductive Toxicology) data was consolidated on the SIS web portal, combining toxicology citations from PubMed and DART Special. Users continue to retrieve only the DART special citations.

FY2007 activities included DTD changes to PubMed/Medline and the NLM Catalog and automated indexing of the 100,000 meeting abstracts using 2007 MeSH, periodic updates of the UMLS and the MeSH mapping file and the HSRProj (Health Services Research) database. The Gateway system now communicates with the latest version of NCBI's Esummary Eutility API. Developers upgraded server hardware and the Java and Visibroker software and are optimizing

the system to support the MedlinePlus use of the Vivisimo search engine. Researchers continue to conduct usability testing.

Communication Infrastructure Research and Tools

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

Advanced Biomedical Tele-Collaboration Testbed

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

In FY2007, the research team completed the substantial infrastructure required to test the scenarios, including the implementation of color algorithms to real time volume rendering of CT and MRI data and stereo display in the AG environment as well as the use of the technology in surgical education and planning. Virtual reality methods have been employed in a haptic environment allowing surgeons to rehearse liver operations. Several additional successful wide area wireless demonstrations of transmitting video and other patient data from ambulances using 3G and mesh cellular technology have been completed.

Scalable Information Infrastructure Initiative

The Scalable Information Infrastructure (SII) Project encourages the development of relevant health applications that are network aware and able to automatically adjust to changing network conditions and resources. Public next generation networks with SII capabilities hold the promise of adding advanced networking capacity to the tools available to healthcare professionals. Virtual reality and home health care may become realizable at reasonable costs based on next generation networking technology. Applications include wireless and geographic information system (GIS) techniques.

In order to carryout its program in Next Generation Networking Research, OHPCC as established the Collaboratory For Interactive Technology to serve as a laboratory, testbed, and demonstration area for NGN, SII, medical imaging, and high performance medical communications and networking applications. This facility is used to develop and test interactive imaging and communications protocols as applicable to telemedicine and distance collaboration under conventional, internet, and NGN/SII conditions. The facility also is configured to act as a hands-on demonstration site for remote interactive imaging, telemedicine, collaboration, and distance learning paradigms. Communications infrastructure connected to the Collaboratory, including an Internet2 Access Node capability, gives research staff the ability to collaborate with distant research colleagues, and at the same time, demonstrate much of the work being sponsored by the NLM Visible Human and NGN/SII programs.

The Collaboratory is used for a variety of research projects; develop and test image analysis; test manipulation and segmentation algorithms for use with the Visible Human data set in particular and in medical imaging in general; and test advanced communication and collaboration technologies that complement those employed in OHPCC sponsored research. Research staff also use Collaboratory resources to demonstrate advances in telemedicine, imaging, and collaboratory techniques, including haptics, and to support ongoing programs of the NLM. Three-dimensional projection capability with access to remote 3D data sets is available.

The Collaboratory includes a fully functioning immersive Access Grid Room Node, with multiple projection capability, camera sources, and audio inputs. Several demonstrations were conducted in FY 2007, including one for the Federal Communications Commission Panel on Rural Telemedicine. Experimental distant learning programs were conducted, working with SIS, NCBI, and remote collaborators at the Charles R. Drew University of Medicine and Science, the University of Puerto Rico Medical Campus and the University of Michigan. Resources were acquired to begin experimentation with IPHDTV, working with the University of Washington and the Internet2 Research Channel Working Group.

Videoconferencing and Collaboration

A new initiative was undertaken to experiment with uncompressed video over IP as well as high definition television. Compressed HD videoconferencing codecs were investigated using the H.264 technology that is compatible with and part of the revised H.323 standard. Digital Video Transport System (DVTS) technology was implemented, both as a standalone technology and as a component of the Access Grid. DVTS was developed by the WIDE consortium in Japan and is used by various Internet2 members to send uncompressed digital video at 30 megabits per second over IP. In addition, the Collaboratory became part of the Research Channel Working Group within the Internet2 and started acquiring components to implement uncompressed HD video at 1.5 gigabits per second. Collaboratory staff developed major enhancements for the Access Grid's (AG) shared browser and presentation tools. The use of open source browsers and presentation software as the basis for making the enhancements is being considered.

A distance learning program in collaboration with SIS, coordinator of the NLM Adopt-A-School Program, continued to provide on-site and distance education about varied health science topics and information sources to students at the King Drew Medical Magnet High School, affiliated with the Charles R. Drew University of Medicine and Science in Los Angeles. The NIH Office of Science Education participated again in the program and conducted several sessions on health science careers. Initial exploratory work was completed in trying to link a second school serving Native Americans in Alaska. Each session was assessed as in previous years. As in the past, a statistical analysis of student ratings of teaching showed students rated the distant presentations higher than those on-site.

Methods for providing application sharing and image manipulation with low latency were identified and methods developed enabling the instructor at NLM to view each remote student's desktop. Successful pilot training sessions have been done with the University of Puerto Rico using the application sharing methods in conjunction with H.323 videoconferencing and with the University of Michigan with Access Grid (AG) technology.

A study of collocation as a factor in synchronous learning was completed with the University of Alabama at Birmingham in which students were tested on lectures delivered by videoconference and asked to collaborate on search tasks before being tested. They also were asked to rate teaching effectiveness of the lectures. Students were either physically collocated in a computer

lab or meeting virtually in a multipoint videoconference. The data collected is currently undergoing analysis to determine if physical presence or absence affects performance, patterns of interaction, and perceptions of learning experiences.

The Center for Public Service Communication (CPSC) completed a successful pilot test of the use of video over IP to provide remote medical interpretation services at public health clinics in Duval County Florida. Valuable information about how the technology was and should be used was obtained, and the CPSC will move the technology to another public health environment in Florida.

Both the Web casts of the bi-monthly Washington Area Computer Assisted Surgery Special Interest Group and videoconferencing added last year continued. There is now two-way interaction between those attending the meeting in the Lister Hill auditorium, where the presentations are made, and those in an auditorium at the Allegheny Hospital System in Pittsburgh. Attendees are able to obtain continuing medical education credits because of this linkage.

Access for Evidence Based Medicine - PubMed for Handhelds

PubMed for Handhelds was publicly released in FY 2005. Developed to facilitate evidenced-based medical practice with Medline access at the point of need via smartphones, wireless PDA's or portable laptops, PubMed for Handhelds requires no proprietary software and reformats the screen display as appropriate for the wireless handheld device being used. In support of evidence-based clinical practice, clinical filters feature easy access to relevant clinical literature. Newly developed resources allow searching Medline through text-messaging.

BabelMeSH

BabelMeSH is a multilanguage and cross-language search tool for healthcare personnel who prefer to search MEDLINE/PubMed in their native languages. Through international collaborations, including WHO Eastern Mediterranean Regional Office in Cairo, users can now search in Arabic, French, German, Italian, Japanese, Portuguese, Russian, Spanish, and English.

PICO Linguist

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

Advanced Networking

Center staff are exploring advanced optical networking techniques involving dynamic allocation of network resources using Quality of Service (QoS) characteristics in protocols such as RSVP. The purpose of the project is to gain experience with means of creating end-to-end channels with defined or reserved bandwidth and other qualities of service. This effort builds on the work of the DRAGON (Dynamic Resource Allocation in GMPLS Optical Networks) project, a collaboration among the University of Southern California Information Sciences Institute, the University of Maryland/Mid-Atlantic Crossroads (MAX), and George Mason University.

DocView Project: Tools for Using and Exchanging Library Information

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

DocView

Originally released in 1998, this Windows-based client software is widely used to facilitate delivery of TIFF documents for interlibrary loan services. More than 18,600 users in 195 countries have downloaded it since it was released, with 600 new users in the last year. Many of these users are patrons of biomedical libraries, who receive electronic journals sent by the libraries via the Internet to DocView running on the patron's desktop computer. The DocView software has seen declining usage in recent years because the PDF file format, rather than TIFF, has become the popular choice for electronic document distribution.

MyDelivery

The goal of this project, seen as a successor to DocView, is to develop a new collaborative tool to improve the delivery and exchange of medical and health information, especially information contained in very large files. MyDelivery is intended to enable biomedical researchers, administrators, librarians, physicians, patients, hospitals, and other health professionals to exchange medical information, regardless of the size of the electronic file in which it resides. This communication method is expected to be fast, easy, reliable, safe, and secure.

The MyDelivery project seeks to overcome three significant obstacles: (1) transmission of large electronic files (e.g., document images, digitized photographs and x-rays, sonograms, CT and MRI scans, and digital video) over the Internet; (2) sending files reliably and securely; and (3) complying with requirements of the Health Insurance Portability and Accountability Act (HIPAA). To solve all three problems, the MyDelivery project focuses on the development of server-based software running on a cluster of Internet-based servers, and the development of client software for use by collaborators. MyDelivery allows two client computers to exchange large files through an intermediary server via a user interface similar to email. Potential applications of this new technology include a secure way for researchers to easily exchange research data of virtually any size. Part of the development of MyDelivery has been to create a method of automatically recovering from communication failures due to reduced signal strength.

DocMorph

The DocMorph system is a Web site that allows the conversion of more than 50 different file formats to PDF, for instance, to enable multi-platform delivery of documents. Since 1999, DocMorph has served both browser-based users (16,600 to date: 2200 more than last year) and client-based MyMorph, of which there are more than 8,000 registered users. Many of the users are biomedical document delivery librarians. DocMorph also provides conversion of files to TIFF, text and synthesized speech. By combining text extraction with speech synthesis, DocMorph enables the visually impaired to use library information. It has been used by librarians for the blind and physically handicapped to convert documents to synthetic speech recorded onto audio tapes for patrons. DocMorph is available at http://docmorph.nlm.nih.gov/docmorph.

MyMorph

MyMorph replaces the browser for DocMorph users who have large numbers of files to convert. Version 2 of MyMorph was released in April 2007 with the capability of mass converting PDF documents to TIFF. This became necessary since many document delivery library services,

including that of NLM, use commercial document transmission systems that deliver TIFF, rather than PDF, documents. NLM and many libraries now use our current version of MyMorph to more easily deliver documents to their patrons.

Language and Knowledge Processing

Terminology Research and Services

LHNCBC research staff build and maintain the SPECIALIST Lexicon, a large syntactic lexicon of medical and general English that is released annually with the Unified Medical Language System (UMLS) Knowledge Sources. New lexical items are continually added using a lexicon building tool; the SPECIALIST lexicon contains over 360,000 records. The UMLS Lexical tools, including lexical variant generator (LVG), wordind, and norm are distributed with the UMLS as are text processing tools which analyze documents into sections, sentences, and phrases. The SPECIALIST lexicon, lexical tools, and text processing tools are released as open source resources and available under an unrestrictive set of terms and conditions for their use.

LexBuild is an evolving lexicon building tool designed to aid the lexicon building team by facilitating entry of lexical information and providing real time quality control. The SPECIALIST lexicon release tables are annually generated using the LexBuild tool. The SPECIALIST lexicon and tools are UTF-8 compliant and capable of dealing with non-ASCII characters. MMTx, the Java implementation of the MetaMap algorithm is a major application of the SPECIALIST lexical and text tools.

LHNCBC researchers have created a Java implementation of the Journal Descriptor Indexing (JDI) tool for release as part of the UMLS lexical tools. The JDI tools provide an element of context that can be useful for word sense disambiguation and other natural language processing tasks. LHNCBC research staff also develop and maintain the UMLS Knowledge Source Server (UMLSKS) that provides Internet access to the UMLS knowledge sources through application programs and a user interface. UMLSKS is updated quarterly to accommodate quarterly UMLS releases. A Grid/Web services implementation of the UMLSKS backend and an implementation of the user interface as a portal consisting of user-chosen “portlets” representing different parts and views of the UMLS data are being released with the 2008 UMLS release.

Medical Ontology Research

While existing knowledge sources in the biomedical domain may be sufficient for information retrieval purposes, the organization of information in these resources is generally not suitable for reasoning. Automated inferencing requires the principled and consistent organization provided by ontologies. The objective of the Medical Ontology Research project is to develop methods whereby ontologies can be acquired from existing resources and validated against other knowledge sources, including the Unified Medical language System (UMLS).

This year, the research team focused on biomedical information integration. First, in the domain of oncology, where the resources used to annotate data include the International Classification of Diseases for Oncology (epidemiology data), SNOMED CT (clinical data) and the NCI Thesaurus (research data), we studied the limitations of automatic mapping between these resources and checked their consistency. In a different domain, we enhanced RxNav, transforming it from an interface to the RxNorm drug terminology system to an environment integrating additional drug information, including pharmacologic actions and product label information. Finally, we contributed to a demonstration created by the Semantic Web for Health Care and Life Sciences interest group consisting in the integration of about 20 resources in the domain of neurosciences.

As part of the research on ontology alignment, new methods were developed for validating matches and identifying mismatches between anatomical ontologies, including the Foundational Model of Anatomy and GALEN. We also carried out a study of criteria for evaluating biomedical vocabularies in caBIG on behalf of the National Cancer Institute. The method was applied to the evaluation of the NCI Thesaurus. Finally, the mapping of UMLS concepts to MeSH used in the Indexing Initiative was improved and it was adapted to mapping it to the International Classification of Diseases.

The research team continues to work on the creation of an ontology of relationships as it is one critical element of a repository of biomedical knowledge supporting knowledge discovery and reasoning. We continue to participate in the progress of the Semantic Web for Health Care and Life Sciences and collaborate with leading ontology and terminology centers, including the National Center for Biomedical Ontology and the International Health Terminology Standards Development Organization.

Semantic Knowledge Representation

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

MetaMap was recently improved by the addition of a word sense disambiguation (WSD) feature that chooses the best concept for a text phrase, guided by the surrounding context. The WSD method relies on Semantic Type Indexing and selects the concept having a semantic type most consistent with concepts in its immediate neighborhood.

The development of SemRep is based on viable strategies for effective natural language processing. Extending linguistic coverage is at the core of this research, and recent work focused on assertions about risk factors for disease in MEDLINE citations. Semantic predications produced by SemRep support continued work in biomedical information management. Application areas include automatic summarization and visualization of text from MEDLINE and ClinicalTrials.gov.

As a practical implementation, the Semantic MEDLINE application integrates PubMed searching, advanced natural language processing, automatic summarization, and visualization into a single Web portal. This program helps users manage the results of PubMed searches by normalizing core assertions in the citations retrieved. These normalized forms constitute computable knowledge accessible to further manipulation, including automatic summarization. The summarized output of Semantic MEDLINE is visualized as an informative graph with links to the original MEDLINE text. Convenient access is also provided to relevant knowledge resources, such as Entrez Gene, the Genetics Home Reference, and the UMLS Metathesaurus. In collaboration with the National Heart, Lung, Blood Institute (NHLBI), Semantic MEDLINE is being enhanced to help conduct literature searches for the creation of clinical practice guidelines.

UMLS and Focused Clinical Vocabularies

Unified Medical Language System (UMLS)

The most recent quarterly release of the UMLS Metathesaurus contains over 1.3 million concepts (+18%) and 6.4 million concept names (+20%). The UMLS provides the only way for the U.S. health care community to obtain SNOMED CT, the largest HIPAA standard clinical vocabulary, under the U.S. government license. Center staff have been principals in the successful transition to NLM's Office of Computer and Communications Systems (OCCS) aspects of the UMLS Metathesaurus project relating to a) software and processes for the Metathesaurus Editing and Maintenance Environment (MEME), b) software for the editing and workflow management, and c) merging the Metathesaurus release data with other knowledge sources and packaging the final product for distribution. Researchers continue to work with colleagues in the MEDLARS Management Section (MMS) of Library Operations to ensure the quality of the Metathesaurus and help with knowledge transfer for customer support. Staff continue to be responsible for the development, testing, and quality assurance of the Metathesaurus installation and customization program, MetamorphoSys and continue to enhance the product to support and anticipate the growing needs of the user community, such as investigating different output formats (TREF, XML), improving the search and display tool for customized subsets, performance, and user interface issues such as Section 508 compliance.

UMLS-CORE Project

The goal of this new project is to identify a CORE (Clinical Observations Recording and Encoding) subset of the UMLS to support consistent high level encoding of clinical information. The creation of the CORE subset will be based on real clinical data to ensure adequate coverage of commonly occurring clinical conditions. The CORE subset will promote and facilitate the use of standard clinical terminologies by helping users to identify the most frequently used portion of these terminologies. It will also enhance data interoperability by reducing coding variability. For terminology developers, the subset will help them identify gaps in coverage and focus their qualify improvement efforts. Datasets that will be very helpful in creating the CORE subset have been obtained from the Mayo Clinic, Intermountain Health Care, the Regenstrief Institute, and the Hong Kong Hospital Authority.

Terminology Representation and Exchange Format (TREF)

The goal of TREF is to serve as a standard publishing format for single-sourced terminologies. Its use will facilitate the exchange of terminologies and the sharing of terminology related tools. TREF will also simplify the task of inversion of source terminologies into the UMLS editing environment. The relational and XML specifications of TREF have been finalized. In an experimental version of MetamorphoSys, the functionality of generating TREF output from the Metathesaurus has been developed and tested.

UMLS User and Usage Statistics

The web forms for UMLS users to submit their annual reports were updated in FY2007. The annual report collection process started in February and ended in April. A total of 2,300 users submitted their annual reports. The data will be analyzed and compared with that from previous years. The information obtained will guide future developments of the UMLS.

UMLS Archive

Work has begun on constructing a data repository of all the 35 UMLS releases from 2007AB back to 1990 – the UMLS archive. Both the original file contents (as released to the public) and a database version of the archive are available for research. The archive is expected to grow to about two terabytes. The physical archive consisting of the distributed media, CD-ROM, DVD,

documentation green books, etc. is also being gathered. Both ORF (Original Release Format) and RRF (Rich Release Format) versions of the data since 2004AA are being archived. Data from 1990 and 1991 was in a pre-ORF format and is being converted to ORF.

To ensure comparability, data modifications involve a) consistency in assignment of UMLS identifiers; for example, identical strings should have the same SUI (String Unique Identifier), b) consistent use of column and attribute names (e.g., CODE instead of SCD in MRSO), c) recreating metadata (MRSAB, MRDOC, MRRANK, etc) where possible, d) consistency in use of versionless SAB (source abbreviations) values, e) recomputing index files by using a single, consistent version of the LVG program. These "comparable" releases will exist separately from the original releases so analyses can be done on either version or across versions. Current work is focused on making ORF comparable releases. Creating RRF comparable data will be a greater challenge as the ORF lacks atomic information present and tracked in the RRF. The archive will provide opportunities for research in many areas. A few examples:

- Concept history – ways to visualize the additions, deletions and transformations in a concept's life cycle. A meta-Metathesaurus to capture and identify changes in source level data aggregations (codes, concepts, descriptors) over multiple edit cycles within UMLS concepts.
- Inter- and intra-release gathering of statistical data and visualization. Examples are: characterizing high-level differences between releases, semantic profile of sources, content overlaps between sources, unique content from a source (i.e., areas where there is no overlap with other sources in the Metathesaurus). Some of this work is already underway.
- A prototype viewer is being built in Perl that is archive-aware and understands and highlights cross-release differences.
- Cross-release summary and metadata are being gathered to help with this work. For example, there are tables with keys (CUI, SUI, STR, etc) and accompanying bit strings representing each release that they are present in.

Identifying and separating "core" functionality in MetamorphoSys from add-on or plug-in functionality will also facilitate research in many areas. Such a Software Development Kit (SDK) for MetamorphoSys can be released to the UMLS user community for collaborative research. This would ideally involve collaboration and harmonization with other extant UMLS object models, e.g., MetamorphoSys, MEME, UMLSKS. User and programmer documentation, UML diagrams and reference implementations would be required before this can be made public. Once completed, the core MetamorphoSys maintenance and development can be transitioned to OCCS while research that uses the SDK continues in LHNCBC.

Training Opportunities

Working towards the future of biomedical informatics research and development, the Lister Hill Center provides training and mentorship for individuals at various stages in their careers. The LHNCBC Informatics Training Program (ITP), ranging from a few months to more than a year, is available for visiting scientists and students. Each fellow is matched with a mentor from the research staff. At the end of the fellowship period, fellows prepare a final paper and make a formal presentation which is open to all interested members of the NLM and NIH community. In FY2007, the Center provided training to 46 participants from 13 states and 9 countries. Participants worked on research projects including medical image processing, consumer health informatics, document analysis, grid computing, information retrieval, machine learning, medical illustration, micro-pathology, medical terminology research, natural language processing, medical ontology research, telemedicine, and ubiquitous computing. The program maintains its focus on diversity through participation in programs supporting minority students, including the Hispanic

Association of Colleges and Universities and the National Association for Equal Opportunity in Higher Education summer internship programs.

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