Access Techniques for Document Image Databases

FRANK L. WALKER AND GEORGE R. THOMA

ABSTRACT

IN THE MOST GENERAL SENSE, "access" evokes the paradigm of a seeker of information asking a question of a machine which searches for and retrieves an answer. In a more practical vein, this entails accessing a bibliographic database by entering a query comprising key words or phrases, either free text or terms out of a controlled vocabulary, and receiving citations to the literature.

In a database consisting of images, say bitmapped digital images of documents stored on high density media such as optical disc, automated access actually may be done in several ways. One way is for the user to first search a bibliographic database, after which the system retrieves citations and links these to corresponding document images on optical disc. Another way is to browse a list of stored document titles, to select one and continue the search through another list at a lower level (e.g., a table of contents in a monograph or a list of articles in a journal issue); then, on making a selection from this latter list, to be presented with the document image retrieved from electronic storage. A third way is to perform a "full-text" search of the machine-readable areas of the stored documents and, then have the system retrieve and integrate the text and graphic regions to form composite images that appear similar to the original paper documents.

This article describes the access and retrieval techniques implemented as part of a research and development program in electronic imaging (EI) applied to document storage and retrieval applications at the National Library of Medicine (NLM).

Frank L. Walker, Lister Hill National Center for Biomedical Communications, National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894
George R. Thoma, Lister Hill National Center for Biomedical Communications, National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894
LIBRARY TRENDS, Vol. 38, No. 4, Spring 1990, pp. 751-86

Introduction

As part of a research and development (R&D) program to investigate the role of electronic images in document preservation, a prototype Document Conversion System (DCS) was developed. Its purpose was to capture paper-based material as bitmapped images by means of a document capture workstation (DCW), inspect the images for quality by means of a quality control workstation (QCW), and transfer them from temporary magnetic storage onto digital optical WORM discs on an archiving workstation (AW). These three workstations, all subsystems of the DCS, are self-contained devices, each controlled by an IBM AT-class computer, and networked via a token ring Local Area Network (LAN). The system description, experimental objectives and results, and the genesis of the R&D program have been extensively discussed in the citations following the text of this article.

Once archived, the images have to be accessed, retrieved, displayed, and manipulated by a user. The access to and retrieval of document images stored on optical discs may be accomplished in several ways. The system that accesses, retrieves, and displays images is the image retrieval workstation (IRW). Using the IRW, a user gains access to images by either browsing a list of document titles presented on the screen, or by initiating a search of NLM's MEDLINE ® or CATLINE ®. In both cases the IRW software links with the document image through an index via the document's unique identifier (UI) entered by the document capture operator while the paper document is initially scanned. Once the images have been located, they may be retrieved either from a "local" optical disc mounted in a drive that is part of the IRW, or from a "remote" disc drive that is part of an image server (IS).

Functions to Access, Retrieve, and Use Images

There are many functions required to access and retrieve electronically archived documents. These include functions to: determine what documents are available, locate the desired document, retrieve the document (in compressed form) after it is located, expand each compressed image, and then display it. Once the document has been retrieved, there are additional functions for using it. These functions allow the user to manipulate images and to place electronic bookmarks (icons) on important images for later use.

Methods of Access

The two methods of accessing electronic documents are meant to allow the system to accommodate two types of library users. We may call them the "serious researcher" and the "casual patron." The serious researcher is assumed to be searching for all biomedical

documents related to an area of interest. First, citations to the documents are obtained by searching NLM's databases. GRATEFUL MED ®, an NLM developed user interface to these databases, is integrated into the image retrieval workstation software and provides a convenient means to retrieve the document citations. After the user enters the key search terms, GRATEFUL MED automatically logs into the NLM mainframe and performs the search in MEDLINE or CATLINE. The returning citations are downloaded to the magnetic disc on the IRW. At this point the user may list or print the citations using GRATEFUL MED or exit. If the user continues to run GRATEFUL MED, he can perform additional searches to narrow or broaden the search strategy. Once the user exits from GRATEFUL MED, the IRW software will list the citations received from the most recent search. If the user is interested in a particular citation, the IRW will extract its unique identifier and check its presence in an index of document images. All document images are indexed by a UI to permit easy linkage with the citations. If the search of the document image index reveals that the corresponding document images exist, the user is so notified. If the user wishes to see them, the first page of the electronic document is automatically retrieved and displayed. Subsequent pages of the document are retrieved and displayed upon command from the user. Once the patron is finished with the first document, the next citation is displayed and linked with the corresponding electronic document if it is available on optical disc. This pattern continues for the rest of the retrieved citations.

The second type of library user, the casual patron, is assumed to be less interested in doing database searches than in browsing through a book or journal issue. The paradigm for this kind of use might be to take a volume off a shelf, skim through the table of contents, and read any chapter or article of interest. To accommodate this kind of usage in an electronic archive, obviously a database search function is inappropriate. To meet this need, the image retrieval workstation provides a browsing facility which alphabetically lists the titles of all available electronic documents. From the title list, the user gains immediate access to the corresponding document images. The browsing function handles both monographs and serials. For monographs, the IRW lists the titles with each corresponding to an archived document. The user can search either on the first letter of the book title or on a string of characters which may appear anywhere in the title. Once the desired title is chosen by the user, the IRW automatically retrieves the first page of the electronic document. Subsequent pages of the document are retrieved and displayed upon command from the user. Similarly, in the case of journals, the IRW lists journal titles from which the user may make a selection. The IRW then lists the issues available for the chosen

title. Once the user picks a specific issue, the IRW lists the articles in that issue. After the user chooses an article, the IRW automatically retrieves the first page of the article. As in the case of books, the user controls the retrieval of the remaining pages in the article.

There are tradeoffs in choosing between the database search function or the browsing method to retrieve archived documents. On the one hand, the database search provides all citations which might be relevant to a user's field of interest but, for every citation, an archived document may not actually exist on optical disc. Because the database search function also takes time, the user must be prepared to wait for the citation query to complete. On the other hand, the browsing function guarantees that an archived document is immediately accessible for every listed book or journal title. This function therefore best serves the casual user who simply wants to browse through a collection of books or journals.

Locating the Archived Document

As already mentioned, document images are accessed through the unique identifier which is first associated with the document when it is captured. That is, at the time of capture, the operator enters the UI for that document. When the images are archived to optical disc, the archiving workstation keeps track of the size of each image and its location on the optical disc. After the optical disc has been filled with images, the UI, image size, and location information are written to the disc, these data constituting the disc index.

The next step is to use the index manager software (discussed in a later section) to create indexes of the total optical disc collection and to store these at each retrieval workstation. To make every disc's contents known to all retrieval workstations, the index manager is used to enter the contents of each optical disc in the collection into a master index. The master index is a B-tree index which keeps track of all unique identifiers, the number of page images associated with each UI, the size of each image (in kilobytes), the location of each image on its disc, and the label of the disc on which the images for a particular UI reside. A copy of the master index is then put on the magnetic disc at each retrieval workstation. Later, while displaying citations from a GRATEFUL MED search, the IRW software extracts the UI from the citation and queries its master index. If the document corresponding to that UI is archived on an optical disc, the software will get the label of the disc, the number of pages in the document, and the size and location information for every page image. At this point, the image retrieval workstation can retrieve the images.

In a manner similar to locating images through a GRATEFUL MED search, the browsing function also provides a means for quickly

locating document images. This is performed by searching a series of indexes created for browsing by the index manager. When an optical disc's contents are put into the master index, the index manager also creates browsing indexes for books and journals. This is accomplished by using the GRATEFUL MED search engine feature. The index manager first creates a list of the unique identifiers for all archived documents, then invokes the GRATEFUL MED search engine to get all book titles, journal titles, issues and article titles. It extracts this information from the retrieved citations to create B-tree indexes for this information. The resulting browsing indexes are placed along with the master index at every image retrieval workstation on magnetic disc. Then, as a user browses through a title list and selects a title, the workstation software queries the appropriate browsing index to extract the UI corresponding to that title. Next it queries the master index to get the optical disc starting block number for the chosen document images. Once the document has been located, the document images may be retrieved, displayed, and used.

Retrieving the Document Images

Once the electronic document has been located, it may be retrieved either from a local optical disc or from a disc at a remote image server. There are three options available for document retrieval. First, the image retrieval workstation may have a local optical disc drive for accessing images on discs which the workstation user can physically insert and remove. This provides a means of retrieving images from a local collection of discs. The second option for image retrieval is to use an image server. The image server could have several optical disc drives, each containing an optical disc, for sharing archived documents among several workstations. This second option is suitable for those documents which are frequently accessed and perhaps needed simultaneously by multiple users. Finally, as a third option, it is possible for an IRW to have both a local optical disc drive and a connection to an image server over a LAN. This provides access to both a local and a remote collection of archived documents.

One way to organize the collection of document images among discs that are "local" at an image retrieval workstation and those that reside at a remote image server might be to base the distribution on anticipated use. Those documents that are frequently used and in demand by multiple users simultaneously should be at the IS accessible by multiple IRWs. For documents that are infrequently used, it might be sufficient to have them on discs that are locally available for a user to physically remove from a shelf, say, and insert into an IRW when needed. Considering the level of usage of much

of the older biomedical literature, the latter might be sufficient. The system has, however, been designed to accommodate more frequently used material as well.

If an image retrieval workstation needs to access documents on optical discs controlled by an image server, a communications protocol is required for quick retrieval over a LAN. A protocol has been developed to permit an IRW to query an IS to find out which optical discs are available for online access over the LAN. The protocol also permits an IRW to reliably retrieve documents from the IS over the LAN. Research into prototype servers and workstations has allowed the design of an image communications protocol that permits images to be retrieved quickly.

Using the Electronic Document

Once the electronic document has been retrieved, the image retrieval workstation provides the user with a variety of functions that promote easy and flexible use. First, the IRW displays the document images in either soft copy form on a high resolution image display device or in hard copy form on a laser printer. The user may "flip" page images forward or backward through the document or go directly to any arbitrary page in the document. The IRW has an image manipulation function to allow the user to zoom into images to see fine detail. The user can also rotate an image 90 degrees right or left to display pages printed in landscape mode in the original paper document. Finally, the IRW provides an electronic "bookmark" function which allows placement of an icon representing a bookmark on an image. Up to ten page images in every document may be marked with no limit on the number of marked documents. The electronic bookmark function permits a user to track important sections in the electronic document in a manner analogous to using a bookmark in a paper book. It permits a direct jump to marked pages in a document, retrieval of the most recently marked document, and movement between the marked section in one document and marked section in another document.

Image Retrieval Options: Discussion

As mentioned earlier, images may be retrieved either from an optical disc drive locally connected to a workstation or from optical disc drives located at a remote image server. Having a local optical disc drive connected to a retrieval workstation permits a user to choose a disc from a local collection of discs. This self-contained configuration is adequate when the disc collection is not likely to be simultaneously shared by more than one user. The second method, based on an image server, is the best alternative for handling several users needing to simultaneously share a common archive of images

stored on multiple discs. With a remote store of images, however, in addition to accessing the image data on disc and its retrieval, there is also the problem of image transmission to the image retrieval workstation possibly located at a distance.

An earlier "baseline" prototype system (Thoma et al., 1985), based on a centralized architecture in which a DEC PDP 11/44 served as system controller, relied upon high-speed point-to-point modems to deliver uncompressed page images over the NLM's Broadband Cable Network. The effort, though technically successful in terms of speed, did not offer a reliable path to scale up the system to a larger number of display terminals. Also, the measured bit error rate was high enough to cause concern over the prospect of transmitting compressed images since errors that are tolerable in uncompressed image transmission could seriously affect the quality of compressed images. With the subsequent refinement of the system concept to a set of distributed IBM AT-based workstations implementing the key functions of capture, quality control, and archiving in the Document Conversion System, the stand-alone image retrieval workstation in which images were retrieved locally from the workstation's optical disc drive was developed. This approach works well for a single user, but does not scale up economically. Not only does each workstation require a relatively expensive optical disc drive (approximately $10K), but the set of optical discs would have to be shared by different users resulting in a wait for the more popular discs. Alternatively, multiple copies of frequently used discs would have to be made available. The solution selected is a document image retrieval network that would allow several image retrieval workstations simultaneous access to a database of document images.

For the transmission of images from an image server to the requesting image retrieval workstation, the point-to-point modem approach did not have the necessary features. In addition to high speed—a feature of point-to-point modems—it was also desirable to have low transmission error rate, built-in error checking, support for multiple stations, and off-the-shelf availability. These requirements led to investigations of local area network technology which has become widely available for all types of computers including machines in the personal computer class. The cost of intercomputer communication varies with bandwidth, buffer size, and software sophistication, but there is a wide and growing selection of reasonably priced LAN interfaces for AT-class computers. It was therefore logical to explore the use of this technology to support document image retrieval.

Furthermore, since the document image retrieval functions are independent of the document conversion functions, it is not necessary that a single LAN or type of LAN support both applications. In

the prototype systems developed in the laboratory to evaluate electronic imaging for document preservation, the document image retrieval LAN is completely different from the LAN supporting document conversion: they differ in terms of topologies, physical layers, and protocols. The only point of contact between document conversion and image retrieval is the database of images stored on optical discs created by the document conversion system and utilized by the document image retrieval system. This approach allowed the design and selection of the optimum LAN for each application.

One way to integrate standalone workstations into a networked system would be to connect several standalone workstations to a LAN and to let each access the other's image files. This would require each workstation to serve as both a personal image retrieval workstation and as an optical disc image file server. Controlling these two operations under DOS, a single-tasking operating system, would be difficult. Managing index data in such a configuration would also be awkward.

In light of these factors, it was found that the document image retrieval network is best configured as a file service system similar to the document conversion system. There are key differences, however: the image files to be retrieved are permanently stored on optical discs and are only available for reading; since file servers with drivers for optical discs are not available off-the-shelf, a special purpose image server with an efficient communications protocol was designed for this application.

Image Retrieval Workstation: Functions

The image retrieval workstation permits a user to access, retrieve, and use document images archived on optical disc. The optical disc drives may be local—e.g., connected to the IRW (see Figure 1), or remotely located on a local area network and connected to an image server (see Figure 2). The "local" option is illustrated in Figure 1. Here, the IRW may have one or more optical disc drives. It is controlled by an IBM AT-class personal computer, and it is equipped with a high resolution soft copy image display device for viewing document images and an alphanumeric display for the user interface. It has a laser printer for obtaining hard copies of the document images, and a mouse for manipulating images on the soft copy display. Finally, it has a telecommunications link to NLM's mainframe resident databases MEDLINE and CATLINE.

The remote retrieval option is shown in Figure 2. There may be one or more image servers and several retrieval workstations. Each retrieval workstation has an Ethernet connection to a baseband Ethernet network in place of the local optical disc drive. Here also each IRW has a telecommunications link to NLM's databases.

The local option allows the user to maintain a collection of optical discs, and to use one disc at a time in the drive. The networked

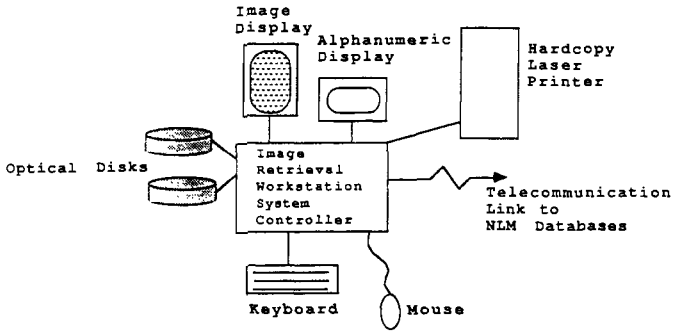

Figure 1. Image Retrieval Workstation (IRW) Accessing Images on a Local Disc

case allows the user access to those disc drives and media connected to the image server. The advantage of a local optical disc drive is that it is suitable for an environment where there are a small number of users, none of whom would need to simultaneously share the same archived collection. The advantage of having a network of one or more image servers and several retrieval workstations is that multiple users can simultaneously access a single archived document collection.

A third case is also possible: where the image retrieval workstation is both networked and "combined" equipped with a local optical disc. This is the most flexible option since it offers the advantages of having a permanent, centrally located collection of archived documents while also offering a user the choice of using discs from a local collection. This case is most useful if the document collection can be classified by different degrees of demand so that those documents frequently accessed might be located at the image server, and those that are infrequently accessed at each user's workstation.

The image retrieval workstation has the following basic functions:

1. A *database search* function which allows a user to perform a bibliographic search in MEDLINE or CATLINE via GRATEFUL MED. Once the database search is complete, the user may view the results of the search—i.e., the retrieved citations—after which the system links each citation to the corresponding document images on optical disc.
2. A *browse* function which presents a user with a list of titles of books or journal articles archived on the optical disc collection.

A user selection of an item from the title list activates the image retrieval workstation to provide an automatic link to the archived document images.

3. A *display* function which permits the document images obtained

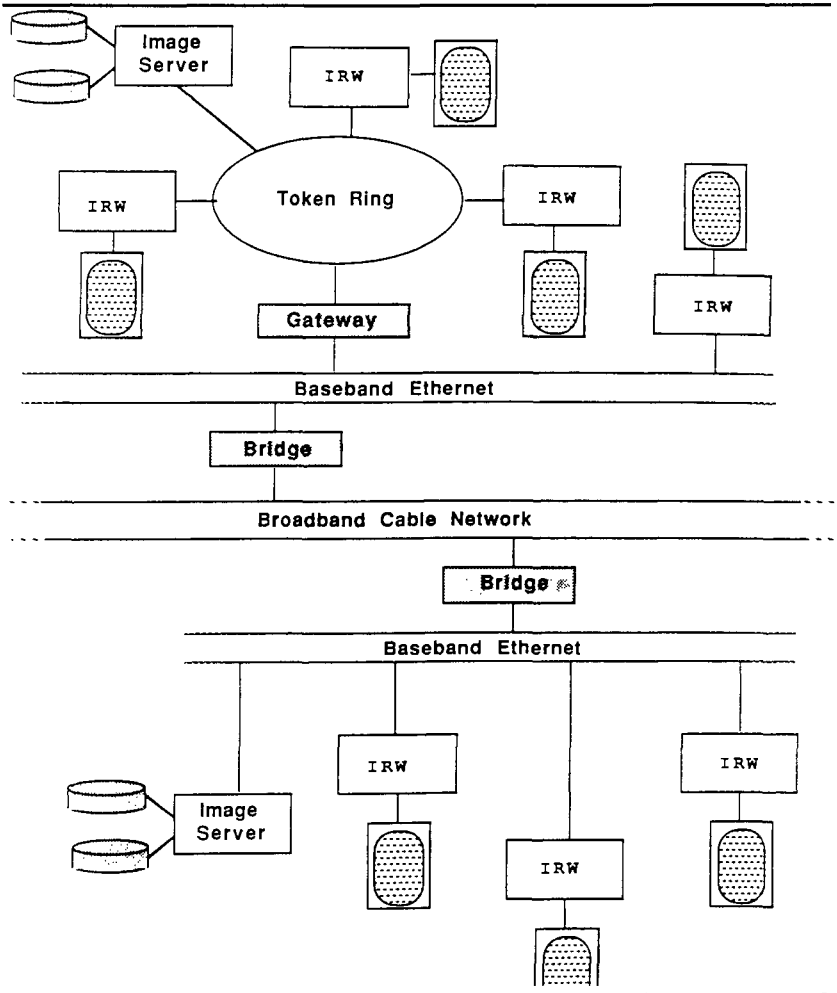

Figure 2. Connecting Token Ring and Ethernet LANs via Gateways and Bridges

from the linkup in 1 or 2 to be retrieved and displayed in soft copy form on the high resolution image display device or in hard copy form on a laser printer.

4. An *image manipulation* function which permits an image displayed on the soft copy display device to be zoomed, shrunk, rotated, panned, or scrolled.

5. An *electronic bookmark* function which permits a user to place bookmarks (visual icons marking the displayed page) on up to ten pages in each document with no limit on the number of marked documents. This function permits a user to move from the marked section in one document to the marked section in another, and also to retrieve the most recently marked document.
6. A *list* function displays a list of all optical discs indexed and from which the IRW may access the archived documents.

Image Server: Functions

The image server is essential for remote access and in prototype form uses a 10 Mbits/sec. Ethernet LAN. Ethernet is chosen partly because it is an industry standard with inexpensive and readily available interfaces. It is attractive also because it is available either as a baseband or a broadband LAN. The only difference in equipment is the transceiver; the computer interface hardware and all of the software are suitable for both modes. However, other physical layer protocols need not be excluded from consideration as there are many bridges and gateways currently available to support connecting "local" LANs to an Ethernet backbone. Figure 2 also shows how token-ring LANs and Ethernet LANs could coexist in a document image retrieval application connected by a backbone of broadband and baseband Ethernet. The price of connectivity is the cost of the equipment and some degradation in throughput associated with each bridge or gateway.

The functions of the image server are to:

1. Interface one or more optical disc drives to a 10 Mbits/sec. Ethernet LAN.
2. Retain an index of documents archived on the set of optical discs mounted on those drives. The index is designed to permit fast retrieval of any data record.
3. Respond to a request from an information retrieval workstation to provide a list of disc labels for the optical disc drives connected to the server. Upon initialization, it is necessary for a workstation to be able to determine the total set of optical discs available for online access. This set of discs includes those available at an image server or from a local disc drive connected to the workstation. It is assumed that any indexed disc not online needs to be inserted into the local disc drive. With knowledge of the location of each optical disc, the workstation can retrieve images from the correct source.
4. Respond to a request for an image from a workstation.
5. Give status data to the requesting workstation as to whether there are errors detected in image retrieval, whether excessive delays are expected due to heavy user demand, or if image transmission has

been completed.

Image Retrieval Workstation: Hardware Description

While at various developmental stages it was necessary to design and fabricate individual interface or controller boards, the monitoring of parallel developments in the electronics industry served to identify commercially available alternatives. For example, the high resolution display interface and the compression/expansion subsystem were originally in-house developed board-level prototypes, but they may be replaced by currently available off-the-shelf alternatives. The following list gives the hardware components of the information retrieval workstation as designed and implemented in the laboratory. Where appropriate, optional items are mentioned for equivalent performance.

1. The system controller is the IBM AT (or compatible) personal computer with a CPU having a minimum clock rate of 8 MHz and 512 KB of main memory.
2. An operating system equivalent to DOS version 3.3 or higher.
3. A magnetic disc drive to be used for storage of index information and operating system software. The disc must have a capacity of 110 MB (or sufficient capacity for storing the indexes required to access the archived documents); and an average access time of twenty-eight milliseconds or better. A rough estimate of the required disc size comes from test indexes used in the prototype information retrieval workstation. It shows that 110 MB is large enough to index approximately 3 million images. This estimate assumes an average of ten images per document and that half the documents are books and the other half journals. For other applications, the disc size requirement will be different, depending on the average number of images in a document and on whether the documents are monographs or journal articles.
4. An enhanced graphics adapter (EGA) and EGA color monitor serving as the primary man/machine interface.
5. Two RS-232 serial interfaces and a Centronics-compatible parallel printer interface on a single board. One serial interface controls the mouse, which is used for image manipulation. The second serial interface is used to communicate with the NLM databases. The printer interface is used for printing citations from GRATEFUL MED.
6. A Microsoft-compatible mouse used for manipulating images.
7. A Hayes-compatible 1200 baud modem used by GRATEFUL MED for searching NLM's MEDLINE and CATLINE databases.
8. A 300 dot per inch laser printer, used for producing hard copy printouts of images capable of printing text pages at a speed

of at least six pages per minute. Options include the QMS Kiss or Hewlett Packard Laser Jet Series II printers.

9. A printer interface capable of transferring image data via direct memory access from computer memory to the printer controller.
10. A printer controller capable of accepting both image data as well as ASCII text data to be printed on the laser printer. For image data sent to the printer controller, it has a pixel replication algorithm to magnify the images so that an 8.5 × 11 inch image scanned at 200 dots per inch retains the same dimensions when printed at 300 dots per inch on the laser printer. For text data sent to the printer controller, it passes the data straight through to the laser printer without change in appearance. The printer interface, controller, and printer are capable of printing an image in less than twenty seconds after the initiation of image data transfer to the printer interface.
11. An image expander capable of expanding an image compressed by the CCITT Group 4 two-dimensional compression technique. It can expand an image of a typical NLM page in less than two seconds. A 1 MB onboard memory is recommended for this expander; typically, the memory should be large enough to hold both a compressed and uncompressed image.
12. An interface for controlling a local optical disc drive such as the industry standard small computer system interface (SCSI). This component is required for local image access. The controller must be capable of direct memory access and capable of transferring 33 KB of data (an average compressed page image) from the disc drive to computer memory in 0.6 seconds or less.
13. An optical disc drive used for retrieving document images from a local collection of discs. This component is required for workstations to have a local image access capability. The optical disc drive must be capable of transferring a typical compressed NLM image (33 KB) to computer memory in 0.6 seconds or less.
14. A LAN interface connecting the image retrieval workstation to a LAN; this component is required for the workstation to be used in a networked environment for remote image access.
15. A display interface for transferring image data from the computer memory to the high resolution display monitor; it has 2 MB of internal memory for image manipulation functions, is capable of zooming/shrinking the image 2:1, is able to pan and scroll, and is able to rotate an image clockwise and counterclockwise by 90 degrees.
16. An image display monitor capable of displaying the entire image of a page scanned at 200 dots per inch resolution—i.e., have 1,728 × 2,200 pixel display capability such as the Discorp model VMB 2002.

Image Server: Hardware Description

The hardware required for the image server used as a file server consists of the following:

1. A small computer equivalent to the IBM AT personal computer with a CPU having a minimum speed of 8 MHz, and 512 KB of main memory.
2. An operating system equivalent to DOS version 3.3 or higher.
3. Ethernet hardware and software interface to the selected LAN.
4. Hardware and software interface to the optical disc drive including:
   - —SCSI host adapter (or some standard appropriate for the selected drive)
   - —Special hardware and software to ensure maximum transfer rate of the drive when reading data into computer memory
   - —Ability to manage data transfer from multiple drives.
5. Sufficient computer memory to hold one uncompressed page image file (one-half MB for 200 dpi images), in addition to program and other data; requires extended memory if an AT-type computer is used.
6. A magnetic disc drive, with minimum capacity of 60 MB to store the indexes required for accessing the archived documents; average access time of twenty-eight milliseconds or better to be used for storage of index information and operating system software.
7. A monitor and adapter for operator interface; may be color or monochrome.

The requirements for the LAN shall be as follows:

1. 10 Mbits/sec. data rate or faster.
2. Cabling for selected speed/distance; recommend shielded twisted pair (Type 6 or better).
3. Should allow adding or deleting nodes without disrupting operations.
4. Should be expandable over greater area as application grows.
5. Data buffers on the interface boards; recommend at least 128 KB.

The requirements for the upper level protocols and application programs are as follows:

1. Permit multiple servers including a jukebox.
2. Accommodate a separate index node.
3. Able to achieve 3 Mbits/sec. or faster memory-to-memory transfer of page image files from server to display workstation.

The last section of this article gives hardware requirements for a version of the image server when it is used as an index server.

Image Server Communications Protocol: Rationale and Design

A communications protocol designated Image Transmission

Protocol (ITP) was developed in-house to support the transmission of images from an image server to image retrieval workstations. This development evolved from research into techniques for high speed reliable image data transmission over the NLM's dual cable broadband cable network (BCN). Originally a 10 MBps point-to-point link was designed for transmission between a single image transmitting station and a single image receiving station using quadrature phase shift keying (QPSK) modems. While the speed was adequate, this technique had disadvantages such as: multiple stations could not be supported and the system included no automatic method of error recovery. To offset these shortcomings, a set of design goals was developed for an image transmission technique that included: operating preferably with a well defined industry standard; a low effective error rate; built-in error checking; support for multiple stations; modularity; and availability of off-the-shelf components. Broadband Ethernet was found to meet these objectives.

Initial studies with off-the-shelf broadband Ethernet modems showed high reliability since the effective error rate was too low to be measured in the laboratory. For the higher layers, the FTP protocol (part of the industry standard TCP/IP) was acquired and evaluated, partly because it is a standard file transfer protocol used by all systems on the Internet. However, using off-the-shelf FTP for image file transfer resulted in very low throughput—less than 100 KBps.

The demonstrated reliability of Ethernet, with the potential for higher throughput than that available from a widely used standard, suggested a reason to design a protocol suitable for image transmission to deliver both speed and reliability objectives.

The ITP is application driven and application dependent. It relies heavily on the low error rate of local transmission provided by the Ethernet hardware and (low level) protocol, on the finite size of a page image file, and on the fact that image files need only be transmitted from a server to a display workstation. Additional assumptions are that the image requesting node wait until the entire image is received. Because of these design assumptions, very little information needs to be transmitted with each image data packet, and very few packets other than data packets need to be transmitted with each page image file, reducing both bandwidth overhead and data processing overhead.

The ITP is therefore a special purpose protocol. It trades flexibility for speed, a more important commodity for an application involving image files almost exclusively. It is not intended for a general purpose application, for instance, to transfer any generic file from one computer to another. It is designed for one purpose: to transfer image files from an image server to an image retrieval workstation at high speed—approximately 3.2 Mbits/sec.—as measured in the laboratory.

The document image retrieval application is less complicated than the document conversion application in that page image files will only be transmitted from the nodes having optical discs (image servers) to the nodes having high resolution displays (display workstations). Also, the numbers of image and index files are relatively static, and the display workstations need not store the image files on local magnetic media. For these reasons, it was possible to design a private, application-driven protocol to manage the transmission of images from image servers to display workstations that exploits the finite requirements of the transaction to achieve very high page image file throughput over Ethernet. The component of the system that predominantly determines server throughput is the optical disc interface which determines the rate at which page image files are read from the optical disc into the server's memory. Appropriate design of the optical disc interface and driver allows data to be retrieved from the opical disc drive at the maximum transfer rate of data from the optical disc platter. This includes the design of the interface of the optical disc drive to the SCSI bus, the SCSI host adapter for the AT-bus, and the software driver. The optical disc drive should have a large (at least 64 KB) internal buffer to accommodate asynchronous data transfer between the AT and the drive. The Optimum 1000 Optical Disk Drive has a maximum transfer rate of 0.48MB/sec.

As shown in Figure 3, there are four basic layers in the ITP. At the lowest level is the physical layer which consists of either baseband or broadband Ethernet. This is the hardware level which facilitates the bit stream signal transmission along the Ethernet cable between the image server and each image retrieval workstation. Immediately above the physical layer is the data link layer, implemented in the prototype IS and IRW systems by means of an Ethernet data link processor which plugs into the computer. The data link processor is responsible for access control, addressing, and a low level of error detection. Above the data link layer is the application layer. The modules of the application layer use the header field of Ethernet packets for communication and error control between the IS and IRW (see the definition of the ITP following this paragraph). Both the IS and IRW have transmit/receive modules which encode and decode the data fields in the Ethernet packets. They are also responsible for data management, packet sequencing, higher level error detection, and queue management. Above this is the user interface software which converts user requests to network activities and provides high level control over error conditions.

Definition of the ITP

The communications protocol is defined by information placed

into the header of each Ethernet packet by the image server or the image retrieval workstation. All protocol codes are two bytes in length shown as follows:

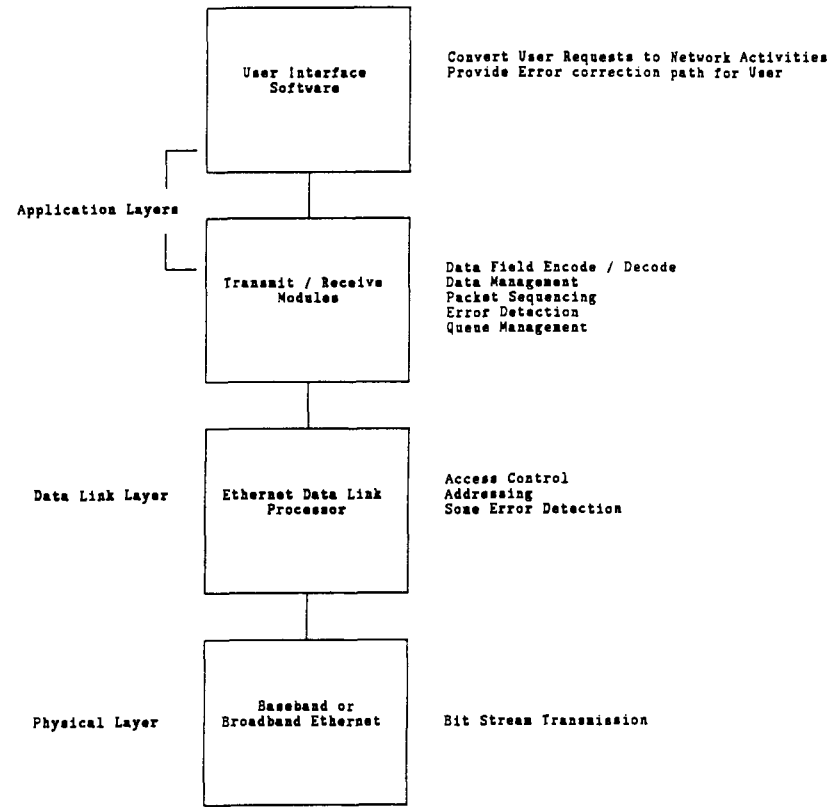

Figure 3. Image Transmission Protocol Layers

The low byte identifies the general form of the data being transmitted or received—e.g., image data, control message, etc. The high byte identifies the specific form of the data being transmitted or received—e.g., first image data packet, number of packets being sent, etc. The following is a description of protocol codes which the image server can transmit:

Low byte = **01** indicates the packet contains image data

High byte = **01** indicates this is the first image data packet
Example:

01	01

High byte = **02** indicates this is an image data packet in the range from 2 through the last packet
Example:

02	01

High byte = **03** indicates the packet contains the number of image data packets transmitted; used for error checking

Low byte =**02** indicates the packet contains nonimage data

High byte = **01** indicates the packet contains labels of discs controlled by the server

High byte =**02** indicates the packet contains queue position data

Low byte = **06** indicates a message sent to the image retrieval workstation

High byte = indicates a nonfatal message; workstation software is to continue

High byte = **02** indicates a fatal message; workstation control software is to be terminated

The following are protocol codes which the image retrieval workstation can transmit:

Low byte = **05** indicates the packet contains a request for an image server

High byte = **01** indicates this is a request for image data

High byte = **02** indicates this is a request to get labels of the discs currently mounted at the image server.

Software for the Image Retrieval Workstation

The prototype image retrieval workstation requires an executable module and a number of index files for operation. The executable module for the prototype image workstation runs under DOS on an IBM AT-class computer and consists of many modules written in the C language as well as assembly language. The C modules were compiled using the Lattice C compiler version 3.1, the assembly modules assembled using Microsoft Assembler version 5.0, and all

were linked using the Microsoft Linker version 3.6. The Lattice Windows software package is used to produce the screen displays for user interface. There are several index files required for the image retrieval workstation; they are B-tree index files created and managed using C-tree by Faircom. None of these products is a specific requirement of the workstation; products from other manufacturers and other operating systems (UNIX, AIX, etc.) could be substituted with minor modifications.

Information Retrieval Workstation Manager

The overall flow control for the image retrieval is illustrated in Figure 4. The manager is the main module which controls the process. Four basic functions are initially provided to the user. The first function allows the system to search two of NLM's databases, MEDLINE and CATLINE. If the user chooses this option, the databases module is invoked which permits the user to enter search terms to GRATEFUL MED. GRATEFUL MED logs onto the NLM mainframe, performs the search in MEDLINE or CATLINE, downloads the citations to the magnetic disc on the computer, and logs off the mainframe. Then, in a standalone manner, the workstation will display each citation. For each citation in which the user is interested, the software will check for the existence of the corresponding document images. If the document images exist, the user can view them in soft copy or hard copy form.

The second function provided by the image retrieval workstation manager is a browsing function, which is performed by the browse module. Browsing permits a user to view a list of document titles available in archived form. For books, titles are listed from which the user may choose. For journals, a list of titles are initially given from which the user may choose followed by a journal issue and article within that issue. After a book title or journal article title has been chosen, the system automatically retrieves the corresponding document images for display in either soft copy or hard copy form.

The third function of the manager is an electronic bookmark function performed by the bookmark module. Once a document is marked, it is possible to directly retrieve it later even if the workstation has been turned off after marking.

The fourth function provided by the image retrieval workstation manager permits the user to see a listing of all optical discs which are indexed. The indexes, invisible to the user, keep track of the location of all images on all optical discs plus the book titles, journal titles, issues and article titles for document browsing.

Finally the workstation manager permits the user to exit from this software. In doing so, all appropriate index files are closed upon exiting to the operating system.

Database Function

The function permits a user to perform a search of either MEDLINE or CATLINE, link the resulting citations to the archived

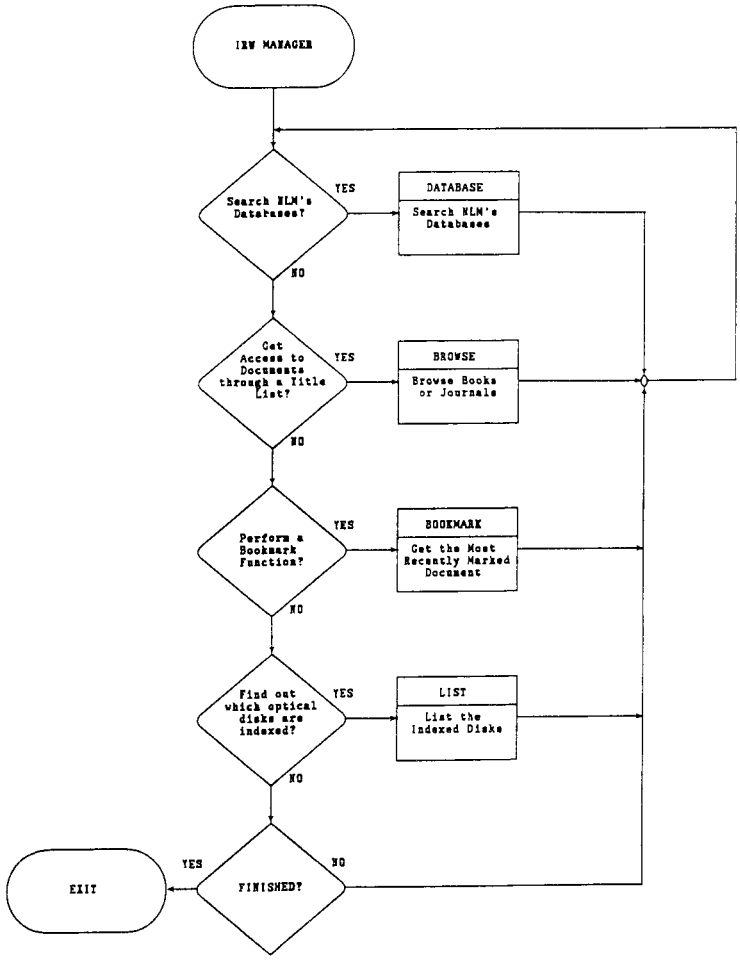

Figure 4. Image Retrieval Workstation Overall Flow Control

store of document images, and retrieve the document images. As illustrated in Figure 5, this function first invokes GRATEFUL MED which permits computer-assisted searches of MEDLINE or CATLINE. Through the assistance of GRATEFUL MED, the user can enter search terms then have the search performed automatically. The search citations are downloaded to the magnetic disc on the workstation, and GRATEFUL MED returns control to the database routine.

The first citation from the database search is displayed on the workstation's color monitor. If the user is interested in the citation, the image workstation software will check the index of unique identifiers to determine whether the document has been archived. If it has been archived and is available for access, the user is notified and is asked whether he wants to see it. If the user answers affirmatively, the display routine is invoked, which permits the user to retrieve the document. After the document has been accessed, or if the user did not want to see the document, the software checks to see if there are more citations that resulted from the GRATEFUL MED search. If more citations are present, they are displayed as before. If there are no more citations, software control returns to the workstation manager.

GRATEFUL MED permits citations to be printed. If the user wants a hard copy of the retrieved citations, a method of switching the hard copy laser printer to text mode is available. This printer is normally used for printing images but can also accommodate text. Once printed, the citations are available for future reference, and the archived documents, if available, can be retrieved by manually entering the unique identifier (part of the citation), as one of three methods for accessing documents from the browse function.

Browse Function

The browse function, invoked by the workstation manager, permits the user to access archived documents without a database search. It is intended for the casual user who wants to browse through the collection of archived documents. By listing all document titles, the browse function provides a one to one correspondence between document titles and archived documents; the user is guaranteed that the document images corresponding to the listed titles are available. This is not the case for the database function, which does not provide such a one to one correspondence. The database function, intended for more serious researchers, lists all citations relevant to a user's search strategy, whether or not they correspond to archived documents. While intended for different purposes, the database and browse functions are both useful and complement each other.

The browse function permits a user to access archived documents either through a list of book titles, a list of journal article titles, or through a unique identifier. If the user is interested in books, the browse function will alphabetically list all available book titles on the color monitor screen. The user may start a search on book titles by entering a single character, which finds the first title beginning with that character. It is also possible to search a string of characters appearing within a title. An example would be to find the next title which contains the word *doctor.* Once a search has

been completed, an arrow appears on the screen at the title of interest. The user can move the arrow about the screen to select any other title if he desires. Once a title of interest has been found, the user presses the enter key on the keyboard, and the browse software checks the title index to extract the unique identifier of the document. Then it goes to an image index to retrieve the document images through the display function.

If instead of books, the user is interested in journals, the browse function will list all journal titles of archived journal articles. Once the user selects a journal title, the issues available for that title are listed. After the user picks a journal issue, the browse function lists all articles available for that issue. At this point the user may search on article titles in the same manner available for searching on book titles either by the first letter of the title or through a string search of the contents of all titles. Once an article is selected, the unique identifier is extracted and the display function is invoked to access the article.

The third method of accessing documents through the browse function is to enter the unique identifier of the document. The UI is available from the GRATEFUL MED search in the database function. If the document has been indexed, it can then be accessed and retrieved through the display function.

List Function

The list function lists all optical discs indexed and available for access, either from an image server or from a local optical disc drive; the discs may or may not be currently mounted. The list function reads the volume index and keeps track of the labels of all indexed optical discs. Then it displays the disc labels on the color monitor screen and returns control to the image retrieval workstation manager.

Bookmark Function

The bookmark function, called by the image retrieval workstation manager, provides the user of the workstation a degree of flexible control over the electronic document in a manner similar to the control a reader has over a paper document. Analogous to the case of paper documents where bookmarks keep track of important sections, the electronic bookmark does the same thing for electronic documents. Up to ten pages in every document can be marked with no limit on the number of marked documents. Even when the IRW is turned off, the bookmark information is not lost since it is kept in a file. The bookmark function permits the most recently marked document to be retrieved quickly by invoking the display function which retrieves the most recently marked pages in this document. The display function then permits the user to view the document, to jump to

other marked pages in the document, or to jump to other marked documents. This last feature is an advantage an electronic document has since there is no analog to this function in a world of paper documents. Once the user is finished with the bookmark functions through the display module, control is returned to the manager.

Display Function

The display function allows the operator to use the electronic document. Usage includes viewing the document, jumping to any page in the document, placing or removing bookmarks, manipulating images, printing the document, or jumping to other marked documents. Normally the flow control begins with the current page number being set to one. However, if the display function is called from the bookmark function previously described, the current page number is set instead to the most recently marked page in the document. This first image is automatically retrieved from either the local optical disc or image server, expanded and displayed on the soft copy display device. If the image has been previously marked, a bookmark icon is displayed on the image. The user is presented with a menu of choices, each of which invokes a specific function, described as follows:

1. *Next Page.* If the user selects Next Page, the current page number is incremented by one with the largest page number being the number of pages in the document. Then the image of the next page is retrieved, expanded, and displayed. If the new page has been marked, the bookmark icon is also displayed.
2. *Previous Page.* If the user selects Previous Page, the current page number is decremented by one, with the smallest page number being 1. Then the image of the new page is retrieved, expanded, and displayed. Here, too, if the new page has been marked, the bookmark icon is displayed.
3. *Page Jump.* If the user selects Page Jump, he can move directly to a specific page in the document, jump to one of ten marked pages, or jump relative to a marked page. The page jump process returns with an updated current page number. Then the image of the new page is retrieved, expanded, and displayed. If the new page has been marked, the bookmark icon is displayed.
4. *Manipulate Image.* The Manipulate Image process permits the image on the soft copy display device to be zoomed 2:1 or shrunk back to normal size. It also permits the image to be rotated left or right 90 degrees, panned, or scrolled. Panning and scrolling are accomplished through the use of a mouse. After the manipulate process is completed, the software returns control to the user, allowing the selection of another display function. Details appear later.

5. *Mark Page.* If a user selects Mark Page, he can mark the image currently displayed on the soft copy display device with a bookmark icon, or remove this icon if present on the image. This feature also permits the user to remove all bookmarks from the current document or remove all bookmarks from all documents if any exist. After the Mark Page process is completed, the software control returns to display which allows the user to select another function. Details appear later in the Mark Page function section following.
6. *Print Page.* If the user selects Print Page, he is given many printing options: either to print the image currently displayed, part of the current document, all of the current document, or all marked pages in all documents. After the print page process is completed, the software control returns to display which allows the user to select another function. Details appear later in the Print Page section following.
7. *Jump to Next Marked Document.* If there are marked documents following the current document, the Jump to Next Marked Document function is available for use. If the user selects it, the software finds the unique identifier of the next marked document and sets the current page number to the most recently marked page in this document. Then this image is retrieved, expanded, and displayed. The bookmark icon is also displayed on the image.
8. *Jump to Previously Marked Document.* If there are marked documents prior to the current document, the Jump to Previously Marked Document function is available for use. If the user selects it, the software finds the unique identifier of the next marked document and sets the current page number to the most recently marked page in this document. Then this image is retrieved, expanded, and displayed. The bookmark icon is also displayed on the image. Once the user exits from the display function, the software returns control to the calling routine.

Page Jump Function

In addition to the features mentioned earlier, if the user wants a specific page, the page number must be entered. The range of valid page numbers varies from one to the maximum number of pages in the document. If the user types a number larger than the highest numbered page, the software automatically resets to the last page. If the user wants to go directly to a marked page, he can choose the page from a list of marked pages presented on the color monitor by using the cursor keys to highlight the desired marked page number, then pressing the enter key. If the user wants to jump relative to a marked page, it is possible to move forward n pages from a marked page or reverse n pages relative to a marked page. The user must enter the number n and also select the marked page with the cursor

keys. Once a page has been chosen from one of these methods, the software returns control to the display function with the updated current page number.

The facility for jumping relative to a marked page helps solve a problem in electronically archived documents—i.e., page number sequencing. It is quite common for the table of contents of a volume to begin with roman numerals (e.g., i, ii, iii), and the first chapter to start with Arabic numerals (e.g., 1, 2, 3). While the workstation permits a direct jump to any page in the document, the page number in the electronic document may not correspond to the real Arabic number in the printed volume. For example, page 50 of the paper book would correspond to the 53rd page of the electronic document if there are three pages (e.g., of the table of contents) preceding Arabic page number 1. Then when the user wants to jump to page 50 the system gives him page 47. This problem is solved through a feature of the workstation termed relative jumping. If a bookmark is placed at page 1 (Arabic) of the document, then by jumping relative to that bookmark, the user is able to move directly to the desired page.

Manipulate Function

The manipulate function permits the existing image on the soft copy image display to be zoomed, shrunk, panned, scrolled, and rotated. A menu permits the choice of zoom/shrink, rotate right 90 degrees, and rotate left 90 degrees. If the zoom/shrink option is chosen and if the existing image is a normal size, the image is zoomed 2:1. If the zoom/shrink option is chosen and if the existing image has previously been zoomed, the image is then shrunk to normal size. A mouse controls the panning and scrolling functions. Here, panning refers to moving the image right and left on the screen. Scrolling refers to moving the image up and down on the screen. The image is panned or scrolled in proportion to the mouse movement and is most effective for images which have been zoomed or rotated. Once the user is finished manipulating the image, software control returns to the display function.

Mark Page Function

The Mark Page Function permits the user to place an electronic bookmark icon on the image displayed on the soft copy image display device. If the user marks the current image, an icon representing the bookmark is overlaid on the image while a file of bookmark data is updated. The bookmark icon may also be removed from the current image if marked. If this option is selected, the bookmark icon is removed from the current image and the original data beneath the bookmark is restored. In addition, all bookmarks may be removed from the current document or all marked documents. If either of

these options are chosen, the bookmark data file is updated appropriately. Once the Mark Page Function is completed, system control returns to the display routine.

Print Page Function

In the Print Page Function, the user is presented with a menu of five choices for printing the document. If he chooses to print the page currently displayed on the soft copy image display device, that image is retrieved from either the local optical disc or remote image server, expanded, and printed on the hard copy laser printer. If the user chooses to print the entire document, the retrieval, expansion, and printing functions are automatically repeated for every page image in the document. If the user chooses to print only part of the document, he can specify the starting and ending pages to be printed. The appropriate images are then retrieved and printed. If the user chooses to print all the marked pages in the current document, the system retrieves, expands, and prints those that are marked. Finally, if the user chooses to print all marked pages in all marked documents, the system will do so. Once any of the print subfunctions have been completed, the software returns control to the display module.

Software for the Image Server

The image server requires an executable module and a number of index files for its operation. As with the software description of the image retrieval workstation a functional description of the IS software will be given from a high-level viewpoint.

As shown in Figure 5, the overall flow control for the image server is simpler than that for the image workstation. Upon initiation, the server manager opens the appropriate index files and reads the labels of all mounted optical discs attached to the image server. Next it stays in one loop which terminates only if any key is pressed on the keyboard. If no keys are pressed, the server checks to see if a request has arrived from a workstation. If a pending request is detected, the software checks to see if it is a request for an image. If it is an image request, the software extracts the requestor's address, page number requested, and unique identifier of the document image requested. If the unique identifier is in the local index file, the image server transfers the image from the appropriate optical disc to memory in the computer, then to the requestor over the LAN. If the unique identifier is not in the local index file of unique identifiers, an error message is sent to the requestor. If the pending request for the image server was for disc information, the image server sends a list of disc

labels for all mounted optical discs to the requesting image retrieval workstation. The software continues to check for pending requests or for a pressed key on the keyboard and responds appropriately.

Index Manager Software

The index manager software package runs separately from the

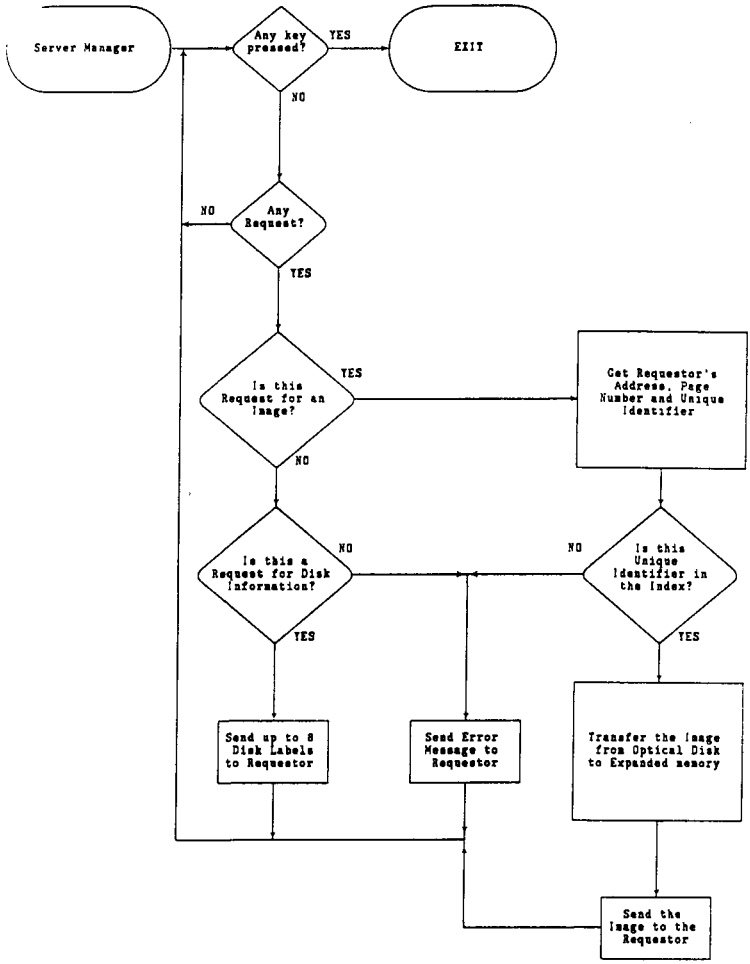

Figure 5. Flow Control for the Image Server

software used in the image retrieval workstation or image server. It is used to create and manage the various index files for the IRW and IS. The index files contain the labels of all indexed optical discs, unique identifiers, images in all documents, titles of archived books, journal titles of archived journal articles, issues of archived journal

articles and titles of archived journal articles. The index manager has three basic functions: to create these index files, delete obsolete index file information, and to list the indexed disc labels. The index files are in the form of B-trees to enable quick access. The advantage of using B-trees is that the time required to access records in a file increases very little as the number of records increases. An evaluation of B-tree indexes for this function found that, if ten optical discs are indexed, with an average of 28,800 images on each disc, the average time to access the index information for the first three images of every document was about .15 seconds. This is a very good result, and for this reason it is recommended that B-tree indexes be used for indexing the optical disc information.

The following is a description of the index files implemented in this design. The first thing to appear in each description is the name of the index file and a summary of its purpose. Next is a list of the software packages which use the file. This is followed by a list of the data contained in the file, then by the key required by the requesting software to access the data. To give an example of what this means, for the volume file, to find out whether a given disc label exists and how many documents have been archived on that disc, the requesting software must supply a key field containing the desired disc label.

Volume File: Keeps track of the labels of all indexed optical discs.
Used by: Index Manager; Image Retrieval Workstation; Image Server.
Data contained: disc label; number of documents on the disc.
Key required for data access: Disc label

Image Index File: Keeps track of the location of all archived images
Used by: Index Manager; Image Retrieval Workstation; Image Server
Data contained: disc label; Unique Identifier; page number; location of the image; length of the image
Key required for data access: Unique Identifier and page number

Book File: Used for browsing book titles; keeps track of the titles of all indexed books. The data may be accessed from one of three keys
Used by: Index Manager; Image Retrieval Workstation
Data contained: disc label; book title; Unique Identifier
First key required for data access: Disc label and first thirty characters of the book title (used to select titles from a specific optical disc)
Second key required for data access: Unique Identifier
*Third key required for data access:*First thirty characters of the book title (used to select titles from all optical discs)

Journal File Number 1: Used for browsing journals; keeps track of all journal titles archived on all optical discs
Used by: Index Manager; Image Retrieval Workstation
Data contained: disc label; Unique Identifier; journal title; journal issue
Key required for data access: Journal title (used for viewing journal titles on all optical discs)

Journal File Number 2: Used for browsing journals; keeps track of all journal titles and issues archived on all optical discs
Used by: Index Manager; Image Retrieval Workstation
Data contained: disc label; Unique Identifier; journal title; journal issue; year and month of issue
Key required for data access: Journal title and issue (once a journal title has been selected, this file is used for viewing journal issues on all optical discs)

Journal File Number 3: Used for browsing journals; keeps track of all journal titles, issues, and article titles archived on all optical discs
Used by: Index Manager; Image Retrieval Workstation
Data contained: disc label; Unique Identifier; journal title; journal issue; year and month of issue; article title
Key required for data access: journal title and issue (once a journal title and issues have been selected, this file is used for viewing the titles of journal articles from the selected title and issue archived on all optical discs)

Journal File Number 4: Used for browsing journals; keeps track of all unique identifiers
Used by: Index Manager
Data contained: disc label; Unique Identifier; journal title; journal issue; year and month of issue; article title
Key required for data access: Unique Identifier (used by the Index Manager for maintenance purposes only to verify accuracy of data records)

Journal File Number 5: Used for browsing journals; keeps track of all unique identifiers for specific optical discs
Used by: Index Manager; Image Retrieval Workstation
Data contained: disc label; Unique Identifier; journal title; journal issue
Key required for data access: Optical disc label and journal title (used for viewing journal titles from a specific optical disc)

Journal File Number 6: Used for browsing journals; keeps track of all journal titles for a specific optical disc
Used by: Index Manager; Image Retrieval Workstation
Data contained: disc label; Unique Identifier; journal title; journal issue; year and month of issue
Key required for data access: Optical disc label, journal title and issue (used for viewing journal issues from the selected title)

Journal File Number 7: Used for browsing journals; keeps track of all optical discs, journal titles, and issues for a specific optical disc
Used by: Index Manager; Image Retrieval Workstation
Data contained: disc label; Unique Identifier; journal title; journal issue; year and month of issue; article title
Key required for data access: optical disc label, journal title and issue (once a journal title and issue have been selected this file is used for viewing journal articles from the selected title and issue on the selected optical disc)

Figure 6 shows the flow control for the index manager. This software will run on any workstation containing a local optical disc drive and requires a communications link to the NLM mainframe. It has three basic functions: to add a new optical disc index to the indexes, to delete an optical disc index from the indexes, or to list the labels of all discs indexed. If the user wants to add a new optical

disc's index content information to the master indexes, he must first mount the optical disc in the workstation's optical disc drive. The software then reads the label of the disc and adds the label to the volume file. Then it reads the index from the optical disc and adds its information to the index file. Next it creates a search strategy for GRATEFUL MED. The purpose of this is to get the appropriate

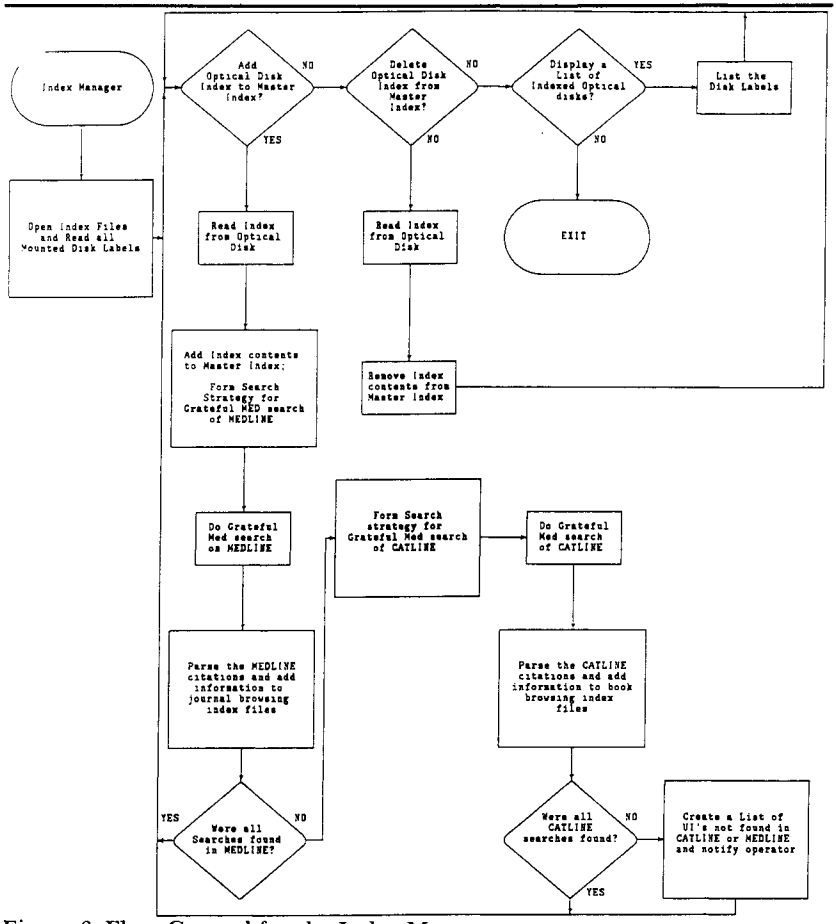

Figure 6. Flow Control for the Index Manager

information to enter into the book and journal index files. The search strategy begins with a list of unique identifiers of the documents from the disc's index file. After the search strategy is formulated, the GRATEFUL MED search engine is invoked, and the MEDLINE file is searched for the list of unique identifiers. After GRATEFUL MED returns control to the index manager, the software parses the citations retrieved and adds the journal title, journal issue, month and year, and article title for each unique identifier into the various

journal index files. Then the software checks to see if all unique identifiers from the search strategy were found. If all were found, this indicates that all were for journal articles. If some were not found, this indicates that some were for books so CATLINE must be searched.

A new search strategy is formulated for CATLINE from the unique identifiers not found in MEDLINE. Then the GRATEFUL MED search engine is invoked to search CATLINE. After GRATEFUL MED returns control to the index manager, the software parses the citations retrieved and adds the book titles into the book index files. Then the software checks to see if all UIs from the search strategy were found. If all were found, then this function has completed successfully. If some were not found, the operator is notified, and the UIs not found are written to a "not found" file on the workstation for later review.

The second function of the workstation, to delete optical disc information from the master index files, permits a user to remove unwanted disc index information from the files. The optical disc whose directory information is to be deleted must first be placed in the optical disc drive controlled by the workstation. Then the software reads the index from the disc and removes all data from all index files containing that index.

The third function is identical to that found in the image retrieval workstation: list indexed disc labels. This serves to verify that the two previous functions have completed successfully.

Operation of the Image Server and Image Retrieval Workstation

The operation of the image server begins with turning on all disc drives controlled by the IS with the appropriate discs mounted in each. Then the software is run by the user. There is no further intervention until the user wants to shut down the IS which is done by pressing any key on its keyboard.

The operation of the image retrieval workstation is more complex since there are many levels of menus available for the user. Each menu was designed to provide an error-free system. The system is error free in the sense that it automatically eliminates those options that are logically impossible, thereby preventing errors that users could cause by inadvertently making those selections. The implications of this become clear with a description of each of the menus. The contents of each menu are listed below with a discussion of how the error-free design prevents mistakes that could be caused by users. Each main menu is also described in previous sections of this article.

MAIN MENU
Search NLM's Databases and Display Documents
Browse Documents

Retrieve the Most Recently Marked Document
Display a List of Optical Discs Known to the System
Exit

The third item in the menu, "Retrieve the Most Recently Marked Document," can be selected by the user only if a document has been marked.

DISPLAY MENU
Display Next Page
Display Previous Page
Jump to Page
Manipulate Image
Mark Page
Page Print
Jump to Next Marked Document
Jump to Previously Marked Document

The "Display Next Page" choice is available at all times except when the last page of the document is being displayed. The "Display Previous Page" choice is available only if the first page of the document is not being displayed. The "Jump to Next Marked Document" choice is available only if there is at least one marked document after the current document in the list of marked documents. The "Jump to Previous Marked Documents" choice is available only if there is at least one marked document prior to the current document in the list of marked documents.

PRINT PAGE MENU
Print Page Currently Displayed
Print Entire Document
Print Part of Document
Print All Marked Pages in this Document
Print All Marked Pages in ALL Documents

The fourth choice in the menu, "Print All Marked Pages in this Document," is available only if the current document contains at least one marked page. The "Print All Marked Pages in ALL Documents" choice is available only if at least one document has been marked.

PAGE JUMP MENU
Enter the Page Number:
Go Forward Pages from Selected Bookmark
Go Reverse Pages from Selected Bookmark

The second and third choices in the menu are available only if at least one page in the current document has been marked.

MARK PAGE MENU
Mark Page Currently Displayed
Remove all Bookmarks from this Document
Remove all Bookmarks from ALL Documents

Remove Bookmark from Selected Page

The "Mark Page Currently Displayed" choice is available only if the image displayed is unmarked. The "Remove all Bookmarks from this Document" choice is available only if the current document contains at least one marked page. The "Remove all Bookmarks from ALL Documents" choice is available only if at least one document has been marked. The "Remove Bookmark from Selected Page" choice is available only if at least one page in the current document has been marked.

MANIPULATE IMAGE MENU
Zoom/Shrink
Rotate Right
Rotate Left

The "Rotate Right" choice is available only if the image is upright or has already been rotated to the left. Similarly, the "Rotate Left" choice is available only if the image is upright or has already been rotated to the right.

BROWSE MENU #1
Browse Available Book Titles
Browse Available Journal Titles
Browse by Unique Identifier

These choices are available at all times.

BROWSE MENU #2
All Optical Discs
Currently-Mounted Optical Disc

The browse function permits the user to browse either all optical discs or only optical discs currently mounted in either the server or local disc drive. It is possible for a large number of discs to be indexed in the master B-tree index files, but it may not be possible to have all simultaneously mounted. If the user browses all optical discs (choice 1), and wishes to view a document on an unmounted disc, the system will give him the label of the appropriate disc to insert in the local disc drive if one is controlled by the image retrieval workstation.

BROWSE MENU #3
Search for Letter
Search for String

The browse function also permits the operator to search for a book title or journal article title either by the first letter of the title or by a string of up to 10 characters in length which may appear in any part of the title. If the search is successful, the list of titles is updated with an arrow pointing to the title found.

Large Image Databases: Issues and Design Considerations

While online access to, and retrieval of, older documents preserved electronically may not be warranted by user demand, some thought was given to a situation requiring online retrieval of images from a large disc collection. The organization of such a collection among multiple servers is considered here.

As suggested in Figure 2, one image server can support several optical disc drives. Many commercially available optical disc drives are compatible with the SCSI interface standard which allows up to eight SCSI controllers on one SCSI bus. At $10K per drive, it becomes uneconomical to have more than a few drives per server or even per network. Having several drives on one server also limits the performance to the extent that it can only retrieve one image at a time; if all drives at a server have discs of "popular" document images, the average queue size at the server and therefore retrieval time will increase in proportion to the number of drives. A possible compromise between cost and performance is illustrated in Figure 7. This system has multiple servers. Those discs most frequently accessed are distributed among the "fast" servers, where they are mounted in drives. The remaining discs are located at the "slow" server which has a jukebox with one or more drives and a collection of discs that are unmounted, but which are mechanically retrieved and mounted when needed. Since jukeboxes are quite expensive, they should be considered only when the anticipated image database is too large to manage economically on a few drives and user requirements permit some delay in file access.

Another issue associated with a large database of image files is location of and access to the database indexes. If the database is small (a few optical discs) and fairly static, it is possible that all servers and retrieval workstations could maintain private copies of all indexes. If the database is large, magnetic storage requirements at retrieval workstations increase the cost of the workstation. And if the database is dynamic and growing, it becomes a significant operational problem to keep indexes at all nodes updated. In commercially available integrated systems, jukeboxes are commonly found attached to a computationally powerful minicomputer performing the functions of host, controller, and index manager. In these systems, the jukebox is generally the only node with optical disc drives, and so it is appropriate that its host/controller maintain the only copy of the indexes.

Figure 7 shows how the indexes might be managed in a system where the image file database is distributed among multiple servers. The index server maintains index data for all optical discs in the system. It also communicates via the network with all servers to

establish which discs are currently mounted at each fast server and which are available at the jukebox. Each server then only needs to maintain sufficient index data to access the images on their mounted discs which should help keep both response time and magnetic disc costs down. Retrieval workstations can utilize the index server at several levels depending on their own memory and magnetic disc capacities. Index data for one or more optical discs can be downloaded

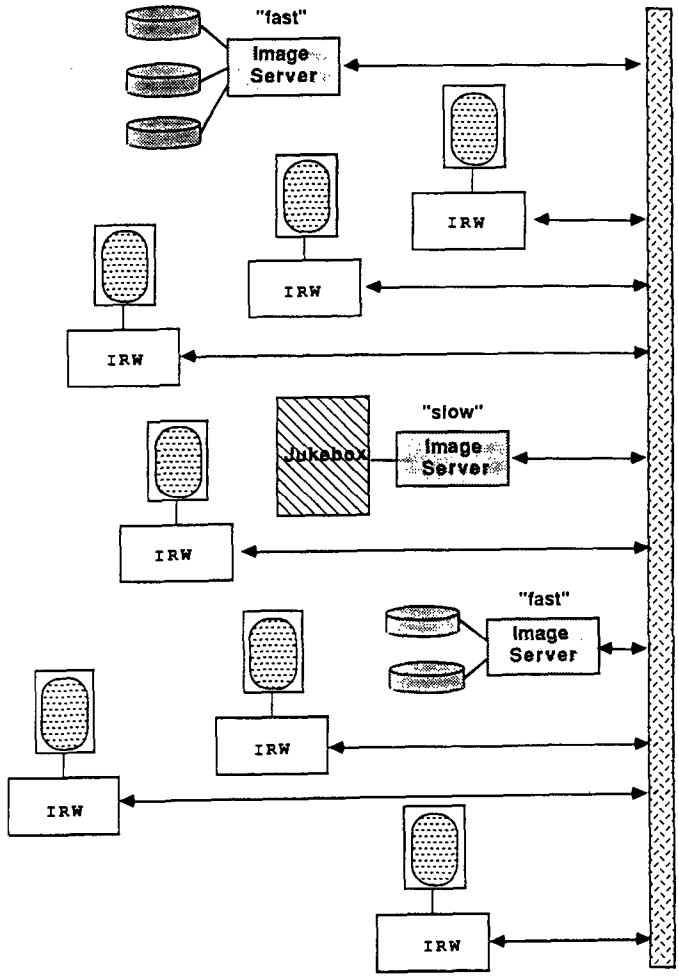

Figure 7. A System with Fast and Slow Image Servers

for local access at the beginning of a session. This would be especially appropriate for the browsing function which uses several index files.

Alternatively, index data can be retrieved on a document by document basis, which might be more appropriate when viewing documents that result from a GRATEFUL MED search.

The hardware requirements for an Index Server are the following:

1. A small computer such as the IBM AT (or compatible) personal computer with an 8 MHz CPU and 512 KB of main memory.
2. An operating system equivalent to DOS version 3.3 or higher.
3. Hardware and software interface to the selected LAN.
4. Sufficient magnetic disc capacity for all programs and utilities plus index files for total system database of document image files.
5. A monitor and adapter for operator interface; may be color or monochrome.

References

Document preservation by electronic imaging: System and experiment description, Vol II. (1989). Technical report of the Lister Hill National Center for Biomedical Communications, National Library of Medicine. Bethesda, MD. Order from NTIS: PB89-266534.

Document preservation by electronic imaging: Experiment plan. (1986). Communications Engineering Branch, Lister Hill Center for Biomedical Communications, National Library of Medicine.

Document preservation by electronic imaging: Interim project report. (1986). Communications Engineering Branch, Lister Hill National Center for Biomedical Communications, National Library of Medicine.

Thoma, G. R.; Suthasinekul, S.; Walker, F.; Cookson, J.; & Rashidian, M. (1985). A prototype system for the electronic storage and retrieval of document images. *ACM Transactions on Office Information Systems, 3*(3), 279-291.

Thoma, G. R.; Suthasinekul, S.; Walker, F.; Cookson, J.; & Rashidian, M. (1987). Design considerations affecting throughput in an optical disk-based document storage system. *Proceedings of the American Society for Information Science, 24,* 225-233.

Walker, F. L. et al. (1987). A distributed approach to optical disk-based document storage and retrieval, (pp. 44-52). *Proceedings of the 26th Annual Technical Symposium of the Washington, D. C. Chapter of the ACM, Gaithersburg, MD.*