Communications Engineering Branch
Annual Report 2006
Submitted October 2006

George R. Thoma

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

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

Image Processing

Biomedical imaging R&D

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

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

WebMIRS (*Web-based Medical Information Retrieval System)*, developed some years ago and still

in active use, continues to provide access to images and text from nationwide surveys conducted by the National Center for Health Statistics. At the current time there are 444 users of WebMIRS in 54 countries. This Java application allows remote users to access data from the National Heath and Nutrition Examination Surveys II and III (NHANES II and III), carried out during the years 1976-1980 and 1988-1994, respectively. The NHANES II database accessible through WebMIRS contains records for about 20,000 individuals, with about 2,000 fields per record; the NHANES III database contains records for about 30,000 individuals, with more than 3,000 fields per record. In addition, the 17,000 x-ray images collected in NHANES II may also be accessed with WebMIRS and displayed in low-resolution form. The NHANES II database also contains vertebral boundary data collected by a board-certified radiologist for 550 of the 17,000 x-ray images. This data consists of *x,y* coordinates for approximately 20,000 points on the vertebral boundaries in the cervical and lumbar spine images. Users may do queries for both radiological and/or health survey data. An example of such a query is: "Find records for all persons having low back pain (health survey data) *and* fused lumbar vertebrae (radiological data)". The boundary data points are displayable on the WebMIRS image results screen and may be saved to the user's local disk.

The Digital Atlas of the Cervical and Lumbar Spine, also developed some time ago, remains available for the public from the CEB Web site either as a Java applet, or downloaded as a Java application. 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. The Atlas has capabilities to display color images, to add extensive text annotations, and to import/export sets of images and annotations as a package.

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

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

The *Multimedia Database Tool* (MDT), designed as the next generation WebMIRS system, provides (1) a software framework for the incorporation of new text/image databases in a much more general way than the current WebMIRS, and (2) new features for the database end user that extend current WebMIRS capabilities. This new system is intended to accommodate the new text/image database currently being created from the collection of uterine cervix images from NCI, but also the existing WebMIRS databases (of x-rays and associated text from NHANES) as well. The framework designed has the goal of accommodating these sets of text and images under a very flexible database schema and GUI approach intended to allow new databases to be incorporated with work done only at the level of the database administrator, and not at the software modification level. New features being incorporated include support for multiple levels of user privileges and capability for users at authorized levels to make new data entries into database fields. Hence, the system will allow not only data dissemination but distributed data collection as well.

Other tools motivated by our work with NCI include:

- *Boundary Marking Tool* (BMT). This provides Web capability to manually mark boundaries on cervicography images, and to manage collected data with a MySQL database. It is in active use by NCI for multiple studies.
- *Virtual Microscope* (VM). Though at an early planning stage, the VM will provide Web capability to view and collect information on histology images from expert observers. The system is based on concepts developed on pilot histology viewing systems already developed by CEB.
- *Teaching Tool* (TT). This system is for training medical personnel in cervix anatomy/pathology. It displays the uterine cervix images and quizzes an observer, and enables an NCI medical expert to tailor exams by specifying images and questions to use on an examination. The first phase is near completion; second phase is in early planning.

For our work with NCI, these systems are interrelated through the data that is used. The MDT will distribute 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.

The current status of work in multimedia databases and tool development is summarized as follows:

(1) A working implementation of the MDT is regularly used for demos, both by NLM and NCI. It operates on a database of a few thousand JPEG cervigrams and associated clinical text data from the Guanacaste Project. It has a WebMIRS-style graphical user interface and allows queries on any of the collected text data items. MDT design has included a Tabular View required to display and allow navigation of query results in a comprehensible and friendly manner, given the many-to-one relationship common in this data. For example, one patient may have multiple visits, each of which has multiple associated images. This latest in-house version also includes capability to query and display images segmented with the Boundary Marking Tool.

(2) The BMT is at a more mature level of development than the MDT. The current BMT allows a user to draw free-form boundaries for the very irregular regions in cervigrams, and to enter detailed, object-specific information that is important to researchers and practicing physicians in uterine cervix oncology. The BMT has supported a 525-image study (involving Dr. Michelle Ross of South Africa) of visual characteristics of patients with precancer undetected by colposcopy. In September, 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. The collaborative work using the BMT, was referenced in a paper, "Colposcopy at a crossroads", now in press for the *American Journal of Obstetrics & Gynecology*. Aside from the colposcopic work, the BMT has been used to mark and label dermatological lesions for the NCI Viral Epidemiology Branch, headed by Dr. Jim Goedert. This group has an immediate interest in conducting a two-observer study with the BMT that will use 926 dermatological images.

(3) For the histology viewing work, we maintain (1) a simple demo system of basic histology image and data collection capability; (2) fully-functional systems to support two on-going studies. This work is the basis for the design of the Virtual Microscope, an open source system based on Java technology and compatible with Internet Imaging Protocol standards.

(4) The Teaching Tool development has resulted in a prototype system available for experimentation by NCI and American Society for Cervical Pathology and Colposcopy (ASCCP) experts. The prototype demonstrates system access methods and provides several questions for an example certification examination. Efforts are under way toward implementing ASCCP's Resident Online Exam (ROE) for experts to exercise and provide feedback. In addition to implementing the Teaching Tool for training and testing colposcopists, we are preparing for an NCI study that will use the Teaching Tool for a multinational evaluation of a new approach for cervical cancer screening and treatment. This study is expected to use 721 images and will involve expert observers in the U.S., Peru, Costa Rica, Nicaragua, and South Africa.

CONTENT-BASED IMAGE RETRIEVAL (CBIR). The first goal of this project is to investigate approaches for query and retrieval of biomedical images by direct use of image data, possibly in association with text related to the biomedical images. Our emphasis is on two-dimensional images, primarily the NHANES spine images, using shape methods on vertebrae in the images, and on NCI cervigrams, using color and texture methods to differentially identify tissue regions and tissue characteristics within these images. The second 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.

Current status of the work is as follows. Our CBIR system currently at the highest level of functionality is CBIR3, which allows search of spine vertebrae by shape and/or descriptive text, using a database of several thousand pre-segmented vertebral shapes and text data from the NHANES II database used by WebMIRS. The key characteristics of this system, developed in MATLAB and Java, are that it can operate in networked or standalone modes, uses XML for reporting, and allows the user to select either a more mature or an experimental version of the system.

Another significant program we have developed is Relevance Feedback (RF), which is currently a standalone MATLAB experiment in utilizing feedback from an expert user for CBIR image retrieval. Work is under way to incorporate the capabilities of the CBIR3 and RF systems into SPIRS, our new, Web-based *Spine Pathology and Image Retrieval System*. In addition, development continues on the Pathology Validation and Collection (PathVa) tool, our Java-based system for segmentation review and editing. This tool will retrieve spine images that have been compressed with methods developed for NLM by Texas Tech, over the Internet, along with the segmented boundary data. Three board-certified radiologists are collaborating with us to review several hundred images and record presence, type, and degree of severity of anterior osteophytes, disk space narrowing, subluxation, and spondylolisthesis. The recorded information is automatically transmitted to a CEB database after validation of the segmented boundaries.

Other work includes the following:

(1) Collaboration with Tel Aviv University concentrated on integrating and improving previous algorithms for the development of a multi-step scheme for segmenting and labeling columnar epithelium, squamous epithelium, and acetowhite regions within cervix images, and on evaluation of multi-observer segmentation data with the Simultaneous Truth and Performance Evaluation (STAPLE) algorithm.

(2) Joint work with Texas Tech Univeristy is aimed at developing methods for discriminating between acetowhite lesions on cervix images that exhibit mosaicism patterns versus those that exhibit punctation patterns.

(3) PathVa, a Java tool for validation of image pathology and segmentation, now includes capability for LiveWire segmentation. We conducted the first remote test of this tool with collaborators at the University of Missouri.

(4) Research toward Level Sets Segmentation for the x-rays is continuing in partnership with Texas Tech University, and work is also under way to create an integrated shape segmentation system.

(5) We have begun a collaboration with the University of Aachen, Germany toward the creation of a shape query system for vertebrae using our NHANES vertebral images and a federated capability that incorporates capabilities of both the Aachen IRMA system and our own MATLAB based CBIR work. The NLM/Aachen interface protocol for this system has now been established and initial experiments in establishing the communications link are expected shortly.

(6) The initial version of a MATLAB-based CBIR retrieval system for uterine cervix images for engineering experimentation has been developed.

Document image analysis and understanding (DIAU)

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

Medical Article Records System (MARS). The MARS production system has evolved through several generations of increasing capability. Its core engine consists of daemons based on heuristic rule-based algorithms that use geometric and contextual features derived from OCR output to automatically segment scanned pages of journal articles, assign logical labels to these zones, and to reformat zone contents to adhere to MEDLINE conventions. About a quarter of the total citations in MEDLINE now are created by MARS, the remaining coming in as XML-tagged data directly from publishers.

Changes continue to be made to the MARS production system to accommodate new requirements from our colleagues in the Indexing section. We modified three MARS software modules (Edit, Reconcile, and Upload) and the validation library to automatically extract ISRCTN clinical trials and GEO (gene expression omnibus) databank numbers, and Wellcome Trust grant numbers. In addition,

changes were made to the Reconcile and Upload modules to accommodate another new requirement: to list author and corporate author (collective) names in the order they appear in the published article.

Rules in our algorithms continue to need changes. For example, our rule-based algorithm that reformats author names into the format required in MEDLINE has been modified to accommodate newly encountered names: e.g., G/Selassie (an Ethiopian name whose "slash" symbol was detected as an invalid character), and OConnor (with adjacent capital letters.)

A new module, SeekAffiliation (SA), was developed and incorporated in the MARS system. This module is intended to correct a long standing problem, viz., the labor involved in correcting the "affiliation" data indicating the institution of the principal author of the article (university, college, department, city, state, country). The reason for this higher manual effort is that this information is often in small print and italics, and hence is seldom recognized correctly by the OCR system. SeekAffiliation uses the first author's name (from OCR output) to query PubMed with Entrez e-utilities to find the institution to which the author had been previously affiliated. The search results are then compared with the affiliation text from the OCR output to find close matches. A closely matched affiliation from MEDLINE (by definition, complete and correct) is then offered to the Reconcile (text verification) operator, together with the affiliation from OCR output, so that either one may be selected or edited. The new system was released in late December, and the operators are already using SeekAffiliation's output. Usage data from January and February suggest that SA found potential affiliations for 24% of the articles processed in MARS, and that Reconcile operators selected one of these suggestions for 14% of those articles. Four of the operators, who use SA affiliations more frequently than other operators, used one of the suggested affiliations for 23.5% of the articles for which there were suggested affiliations. A paper on this technique, "Historical Author Affiliations Assist Verification of Automatically Generated MEDLINE Citations" was accepted by the 2006 AMIA Symposium.

WebMARS. Efforts continue toward meeting goals of the Indexing 2015 Initiative through the continuing development of two systems relying on WebMARS to assist both operators and indexers. The initial versions of both systems, WebMARS Assisted Indexing (WAI) and Publisher Data Review (PDR) are currently under test. The PDR system has been installed in the Indexing Section for beta testing by staff. This system will be augmented by a recently developed module which automatically extracts PubMed IDs for articles that comment on the article being processed. The technique is based on Support Vector Machine technology followed by a PubMed search for the PMIDs, and has shown an accuracy exceeding 95%.

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

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 search through the text and highlight these terms for the indexer to simply confirm and select, thereby reducing manual effort. An initial prototype was demonstrated to indexers who provided feedback for improvement. A pilot version of this system was delivered to the Indexing Section, and following a period of testing by indexers, and an analysis of their comments and suggestions for modifications, the system will be finalized for production.

ACORN. This system is intended to extract bibliographic information from 60 volumes of the printed Quarterly Cumulative Index Medicus (QCIM) from 1927 to 1956 to populate the OLDMEDLINE database. The design of the system is rooted in research in document image analysis and pattern matching techniques.

With the help of NLM's Preservation and Collection Management Section, the microfilm version of a particular volume (Vol. 59, Jan – June 1956) was scanned and the TIFF images subjected to OCR conversion. Though the OCR error rate is about 3% from microfilm compared to 2% from scanned paper, we are leaning toward microfilm scanning since it is faster and cheaper than scanning paper. Currently, a module is being created to extract journal name abbreviations from the DCMS database to compare against the abbreviations in the microfilm images. Also, work continues in developing and debugging several VB.NET class libraries to complete ACORN.

Ground truth data for document image analysis. By the end of September 2006, the Medical Article Records Groundtruth (MARG) database had 9688 unique IP visits from 96 countries. That is an increase of 3000 visits over last year. MARG provides TIFF images of biomedical journal articles and corresponding page segmentation and labeling results. This data set is used by designers to validate their own zoning and labeling algorithms.

AnatQuest: A window into the Visible Human

The two goals of this project are (a) to bring the *high resolution* Visible Human images to the lay public in an effective way (AnatQuest system); (b) to link text documents received from Web sources to relevant anatomic objects (TILE). In June 2006 the most-downloaded publication from the CEB Web site was the report on this project to the Board of Scientific Counselors (8.5% of total downloads.) Also, the monthly number of unique visitors to the AnatQuest Web site ranged between 8403 and 10,962.

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

prior to shipment back to the user. The third tier is the object-oriented database of high resolution VH images and rendered 3D anatomic objects. Low bandwidth connections are accommodated by a combination of adjustable viewing areas and image compression done on the fly as images are requested. Users may zoom and navigate through the images.

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

Research in TILE seeks the best alternatives for the functions needed to accomplish this linkage. These functions are: *identifying biomedical terms in a document, identifying the relevant anatomical terms, identifying the images in the image database, and linking the identified terms to the images.* Our main research focus is on the second function, the Term Mapper, which associates the biomedical terms (which are more likely to be disease terms rather than explicitly anatomic ones) in the document to appropriate anatomic concepts through the Metathesaurus concept relation table, and ultimately to images. Since this table typically yields several relationships that can potentially map a biomedical term to multiple anatomical concepts, relevance ranking is called for. Three ranking strategies are considered. The first, an image-label based ranking strategy clusters mapped concepts around the labels on images, assigning higher rank to the bigger cluster. The main problem with this technique is that ranking depends on how judiciously the images have been labeled. If the labels are too sparse or too diverse, ranking may not identify the most relevant image.

The second relevance ranking strategy, heuristics-based ranking, depends on the *number* of intermediate concepts and relationships linking the biomedical term to the anatomic concept. We hypothesize that fewer links in this chain indicate a higher ranking. Heuristics are obtained by a review of ranking results, and elimination of irrelevant concepts or relationships, as well as certain combinations of these. For instance, a combination of part_of and has_part results in sibling relationships among concepts belonging to different anatomical structures, thereby yielding a structure that is not relevant to the biomedical term.

The third strategy, model-based ranking, groups mapped concepts that are closely related according to the UMLS Metathesaurus and the semantic network. This is an attempt to offset the subjective nature of assigning labels to images. Following an investigation of several clustering software packages, Multivariate Data Analysis (from Univ. of Louis Pasteur, France) was selected, although further work is needed to overcome speed barriers.

As part of the research in evaluating these different strategies, and thereby arriving at an optimal approach to term mapping, we are developing a tool for a content expert to validate the relevance of the anatomic object retrieved to the text of the document. This Web-based research tool: (1)

displays the document text with biomedical terms highlighted; (2) displays a table of mappings of biomedical terms to anatomic structures and images; (3) and allows a researcher to enter data validating relevance of the anatomic image.

Information Systems

DocView Project: Document imaging for the biomedical end-user

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

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

However, reflecting the declining use of TIFF for distributing document images (compared with PDF), and the age of the software itself, the use of DocView is expected to decrease. Libraries use DocView in tandem with Ariel® software for their interlibrary loan services. Since new versions of the Ariel software issued by the marketer (Infotrieve) are not compatible with DocView (our Web site notifies users of this), the use of our software will drop as libraries change to the new Ariel software. Nevertheless, as evident from current use statistics, this changeover is likely to be gradual especially in foreign countries since their purchase of the new Ariel may take longer.

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 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.

An impetus for MyDelivery was the President's 2004 announcement of a ten-year national goal for computerizing patient health records. It was further boosted this year by the long-range NLM plan, calling for NLM's participation in developing and using electronic health records. Also, in spring 2006 the Lister Hill Center Board of Scientific Counselors advocated the development of tools for distance-independent collaboration between life science and informatics researchers.

The MyDelivery project seeks to overcome three significant obstacles. The first, to discover a way to send *large* electronic files over the Internet. The second, to create a way to send large files

reliably over wireless networks, which tend to be unreliable, but which are becoming ubiquitous. The third is to comply 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.

The transmission of large files over the Internet has always been problematic. Examples of large files include document images, digitized photographs, digitized x-rays, sonograms, CT and MRI scans, and digital video. These range from megabytes to several gigabytes in size. Software tools traditionally used for file communication have problems with large files. At NIH email attachments are limited to 50 megabytes. While FTP servers can be used for large file exchange, they are difficult to set up and administer. Instant messaging can be used for exchanging files, but it tends to be unsafe, and NIH for the most part has banned its use. We have created a novel method for large file exchange for MyDelivery that allows two client computers to exchange large files through an intermediary server via a user interface similar to email. Our proof-of-concept system permits the exchange of files ranging up to gigabytes in size.

Unreliable networks pose a major challenge to large file exchange. Twenty-five years ago at the dawn of the Internet, computer communication was very reliable, and mostly done through land-based wired connections. Today Internet communication is considerably less reliable due to the widespread use of wireless connections: communication is problematic unless there is a strong signal. We expect that many of our users will be using wireless networks. Since large files require a long time for transmission, it is essential that a mechanism be developed to communicate over intermittent networks. Part of the development of MyDelivery has been to create a method of automatically recovering from communication failures due to reduced signal strength. This part of the project has been completed, and successfully tested over unreliable networks.

The third and final stage of MyDelivery development is under way: making the system HIPAA-compliant, a necessary requirement for computer systems that use patient information. HIPAA requires the system to encrypt all communication of electronic health information, and to verify that it transmits the information reliably. The challenge is to create a method of encryption that allows users to roam from computer to computer in a manner that is not only safe and secure, but also easy. For instance, users may want to use the client software at work, at home, or on the road. Techniques used for encryption over the past ten years utilize certificates, which users must procure, install on their computers, distribute to colleagues, export to other computers, and renew annually before they expire. The use of certificates is therefore not easy. Our current work centers on using public domain software for certificate generation and use. A technique is being developed that will allow user roaming in an easy manner, without conflicting with existing patents in this area.

This third stage of MyDelivery development will be followed by an extensive in-house alpha test of the system, during which bugs will be discovered and fixed. The system will be load-tested to determine how it performs under heavy usage, and its design will be refined. This alpha test will be followed by beta testing, initially within NLM, then with selected outside groups (research, administrative, library and clinical) to evaluate both the user interface and the system to ensure that it meets our project goals.

DocMorph and MyMorph. The DocMorph system continued to serve both browser-based users (14,400 to date: 1900 more than last year) and MyMorph users (6500 users) this year. Most of the registered users are biomedical document delivery librarians. DocMorph allows the conversion of more than 50 different file formats to PDF, for instance, to enable multi-platform delivery of documents. Also, by combining OCR with speech synthesis, DocMorph enables the visually impaired to use library information. It has been used by librarians for the blind and physically handicapped to convert documents to synthetic speech recorded onto audio tapes for blind patrons. Most users continue to use it to convert files to PDF to enable multi-platform delivery of documents. DocMorph is available at http://docmorph.nlm.nih.gov/docmorph.

MEDLINE Database on Tap (MDoT)

This project (formerly called PubMed on Tap) seeks to discover and implement systems and techniques to assist mobile clinicians in quickly finding relevant, high quality information addressing clinical questions that arise at the point of care. By combining what we learn about the habits and preferences of our targeted users with our understanding of the data available from MEDLINE citations, we present information to users so that they can quickly find the most pertinent parts, despite the limitations placed by the small screen and restricted bandwidth of handheld computers. We explore display and navigation techniques, as well as information organization and content. We have also incorporated tools and systems from other LHC projects, such as MetaMap (to identify biomedical concepts) and the Essie search engine. Essie and Google are offered as options, while the primary search engine is PubMed. We also seek to integrate semantic data from the search query and found citations to optimally rank results while maintaining real time response. As our primary method of discovery, we developed a testbed system that supports MEDLINE search and retrieval from a wireless, Internet-connected PDA. Our client software for Palm OS and Pocket PC OS are freely available from our Web site which experiences between 5000 and 6000 hits every month. This Web site provides information about the project as well as the software, and allows us to solicit feedback from our users and monitor aggregate user behavior. There are over 500 registered users of MDoT, and an unknown number of unregistered ones.

In September 2005, the Board of Scientific Counselors cited the project as an important one for NLM as it expands the library's role in the area of clinical decision support. They commended the team for its approach to rapid prototyping and design refinement, and for the use of other products, tools and research approaches.

From the Fall of 2005 through 2006, MDoT was evaluated in clinical settings at two institutions, the first at the University of Hawaii's John A. Burns School of Medicine (JABSOM) and the other at the VA Medical Center (VAMC) in Washington, D.C. At JABSOM, medical residents enrolled in Dr. Jacobs' Medical Informatics elective conducted fieldwork between December 2005 and March 2006. Each resident accompanied medical teams on morning rounds for approximately 4 weeks, using MD on Tap to seek answers to clinical questions that arose at the point of care. They submitted daily summaries to MD on Tap briefly describing each scenario and question and noting which, if any, citations found with MD on Tap were relevant to

answering the question. Together, they submitted summaries for 44 rounds of about one hour each, with 187 clinical questions. Using a variety of MD on Tap options, they found relevant citations for 153 (82%) of the questions.

To observe the evaluation, three CEB members of the MDoT team conducted a site visit of JABSOM in February. They presented an overview of the system for Grand Rounds at Queens Hospital, including a brief description of the evaluation study. They also conducted workshops for the Medical Education Fellows, the medical students, medical librarians from all over Oahu, and the Geriatric Fellows. They took a brief tour of the ICU at Kuakini Hospital where the residents conducted most of their evaluations, and attended a seminar for 3rd year medical students on Internet resources and tools for EBM. In all these encounters, they used three PDA/cell phones to work individually with participants, give them hands-on experience and observe their reactions.

The second evaluation was conducted at the VA Medical Center (VAMC) in Washington, DC in the summer of 2006. This was a three-part collaboration among NIH/NLM/LHNCBC, the VAMC and Universidade da Beira Interior, Portugal. Sara Rocha, a sixth year medical student at that university, was recruited to conduct the fieldwork. She rounded for 20 days during August with four different teams, using MDoT to search for answers to questions that arose on rounds. She recorded 144 clinical questions that were asked in context of 78 in-hospital clinical scenarios and 17 topic reviews. Our server recorded 1626 transactions at the VAMC, including 942 queries and 670 citation fetches. She identified 320 citations as being relevant to the clinical question being asked. This information along with her daily summaries and server transactions will be analyzed for search behavior patterns and relations between MDoT features and time to find useful information. Because the VAMC is equipped with WiFi throughout, a significant difference between this study and the Hawaii study, in which residents used PDA/cell phones, is network data rate. One question to address is whether this higher data rate has a notable effect on the ability to find useful information at the point of care. Ms. Rocha presented an overview of her study and results to LHNCBC staff, reporting that answers were found in MEDLINE for 73% of the questions that arose on rounds, an unexpectedly high figure. On average she initiated 5.2 queries per question, spending less than 4 minutes per question finding relevant citations. Her results show that MDoT is fast and effective.

Papers on this project accepted by the 2006 AMIA Symposium were: (1) “MEDLINE as a Source of Just-in-Time Answers to Clinical Questions”; (2) “Preliminary Comparison of Three Search Engines for Point of Care Access to MEDLINE® Citations.”

The MDoT evaluation plan calls for a “second opinion” of the selected citations by a senior investigator and expert MEDLINE indexer at LHNCBC. Based on her review of the scenario and clinical question, she assigns each selected citation a score of **A** (the abstract answers the question), **B** (the abstract contains a partial answer or is topically relevant and clearly indicates that the full text might answer the question), or **C** (the abstract does not answer the question and was selected by the resident for some other reason.) Analysis of this data is planned for the next reporting period, in addition to writing a final report to the NIH Office of Evaluation and preparing a paper for publication.

Outcomes Research. Currently as part of the MDoT project, but with applications beyond, 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 our tools both to find outcomes in general, and to find high quality outcomes that answer specific clinical questions. The *ability of our system to find outcomes* was measured by comparing outcomes extracted from MEDLINE citations to the outcomes manually annotated in these citations. Of outcomes extracted by the system 87.7% to 92.8% (depending on the clinical task) were also annotated as such by a group of experts.

The *ability of our system to select highly relevant outcomes* as answers to specific clinical questions was tested on 24 clinical questions evaluated by a board certified family physician. Our system extracted five top outcomes to answer questions, such as "Can selective serotonin reuptake inhibitor (SSRI) use cause impulsive suicidal or homicidal behavior?" As an answer to this question, a highly ranked outcome extracted by our system is the following meta-analysis published in BMJ in February 2005: "We found weak evidence of an increased risk of self harm (1.57, 0.99 to 2.55). Risk estimates for suicidal thoughts were compatible with a modest protective or adverse effect (0.77, 0.37 to 1.55). When prescribing SSRIs, clinicians should warn patients of the possible risk of suicidal behaviour and monitor patients closely in the early stages of treatment". Extracted outcomes were evaluated on a three point-scale: 1) it answers the question; 2) it contains useful information that could lead to an answer; and 3) it is not useful. For the 24 questions at least one of the extracted outcomes was judged as providing an answer to 20 questions. For the remaining four questions, at least one extracted outcome contained useful information for three, and no useful information was found for one question.

We propose to use our extraction system to create a repository of patient outcomes and strength of evidence extracted from MEDLINE citations. Such a repository should be useful for the practice of evidence-based medicine. Our preliminary development and evaluation strongly suggests that it is feasible to create a database of patient oriented outcomes automatically extracted from the medical literature. Once the prototype repository is created, its usefulness in providing patient oriented outcomes needs to be evaluated. Application areas anticipated might include clinical trials design, EMR, and a patient-oriented service.

Interactive Publications Research

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

HTML and PDF. The IPs in these formats were compared in terms of ease of use and development effort. A paper describing the development process was published in 2006:

Thoma GR, Ford G, Chung M, Vasudevan K, Antani S. Interactive publications: creation and usage. Proc. IS&T/SPIE Electronic Imaging 2006: Digital Publishing. San Jose, CA. Jan 2006. SPIE Vol. 6076: 607603 (1-8).

While using such a document, the reader is able to: (a) view any of these objects on the screen; (b) hyperlink from one object to another; (c) interact with the objects in the sense of exercising control over them (e.g., start and stop video); (d) and importantly, reuse the media content for analysis and presentation.

This project was presented to the Board of Scientific Counselors in September 2005, including a demonstration of a prototype Interactive Publication. Functions demonstrated were converting tables to graphs, zooming into graphs, creating subsets of the tabular data, zooming into images and changing contrast in DICOM images. Details of the research appear in the report written for the Board. The Board commended the team for developing the project specifications, investigating current multimedia standards, and developing the ITAG tool in a short period of time. They also felt that the ongoing work is clearly relevant to the broader goal of improving access to published information.

In light of the large sizes of such publications possibly in the range of hundreds of megabytes, research is ongoing toward identifying techniques and protocols for rapid progressive download of the publications, and the development of a Download Manager based on this research.

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

Digital Preservation Research

The goal of this project is to investigate key issues related to the long term preservation of digital material, both digitized documents and video. Our work in document preservation is more mature, and focuses on two functions of an economical and robust digital preservation system: automated metadata extraction and file migration.

For document preservation, a prototype *System for Preservation of Electronic Resources* or SPER was developed. SPER is a flexible, modular system that demonstrates key functions such as ingest, automated metadata extraction (AME) and bulk file migration. AME is implemented for the extraction of descriptive metadata from scanned and online journal articles as well as NLM’s obsolete Web pages. Bulk file migration is implemented through an existing CEB system,

DocMorph. While these functions are developed in-house, for the necessary infrastructure capabilities in SPER we have incorporated into the system, and customized, the latest version (1.4) of MIT's open source DSpace software. The Java client GUI for SPER was enhanced to incorporate batch metadata extraction and ingest for journal article TIFF pages, online journal articles and NLM Web pages (HTML). The GUI was also redesigned to display Web pages and online articles through Java Swing components.

SPER, in an abbreviated form, is being used in the preservation of a new collection at NLM consisting of over 65,000 historical Food and Drug Administration court records. 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. Our approach consists of: scanning the paper documents, auto-zoning the TIFF files using OCR output from the scanned documents, feature extraction, optimal feature selection, feature classification using a Support Vector Machine (SVM) classifier, multi-class probability estimation, and statistical parsing using the Stolcke-Earley parsing algorithm.

Presentations of ongoing research in digital preservation were made to staff in 2006 focusing on overall SPER design, an end-to-end design of the automated metadata extraction subsystem based on learning methods, and experimental results. Also, a paper describing the application of SPER to the preservation of the FDA collection was published in the proceedings of the 10th European Conference on Advanced Technologies for Digital Libraries (ECDL 2006).

While document preservation was our main focus, 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. Following 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, problems to be solved were identified, including: roadblocks to synchronizing metadata with video; pertinent standards bodies to consider certain specific improvements for the MXF and MJ2 file formats; tools to make the widespread adoption of disk-based lossless compression possible. Information about this meeting appears at http://archive.nlm.nih.gov/VideoArchivists2005/.

Multimedia Visualization

Turning The Pages Information Systems (TTPI)

The goal of this project is to bring the magnificent rare books at the Library to public view in a compelling way: as photorealistic volumes whose pages may be 'touched and turned'. Visitors to the Library may experience this on kiosks, and those offsite may view the books online.

Originating as a collaboration with the British Library in producing two virtual books, Blackwell's 18th century *A Curious Herbal* and Vesalius' 16th century Anatomy book, in TTP form, we have since made significant improvements on the original process. Our process consists of scanning the pages, enhancing these high quality color images by Adobe Photoshop, creating animated 3D

wireframe models of the pages and book cover using Alias Maya (an innovation in our approach), run on a computer by Macromedia Director software, and displayed on a touchscreen monitor in kiosks. The library patron may 'touch and flip through' each of these books in an intuitive manner that evokes the feel of a 'real' paper volume.

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

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

To create a Web version of TTP (*TTP Online*) several changes were made to the images and the mode of delivery. First, the page image resolution was decreased from 1152 x 870 pixels (kiosk version) to 800 x 600. To provide interactivity with the system and page sequencing, the platform was changed from Macromedia Director to Flash since the latter uses less memory (being vector-based), has more compression alternatives, and is more widely available in Web browsers.

In TTP Online, each image of the page animation imported into Flash is compressed to 15% quality (a compression ratio of about 6). The end pages are compressed to 40% quality (as the book will stop on these end pages of each page spread). The zoom modes (which require less compression) are generally 50 – 60% quality. This reduces the overall size of each page flip animation.

To offset the Internet bandwidth limitation that many users will experience we use progressive transmission. In other words, while the animation on page 2 is playing, the animations on pages 3 – 5 are loading in the background. We also preload the first five page flip animations to keep loading times down to a minimum. Generally, users will not notice any loading times aside from the initial load period on DSL or cable internet connections. On the Web version all pieces of TTP (audio, text, zoom, page flip, instructions) are separate files. Thus they all load separately on demand rather than all at one time, making the delivery as efficient as possible.

Extensions were made to the TTP versions of Blackwell's Herbal and Vesalius. Blackwell's Herbal was redesigned to retain 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, one can go to a PubMed search and get citations, or link to ClinicalTrials.gov and get information on clinical trials of this drug. A design strategy similar to that followed for Blackwell was undertaken for Vesalius, such as: a menu button invoking a table of contents, animated page flipping, timeouts and countdown warnings. In addition however, the page images from Vesalius were interlinked to images from other sources (e.g., rendered Visible Human

images, pictures of Italian cities, etc.) to present the patron with several multimedia 'stories,' e.g., "Man of Padua," "Modes of portraying anatomy." By incorporating explanatory and current online information, our TTP versions of both Blackwell and Vesalius deliver services useful for the information-seeking user.

Future goals of the TTPI project are to continue the search for efficiency in producing the TTP books, as more historical books are selected for the library's constituencies, and to investigate the tradeoffs in distributing them through the Web while maintaining high quality.

In 2005 we collaborated with Library of Congress staff to create TTP at their institution. They were invited to a demonstration of our kiosk version of TTP, and a technical discussion of our process. In addition to using their own resources and knowledge to accomplish the scanning, image enhancement and 3D modeling, they needed our templates for the final stage (to produce the software providing the interactivity). At our suggestion, they selected books in the life sciences: one by a Dutch surgeon who spent a decade with pirates in the Caribbean (1678); the other an encyclopedia of flora and fauna in the New World (1635). The first book was shown at the opening of the Kislak Collection, an event at the Library of Congress in 2005.

Engineering Laboratories and Resources

The R&D conducted by the Communications Engineering Branch relies on laboratories designed, equipped and maintained by the Branch, as well as content resources that support research.

Image Processing Laboratory. The CEB Image Processing Lab is equipped with a variety of high end servers, workstations and storage devices connected by a mix of 100 and 1000 Mb/s Ethernet. The laboratory supports the investigation of image processing techniques for both grayscale and color biomedical imagery at high resolution. In addition to computer and communications resources and image processing equipment to capture, process, transmit and display such high-resolution digital images, the laboratory also archives a variety of image content.

The equipment includes a Sun Enterprise 4500 server with dual 400 MHz CPUs, and 1.5GB memory, and a SunFire 280R server with dual 1.2 GHz CPUs, 3 GB memory, and two internal 73 GB SCSI disks. Additional computers in the lab include two Sun Ultra 10 workstations, each with a 440 MHz CPU, 512 MB memory, and an external 36 GB SCSI disk; and two Sun Ultra 10s, each with a 300 MHz CPU and 512 MB memory. All of these machines run the Solaris 9 operating system. Desktop computers for the research staff are largely high end PCs running both Windows and Linux.

Large-scale magnetic storage is provided by a Network Appliance FAS960 which is a network-attached storage (NAS) device connected by redundant Gb/s Ethernet connections and provides 24TB of RAID storage.

For the ultra-high-resolution display of x-ray images, two E-systems Megascan monitors provide image display at a spatial resolution of 2048x2560 pixels.

The laboratory also contains specialized equipment and software for device calibration and color profile creation. This includes a USB-interfaced MonacoOPTIX colorimeter, capable of color measurement from emissive sources, for CRT and LCD monitor color calibration, and used with MonacoOPTIX software; and a USB-interfaced GretagMacbeth Eye-One spectrophotometer, which measures color in the 380-730 nm range, with resolution of 10 nm, from both emissive and reflective sources, used with MonacoProof software, for the creation of standard color profiles which characterize the color I/O of devices such as scanners, monitors, and printers using the International Color Consortium (ICC) standard.

Image Processing Lab content resources. A large part of the NHANES II data has been put into the WebMIRS database tables. All of the NHANES II demographic, anthropometric, physical examination, and adult health questionnaire data is available through WebMIRS, as well as the statistical weighting and sampling strata variables required for analysis of the data. This data covers a nationwide sample of approximately 20,000 survey participants. In addition, the 17,000 NHANES II cervical and lumbar spine x-ray images are available for viewing through WebMIRS, in one-quarter spatial resolution format. These 17,000 images are stored in a magnetic RAID system and are available for public downloading via FTP, in their original digital 12-bit format; in addition, 1,000 of the images are available in TIFF 8-bit format, for compatibility with widely available image display and processing software.

Similarly, a large part of the NHANES III data has been put into the WebMIRS database tables. All of the NHANES III demographic, physical examination, health questionnaire, and laboratory data are available through WebMIRS, as well as the statistical weighting and sampling strata variables required for analysis of this data. The NHANES III data covers a sample of approximately 30,000 survey participants.

Currently, 100,000 images of the uterine cervix from a large National Cancer Institute (Guanacaste) study are being scanned for Web distribution. In addition to these are pap smear and histology images, also from this study.

In addition to the data above, the Image Processing Lab also contains a selection of History of Medicine color images digitized at high resolution from the Library's Arabic and Persian medical manuscript collection.

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

The laboratory also contains rack-mounted, networked processors running all recent versions of Windows-based operating systems to support the DocView, DocMorph, MyMorph and MyDelivery projects. This provides an easily-configurable test platform for simulating a variety of potential user environments, including those with firewalls, for testing, modifying and improving software developed in these projects.

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

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

CEB Publications 2006

Thoma GR, Mao S, Misra D, Rees J. Design of a digital library for early 20th century medico-legal documents. Proc. ECDL 2006. Eds: Gonzalo J et al. Berlin: Springer-Verlag; LNCS 4172: 147-57.

Long LR, Antani S, Jeronimo, Schiffman M, Bopf, Neve L, Cornwell C, Budihas SR, Thoma GR. Technology for medical education, research, and disease screening by exploitation of biomarkers in a large collection of uterine cervix images. Proc. 19th International Symposium on Computer-Based Medical Systems (CBMS 2006), June 2006, Salt Lake City, Utah; 826-31.

Xiaoqian X, Lee DJ, Antani S, Long LR. Pre-indexing for fast partial shape matching of vertebrae images. Proc. 19th International Symposium on Computer-Based Medical Systems (CBMS 2006), June 2006, Salt Lake City, Utah; 105-10.

Yao J, Antani S, Long R, Thoma G, Zhang Z. Automatic medical image annotation and retrieval using SECC. Proc. 19th International Symposium on Computer-Based Medical Systems (CBMS 2006), June 2006, Salt Lake City, Utah; 820-5.

Jeronimo J, Massad S, Wheeler C, Neve L, Long R, Schiffman M. Colposcopic appearance of the cervix of HPV-infected and HPV non-infected women. 23rd International Papillomavirus Conference and Clinical Workshop, 1-7, 2006, Prague, Czech Republic. (Abstract)

Zou J, Le DX, Thoma GR. Combining DOM tree and geometric layout analysis for online medical journal article segmentation. Proc. Joint Conference on Digital Libraries (JCDL), June 2006, Chapel Hill, NC; 119-28.

Kim J, Le DX, Thoma GR. Automated Extraction of Bibliographic Information from Biomedical Online Journal Articles Using a String Matching Algorithm. Proc. 19th IEEE International Symposium on Computer-Based Medical Systems, June 2006, Salt Lake City, Utah; 905-10.

Gordon S, Zimmerman G, Long R, Antani S, Jeronimo J, Greenspan H. Content analysis of uterine cervix images: initial steps toward content based indexing and retrieval of cervigrams. Proc. SPIE Medical Imaging 2006. Eds: Reinhardt JM, Pluim JP. San Diego, CA; Feb 11-16, 2006; SPIE vol. 6144: 1549-56.

Castle PE, Jeronimo J, Schiffman M, Herrero R, Rodriguez AC, Bratti MC, Hildesheim A, Wacholder S, Long LR, Neve L, Pfeiffer R, Burk RD. Age-related changes of the cervix influence human papillomavirus type distribution. Cancer Research 2006; Jan 15, 2006; 66 (2): 1218-24.

Jeronimo J, Long LR, Neve L, Bopf M, Antani S, Schiffman M. Digital tools for collecting data from cervigrams for research and training in colposcopy. Journal of Lower Genital Tract Disease. Jan 2006; 10 (1): 16-25.

Jeronimo J, Long R, Neve L, Ferris D, Noller K, Spitzer M, Mitra S, Guo J, Nutter B, Castle P, Herrero R, Rodriguez AC, Schiffman M. Preparing digitized cervigrams for colposcopy research and education: determination of optimal resolution and compression parameters. Journal of Lower Genital Tract Disease. Jan 2006; 10 (1): 39-44.

Thoma GR, Ford G, Chung M, Vasudevan K, Antani S. Interactive publications: creation and usage. Proc. IS&T/SPIE Electronic Imaging 2006: Digital Publishing. San Jose, CA. Jan 2006. SPIE Vol. 6076: 607603 (1-8).

Antani S, Cheng J, Long J, Long LR, Thoma GR. Medical validation and CBIR of spine x-ray images over the Internet. Proc. IS&T/SPIE Electronic Imaging 2006: Internet Imaging VII. San Jose, CA. Jan 2006, SPIE Vol. 6061: 60610J (1-9).

Demner-Fushman D, Few B, Hauser SE, Thoma GR. Automatically identifying health outcomes in MEDLINE records. J Am Med Inform Assoc. 2006 Jan-Feb;13(1):52-60.