# Communications Engineering Branch
# Annual Report 2005
**Submitted October 2005**
**George R. Thoma**

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

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

## Image Processing

### Biomedical imaging and multimedia database 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), the National Institute of Arthritis, Musculoskeletal and Skin Diseases (NIAMS), and the National Cancer Institute (NCI).

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.

Developed some years ago and still in active use, WebMIRS (*Web-based Medical Information Retrieval System)* continues to provide access to images and text from nationwide surveys conducted by the National Center for Health Statistics. This Java application allows remote users to access data from the National Heath and Nutrition Examination Surveys II and III (NHANES II and III), carried out during the years 1976-1980 and 1988-1994, respectively. The NHANES II database accessible through WebMIRS contains records for about 20,000 individuals, with about 2,000 fields per record;

the NHANES III database contains records for about 30,000 individuals, with more than 3,000 fields per record. In addition, the 17,000 x-ray images collected in NHANES II may also be accessed with WebMIRS and displayed in low-resolution form. The NHANES II database also contains vertebral boundary data collected by a board-certified radiologist for 550 of the 17,000 x-ray images. 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. At the current time there are about 400 users of WebMIRS in 44 countries.

Another tool developed some time ago, the Digital Atlas of the Cervical and Lumbar Spines, 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 our 17,000 digitized spinal x-rays continues to be very active, with about 400 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.

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

Other tools include:

- Boundary Marking Tool (BMT)

    - Provides Web capability to manually mark boundaries on cervicography images
    - Provides capability to manage collected data with MySQL database
    - In active use by NCI for multiple studies
- Virtual Microscope (VM)
    - Will provide Web capability to view and collect information on histology images from expert observers
    - Will be created from concepts developed on pilot histology viewing systems already developed by CEB
    - In early planning stage
- Teaching Tool (TT)
    - Provides capability to display NCI uterine cervix images and quiz observer for training medical personnel in cervix anatomy/pathology
    - Provides capability for NCI medical expert to tailor exams by specifying images and questions to use on exam
    - 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) Our working implementation of the MDT is regularly used for demos, both by NLM and by 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 development has focused on software infrastructure work to provide enhanced data views to the user, streamline memory and transmission requirements, and to analyze expected use cases for viewing collections of images of varying image modalities, such as the NCI cervigrams, histology images, and PAP smear images.

(2) The BMT, in its current version, allows a user to draw free-form boundaries for the irregular regions in cervigrams, and to enter detailed, object-specific information important to researchers and practicing physicians in uterine cervix oncology. Developmental work includes making the BMT easier to install through the use of Java Web Start, adding more online documentation, better error reporting, and the incorporation of a JDBC Type III driver called JDBTunnel, which frees BMT clients from the need to connect directly to the MySQL database server. This latter requirement in a previous version of the BMT was the source of some user problems, due to the firewall configurations at their installations. A significant NCI study is now being planned that will have 20-25 expert observers use the BMT to mark boundaries. For this effort, a new BMT capability was incorporated that allows the BMT to be configured to collect only items of interest for a particular study, with a workflow that is suitable for that study, meaning that all unnecessary warnings and error messages are turned

off.

(3) The histology viewing work has advanced significantly; we now have (1) a simple demo system of basic histology image and data collection capability, and (2) a fully-functional histology viewing system for a specific study on a set of several hundred patients that includes capability for our NCI collaborators to graphically annotate areas of interest, for observers participating in the study to view the annotated images and record cell counts within the regions of interest, and the aggregation of the collected data in a systematic fashion for analysis. A third histology viewing system is currently being developed for an upcoming NCI study that will include 600 patients and five expert observers.

(4) The histology viewing work will now be extended to create the Virtual Microscope. We are now on the second version of the concepts document for that system, following a technical meeting with NCI collaborators. The VM is being designed with the goal of meeting the requirements of additional histology researchers, including two groups at NCI.

(5) The Teaching Tool is nearing completion of the first phase of its development and has been demonstrated. In addition, an operational version has been installed for further evaluation and feedback. We are now in the early planning stages for the further development of this tool through next year.

(6) Our current image data acquisition work includes the ongoing digitization of the 60,000 35mm color cervigram slides from the Guanacaste Project. This work includes slide cleaning, scanning, quality control viewing, logging, and recording to both DVD storage and storage on our NetApp equipment. These slides are being digitized by a Nikon Coolscan scanner at 2000 dpi. We now have over 42,000 of these slides digitized.

Content-Based Image Retrieval (CBIR). The goals of this work are: (1) To research and implement the latest technological approaches for indexing and retrieving biomedical images by direct use of image data, and also 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. (2) To develop effective CBIR methods to be incorporated into our multimedia database programs (such as the MDT) or into separate, prototype systems for use by the biomedical research and/or clinical communities.

Our approach is based on developing prototype systems that enable CBIR. The system which has the best integrated set of functionality is CBIR2, 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. Work is proceeding on the third version of the prototype, CBIR3. 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.

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

Level Sets Segmentation methods for the spine x-rays; and the development of an integrated shape segmentation system.

Ongoing work in CBIR is summarized as follows:

(1) Algorithm results on a 120 image truth set have shown promising capability to satisfactorily remove specular reflections (from camera flash), correct for illumination variance, segment acetowhitened regions, and locate the os, as verified by our NCI collaborators. Our collaborators are preparing a second set of truthed images for further testing.
(2) As part of our ongoing activity to create "truth sets" for biomedical images that we are working with, all sets of vertebrae "9-point" segmentation data (that were planned for 2005) have now been created by University of Missouri radiologists and received at NLM; these 9-point, simple segmentations will be used to initialize our automatic segmentation process that will output more complex, 36-point segmentations; these 36-point segmentations will then go back to the experts in Missouri for checking/editing.
(3) Research toward Level Sets Segmentation for the x-rays is continuing at Texas Tech University and work is also under way to create an integrated shape segmentation system.
(4) A relevance feedback technique for shape-based image retrieval was completed. To evaluate this work, a truth set was created with about 1000 shapes. Our relevance feedback system classified these shapes for severity and type of anterior osteophytes. In most test cases, the system was able to attain 100% relevance of the top 10 retrieved shapes, by using two feedback iterations.
(5) Work to develop a new image segmentation reviewer/editor is continuing. This system is intended for use by groups such as the University of Missouri radiologists who will review outputs from our automatic segmentation algorithms, edit them, and record expert evaluations of the pathology as well as biomedical features in the images.
(6) An exploration was done of *global* image classification using techniques similar to those of the European ImageCLEF community using the Gnu Image Finding Tool (GIFT). This work classified medical images into 11 categories, including anatomy, orientation, and view. The algorithm was trained on 9000 images with known truth, and evaluated on 1000 test images. The results showed an average classification error of 15%, across the various categories, and are among the best, when compared with the ImageCLEF 2005 results.
(7) One of our Yale collaborators provided and installed a shape space indexing method for organizing segmented spine shapes into a tree for efficient search and retrieval. Initial tests indicate that we get a 5-fold speed-up in our searches.
(8) A prototype CBIR client/server application was developed. This application allows remote users to pose shape queries to shapes that have been indexed with the shape space indexing method developed by our Yale collaborators. The system incorporates a Java image server that delivers images using Texas Tech HVSQ compression. It is a loosely coupled set of modules developed in MATLAB and Java and is CEB's first demonstrable prototype for CBIR over the Web.

**Document image analysis and understanding**

Our DIAU research 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. Earlier, we had made several improvements to the production MARS system including a software suite for accommodating foreign language journals, Meeting Abstracts Extensions (MAX), and many others.

*Medical Article Records System* (MARS) in 2005 progressed sufficiently to enable NLM to discontinue the manual keyboarding activity entirely. A key element in allowing NLM to eliminate its keyboarding contract is the capability designed in MARS to accommodate foreign language journals (that account for 11% of MEDLINE citations). This requirement introduced new rules to extract vernacular titles (requied in Roman script languages but not in others), and process the second pages of articles (to accommodate abstracts that spill over to a second page). These goals have been achieved by our FLEX software suite that is incorporated in several MARS workstations.

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

New requirements set by NLM's Index Section continue to require modifications in the MARS system.

(a) New *Research Support* check tags: To expand the "research support" information that identifies the source of funding for the work reported in an article, several modules were modified and incorporated in the production MARS system, e.g., Edit, EditDiff, Reconcile, and Upload. These new check tags indicate whether the work was supported by intramural or extramural programs at NIH, as well as the Wellcome Trust.
(b) Citation format validation library: To ensure compatibility with the Indexing Section's format for MEDLINE citations, we created a document which was reviewed by them. This Citation Format Validation document was created by the MARS team by analyzing documents and technical memorandums from the Index Section. The rules (as we understand them) were confirmed by the Indexing Section. Based on this review, a software library was implemented and is being tested. After testing, it will be incorporated into the Reconcile and Upload modules. This is expected to reduce manual effort by the MARS operators since they will not have to manually enforce these rules.
(c) A new requirement is for Modern Greek to be handled in the same way as Russian, Bulgarian, Ukrainian, and Serbo-Croatian, all languages in Cyrillic, which do not require vernacular titles to be uploaded. The modules to be changed to accommodate this are Edit, Reconcile, and Upload.
(d) To accelerate our testing process for new or modified software modules, an automated testing package has been acquired.

*WebMARS*. This system was created to complement MARS by enabling the extraction of bibliographic citations from online journals. Since a majority of citations now come directly from publishers in XML format, WebMARS functions have been used to develop two other systems that

will serve to increase the efficiency of creating citations for MEDLINE so that the expected doubling of the citation rate in a few years can be accommodated through automation, a goal of NLM's Indexing 2015 Initiative. One of the advantages of WebMARS is that all of the bibliographic data contained in the online article may be extracted.

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

An initial version of the PDR system was completed after testing the software to handle databank accession numbers and grant numbers using a training set of several hundred online articles. Currently efforts are under way to discover discrepancies between those detected and actual data in MEDLINE. In collaboration with the Indexing Section, rules are being developed to eliminate these discrepancies, so that PDR can reliably provide operators data missing from the XML citations sent in directly by publishers. PDR will also correct incorrect data sent in by publishers. These features will help reduce the manual effort in creating citations for MEDLINE.

The second system, the WebMARS Assisted Indexing (WAI) system is for the indexers; it will help them search for terms in an article that correspond to biomedical terms in a predefined list. Again, indexers currently have to read through the entire article to confirm the occurrence of these terms, a labor-intensive process. WAI will automatically search through the text and highlight these terms for the indexer to simply confirm and select, thereby reducing manual effort. An initial prototype was demonstrated to indexers who provided feedback for improvement. A pilot version of this system was delivered to the Indexing Section in July. 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.* The goal of this system, rooted in research in document image analysis and pattern matching techniques, is 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 Scan or Import/Display/Quality Control module and the Image Segmentation and Labeling module have been completed. Work is proceeding on the development of the Zone Identification and Label Confirmation module. Also, an error handling module is being designed to maintain the production system by reporting software bugs and identifying software error sources. The last two modules to be implemented are the OLDMEDLINE Citation Record Creation module and the OLDMEDLINE Citation Record Reconciliation module.

Since NLM possesses a complete set of QCIM volumes on microfilm, we are investigating the possibility of getting the microfilm scanned, as an alternative to scanning the paper volumes. This approach is expected to be faster, but whether the OCR performance from scanned microfilm is

equivalent to converting from paper is unknown. A test is planned to answer this question since the quality of OCR output is important for the segmentation and labeling stages. If the test reveals equivalent OCR performance, the first stage in the ACORN process will be to import TIFF files from the microfilm scanning operation rather than from a paper scanner.

*Ground truth data for document image analysis*. By the end of September 2005, the Medical Article Records Groundtruth (MARG) database had 6364 unique IP visits from 92 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 goal of this project is to bring the *high resolution* Visible Human images to the lay public both directly as well as by linking text documents received from Web sources to relevant anatomic objects. This is achieved by two systems: AnatQuest and TILE.

AnatQuest is a Web-mediated system designed to provide widespread access to the Visible Human images for a broad range of users, including the lay public frequently limited to low speed Internet connections. This system is based on a 3-tier architecture in which the first tier consists of Java applets for displaying thumbnails of the cross-section, sagittal and coronal images of the Visible Human Male, from which detailed (full-resolution) views are accessed. The second tier is a set of servlets that process user requests and compress the requested images prior to shipment back to the user. The third tier is the object-oriented database of high resolution VH images and rendered 3D anatomic objects. Low bandwidth connections are accommodated by a combination of adjustable viewing areas and image compression done on the fly as images are requested. Users may zoom and navigate through the images. Since its release in 2003, it has averaged about 60,000 hits per month, about 5 times the number of hits for the AnatLine system developed in an earlier project intended mainly for the scientific and visualization communities.

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

TILE (*Text to Image Linking Engine*) is designed to transparently link the print library of functional-physiological knowledge with the image library of structural-anatomic knowledge into a single, unified resource for health information, a long term NLM goal. We interpret this goal as adding value to text resources such as PubMED and MedlinePlus by linking to anatomic images. A modular prototype TILE system is being developed to serve as a testbed to investigate the alternatives in the functions needed to accomplish this linkage. These functions are: identifying biomedical terms in a document, identifying the relevant anatomical terms, identifying the images in the image database, and linking the identified terms to the images. For the first (Document Analyzer) function, the Lister Hill Center's MetaMap system, specifically its

Java implementation (MMTx), is employed to analyze document text to identify biomedical terms.

Currently, 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, with the objective of identifying the most relevant anatomic concepts. 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.

Recently, steps were taken to expand the scope of documents that may be processed by TILE through the image-label based ranking method. This is necessary because the image database contains only about 205 rendered images of the Visible Human, and that too only from the thorax region, thereby restricting the range of documents which can be processed by TILE. To overcome this limitation we populated the database with *surrogate image records*, one for every anatomical structure in the body, for a total of 24,701. Each record contains an image title which is the name of the structure and a set of image labels corresponding to the names of substructures. The records are generated from information in UMLS Metathesaurus files. In addition, each record contains a pointer to its image if one exists. Of the 24,701 records, only 205 point to actual images, the rest pointing to null images. These surrogate image records facilitate the testing of the TILE system by identifying the most relevant image *names* for each document.

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 need a tool a content expert can use to validate the relevance of the anatomic object retrieved to the text of the document. Toward this end a Web-based research

tool is being developed that: (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 input data validating relevance of the anatomic term (or image) to the biomedical term(s) in the document.

## Information Systems

### DocView Project: Document imaging for the biomedical end-user

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

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

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

*MyDelivery*. Seen as a successor to DocView, the goal of the client/server MyDelivery communications system is to enable reliable and secure delivery of very large (gigabyte-sized) files, and large numbers (hundreds) of attachments in a single delivery, especially over unreliable links such as wireless networks. Potential users are medical researchers, clinicians, administrators, librarians and many other health professionals who need to securely exchange electronic medical information residing in a wide variety of file sizes. Examples of large files include document images, digitized color photographs, digitized x-rays, FDA drug applications, and clinical images such as PET scans, MRI scans, CT scans, sonograms, and digital videos.

Recent developments have been in six critical areas in the design of MyDelivery: Log Shipping, Network Load Balancing, operation of the client behind proxy servers, operation of the client on all target operating systems, user interface, and managing the client tendency to hog CPU cycles and bandwidth. A brief discussion follows:

1. Log shipping. To enable manual recovery in case of failure of the primary SQL server of the two that are part of the system, a procedure was created for setting up log shipping, which includes establishing a Monitor server on a third computer to conduct the log shipping. Log

shipping allows MyDelivery to have two database servers, one of which is a backup to the primary server. Through log shipping, the secondary database server will mirror the primary server every 10 minutes. If we should lose the primary server, we have established and tested a procedure for manually switching operation over to the secondary server, and making it the primary server. Our procedure includes a way of reversing the roles back to the original primary server, once it is restored to service.

2. Network load balancing (NLB). This technique distributes incoming user requests across servers in each of the two server clusters, Status Server and Document Exchange Server, allowing the system to handle communications evenly across all servers without overloading any one. NLB also provides a clean path for system expansion in the future. The system has been successfully tested using this technique.
3. Operation behind a proxy server. Many potential MyDelivery users are expected to have proxy servers in their organizations. These proxy servers often serve as parts of firewalls. The client has been designed (and successfully tested) to automatically detect a proxy server, and obtain from the user an ID and password for using the proxy server for all network communication.
4. Cross platform capability. The MyDelivery client has been designed and successfully tested to install and run on all target Windows operating systems: Windows 98, Windows ME, Windows NT Workstation, Windows NT Server, Windows 2000 Professional, Windows 2000 Server, Windows XP Professional, Windows XP Home Edition, Windows XP Media Center Edition, and Windows Server 2003. An attempt to install the beta version of Windows Vista on a computer was unsuccessful, but this will be a target platform for testing in the future when Microsoft releases a more reliable beta edition.
5. User interface. The user interface was improved to maintain positions of all four sub-windows in the client interface from one user session to the next.
6. Performance. When the client sends or receives large files, there are potentially two undesirable effects: it slows down both the CPU and data communication for other processes. A technique being developed to address this problem involves detecting computer usage via keyboard and mouse and governs the speed of the client in both processing and communicating.

*DocMorph and MyMorph.* The DocMorph system continued to serve both browser-based users (over 12,500 to date: 2000 more than last year) and MyMorph users (over 5,200 users) this year. Of the more than 12,000 registered users, many are biomedical document delivery librarians. DocMorph allows the conversion of more than 50 different file formats to PDF, for instance, to enable multi-platform delivery of documents. Also, by combining OCR with speech synthesis, DocMorph enables the visually impaired to use library information. It has been used by librarians for the blind and physically handicapped to convert documents to synthetic speech recorded onto audio tapes for blind patrons. Most users continue to use it to convert files to PDF to enable multi-platform delivery of documents. DocMorph is available at http://docmorph.nlm.nih.gov/docmorph.

Recently two long standing problems were solved in the DocMorph project. The first problem was the inability to run when directly installed on a Windows 2000 server. A workaround was to install Windows NT Workstation, upgrading to Windows NT Server, and finally upgrading again to

Windows 2000 Server. This proved to be a very time-consuming process, taking up to two days to create a single server. Because of limitations on disk partition size of Windows NT, this procedure frequently led to low free disk space on the computer's c: drive that is a consequence of adding security patches to the operating system. A technique was developed that involves relaxing security privileges on the folder containing the Access database, and permitting the system to run both ASP and SOAP3. This investigative work will facilitate the easy migration of the DocMorph production system to faster platforms running a future operating system. For example, when Microsoft releases IIS version 7 and Windows Vista next year, our new procedure will allow the DocMorph server to be readily upgraded to this environment.

The second problem was that when the DocMorph Web site was modified two years ago to be consistent with other NLM Web sites, a new system template was integrated into 35 static Web pages. This left over 70 dynamically generated screens to appear different from this standard template. To provide a consistent look and feel for users moving from pages using the template to those that did not, a major rewrite was done of the Visual Basic code that generates all dynamic web pages to include the template. After testing this upgrade, this code has been successfully moved to the production system.

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

**MEDLINE Database on Tap (MDoT)**

This project, known earlier as PubMed on Tap, was presented to the Board of Scientific Counselors in September 2005 as one that seeks to discover and implement systems and techniques to assist mobile clinicians in quickly finding relevant, high quality information addressing clinical questions that arise at the point of care. An objective is to understand how to display data so that users can quickly find the most pertinent information, while limited by the small screen and restricted bandwidth of handheld computers. Presented were techniques for display and navigation, as well as those for information organization. In addition, the prototype MDoT system was demonstrated.

As our primary method of discovery, we have developed a system that supports MEDLINE search and retrieval from a wireless, Internet-connected PDA. PDA client software for both Palm OS and Pocket PC OS have been developed and are freely available. Also, the MDoT Web site

continues to be accessed at the rate of about 5,000 hits every month. This Web site provides information about the project as well as the software, and allows us to solicit feedback from our users and monitor aggregate user behavior.

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

The Board 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. They cited as an example the use of Essie. The Board noted that relevant research efforts (e.g., the clustering by strength of evidence based on EBM recommendations as implemented in the strength of evidence taxonomy, or SORT) have added another level of research interest and sophistication to a set of already complex interface design issues and project challenges.

The Board also noted that the delivery of clinical information via handheld devices can be seen not only as an end in itself but also as a means for exploring new and more flexible and intuitive approaches to searching, summarization and presentation of bibliographic information that may eventually feed back into the re-design of MEDLINE and other NLM resources.

The recommendations of the Board included:

1. In the utility evaluation of MDoT include a study of its usefulness not only for physician decision support, but also its role in support of decision making by other health professionals, including nurses and pharmacists.
2. The work on summarization and efficient presentation of clinical findings is very important and could be of value to other NLM products (e.g., improved representation of clinical content in the MEDLINE database itself.)
3. In light of differential quality of abstracts it would be helpful to consider extraction of clinical findings from available full-text documents (methods and outcomes sections). At a minimum, a study that compares the accuracy of abstracts with the full-text document would be useful.

4. MDoT would be a great tool for teachers and teaching and more emphasis could be placed on evaluating it in a variety of teaching settings including problem-based curriculum contexts.
5. Another aspect of evaluation that needs to be undertaken is to better understand the reality of the user community (ies) – e.g., the information needs and strategies currently used. Segmenting these studies by health professional type – physician, nurse or pharmacist – would provide critical insights into design and implementation issues and drive development of the product.

**Interactive Publications Research**

A project was started this year to create a comprehensive, self-contained and platform-independent multimedia document that is an "interactive publication." Following a study of existing open source formats and standards, a prototype document was created containing many media objects: text, dynamic tables and graphs, a microscopy video of cell evolution, an animated spine in Flash, digital x-rays, and clinical images (CT, MRI, ultrasound) following the DICOM standard. Both self-contained (embedded) and folder-type (linked) documents using all these media types were created in four formats: MS Word, Flash, HTML and PDF. A comparison of these in terms of ease of use and development effort was done.

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

This project was presented to the Board of Scientific Counselors in September 2005, including a demonstration of our 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. In light of the large sizes of such publications possibly in the range of hundreds of megabytes, current work proceeds toward identifying techniques and protocols for rapid progressive download of the publications.

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. The Board had the following observations and recommendations:

1. Work needs to be undertaken to ensure that the MEDLINE citation includes the complete record of all the components of published articles including the so-called "supplementary material."
2. Industry is addressing some of the issues regarding IP and such work could be used to advantage.
3. A survey needs to be conducted with authors as to what tools will be needed to look at multimedia documents, e.g., the tool set could be expanded to include data visualization

tools such as generating scatter plots, scaling plots, etc. A sample population for such a survey could be authors within NLM.

4. A survey needs to be conducted with publishers as to what incentives need to be provided to authors for providing the underlying data, e.g., can it become a requirement for publication?
5. The ubiquity of the PDF format as an existing standard for most journal publications must be considered. Any new IP framework developed by the NLM needs to utilize PDF as the primary document format.
6. An assessment needs to be made of the cost of publication from the perspective of the publisher particularly since storage and transmission costs of large files, e.g., hundreds of megabytes, will be involved.
7. Transmission of huge files to the reader would also necessitate appropriate methods for analyzing IP for progressive transmission.
8. The long-term archiving of IP materials is clearly a key aspect of this work and should remain an important focus. Some of the software that exists today may not be present in the long-term.

**Digital Preservation Research**

This project falls into two broad categories, one concerned with the preservation of documents and the other with video. In each area we focus on some of the key functions of an economical and robust digital preservation system. Two in particular are automated metadata extraction and file migration.
Extracting metadata automatically from the contents of material that need to be preserved, rather than relying on manual entry, is probably the only way large collections can be economically preserved. Techniques are also being developed that automate the migration of files in bulk. This is important for the conversion of files in formats that face obsolescence, largely because they are no longer supported by newer software and modern computers, and will be inaccessible as time passes.

For document preservation, a detailed design was done and a prototype *System for Preservation of Electronic Resources* or SPER was developed. SPER is a flexible, modular system that demonstrates key functions such as ingest, automated metadata extraction (AME) and bulk file migration. AME is implemented for the extraction of descriptive metadata from scanned and online journal articles as well as NLM's obsolete Web pages. The module for metadata extraction from Web pages is internal to SPER while those for TIFF and online journals are implemented via a SOAP interface, so that SPER can access a remote AME system and retrieve extracted metadata in XML format as defined in the METS standard. In addition, a module for the extraction of technical metadata from TIFF file headers is implemented to include many of the items listed in the NISO Z39.87 standard for digital still images.

For the necessary infrastructure capabilities in SPER we have incorporated into the system the latest version (1.2) 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.

The experience of designing and developing the SPER prototype resulted in a 37-page design plan for an NLM Digital Preservation System. This plan, distributed to collaborators in other NLM divisions, is for the creation of an archive to preserve NLM's digital resources, both those born-digital as well as those digitized. Among such items are NLM's Web pages retired from active service but needed as a historic record, and scanned document images ('master files') from which lower quality surrogates have been created for efficient public Web access. The latter are part of collections such as Profiles in Science as well as others in HMD's Digital Manuscripts Program. An estimate of the types and extent of digital resources to be preserved at NLM was made on the basis of a survey conducted by NLM's Preservation staff.

The plan gives the design approach and describes tools used in the creation of the essential functions including ingest, automated metadata extraction from the items as well as from databases containing metadata already created manually, and bulk file migration. The plan also addresses the physical storage requirements, focusing on available hardware specifications, cost and dual storage to assure long term availability.

In summer of 2005, an investigation was begun to address the preservation of a new collection at NLM consisting of historical Food and Drug Administration court records. Our focus is on the extraction of descriptive metadata from these records. In collaboration with the curator for this collection, we identified more than a dozen metadata items which might be extracted automatically. Our approach includes auto-zoning 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.

The SPER prototype was modified for this collection to include the automated metadata extraction function, but not the archiving functions since these are to be handled by other systems at NLM.

As an initial test, about 170 pages consisting of about 90 FDA court cases were collected, scanned, and groundtruthed for training our algorithms. An OCR tool was used to generate text-lines each of which yielded fourteen features. The SVM classifier was used to classify the text-lines into different metadata labels with promising experimental results. Dictionaries have been created for the metadata items that are embedded in free text. A feature selection algorithm called floating search algorithm has been downloaded, modified, and tested. The Java implementation of the Stolcke-Earley parsing algorithm has been downloaded. The relevant Java code is being integrated into a library and will be incorporated into the SPER server.

Presentations of ongoing research in digital preservation were made to staff in 2005 focusing on overall SPER design, an end-to-end design of the automated metadata extraction subsystem based on learning methods, and experimental results. Also, papers describing SPER design as well as bulk file migration were published and presented at SPIE and IS&T conferences.

Research into video preservation focused on identifying an open file format such as Motion JPEG 2000 (MJ2) for archiving digitized video on disk media. Toward this end, a one-day invitational meeting was organized with about 50 archivists and technologists involved in the long term preservation of video and film. Participants considered the potential of lossless, on-disk video storage in light of the "twilight of tape" as a cost-effective storage medium. Barriers to adopting lossless algorithms were identified, and specific directions to overcome them suggested. Also discussed were current video metadata standards, and recent work in automatic extraction of metadata from video. Highlights of the meeting included:

- The first public demonstration by Media Matters, Inc. of real-time, full-screen, mathematically-lossless video compression and decompression based on the Motion JPEG 2000 standard.
- Initial work in using this capability at Yale University.
- A description of current lossy digital archival workflows at New York University and PBS.
- A review of the advantages of scalable JPEG 2000 image compression, as well as the color spaces needed for quality masters by Xerox.
- A survey of the dominant metadata methodologies, such as MPEG-7, by Rutgers which conducts the Moving Images Collection project.
- A demonstration of recent "Informedia" work in automatic extraction of metadata from video, applicable to institutionalized patient management by Carnegie Mellon.
- CEB's contributions to the Open JPEG project, and discussion of the metadata capabilities and shortcomings of Motion JPEG 2000 files.

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

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

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

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.

Another improvement exploits the characteristics of the wireframe model. This model has node attributes that can be adjusted, allowing different rates and styles of curvature to be expressed during page flipping. For example, there is a choice of three flip behaviors, depending on where the finger is placed to start flipping a page. E.g., for a finger at the top right of a page, just this corner would fold over while the page is turned over to the left, providing a more natural feel to the act.

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

Our Web version of TTP (*TTP Online*) required several changes 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 than Director.

In this Web version of TTP, each image of the page animation is imported into Flash and compressed to 15% quality (a compression ration 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 get around 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. This keeps 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. However, the page images from Vesalius and images from other sources (e.g., rendered Visible Human images, pictures of Italian cities, etc.) were interlinked 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 the steps required to create it. In addition to using their own resources and knowledge to accomplish the scanning, image enhancement and 3D modeling, they needed our templates for the final stage (to produce the software providing the interactivity). At our suggestion, they selected books in the life sciences: one by a Dutch surgeon who spent a decade with pirates in the Caribbean (1678); the other an encyclopedia of flora and fauna in the New World (1635). The first book was shown at the opening of the Kislak Collection, an event at the Library of Congress in April 2005.

## 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. 1,000 of the images are available in TIFF 8-bit format, for wide compatibility with image display and processing software.

Currently, 60,000 images of the uterine cervix from a large National Cancer Institute (Guanacaste) study are being scanned for Web distribution. In addition to these are pap smear and histology images, also from this study.

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

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

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 spell-check by pattern recognition techniques, and other key elements of MARS. These techniques are fundamental to the automated extraction of descriptive metadata for the long term preservation of document images. Besides real time performance data, also collected and archived are large numbers of bitmapped document images, zoned images, labeled zones, and corresponding OCR output data. This collection serves as ground truth data for research in document image analysis and understanding.

*Ground truth data for document image analysis*

For research in document image analysis and understanding techniques by the computer science and informatics communities, we provide a database named Medical Article Records Groundtruth (MARG). The data consists of over 1,000 bitmapped images of the first pages of articles from biomedical journals indexed in MEDLINE falling into 9 layout types encountered in MARS production. Included in addition to the page images are the corresponding segmented and labeled zones, OCR-converted and operator-verified data at the zone, line, word and character levels, all in XML format. Also available from this Web site (marg.nlm.nih.gov) is Rover, an analytic tool that may be used to compare the results of a researcher's program with the ground truth data. Rover has been enhanced to allow a visual comparison of researchers' algorithmic results with the ground truth data, as well as some statistical metrics. The MARG server has had over 6,300 unique IP visits from 92 countries.