Application of UMLS Indexing Systems to a WWW-Based Tool for Indexing of Digital Images

Charlie Hatton, Dept. of Pathology, University of Pittsburgh Medical Center (UPMC); Dr. James Woods, Cognitive Science Branch, National Library of Medicine; Dr. Rajiv Dhir, Dept. of Pathology, UPMC; Dr. Sheldon Bastacky, Dept. of Pathology, UPMC; Dr. Jonathan Epstein, Dept. of Pathology, Johns Hopkins University (JHU); Dr. Gary Miller, Dept. of Pathology, University of Colorado; Dr. Joel Greenson, Dept. of Pathology, UPMC; Dr. Kirk Wojno, Dept. of Pathology, University of Michigan; Dr. Michael Becich, Dept. of Pathology, UPMC

INTRODUCTION

The increased use of the Internet and the World Wide Web (WWW) has created new opportunities for development in health care, medical education, and clinical research applications. The WWW in particular offers health care practitioners and medical staff the ability to communicate readily and effectively with colleagues across the world, and to document the exchange of information for later use in clinical reporting, patient care, and education.

One important application of WWW-based technology in medicine is the construction of a comprehensive searchable image database. Such a database will be useful in comparing the features of images with unknown diagnoses to images residing in the database for identification. A distributed online database will also be useful for medical and patient education, allowing users to view specific examples of images of normal tissues, disease processes, and varying grades of disease simply by entering the appropriate keywords.

Two components of WWW-based clinical systems which are essential for the effective use of a clinical image database are the ability to share diagnostic images and the use of a robust and standardized structured language system to allow users to unambiguously describe images.

Lowe and his co-workers (1,2) have built an 'Image Engine' that uses the National Library of Medicine's Unified Medical Language System Metathesaurus for indexing and organizing its database. Reid (3) has also used the Metathesaurus to index clinical images and Browning and co-workers (4) and Carmella (5) have used the WWW to distribute radiology images.

We have developed a WWW-based system which allows clinicians and researchers to share clinical images, and to index those images based on content also using output from the UMLS Metathesaurus at the National Library of Medicine. The1996 Metathesaurus (6) used in this experiment is a repository of roughly 250,000 concepts, with definitions, lexical variants, etc. It contains information from more than 30 biomedical vocabularies including MESH, SNOMED, CPT, ICD-9, etc. Users may submit terms to the Metathesaurus to view related concepts and terms contained in the repository. An example of Metathesaurus output in HTML format is shown in Figure 1.

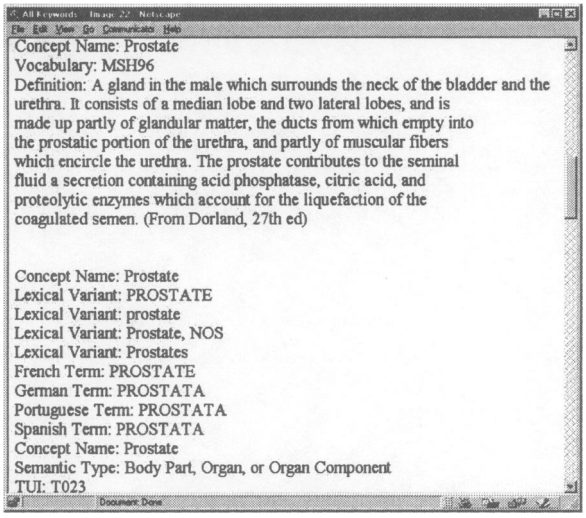

Figure 1. Example Metathesaurus output for keyword 'prostate'

We are currently recruiting partner institutions and individuals to establish a comprehensive relational database of pathology images which may be searched using a variety of methods based on Metathesaurus output. The image upload, indexing, and search tools will eventually be made available to the health care community, and the image repository presented as a resource which can be shared and utilized via the WWW by all interested individuals. We expect the

1091—8280/97/$5.00

comprehensive pathology resource to contain roughly 600,000 images of all types, and plan to complete the acquisition and indexing of all clinical images which currently exist on the WWW (roughly 30,000 - 60,000) within two years.

PROJECT DESCRIPTION

In mid-1996, the first prototype of the WWW-based image indexing system was completed. This version is located online at http://path.upmc.edu/nlm/index.html and allows users to index any image associated with the University of Pittsburgh Medical Center (UPMC) Department of Pathology online case database (http://path.upmc.edu/cases.html), as well as a group of prostate microscopy images provided by the NLM. This version supports a simple keyword search for images and 'flat file' database format, but lacks a mechanism to allow the contribution of new images via the WWW. The prostate image set became the basis for the first tests of the image indexing system, described below. This version of the image indexing system originally utilized the 1996 version of the UMLS Metathesaurus for keyword identification.

Before the completion of this version of the image indexing tool, the NLM, in conjunction with the Agency for Health Care Policy and Research (AHCPR), conducted a Large Scale Vocabulary Test (LSVT) to gauge the utility of the existing Metathesaurus and to determine whether additional dictionaries and information sources should be included in the next generation of indexing tool.The test vocabularies included some 750,000 terms from (1) the 30 vocabularies fully or partially represented in the 1996 version of the UMLS Metathesaurus, (2) the portions of SNOMED International not in the 1996 Metathesaurus, (3) the Read Clinical Classification, and (4) the LOINC (Logical Observations Identifiers, Names, and Codes) vocabulary.

EXPERIMENTAL METHODS

The initial tests of the image indexing system were based on the indexing activity of a group of six genitourinary pathologists from the UPMC, the University of Michigan Medical Center (UMMC), the University of Colorado, and The Johns Hopkins University (JHU). Each participant was asked to index the eleven prostate images in the test image set with up to twenty-five descriptive terms per image. Each image was displayed separately at full resolution, along with the keyword input area for that image. A sample screenshot of the original image indexing interface is shown in Figure 2. The terms entered by each indexer were captured using a Common Gateway Interface (CGI) script, and used as keywords in a Metathesaurus search. The output of the Metathesaurus searches was captured by the same script and associated in a series of HTML files with the appropriate images. An example of the display HTML page, containing keywords and selected indexing output, for an indexed image appears in Figure 3. Both the 1996 version of the Metathesaurus and the LSVT were used in this experiment.

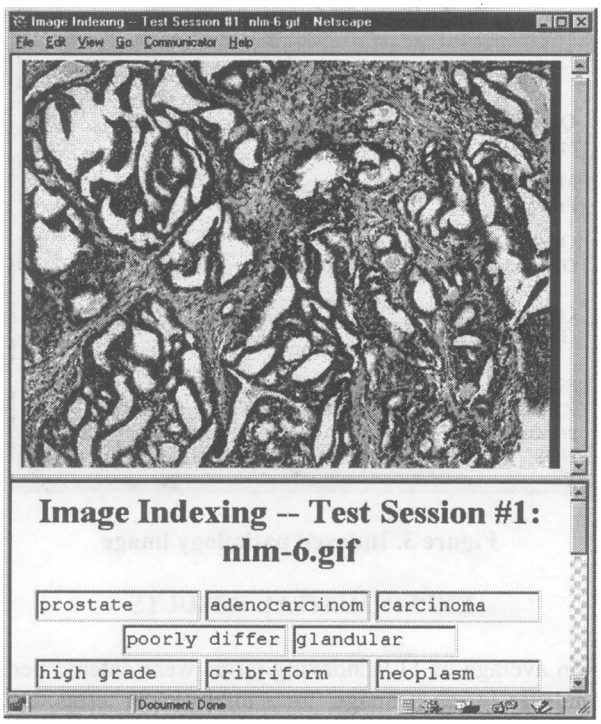

Figure 2. WWW-based image indexing interface

The prostate microscopic images used in this experiment were provided by the NLM, and represent diagnoses ranging from low grade to very high grade prostatic adenocarcinoma. Each image is 500 pixels by 378 pixels in size, in CompuServe GIF format. The indexers were not informed prior to participation in this experiment of the diagnosis associated with each image. Indexers were allowed the option of viewing the Metathesaurus output generated from their indexing terms after indexing each image.

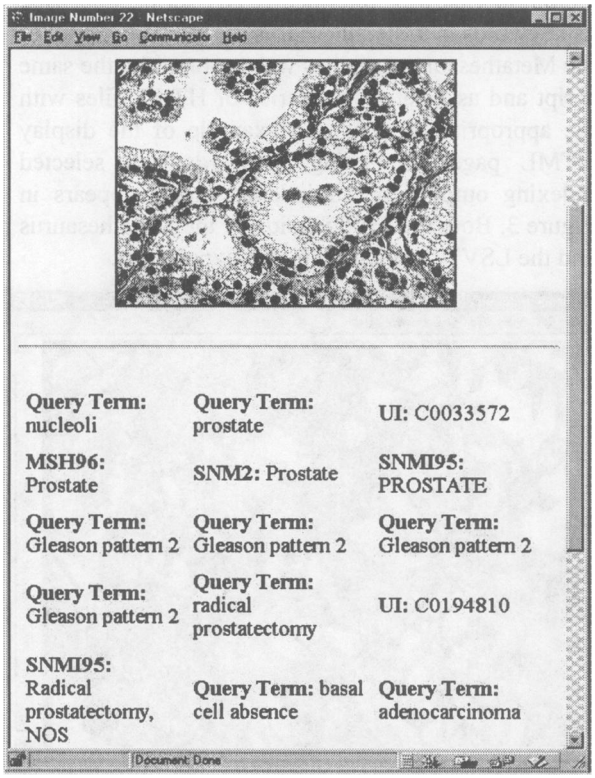

Figure 3. Indexed pathology image

EXPERIMENTAL RESULTS

An average of 11.9 indexing terms were entered per indexer for each image, for a total of 784 terms. Of these, approximately 38% (300 terms) were found in the 1996 version of the Metathesaurus and produced valid indexing output. Of the 784 terms entered by all indexers, 212 were unique, or were combinations of unique terms (e.g., prostatic carcinoma was logged as 'prostatic carcinoma', 'prostatic', and 'carcinoma'). Of these unique terms, only 30 % (64 terms) were found in the 1996 version of the Metathesaurus, suggesting that many of the terms found in this version of the Metathesaurus had been used by several indexers and/or had been used to describe several of the images. Indeed, the three most common terms used for indexing the test image set, 'prostate', 'carcinoma', and 'adenocarcinoma', were found to comprise 14 % (112 entries) of the 784 total indexing events.

We found no indication in this study that the indexers were influenced while indexing images at the end of the test set by Metathesaurus output from earlier images. That is, the indexers did not refrain from entering terms which were not found in the Metathesaurus during previous indexing events. There was considerable variation, however, in the proportion of keywords entered by each indexer which were found in the 1996 version of the Metathesaurus. This value ranged from 22 % found in the Metathesaurus to 81 % found, with a median of 38 % found. The proportion of found terms in the LSVT was considerably higher, ranging from 66.7 % found to 83.7 % found, with a median of 74 %.

The unique terms generated by the indexers that were not recognized by the 1996 Metathesaurus were submitted (as batch files) to the LSVT. The terms found by the 1996 Metathesaurus plus those found by the LSVT ranged from 66.7% to 83.7% with a median of 74%. That means that the 1977 version of the Metathesaurus should recognize more than twice the number of prostate pathology terms recognized by the 1996 version.

CURRENT AND FUTURE DIRECTIONS

The second version of the WWW-based image indexing system has now been developed and is currently being evaluated. This version supports image upload to a centralized WWW server, indexing using a 1997 Metathesaurus keyword list with the ability to enter a term or terms if no appropriate term is found in the list from the Metathesaurus, and enhanced search capabilities, including Boolean searches and search term selection from a keyword list generated from Metathesaurus terms and concepts. This version of the image indexing system is supported by an Oracle 7.3 database which holds all index keywords entered, as well as Metathesaurus output for indexing terms and pointers to all indexed images. This version uses and fully supports the 1997 Metathesaurus format. This phase of development is now complete and is undergoing extensive evaluation and testing.

We are currently designing and constructing the third generation of online image indexing and searching tools. This implementation will allow registered users to deliver images in any of several supported image file formats directly to the image database and to fully index each image and search for groups of images using Java-based tools on the WWW. These tools will allow indexers to 'drag and drop' terms

from a series of Metathesaurus keyword lists onto an image or indexing area to allow comprehensive indexing in a structured entry environment. Users searching for images will similarly be able to select keywords from a set of Metathesaurus lists to locate images in the database. The Metathesaurus keyword lists will be generated dynamically from the image indexing database, which will contain all of the information contained in the 1997 Metathesaurus. For experimental purposes, the system will also support automatic import of new Metathesaurus-supported keywords, use of keywords not present in the Metathesaurus, and possibly use of other UMLS medical language resources. This next and final phase of the WWW-based image indexing project will be completed in stages over the next twelve to eighteen months.

CONCLUSION

The development of a comprehensive distributed repository of pathology images will serve as a valuable resource for a wide range of health care and clinical research applications. Primary care physicians, pathologists, medical students, patients and their family members can benefit from a distributed and comprehensive collection of instructional images. The utility and success of a repository such as this is based on several factors, however.

First, contributors must have access to a facility for uploading and indexing images from their local machines to a centralized image database and server. This facility must be available at all times to indexers and contributors, must have the capability to distinguish between indexers, and must above all be easy for indexers to use. Another important consideration in constructing an indexed image database is the ability to easily search for images or image types, even with many thousands of images and indexing terms present in the database. This requires the use of a robust database server for image and indexing term storage, with a mechanism for rapid retrieval of images and indexing material upon request. Finally, the indexing tool should be based on and be consistent with a uniform language set to allow for unambiguous indexing, or identification, of image contents.

The WWW-based tool and relational database which we have developed for image storage and indexing, in conjunction with the use of the UMLS Metathesaurus, meets the requirements for an online image indexing tool and repository outlined above. The Oracle database design and server provide a robust and powerful 'store and retrieve' capability for images and associated indexing concepts, while the CGI – and eventually Java – WWW tools provide a mechanism for timely delivery of images to researchers or others searching for or indexing images. Finally, use of the Metathesaurus provides indexers with a nearly comprehensive set of clinical keywords in many areas of medicine and research from which to select indexing terms. With these tools in place, we feel that the WWW-based image indexing and search facility will become and remain a significant and valuable resource for clinicians, researchers, and laypersons across the world.

REFERENCES

1. Lowe HJ Image Engine: an object-oriented multimedia database for storing, retreiving and sharing medical images and text. In Safran C (ed.) Proc. 17th Annual Symposium on Computer Applications in Medical Care, 1993 839-843.
2. Lowe HJ; Buchanan BG; Cooper GF; Vries JK Building a medical multimedia database system to integrate clinical information: an application of high-performance computing and communications technology. Bull Med Libr Assoc, 1995: 83(1): 57-64.
3.. Reid JC. Private communication.
4. Browning GC; Liang Y; Buckwalter KA; Kruger RA; Aisen A. Worl Wide Web interface to digital imaging and communication in medicine-capable image servers. J Digit Imaging, 1996, 9(4):178-84.
5. Caramella D; Neri E; Del Sarto M; Lencioni R; Bartolozzi C. Implementation of a server World Wide Web of radiology accessible by Internet. Radiol Med (Torino), 1996, 91(5):622-626.
6. McCray AT; Razi, AM; Bangalore, AK; Browne, AC; Stavri, PZ The UMLS Knowledge Source Server: A Versatile Internet-Based Research Tool, In Cimino, JJ (ed) Proceedings AMIA Annual Fall Symposium, 1996: 164-168.

ACKNOWLEDGEMENTS

This work was supported in part by the National Library of Medicine and the Pathology Education Research Fund (PERF).