A. T. McCray

National Library of Medicine
Bethesda, Maryland
USA

Research and Education

Informatics Research, Development, and Training at the Lister Hill National Center for Biomedical Communications

The Lister Hill National Center for Biomedical Communications is a research and development division of the United States National Library of Medicine (NLM). The Center conducts and supports research, develops research tools and systems, and provides training opportunities to individuals at various stages of their careers. The Center has been in existence since 1968, when it was established by a joint resolution of the United States Congress, with the mandate to conduct research supporting the mission of the NLM. The Center's research programs are reviewed biannually by a Board of Scientific Counselors, an external advisory group of researchers from the informatics community. The most current information about Lister Hill Center research activities can be found at http://lhncbc.nlm.nih.gov/.

The Center's research staff are drawn from a variety of disciplines, including medicine, computer science, library and information science, linguistics, engineering, and education. Research projects are generally conducted by teams of individuals of varying backgrounds and often involve collaboration with other divisions of the NLM, other institutes at the National Institutes of Health (NIH), and other academic partners. Center staff publish in the medical informatics, computer and information science, and engineering communities (See [1-40] for a sample of recent publications by Center staff). The Center is often visited by researchers from academic centers around the world, and our ongoing lecture series features presentations from many invited outside speakers.

Lister Hill Center research activities fall into several broad categories, and each of these is discussed in turn below. Our training program has grown significantly in the last few years and has brought many talented individuals to the Center to learn from and collaborate with our research staff. Our language and knowledge processing research involves basic research in medical language processing and medical knowledge representation, and image processing research involves the development of algorithms and methods to effectively process biomedical images of all types. We have developed and continue to support a number of information systems, all of which are informed by our basic research activities. In addition, Lister Hill Center staff are involved in a number of activities that define and support the Research infrastructure for next generation information systems.

Training Opportunities at the Lister Hill Center

The Lister Hill Center provides training and mentorship for individuals at various stages in their careers. Fellowship programs may be as short as eight weeks or as long as one year, with possible renewal for a second year. Each fellow is matched with a mentor from the research staff who works closely with the fellow throughout the fellowship program. In all cases, fellows define a research project early in their stay and then give a mid-term progress report. At the end of the fellowship period, fellows prepare a final, often publishable, paper and make a formal presentation which is open to all interested members of the NLM and NIH community.

This past year, we provided training to 46 participants from 16 states and 9 countries. The participants included nine undergraduate students, 15 graduate or medical students, 15 postdoctoral or post-MD fellows, and seven visiting faculty scholars. Participants worked on projects in the areas of biomedical knowledge discovery, consumer health systems,

In: Haux R, Kulikowski C (eds.) Yearbook of Medical Informatics 03: Quality of Health Care,
The Role of Informatics. Stuttgart: Schattauer Verlag, 2003; 193-203.

history of medicine, image processing, information retrieval research, just-in-time systems, knowledge based research, natural language processing, ontology research, palm technology, semantic web research, text mining, distance education, and visualization.

We again offered the Clinical Elective in Medical Informatics for third and fourth year medical students in March and April, and we continue to participate in programs supporting minority students including the Hispanic Association of Colleges and Universities (HACU) and the National Association for Equal Opportunity in Higher Education (NAFEO) summer internship programs.

In the summer of 2001 we initiated the NLM Rotation Program. This program provides an opportunity for trainees in NLM supported Medical Informatics Training programs to spend eight weeks at the Lister Hill Center learning about our programs and collaborating with our research scientists. Trainees from any NLM sponsored training program are eligible. The rotation includes a series of lectures and the opportunity for trainees to work closely with established scientists conducting research at the Center. Trainees who participated in the summers of 2001 and 2002 were members of informatics programs at Columbia University, Duke University, Oregon Health and Science University, University of Pittsburgh, University of Minnesota, University of Missouri, and University of North Carolina.

Additional information about our training and visiting faculty programs is available at our web site (http://lhncbc.nlm.nih.gov/) under “Training Opportunities”. Interested individuals will find descriptions of each of the training programs including specific application procedures.

Language and Knowledge Processing

Natural Language Processing Research

The Natural Language Systems research team investigates the contributions that natural language processing techniques can make to the task of mediating between the language of users and the language of online biomedical information resources. The successful integration of these techniques with other information retrieval strategies has the potential of contributing to the resolution of some of the most difficult problems underlying biomedical information management.

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

The Lexical Systems project builds and maintains the SPECIALIST lexicon, a large syntactic lexicon of medical and general English that is released annually with the UMLS Knowledge Sources. New lexical items are continually added using a lexicon building tool, and the lexicon currently contains over 180,000 lexical items. Lexical access tools, including LVG, a set of algorithms for morphological variation in English, are also distributed with the UMLS. Several additional modules have recently been completed and are now available independently as tools for a variety of natural language processing projects. These include a tokenizer, a lexical look-up utility, and a noun phrase extractor.

Innovative methods for providing more effective access to biomedical information depend on reliable representation of the knowledge contained in text. The Semantic Knowledge Representation project develops programs that extract usable semantic information from biomedical text by building on existing NLM resources, including the UMLS knowledge sources and the natural language processing tools provided by the SPECIALIST system. Two programs in particular, MetaMap and SemRep, are being developed, enhanced, and applied to a variety of problems in biomedical informatics. MetaMap maps noun phrases in free text to concepts in the UMLS Metathesaurus®, while SemRep uses the UMLS Semantic Network to determine the relationships asserted between those concepts. Project resources are being applied in a variety of research initiatives aimed at identifying specific biomedical information in MEDLINE® citations, including semantic predications asserting a treatment relationship between drugs and diseases. Several projects focus on molecular biology. One seeks to identify genes, gene products, and gene functions in abstracts and compares this information to that found in the Gene Ontology, and another supports comparison of protein function by identifying protein-protein interactions in text.

The Indexing Initiative project explores concept-based indexing methods. Project members have developed a system, Medical Text Indexer (MTI), that is being applied to automated and semi-automated indexing environments at the NLM. We recently conducted experiments to evaluate the effectiveness of MTI terms for NLM indexers, and these recommendations are now available to all indexers. In addition, results of the MTI system have been assigned as keywords for collections of meetings abstracts in a fully automated mode.

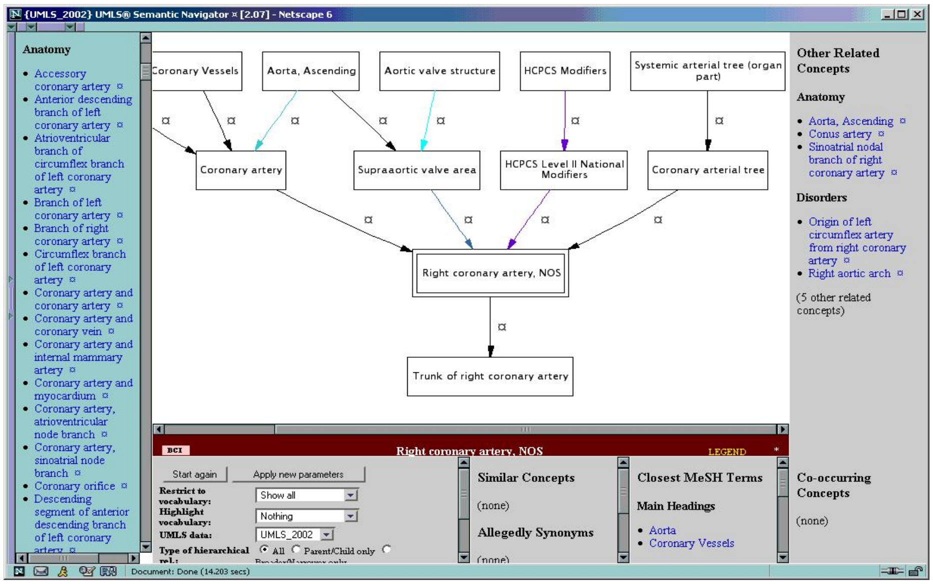
As part of the Indexing Initiative we have also conducted experiments using journal descriptors corresponding to biomedical specialties that are used to index journal titles. The goal of the work is to contribute to the reduction of ambiguity inherent in biomedical language. We have designed an experimental system that associates journal descriptors with words in titles and abstracts in a training set of about 435,000 MEDLINE records. The more words in the document co-occurring with a particular descriptor, the more likely it is that the document falls into that particular biomedical specialty.

Unified Medical Language System®

NLM regularly distributes a set of Unified Medical Language System (UMLS) knowledge sources to the research community. These include the Metathesaurus, Semantic Network and SPECIALIST lexicon. The Metathesaurus is a knowledge source representing multiple biomedical vocabularies organized as concepts in a common format. It thus provides a rich terminology resource in which terms and vocabularies are linked by meaning (See Figure 1).

Vocabularies proposed as standards by the Department of Health and Human Services in the rulemaking accompanying the Health Insurance Portability and Accountability Act (HIPAA) continue to be added and maintained in each release. The most recent release contains over two million names for almost 900,000 biomedical concepts in approximately 60 families of vocabularies or thesauri.

For a decade, the Metathesaurus had been released annually early in the year. Beginning with calendar year 2002, the Metathesaurus is being released quarterly. The systems and software used to create, manage, edit, and release the Metathesaurus are increasingly sophisticated, more automated, and better documented. This makes it possible to continue the schedule of more frequent releases and helps ensure the timely addition of content as required to keep the Metathesaurus current.


{UMLS_2002} UMLS® Semantic Navigator ¤ [2.07] - Netscape 6
Coronary Vessels
Aorta, Ascending
Aortic valve structure
HCPCS Modifiers
Systemic arterial tree (organ part)
Coronary artery
Supraaortic valve area
HCPCS Level II National Modifiers
Coronary arterial tree
Right coronary artery, NOS
Trunk of right coronary artery
BCI
Right coronary artery, NOS
LEGEND
Start again
Apply new parameters
Document: Done (14.203 secs)

Anatomy

- Accessory coronary artery ¤
- Anterior descending branch of left coronary artery ¤
- Atrioventricular branch of circumflex branch of left coronary artery ¤
- Branch of left coronary artery ¤
- Branch of right coronary artery ¤
- Circumflex branch of left coronary artery ¤
- Coronary artery and coronary artery ¤
- Coronary artery and coronary vein ¤
- Coronary artery and internal mammary artery ¤
- Coronary artery and myocardium ¤
- Coronary artery, atrioventricular node branch ¤
- Coronary artery, sinoatrial node branch ¤
- Coronary orifice ¤
- Descending segment of anterior descending branch of left coronary artery ¤

Restrict to vocabulary: Show all
Highlight vocabulary: Nothing
UMLS data: UMLS_2002
Type of hierarchical rel.: ● All ○ Parent/Child only ○ Broader/Narrower only

Similar Concepts

(none)

Allegedly Synonyms

(none)

Closest MeSH Terms

Main Headings

- Aorta
- Coronary Vessels

Other Related Concepts

Anatomy

- Aorta, Ascending ¤
- Conus artery ¤
- Sinoatrial nodal branch of right coronary artery ¤

Disorders

- Origin of left circumflex artery from right coronary artery ¤
- Right aortic arch ¤

(5 other related concepts)

Co-occurring Concepts

(none)

Fig. 1 The Semanitc Navigator is a browser developed for visualizing and navigating biomedical concepts from the Unified Medical Language System. It displays the semantic space surrounding an arbitrary UMLS concept. Clicking on any concept dynamically generates a new display in which this concept becomes the center of the new semantic space. The concept entered above is “right coronary artery.”

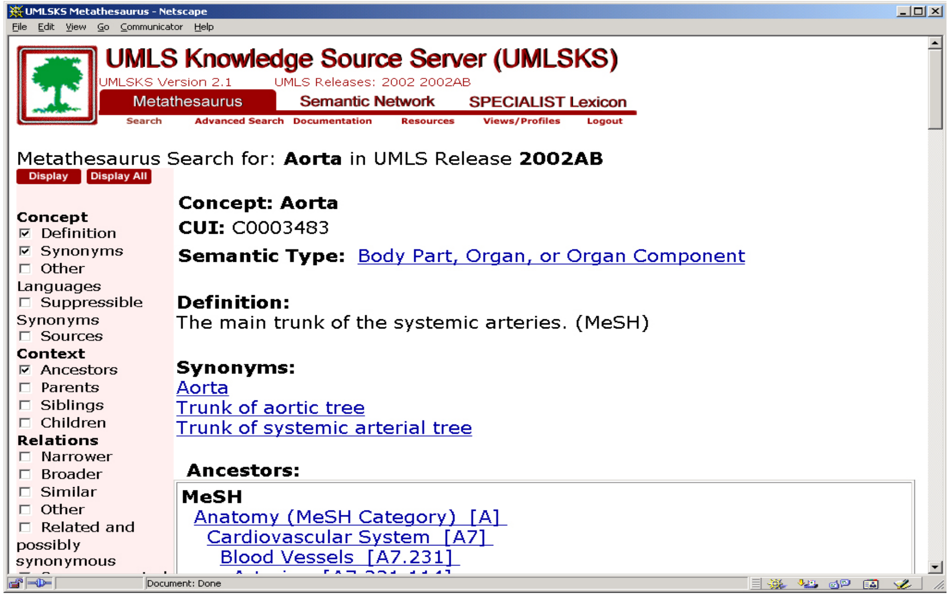
The UMLS Semantic Network provides a semantic framework for the Metathesaurus vocabularies. Each concept in the Metathesaurus is assigned to at least one semantic type, chosen from the set of 135 available types. There are semantic types for organisms, biologic and pathologic function, anatomy, chemicals and drugs, and concepts and ideas. More than 50 relationships bind the semantic types to each other. Thus, for example, a possible assertion might be “Disease or Syndrome” has_location “Anatomical Structure”. As we enhance the Metathesaurus with genomic terminology, we are also reviewing the Semantic Network for its coverage in this domain.

The UMLS data are made available over the Internet through the UMLS Knowledge Source Server, which provides direct access to each component of the UMLS
(See Figure 2).

Using the Knowledge Source Server, users can request information about a particular concept in the Metathesaurus, including definitions, semantic types, and synonyms as well as other concepts that are related to the input term. The Knowledge Source Server also accommodates navigation in the Semantic Network, allowing users to investigate relationships among semantic types to retrieve a list of Metathesaurus concepts assigned to a particular semantic type. Finally, the data in the SPECIALIST lexicon are also made available, providing the user with the syntactic and morphologic information about each lexica item the lexicon contains.

The most recent release of the Knowledge Source Server incorporates several features designed to enhance performance by allowing faster access to UMLS data, providing flexibility through a rich API set, and facilitating scalability in handling ever-increasing user loads and constituent vocabularies. The redesigned architecture includes a web server implemented as a collection of Java servlets that provide quick and easy access to UMLS data. In addition to enhancements to the user interface, XML has been incorporated into the design of the Knowledge Source Server to provide flexibility in delivering data to users. In order to

In: Haux R, Kulikowski C (eds.) Yearbook of Medical Informatics 03: Quality of Health Care,
The Role of Informatics. Stuttgart: Schattauer Verlag, 2003; 193-203.



UMLSKS Metathesaurus - Netscape

UMLS Knowledge Source Server (UMLSKS)

UMLSKS Version 2.1 UMLS Releases: 2002 2002AB

Metathesaurus Search for: **Aorta** in UMLS Release **2002AB**

Concept
☑ Definition
☑ Synonyms
□ Other Languages
□ Suppressible Synonyms
□ Sources
Context
☑ Ancestors
□ Parents
□ Siblings
□ Children
Relations
□ Narrower
□ Broader
□ Similar
□ Other
□ Related and possibly synonymous

Concept: Aorta
CUI: C0003483
Semantic Type: Body Part, Organ, or Organ Component

Definition:
The main trunk of the systemic arteries. (MeSH)

Synonyms:
Aorta
Trunk of aortic tree
Trunk of systemic arterial tree

Ancestors:

MeSH
Anatomy (MeSH Category) [A]
Cardiovascular System [A7]
Blood Vessels [A7.231]

Document: Done

Fig. 2 The UMLS Knowledge Source Server is a tool for providing Internet access to the UMLS Knowledge Sources. Its purpose is to make UMLS data more accessible to users, and in particular to system developers. This represents a view of the UMLS concept "aorta."

support the customization of UMLS terminologies for individual user needs, we are developing filters to help users select subsets of medical terms. Such tools will allow them to define terminologies relevant to their own domain and to tailor the UMLS data further for specialized needs.

While existing knowledge sources in the biomedical domain may be sufficient for information retrieval purposes, the organization of information in these resources is generally not suitable for reasoning. Automated inferencing requires the principled and consistent organization provided by ontologies. Our Medical Ontology Research project develops methods whereby ontologies can be acquired from existing resources and validated against other knowledge sources. Although the UMLS is used as the primary source of medical knowledge, OpenGALEN, CYC, and WordNet are being explored as well. Recent research has focused on two subdomains of biomedicine: anatomy and molecular biology. In one project, the representation of anatomical concepts in two ontologies, the Foundational Model of Anatomy and GALEN, were compared. In another project, we contributed to the integration of the Gene Ontology in the UMLS by studying the properties of this ontology.

Image Processing

Visible Human Project®

The Visible Human project data sets are designed to serve as a common reference for the study of human anatomy, as a set of common public domain data for testing medical imaging algorithms, and as a test bed and model for the construction of image libraries that can be accessed through networks. The Visible Human data sets are distributed to licensees over the Internet at no cost. The data sets are being applied to a wide range of educational, diagnostic, treatment planning, virtual reality, artistic, mathematical, legal and industrial uses. We continue to maintain two databases to record information about Visible Human project use. The first, to log information about Visible Human Project license holders and record their plans for using the images, and the second, to record information about the applications that are being developed.

With research support from the NLM, the University of Colorado Health Science Center, Center for Human Simulation has developed a first version of a head and neck atlas titled "Functional Anatomy of the Visible Human: Version 1.0 The Head and Neck". The atlas, based on the Visible Human data set, is designed in educational modules covering the topics of mastication, deglution, phonation, facial expression, extraocular motion, and hearing. QuickTime movies have been produced using live human subjects portraying the function of the regional anatomy described from a surface anatomy perspective. Tools include basic anatomic structure identification, a model builder, orthogonal plane browser, and links to the PubMed web site for automatic key word searches of the literature. A Visible Human project inspired initiative, the Insight Toolkit, began beta testing last year.

The toolkit makes available a variety of open source image processing algorithms for computing segmentation and registration on a variety of hardware platforms. Platforms currently supported are PCs Research and Education 197 Yearbook of Medical Informatics 2003 running Visual C++, Sun Workstations running the GNU C++ compiler, SGI workstations, Linux based systems and Mac OS-X. This work is being conducted by a consortium of university and commercial entities.

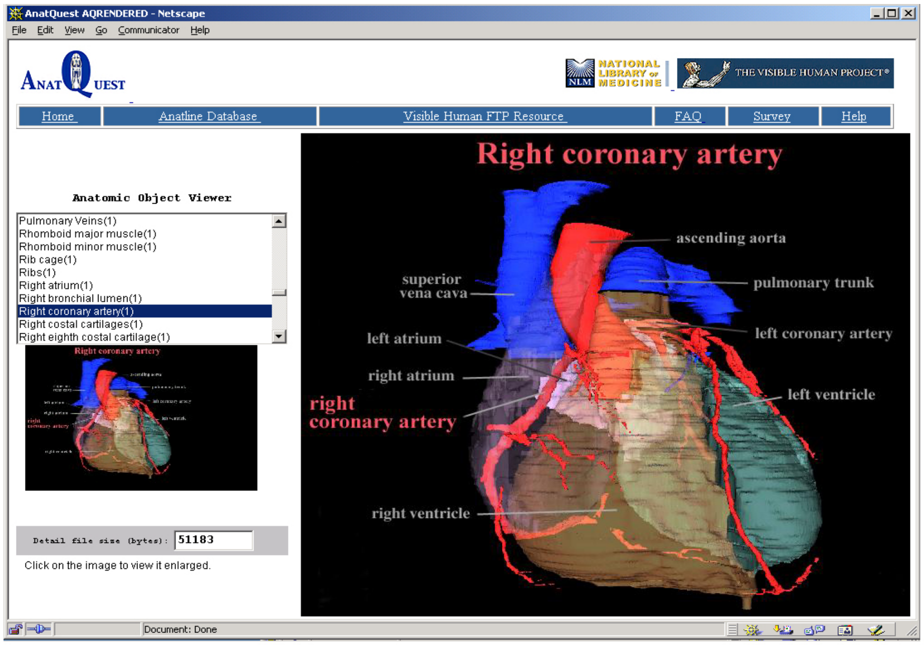
Building on the earlier AnatLine system, the object-oriented database of Visible Human images indexed for the male thorax region, Lister Hill Center researchers created AnatQuest with the goal of providing widespread access to the Visible Human images. AnatQuest offers users thumbnails of the cross-section, sagittal and coronal images of the Visible Male, from which detailed (full-resolution) views are accessed. 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 (See Figure 3).

In: Haux R, Kulikowski C (eds.) Yearbook of Medical Informatics 03: Quality of Health Care,
The Role of Informatics. Stuttgart: Schattauer Verlag, 2003; 193-203.

Medical Image Indexing and Retrieval

The Web-based Medical Information Retrieval System (WebMIRS) is an application that allows remote users to access x-ray and other data from two surveys conducted by the National Center for Health Statistics. These are the National Heath and Nutrition Examination Surveys (NHANES) II and III, carried out during the years 1976-1980 and 1988-1994, respectively. The project investigates fundamental questions that arise in the handling, organization, storage, access and transmission of very large x-ray images. WebMIRS allows a user to control a graphical user interface to construct a query for the data. A sample query might be equivalent to the English statements: “Find records for all individuals who reported chronic back pain. Return their age, sex, race, age when the pain began, and longest duration of pain. Also, return the record data required for statistical analysis and display their x-ray images.” WebMIRS allows the user to save the returned data to the local disk drive, where it may be analyzed with appropriate statistical tools.

The Content Based Image Retrieval project develops methods for effective extraction of biomedical information from digital images of the spine. This work has implications both for indexing of image data and for retrieval of those data. For example, for the NHANES II images, the only indexing data available is the collateral (alphanumeric) data collected in the questionnaires and examinations; no indexing information derived directly from the images is available, and the high cost of employing radiological experts to compile such data by physical viewing and interpreting each image makes it unlikely that such information will ever be acquired by purely manual means. These circumstances could be reversed if reliable, biomedically-validated software could produce image interpretations automatically, or even semiautomatically (See Figure 4).



Anatomic Object Viewer

Pulmonary Veins(1)
Rhomboid major muscle(1)
Rhomboid minor muscle(1)
Rib cage(1)
Ribs(1)
Right atrium(1)
Right bronchial lumen(1)
Right coronary artery(1)
Right costal cartilages(1)
Right eighth costal cartilage(1)

Detail file size (bytes): 51183

Click on the image to view it enlarged.

Document: Done

Fig 3. AnatQuest is a web-based interface for viewing high resolution Visible Human anatomical images. The system provides viewers for rendered anatomical organs and cut away slices of the body. AnatQuest allows browsing over images through an adjustable size viewing area, downloading portions of the 33-meg image slices to the browser as needed, thereby making it feasible to view the images.

Computer-assisted image searching is a potential enabler of enhanced information extraction from a database that has already been indexed. The most popularized form of this type of search is “query by example” or a variant, “query by sketch”. In query by example, the user inputs an image, perhaps by selecting from a set of choices provided by the system, or by providing a completely new image, and queries the database by asking, in effect, “Find records with images like this one”, usually with respect to one or more characteristics of the example image, such as shape, histogram, or texture. In query by sketch, the input image is replaced by a sketch by the user, using drawing tools provided by the system. In either case, the system analyzes the input into component features, then searches the images in the database for those with similar features.

Multimedia Research and Development

Our multimedia research and development efforts concentrate on the engineering of technical improvements applied to media issues such as image quality and resolution, color fidelity, transportability, storage, and visual information communication. In addition to the development by the staff of new methods and processes, the facilities and hardware infrastructure must reflect state-of-the-art standards in a very rapidly changing field. High definition video is a technology area being developed that represents the future for improved electronic image quality. Multimedia systems, scientific visualization and networked media are being pursued for the performance, educational, and economic advantages that they offer. Three dimensional computer graphics, animation techniques, and photorealistic rendering methods have changed the tools and products of the graphic artists in the Center. Digital video and image compression techniques are central to projects requiring storage of large images and rapid transmission.

One of our projects, the “Breath of Life Virtual Tour” was the centerpiece

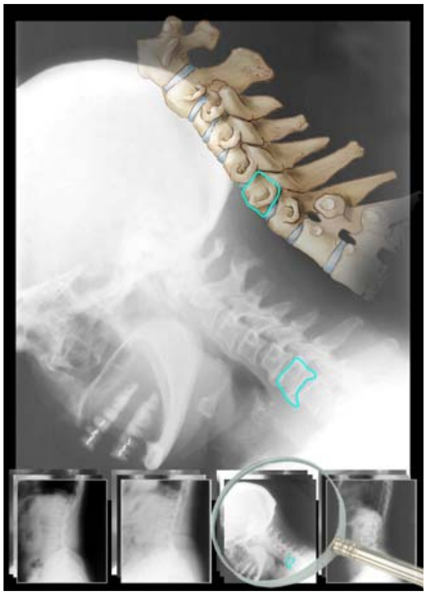


Fig. 4 Content-Based Image Retrieval of Biomedical Images comprises both indexing and retrieval. Indexing, the computer-assisted data reduction of images into mathematical features, may be subdivided into: segmentation, feature extraction, feature vector organization, and classification. Retrieving desired images from the databases comprises user query formulation, user query feature extraction, query search space strategy, and similarity matching method.

of the opening activities on World Asthma Day, and continued as part of program events on the CDC campus throughout that month. Another project, the "Movement Disorders" video database project is a collaborative project with the Yale University School of Medicine. This pilot effort established a digital video database of high quality, full-motion video clips of neurologically based movement disorders. The video database of patients with a variety of clinically diagnosed movement disorders, collected from the Yale University Movement Disorders Clinic underwent updated editing and compression to capitalize on advanced digitization and compression technology. "Expanding the Medical Universe", a new video shown daily in the NLM Visitors Center, is the first NLM video to be produced in the High Definition format. The new video also is presented in 'surround sound' and is delivered on a DVD which has both non-captioned and captioned versions.

Information Systems

Digital Library Research and Development

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

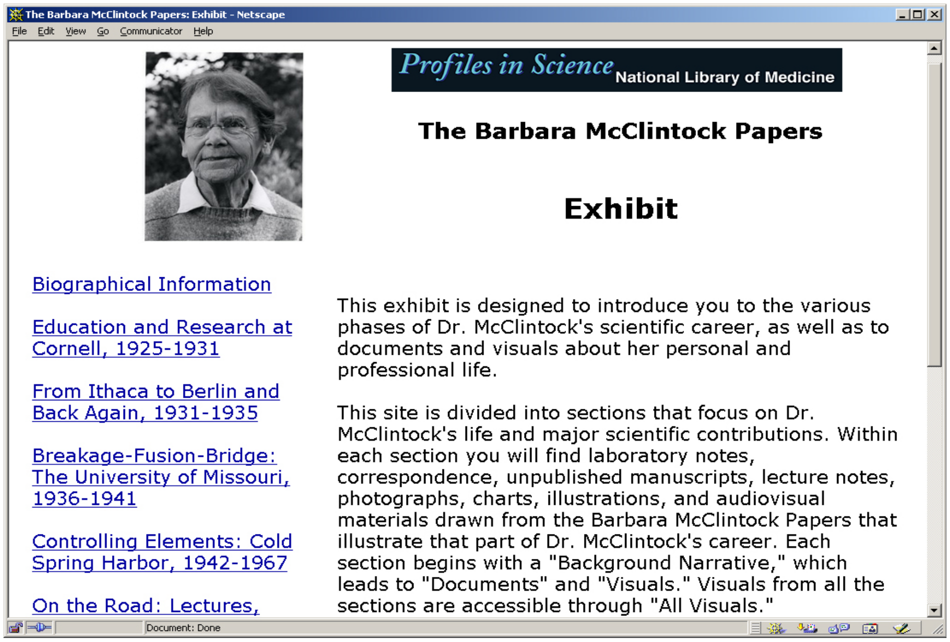
The *Profiles in Science®* web site makes the manuscript collections of prominent biomedical scientists available. The content of the database is created in collaboration with NLM's History of Medicine Division, which processes and stores the physical collections. The documents have been donated to NLM and contain published and unpublished materials, including books, journal volumes, pamphlets, diaries, letters, manuscripts, photographs, audio tapes and other audiovisual resources (See Figure 5).

Automatic data entry continues to be an active area of research and development in our digital library research program. MARS (Medical Article Records System) is a system that automates the production of MEDLINE records from biomedical journals. From bitmapped images of the first page of the articles, this system is designed to automatically extract the article title, author names, affiliations and the abstract. Our current research centers on the identification of rules for page segmentation, zone labeling, optical character recognition (OCR) error correction, and affiliation ranking. Operators enter fields (other than the ones automatically extracted), as well as perform text verification before the records are made available to indexers. The MARS system relies on image analysis and lexical analysis algorithms to correctly extract bibliographic data from images. These algorithms are based on rules constructed from features extracted from the layout geometry and OCR output.

The DocView project facilitates the delivery of library documents directly to the patron via the Internet in multiple ways. Once documents in bitmapped image form are received, the user may use DocView to retain them in electronic form, view the images, organize them into "folders" and "file cabinets", electronically bookmark selected pages, manipulate the images (zoom, pan, scroll), copy and paste images, and print them if desired. Users may receive document images either via Ariel FTP or Multipurpose Internet Mail Extensions protocols. DocMorph provides additional functionality for DocView users. DocMorph enables online users to convert files from one format to another for easier exchange or delivery.

NLM Gateway

The National Library of Medicine offers an increasing number of Internet-based information resources, each with its own user interface. We have created the NLM Gateway to let users initiate searches in multiple retrieval systems from a single web interface. The target audience for the new system is the Internet user who comes to NLM not knowing exactly what is available or how best to search for it. Results from the systems searched are presented in categories (for instance, journal article citations; books, serials and audiovisuals; consumer health information; meeting abstracts) rather than by database. The basic Gateway search interface is



The Barbara McClintock Papers: Exhibit - Netscape
File Edit View Go Communicator Help
Document: Done

Profiles in Science National Library of Medicine

The Barbara McClintock Papers

Exhibit

Biographical Information

Education and Research at Cornell, 1925-1931

From Ithaca to Berlin and Back Again, 1931-1935

Breakage-Fusion-Bridge: The University of Missouri, 1936-1941

Controlling Elements: Cold Spring Harbor, 1942-1967

On the Road: Lectures,

This exhibit is designed to introduce you to the various phases of Dr. McClintock's scientific career, as well as to documents and visuals about her personal and professional life.

This site is divided into sections that focus on Dr. McClintock's life and major scientific contributions. Within each section you will find laboratory notes, correspondence, unpublished manuscripts, lecture notes, photographs, charts, illustrations, and audiovisual materials drawn from the Barbara McClintock Papers that illustrate that part of Dr. McClintock's career. Each section begins with a "Background Narrative," which leads to "Documents" and "Visuals." Visuals from all the sections are accessible through "All Visuals."

Fig. 5 *Profiles in Science* is a digital library site that makes the archival collections of prominent twentieth century biomedical scientists available to the public through modern digital technology. This is the introductory exhibit page for the Barbara McClintock collection.

simple. It takes advantage of the capabilities of the retrieval systems it links to, for example, the advanced query parser and the “Related Articles” and “LinkOut” functions in MEDLINE/PubMed. Users may set preferences to adapt the interface to their needs, such as specifying which elements of a record they wish to see in the display of results.

Consumer Health Informatics Research

Consumer Health Informatics research projects explore the needs, information seeking behavior, and cognitive strategies of health care consumers. The goal is to use medical informatics and information technologies to study ways to develop, organize, integrate, and deliver accessible health information to members of the public. We are currently conducting research that investigates readability metrics and algorithms with the goal of providing health information to the public at all levels of health literacy.

ClinicalTrials.gov provides members of the public with comprehensive information about clinical trials. The site not only simplifies access to research protocols but also directs visitors to appropriate background information, such as health topics on MEDLINEplus® and the biomedical literature on PubMed (See Figure 6).

Currently, *ClinicalTrials.gov* has thousands of protocol records sponsored by the Federal government, the pharmaceutical industry, and nonprofit organizations in tens of thousands of locations, mainly in the United States and Canada. We recently introduced several new search features in *ClinicalTrials.gov* to help users find relevant studies more easily. For example, “Search Within Results” enables users to narrow their search results with additional criteria and “TryIt” automatically suggests alternative queries when no studies are found. In addition, following the release of the Food and Drug Administration’s “Guidance for Industry: Information Program on Clinical Trials for Serious or Life-Threatening Diseases and Conditions”, *ClinicalTrials.gov* now regularly receives protocol information from pharmaceutical industry sponsors.

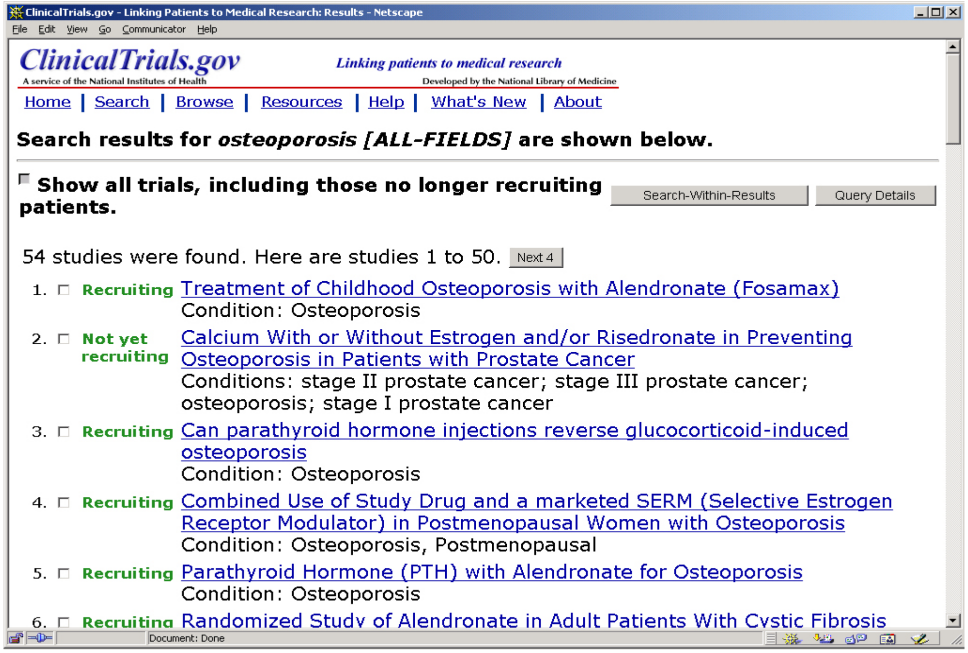
The Genetic Disease Home Reference is a new project that seeks to provide information about genes and diseases to members of the public. When completed, this resource will focus on diseases that are caused by single genes and, in turn, on the genes that cause these diseases. As knowledge of genetics expands, the interrelationships between genes and diseases will continue to unfold. Our goal is to provide a bridge between the clinical questions of the public and health professionals and the richness of data emanating from the Human Genome Project. Other resources will delve more deeply into the clinical aspects of the diseases and the details of the genes. This system is meant to serve as a guide into those other resources.

Research Infrastructure and Support

High Performance Computing and Communications

We are working to define and support Next Generation Internet (NGI) capabilities that will allow the NGI to be used routinely in health care, public health and health education, as well as biomedical, clinical and health services research. These capabilities include quality of service, security and medical data privacy, nomadic computing, network management, and infrastructure technology as a means for collaboration. In 1998, NLM began a three-phase effort to support NGI capabilities in health care. The goal of the project is to gain a better understanding of the impact of NGI on health care, health education, and health research in the areas of cost, quality, usability, efficacy, and security. Phase 1 was a planning effort, and Phase 2 supported the implementation of these plans within a limited geographic scope. Phase 3 was to be a test of scalability of applications to a national scope on public networks featuring quality of service. Such a network is, however, not available because of the rapid

In: Haux R, Kulikowski C (eds.) Yearbook of Medical Informatics 03: Quality of Health Care,
The Role of Informatics. Stuttgart: Schattauer Verlag, 2003; 193-203.


ClinicalTrials.gov - Linking Patients to Medical Research: Results - Netscape
File Edit View Go Communicator Help
ClinicalTrials.gov
A service of the National Institutes of Health
Linking patients to medical research
Developed by the National Library of Medicine
Home | Search | Browse | Resources | Help | What's New | About
Search-Within-Results
Query Details
Next 4
Document: Done

Search results for *osteoporosis [ALL-FIELDS]* are shown below.

□ Show all trials, including those no longer recruiting patients.

54 studies were found. Here are studies 1 to 50.

1. □ Recruiting Treatment of Childhood Osteoporosis with Alendronate (Fosamax)
   Condition: Osteoporosis
2. □ Not yet recruiting Calcium With or Without Estrogen and/or Risedronate in Preventing Osteoporosis in Patients with Prostate Cancer
   Conditions: stage II prostate cancer; stage III prostate cancer; osteoporosis; stage I prostate cancer
3. □ Recruiting Can parathyroid hormone injections reverse glucocorticoid-induced osteoporosis
   Condition: Osteoporosis
4. □ Recruiting Combined Use of Study Drug and a marketed SERM (Selective Estrogen Receptor Modulator) in Postmenopausal Women with Osteoporosis
   Condition: Osteoporosis, Postmenopausal
5. □ Recruiting Parathyroid Hormone (PTH) with Alendronate for Osteoporosis
   Condition: Osteoporosis
6. □ Recruiting Randomized Study of Alendronate in Adult Patients With Cystic Fibrosis

Fig. 6 *ClinicalTrials.gov* provides easy access to information on clinical research studies for a wide range of diseases and conditions. Abstracts of the clinical study protocols include a summary of the purpose of the study, recruiting status, criteria for patient participation, location of the trial, and specific contact information. A list of trials for ”osteoporosis” is shown above.

increase in available bandwidth and a decrease in the price of that bandwidth. In lieu of Phase 3, we began a new initiative in Scalable Information Infrastructure. The purpose of this initiative is to encourage development of health related applications of scalable, wireless, geographic information systems, and identification technologies in a networked environment. The initiative focuses on situations that require or greatly benefit from the application of these technologies in health care, medical decision-making, public health, large-scale health emergencies, health education, and biomedical, clinical and health services research.

The Collaboratory for High Performance Computing and Communication investigates innovative means for assisting health science institutions in their use of online distance learning technologies and explores advanced computer and network technologies for distance interactivity, including wireless technology and virtual reality research. We recently conducted a series of experiments combining wireless and streaming video technology offsite to do live webcasts from NLM that originated at remote sites. Video encoding software on a laptop roaming the scientific poster session at the annual meeting of the American Medical Informatics Association (AMIA) sent the video stream through a wireless network installed onsite back to NLM for broadcast via the Collaboratory streaming server. We established a 802.11b wireless network at the AMIA NLM booth for the duration of the exhibits.

System Security and Advanced Networks

Our research and development activities depend on advanced network capabilities as well as state of the art security systems. Lister Hill Center staff collaborate with other NLM divisions in the development of access controls and security classifications. A secure subnets working group last year developed a classification of NLM systems used to categorize different levels of required network access between each system and the Internet. The first phase of the secure subnets initiative has been implemented, with most desk-based systems placed on subnets that are not accessible from outside NLM. These systems can themselves access sites outside NLM, but transmissions originating outside of NLM cannot access them. The effect of this grouping has been to make these systems less vulnerable to external security attacks. A new network performance testing system generates live traffic for analyzing and tuning the performance of the NLM network. A very high capacity tape library system based on LTO tape technology is used for backups of the Center’s computer systems. Most backup media are online at all times, available for restoration of older data and programs. Backup volumes are created in duplicate, with one set for offsite storage.

The Lister Hill Center is connected to two NGI networks, vBNS (very high speed Backbone Network Services) and Abilene, with connections to the Federal NGI network DREN, the Department of Defense Research Network. The Abilene network supports full IP (Internet Protocol) multicast. We use that mode to receive and transmit multicast voice and video sessions. We recently worked with the Uniformed Services University and its Medical Simulation Center in connecting to the Abilene network. Dark fiber was used to connect the institutions, and we arranged for connectivity to Abilene through the router at NLM. For a cross country test using large images, we conducted memory-to-memory tests between the Armed Forces Institute of Pathology in Maryland through NLM to the San Diego Supercomputing Center in California.

Organizational Structure

The Lister Hill Center has five major components, each of which is listed below, together with its Branch or Office Chief. Many of our research activities involve collaboration not only across Lister Hill Center branches, but also across divisions of the NLM, the NIH, federal agencies, and university departments.

Communications Engineering Branch (George Thoma, PhD)

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, archiving, automated production of MEDLINE records, Internet access to biomedical multimedia databases, and imaging applications in support of medical educational packages employing digitized radiographic, anatomic, and other imagery. In addition to applied research, the branch also develops and maintains operational systems for production of bibliographic records for NLM's flagship database, MEDLINE. Research areas include: content-based image indexing and retrieval of biomedical images, document image analysis and understanding, image compression, image enhancement, image feature identification and extraction, image segmentation, image retrieval by "query by image content", image transmission and video conferencing over networks implemented via asynchronous transfer mode and satellite technologies, optical character recognition and man-machine interface design applied to automated data entry. The branch also maintains archives of large numbers of digitized spine x-rays and bit-mapped document images that are used for intramural and outside research purposes.

The most current information about the Communications Engineering Branch can be found at http://lhncbc.nlm.nih.gov/ceb/.

Cognitive Science Branch (Alexa T. McCray, PhD)

The Cognitive Science Branch conducts research and development in computer and information technologies. Important research areas involve the investigation of a variety of techniques, including linguistic, statistical, and knowledge-based methods for improving access to biomedical information. Branch members actively participate in the Unified Medical Language System project and collaborate with other NLM research staff in the Indexing Initiative project, whose goal is to develop automated and semi-automated techniques for indexing the biomedical literature. The branch also conducts research in digital libraries and collaborates with NLM's History of Medicine Division on Profiles in Science, a project to digitize collections of prominent biomedical scientists. Several branch projects address the challenges involved in providing health information to consumers. *ClinicalTrials.gov* is a resource developed by the branch, and it affords an excellent testbed for conducting consumer health informatics research. The branch is currently developing a system designed to provide information about genes and diseases to the public.

The most current information about the Cognitive Science Branch can be found at http://lhncbc.nlm.nih.gov/cgsb.

Computer Science Branch (Lawrence C. Kingsland III, PhD)

The Computer Science Branch applies techniques of computer science and information science to problems in the representation, retrieval and manipulation of biomedical knowledge. Branch projects involve both basic and applied research in such areas as intelligent gateway systems for simultaneous searching in multiple databases, intelligent agent technology, knowledge management, the merging of thesauri and controlled vocabularies, data mining, and machine-assisted indexing for information classification and retrieval. Research issues include knowledge representation, knowledge base structure, knowledge acquisition, and the human-machine interface for complex systems. Important components of the research include embedded intelligence systems that combine local reasoning with access to large-scale online databanks. Computer Science Branch research staff include the teams that developed NLM's Gateway, Internet Grateful Med and HSTAT programs and the team that annually produces the UMLS Metathesaurus. Branch staff coordinate the NIH Clinical Elective in Medical Informatics for third and fourth year medical students.

The most current information about the Computer Science Branch can be found at http://lhncbc.nlm.nih.gov/csb/.

Audiovisual Program Development Branch (James Main)

The Audiovisual Program Development Branch supports the Lister Hill Center's research, development, and demonstration projects with high quality video, audio, imaging, and graphics materials. From initial project concept through project implementation and final evaluation, a variety of forms and formats of visuals are developed, and staff activities include content creation, editing, enhancement, transfer and display. Included in this effort is the production of a series of video modules, documenting the progress of Lister Hill Center research projects. Consultation and materials development are also provided by the branch for NLM's information programs. From applications of optical media technologies and teleconferencing to support for web design, the requirement for graphics, video, and audio materials has increased in quantity and diversified in format. Included within the branch is the Office of the Public Health Service Historian. This office provides information about the history of Federal efforts devoted to public health, preserves and interprets the history of PHS, and promotes historically oriented activities across the U.S. Department of Health and Human Services.

The most current information about the Audiovisual Program Development Branch can be found at http://lhncbc.nlm.nih.gov/apdb/

Office of High Performance Computing and Communications (Michael J. Ackerman, PhD)

The Office of High Performance Computing and Communications serves as the focal point for the NLM's High Performance Computing and Communications (HPCC) activities. It coordinates NLM's HPCC planning, research and development activities with federal, industrial, academic, and commercial organizations, and it collaborates with Lister Hill Center research branches and NLM Divisions in the development, operation, evaluation and demonstration of HPCC research programs and projects. In addition, it plans, coordinates, and administers the Inter-Agency HPCC research and development program. Office staff serve as NLM's liaison to scientific organizations at all levels of national, state and international government on planning and implementing research in HPCC. The major research activities of the office center around the Visible Human Project, NLM's Next Generation Internet Program, including telemedicine, the Collaboratory for High Performance Computing and Communications, and the 3D Informatics research program.

The most current information about the Office of High Performance Computing and Communications can be found at http://lhncbc.nlm.nih.gov/ohpcc/.

Conclusion

Lister Hill Center research and development is driven by the needs of an increasingly information based society. We work at the intersection of computer and information science and medicine, addressing the information problems faced by biomedical researchers, health care professionals, and, increasingly, members of the public. The health information needs of patients, families, and other members of the public bring new research challenges, especially as the complexity of biomedical knowledge continues to increase. We conduct basic and applied research to support open access to high quality biomedical information, and we develop information systems that are strongly informed by informatics principles and methods. As we look forward, we see a host of exciting research opportunities, and, through our training program, we continue to be committed to the development of the next generation of outstanding informatics researchers.

Acknowledgments

I would like to thank all Lister Hill Center research staff for their contributions to the content of this report.

References

1. Ackerman MJ, Banvard RA. Imaging outcomes from the National Library of Medicine's Visible Human Project. Comput Med Imag Graph. 2000; 24(3):125-6.

2. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc AMIA Symp. 2001; 17-21.

3. Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindflesch TC, Wilbur WJ. The NLM indexing initiative. Proc AMIA Symp. 2000; 17-21.

4. Bodenreider O, Rindflesch TC, Burgun A. Unsupervised corpus-based method for extending a biomedical terminology. Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, North American Chapter of the Association for Computational Linguistics. 2002; 53-60.

5. Bodenreider O. Experiences in visualizing and navigating biomedical ontologies and knowledge bases. Proceedings of the ISMB 2002 Bio-ontologies SIG meeting. 2002; 29-32.

6. Bodenreider O. Using UMLS semantics for classification purposes. Proc AMIA 2000; 86-90.

7. Burgun A, Bodenreider O. Aspects of the taxonomic relation in the biomedical domain. In: Welty C, Smith B (eds.). Collected papers from the Second International Conference on Formal Ontology in Information Systems. New York: ACM Press. 2001; 222-33.

8. Fan Y, Hwang K, Gill M, Huang HK. Some connectivity and security issues of NGI in medical imaging applications. Journal of High Speed Networks. 2000; 9:3-13.

9. Ford G, Hauser SE, Le DX, Thoma GR. Pattern matching techniques for correcting low confidence OCR words in a known context. Proc SPIE, Document Recognition and Retrieval VIII. 2001; 241-9.

10. Hauser SE, Le DX, Thoma GR. Automated zone correction in bitmapped document images. Proc SPIE, Document Recognition and Retrieval VII. 2000; 248-58.

11. Humphrey SM, Rindflesch TC, Aronson AR. Automatic indexing by discipline and high-level categories: methodology and potential applications. In: Soergel D, et al. editors. Advances in classification research. Proceedings of the 11th ASIST SIG/CR Classification Research Workshop; 2001.

12. Kim J, Le DX, Thoma GR. Automated labeling in document images. Proc SPIE, Vol 4307, Document Recognition and Retrieval VIII. 2001; 111-22.

13. Le DX, Straughan SR, Thoma GR. Greek alphabet recognition technique for biomedical documents. Proc 6th World Multiconference on Systemics, Cybernetics and Informatics, Vol. III. 2002; 86-91.

14. Le DX, Tran LQ, Chow J, Kim J, Hauser SE, Moon CW, Thoma GR. Automated medical records citation records creation for web-based online journals. Proc 14th IEEE Symposium on Computer-Based Medical Systems. 2001; 315-20.

15. Locatis, C. Instructional design and technology in healthcare. In Reiser R, Dempsey J (eds.) Trends and Issues in Instructional Design and Technology, Upper Saddle River, New Jersey: Merrill/Prentice-Hall. 2002; 225-38.

16. Long LR, Thoma GR. Identification and classification of spine vertebrae by automated methods. Proc SPIE Medical Imaging 2001: Image Processing. 2001; 1478-89.

17. Long LR, Thoma GR. Use of shape models to search digitized spine x-rays. Proc. IEEE Computer-Based Medical Systems. 2000; 255-60.

18. Long LR, Thoma GR. Landmarking and feature localization in spine x-rays. Journal of Electronic Imaging. 2001; 10(4):939-56.

19. Lu CJ, Bangalore A, Tse T. Developing web browser recording tools using server-side programming technology. In: Proceedings of WebNet 2000, World conference on WWW and Internet. Association for the Advancement of Computing in Education. 2000; 372-7.

In: Haux R, Kulikowski C (eds.) Yearbook of Medical Informatics 03: Quality of Health Care,
The Role of Informatics. Stuttgart: Schattauer Verlag, 2003; 193-203.

20. Marcelo AB, Fontelo PA. A pathology report metadata registry: Framework for semantic interoperability across disparate systems. Archives of Pathology and Laboratory Medicine. 2001; 125(8):1014-15.

21. McCray AT, Gallagher ME. Principles for digital library development. Communications of the ACM. 2001; 44(5):48-54.

22. McCray AT. Better access to information about clinical trials. Annals of Internal Medicine. 2000; 133(8):609-614.

23. McCray AT, Bodenreider O, Malley JD, Browne AC. Evaluating UMLS strings for natural language processing. Proc AMIA Symp. 2001; 448-52.

24. Nishinaga N, Tatsumi H, Gill M, Akashib A, Nogawa H, Reategui I. Trans-Pacific demonstration of Visible Human. Space Communications 2002; 17(4):303-11.

25. Parascandola, JA. From germs to genes: Trends in drug therapy, 1852-2002. Pharmacy in History 44. 2002; 3-11.

26. Parascandola J. The pharmaceutical sciences in America, 1852- 1902. Journal of the American Pharmaceutical Association. 2000; 40:733-735.

27. Pearson G, Moon CW. Bridging two biomedical journal databases with XML: A case study. Proc 14th IEEE Symposium on Computer-Based Medical Systems. 2001; 309-14.

28. Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pacific Symposium on Bio-computing. 2000; 517-28.

29. Rindflesch TC, Rajan JV, Hunter L. Extracting molecular binding relationships from biomedical text. Proceedings of the 6th Applied Natural Language Processing Conference. 2000; 188-95.

30. Rodgers RPC, Sherwin Z. A management system for network-sharable locally installed software: Merging RPM and the depot scheme under Solaris. LISA XV: Proceedings of the Fifteenth Systems Administration Conference. 2001; 267-72.

31. Thoma GR, Ford G. Automated data entry system: performance issues. Proc SPIE Document Recognition and Retrieval IX. 2002; 181-90.

32. Thoma GR, Ford G, Le DX, Li Z. Text verification in an automated system for the extraction of bibliographic data. Proc 5th International Workshop on Document Analysis Systems. 2002; 423-32.

33. Tran LQ, Moon CW, Le DX, Thoma GR. Web page downloading and classification. Proc 14th IEEE Symposium on Computer-Based Medical Systems. 2001; 321-6.

34. Walker FL, Thoma GR. A SOAP-enabled system for an online library service. Proc InfoToday 2002. 2002; 320-9.

35. Walker FL, Thoma GR. Web-based document image processing. Proc SPIE: Internet Imaging. 2000; 268-77.

36. Weeber M, Mork JG, Aronson AR. Developing a test collection for biomedical word sense disambiguation. Proc AMIA Symp. 2001; 746-750.

37. Xiaocheng L, Prettyman, M., Antonucci, R. System expansion and integration with agents in HSTAT. Proceedings of the World Multiconference on Systemics, Cybernetics, and Informatics. 2000.

38. Yoo TS, Morris J, Chen DT, Burgess J, Richardson AC. Template guided intervention: Interactive visualization and design for medical fused deposition models. Proceedings of the Workshop on Interactive Medical Image Visualization and Analysis. 2001; 45-48.

39. Yoo TS. Toward validation databases for medical imaging: Engineering a scientific rendezvous. Proceedings of VISIM Workshop on Information Retrieval. 2001; 7-10.

40. Zamora G, Sari-Sarraf H, Mitra S, Long R. Analysis of the feasibility of using active shape models for segmentation of gray scale images. Proceedings of SPIE Medical Imaging 2002: Image Processing. 2002; 1370-81.

Address of the author:
Alexa T. McCray
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894
USA
Tel.: +1 301 496 4441
Fax: +1 301 435 3146
E-mail: mccray@nlm.nih.gov

In: Haux R, Kulikowski C (eds.) Yearbook of Medical Informatics 03: Quality of Health Care,
The Role of Informatics. Stuttgart: Schattauer Verlag, 2003; 193-203.