OPTICAL DISK STORAGE AND RETRIEVAL IN THE NRL LIBRARY

Doris R. Folen
Kestrel Associates

Laurie E. Stackpole
Chief Librarian
Naval Research Laboratory

INTRODUCTION

This paper addresses the Optical Disk Storage capability at the Ruth H. Hooker Research Library and Technical Information Center at the Naval Research Laboratory. Optical disk technology is a viable, sensible solution to many of the serious recurrent problems that plague the Library community. The most striking feature is that optical storage is a permanent solution to The Space Problem. No matter how you treat them, index them or catalog them, paper products, film products, etc. have to be somewhere taking up space. Optical storage solves not only the space problem but also the problems of the retrieval and the refiling of the product. A paper copy of an item can be printed at the touch of a key and then can be destroyed rather than re-filed when the patron is finished with it. This technology presents a dream scenario in which a librarian can sit in a pleasant environment, identify a paper, document or picture and retrieve the item in minutes without leaving the area or the patron.

The Optical Disk System in the NRL Library has been fully established for over two years. It was one of the first of its kind and is still the largest currently operational in a research library.

PLANNING FOR AN OPTICAL DISK SYSTEM

Planning for an optical disk system is no easier now than it was years ago although the problems are different. The major hurdle three years ago was collecting information and advice since there were comparatively few people who were authorities in this field and very few facilities that had any experience as the technology was so new at that time. Present day problems are due largely to the proliferation of new equipment and the burgeoning of technology.

Optical disk systems are expensive. They require right from the beginning a commitment of time and money. The Library's original budget for the first part of the operation was a million dollars to be spent over two years. Many uncounted hours of staff time went into the planning and preparations.

First, thought must be given to who is going to use the system, who is going to run the system, who is going to be responsible for its maintenance and future development. An optical disk system is a project that once started cannot easily be abandoned. It is going to be with you for a long time and plans for the future have to be considered.

Time must be budgeted. It will take time to plan the system, buy it, install it, run it, maintain it, scan documents into it, etc. Staff and patrons must be trained to use it and that will also take time.

Money must be budgeted. This is a "chicken and egg" proposition. The amount of money available will determine the design. On the other hand, a study of what a system could accomplish for the organization might justify a higher budget.

A study must be made of why an optical disk system is needed, what problems is it going to solve, where it is going to be put, how is it going to be used, how will staff and patrons react to it, etc. Decisions have to be made as to where to concentrate on design such as: is speed of retrieval more important than cost? or how much time and money is available to spend on the scanning process versus the quality of the retrieved image.

Vendors must be considered and selected. Once one or more vendors can be singled out as likely candidates discussions with them will highlight the equipment most suitable to be incorporated into a system to fit your needs. Usually at least three or four configurations slowly emerge for consideration.

VIEWGRAPH NO.1

THE LIBRARY'S OPTICAL DISK SYSTEM

The Library chose On-Line, Inc. as the OEM (Original Equipment Manufacturer). On-Line took the Library's choice of basic equipment and working with the library staff on the design, put together one of the first optical disk systems of major proportions to be used in a research library.

Proper selection of equipment is of immense importance. Just the cost alone necessitates a carefully considered decision. Once purchased the major components become the platform for future developments. Replacement of major components are not easily made. Many hours were spent at conventions and shows looking at equipment and talking to vendors before the equipment forming the nucleus of the Library's Optical Disk System was selected. Dreamware and fantasyware had to be identified as such.

Since the system was to be an in-house stand alone archive, standards were not a primary consideration in the selection of equipment. Standards for the 12" optical disk were in the talking stage at that time as they are still. The library worked around this problem by trying to project standards and by picking leaders in the field as vendors.

VIEWGRAPH NO.2

A Sony Writable Optical Disk Autochanger Model WDA-610 and a SUN minicomputer and its work stations form the nucleus of the system. To be consistent with the market Sony now calls its autochanger a "jukebox" which it somewhat resembles in operation. It is designed specifically for use with Sony's 12" optical disks. In 1989 SONY declared that these disks have a lifetime of 100 years. The jukebox has a footprint of eight square feet and looks from the front remarkably like a two door refrigerator.

With the ability to accommodate 50 disks of 6.4 gigabytes one autochanger provides the equivalent of more than 2,000 reels of magnetic tape and up to 600 file cabinets. Up to seven autochangers can be daisy chained to expand storage capacity to 2.3 terabytes of on-line data on a single SCSI interface. Through the synchronized use of two writable disk drives, the autochanger offers an average disk-to-disk access time of just 5.0 seconds.

Connected to the Sony autochanger by a SCSI interface is a SUN 3/280. It has a console, four work stations, a 892 MB hard disk, 16 MB ECC RAM, online ethernet transceiver and a Ciprico SCSI controller. It also has a 9-track tape drive.

The in-house scanner is a TDC Docuscan DS-2600. It is capable of scanning pages of different sizes and thicknesses and can scan two sides of an 8 1/2" x 11" page in less than two seconds at 200 dpi. It is of compact design, having a footprint of 2' x 2' and weighs 100 lbs.

In using the scanner the operator slides the sheet to be scanned along the alignment guide. The transport system grips the sheet and carries it on a straight path past the scanning area and into the receiver tray. Thus paper jams are virtually eliminated.

The scanner is supported by an AST 386/33 microcomputer containing an Xionics XIP-B Scanner Card. Here the scanned images are compressed and stored on a 300MB hard disk before being transferred to the Sun 3/280 where the images are processed and sent to the Sony autochanger.

For retrieval, images are retrieved on the Sun 3/60 then decompressed and queued on a PC/AT and printed out on an HP Laser-Jet printer.

The equipment is networked on an ethernet, Inmac Multiport Ethernet Repeater and could be networked on the Lab-Wide/Nation-Wide Ethernet/Internet. However, currently access is password controlled to programmers and technicians having a need to work on the equipment. Restrictions on distribution of even the unclassified documents precludes general access to the information.

The report identification configuration consists of the Cuadra STAR program, and Alphamicro minicomputer, 14 TeleVideo 955 terminals and four laserjet printers.

VIEWGRAPH NO.3

A heavy duty paper cutter is necessary. The Library purchased a Challenge Model 20. It has a height of 54 inches, a footprint of 3 x 4 feet, and weighs 530 pounds. It requires two hands on the control panel for operation to keep the operator out of harm's way.

THE CONVERSION PROCESS

VIEWGRAPH NO. 4

The Library's unclassified document collection consists of roughly 100,000 documents averaging 65 pages/document. Before the conversion process began it occupied a space of 3600 sq. ft. It is considered invaluable and irreplaceable, dating back to the early 40's with some reports even older. A large percentage of the older documents are not in very good shape as time, insects, dirt, water and other calamities have taken their toll.

The reports that are scanned take a one way trip to the scanner. The quality of the images is consistently good enough to allow the destruction of the reports after they have been scanned. Nothing is gained if the reports are scanned and then returned to their place on the shelf to continue to take up space. There are exceptions. Occasionally there is a document that has actual historical value and the original is considered valuable. Those few documents are scanned and then saved in an historical file.

Because the entire unclassified collection is to be scanned, no particular care has been taken as to what is going to be scanned first. Documents that have come back from circulation are put aside to scan because there is no sense in filing them and then taking them off the shelf to scan. New NRL originated reports are scanned as soon as the Library receives a copy. Documents are taken from the shelves in sequence but are not checked for

misfiling as they are prepared for scanning because unlike full size paper documents which are filed on a shelf, images do not have to be placed on the disk in any particular order.

Time spent in preparation is time well invested. Each document is reviewed as a candidate for scanning. Duplicates and documents without technical substance are weeded out. Documents that are oversized or of poor quality are put aside for later consideration.

A document is prepared for the scanner by:

a. Removing all blank and extraneous pages especially covers when possible. Usually all of the information on the cover is repeated on the title page. The back covers seldom have anything on them at all.

b. For in-house scanning the documents are then "exploded" by a quick cut by the paper cutter. This removes all binding, staples, etc. If necessary, additional trimming of the sides is done also.

c. Oversize pages and fold-outs are cut to 8 1/2" x 11" or less.

d. Loose color pictures are removed and filed.

It is important to establish a routine for the work flow if there are a lot of documents to be scanned. An occasional report to be scanned is one thing. When there are 100 thousand of them to be reduced to images, efficiency becomes the watchword.

The actual scanning operation is prompted by menu driven cues on the AST 386 terminal screen. The operator keys in the accession number which is the only indexing done by the operator and follows the four or five instructions given on the screen as the document is fed page by page through the scanner. An automatic feeder is an option, but so far the automatic feeders require perfectly cut and positioned sheets for trouble free performance. Documents once scanned and their image checked for quality are destroyed.

The Library also uses off-base scanning facilities. Documents are screened and prepared for scanning much the same as if they were going to be scanned in-house. Then they are shipped in boxes to Docucon Corporation in San Antonio Texas.

Docucon scans the documents at 300 dpi and records the images to nine track tape which is then sent back to the Library to be mounted on the Sun 3/280 for the images to be sent to disk.

VIEWGRAPH NO. 5

RETRIEVAL

The actual retrieval of the document from the optical disk is done by typing the report's six digit accession number into a user friendly menu on a SUN work station. The document can then be viewed page by page on the screen or all or part of it can be printed out by a simple command. More than one document can be viewed on a screen and the same document can be viewed simultaneously at more than one work station at one time.

The ability to identify the report or reports to be retrieved has become a stumbling block for many system designers. Hundreds of thousands of pages of records placed on optical disk can rapidly submerge individual items in a ocean of data. Some designers solve this problem by superimposing a indexing system directly on the optical disk but this takes a lot of time and disk space. Some designers superimpose ASCII characters that can be searched word by word directly on the disk but this also takes a lot of disk space.

Selecting an indexing system can take as much or more time as selecting the hardware for the Optical Disk System. An indexing program can be expensive and if the wrong one is chosen very costly indeed, because if the indexing system is superimposed on the optical disk you are going to have to live with it as it will be almost impossible to change.

The staff of the Library solved this retrieval dilemma by having the retrieval system entirely separate from the Optical Disk System.

The reasoning was:

1. A retrieval system was already in place that the Library staff had developed
2. There was a need for the quickest and the most expedient means of getting the documents to optical disk and a separate retrieval system supported that goal
3. A separate system is less expensive, easier to manage, input can be done at its own pace, and anything on the system can be changed at any time with little fuss
4. At the beginning of the planning stage the optical disk system was thought of as simply another way of storing documents just as if they were still on the shelf
5. These documents do not have to be retrieved in tenths of a second. Retrieval of a document in less than half a minute is considered satisfactory. Compared with the time necessary to retrieve a dusty document from a real and usually dustier shelf, a half minute seems instantaneous.

Since August of 1987, the Library has had in place the Cuadra STAR retrieval system. This system provides an almost infinite number of fields so that a report can be indexed by every conceivable way that someone might think of to identify it; i.e. accession number, title, author, subject, contract, words in an abstract, size, etc. Searching is by every field separately or combined. The fields may be searched in toto or by words, by subfields or by masking. Boolean searching of combination of fields or search results is also possible. The results of all searches can be displayed or printed out by any or all of the fields used in the indexing in any order or in any combination

VIEWGRAPH NO.6
VIEWGRAPH NO.7

The flexibility of the system allows for the addition of a field or subfield if it is discovered that there is a need for one that hadn't been thought of before. Global changes allow for sweeping changes in indexing procedures if that should happen to be deemed necessary. Authority lists and lookup tables are used for tight subject and descriptive cataloging but fields are also available for the cataloger to use flights of fancy if desired. The Library enjoys the reputation of being able to identify and provide any document that the patron needs.

USING THE OPTICAL DISK SYSTEM

The system is available for both staff and patrons. Security is not a problem as the whole configuration is in a secure area and at the current time is not networked outside of the NRL Library. The patron or the staff member sits down on a comfortable chair, identifies the needed reports on the STAR system at the Televideo 955, keys in the accession numbers of the desired reports into the adjacent SUN work station and views and/or prints out any needed pages or reports in a pleasant atmosphere.

COST

The cost of archiving a large collection to optical disk is formidable. The task of figuring out exact cost even more so. Costs for individual aspects vary constantly as the technology changes. In 1988 when the staff of the Library launched into this effort the cost of equipment, design, programming and scanning was combined because we acquired the total system from a vendor. Roughly the cost of the system can be said to have been $250,000 and the cost of scanning the original 11,000 document to optical disk to have been $.35 cents a page.

However, there is no way to use this figure in determining present day costs. An organization considering using optical disk technology would have to consider tangibles such as cost of labor, equipment, space and time, factoring in intangibles such as the value of space saved and the value o the documents archived.

The cost of archiving the unclassified document collection of the NRL Library which consists of 100,000 reports or 7,000,000 pages is estimated at $1,953,500. This figure includes the cost of the original equipment, documents scanned, new equipment acquired and equipment on order, as well maintenance of equipment and salaries of personnel involved. Since this is an ongoing project not yet completed, this figure will change. Also, as additional Library collections are scanned to disk, the cost of equipment will be amortized. Also, this collection occupied 3600 square feet of floor space. At $28.68/sq.ft/yr a tangible savings of $103,248 per year is realized.

Estimated cost of scanning the Library's classified document collection which will be done entirely in house is between $850,000 and $1,000,000. This does not include cost of new equipment.

The estimated cost for scanning the NRL originated report collection both classified and unclassified (a total of about 750,000 pages)is $100,000.

At this time 100 reports are scanned in-house per week and 600 are prepared and sent to an off-site contractor for scanning. Weekly expenses are estimated to be $8500 including the cost of off-site scanning.

A report explaining costs in detail has been prepared and is available.

PLANNING FOR THE FUTURE

Looking to the future gives one the feeling of dabbling with science fiction. Technology in this area moves ahead so fast that it seems that anything is going to be possible. At the present time there is a fantastic quantity of newly developed equipment on the market. The Library has already upgraded its SUN with a SUN 490/Server and a Sun SPARC work station. This is presently being integrated into the system.

Even though the system used to identify reports is separate from the Optical Disk System itself, plans are now being made to merge these systems. The indexing system would still be a separate entity but would have incorporated into it the "hooks" used to retrieve the document from the optical disk once the accession number is

known.

When all of the unclassified documents are scanned the classified collection will be put to disk. This will mean another 10 million pages of images stored on a separate autochanger to be daisy chained to the current one.

Scanning in color will be considered as the technology becomes more affordable

Plans are being made for the networking of the system to the offices of the Naval Research Laboratory scientists. Very soon the images of selected documents will be networked. In the near future even the images of classified documents will be sent over encrypted lines to be decompressed and printed out remotely.

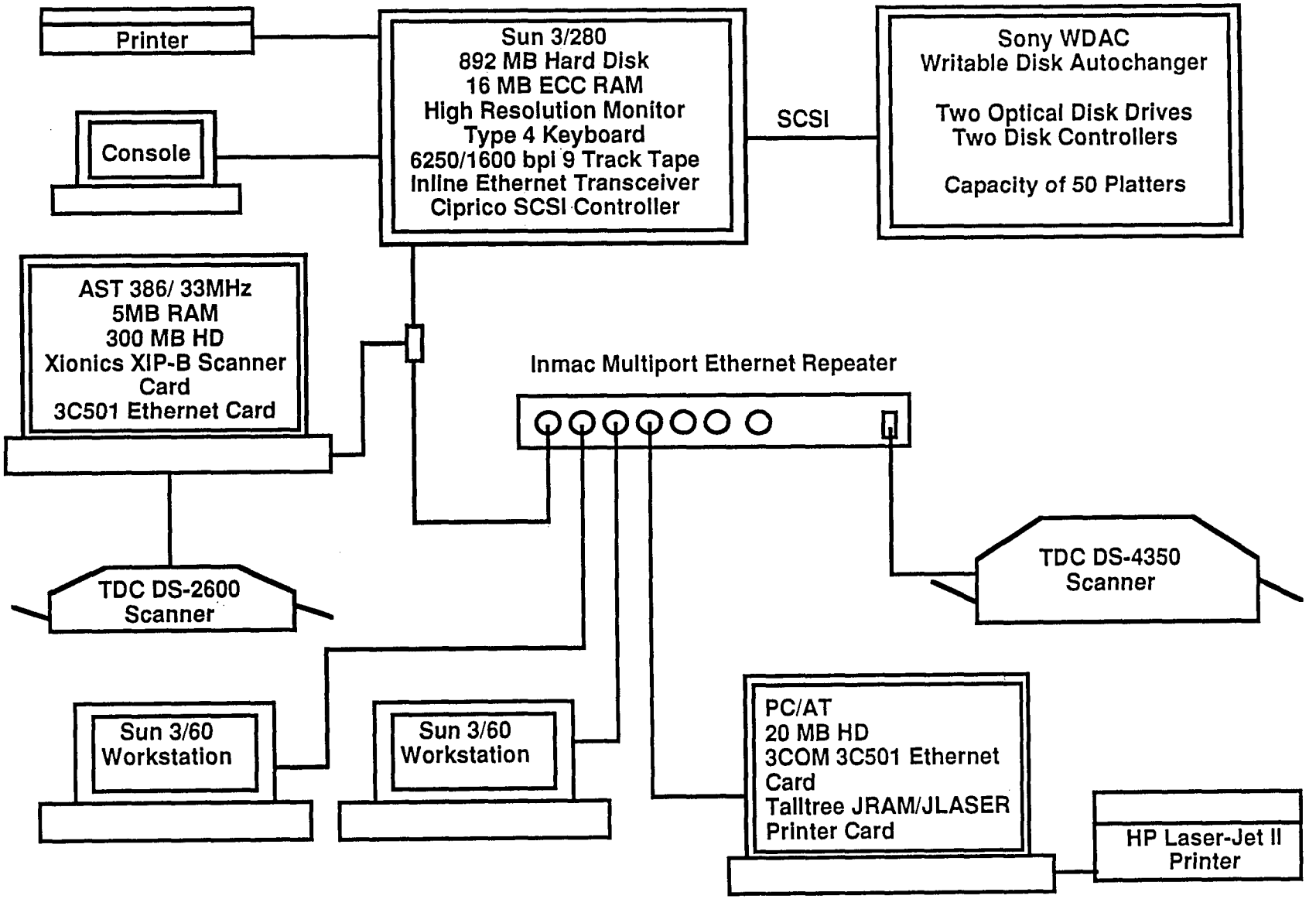
Downloaded images will be shared with other facilities by means of magnetic tape and/or 5.25 optical disks.

Printing speeds will be improved. Currently 8 pages/minute is considered very good indeed. Soon 16 pages will be considered barely satisfactory.

Scanners that not only scan in optical images but provide some OCR are being considered with a view to enlarging the retrieval data base.

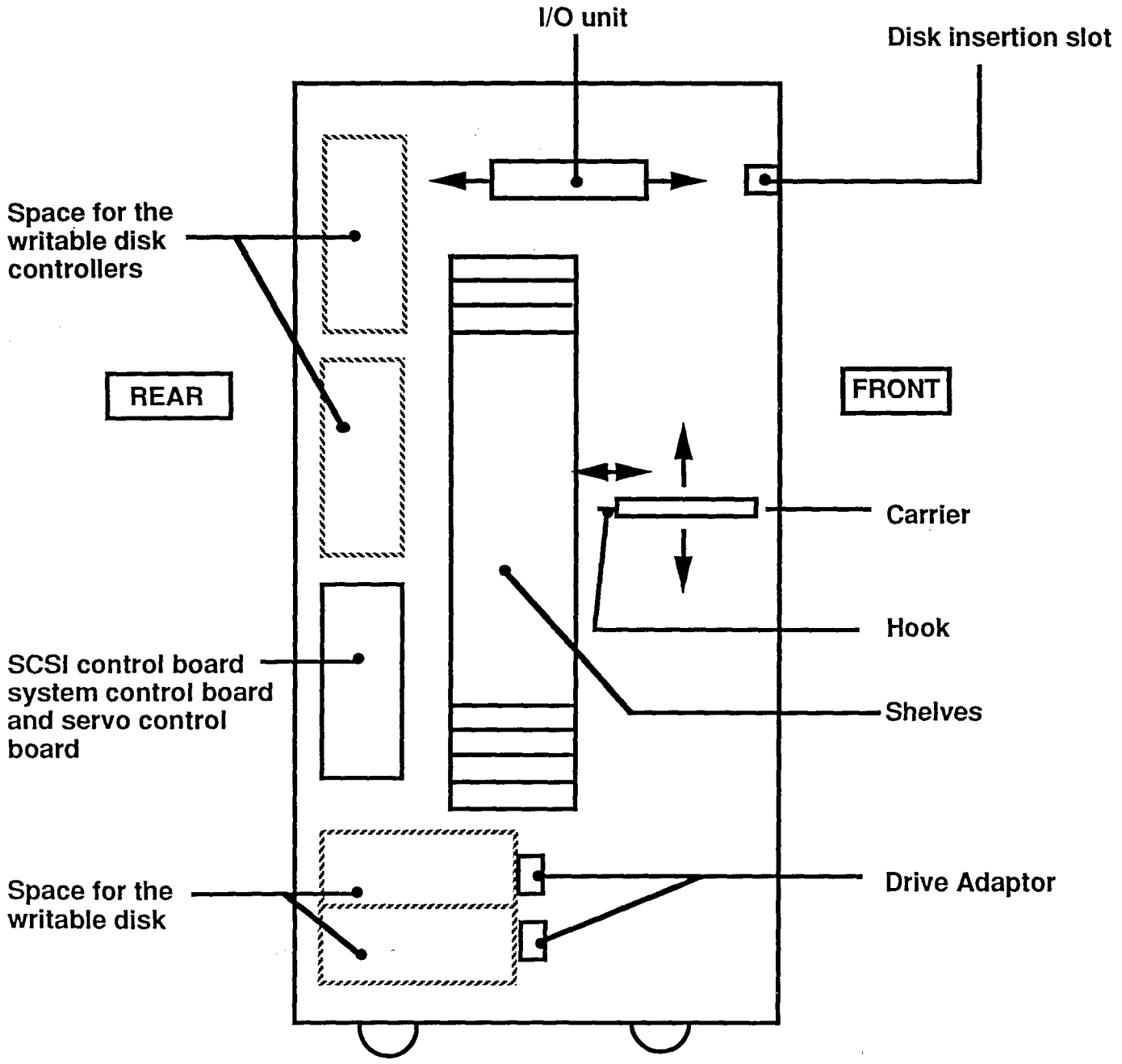
The retrieval of images on the screen will be made even more user friendly and easier to read.

Once equipment like the Optical Disk System is in place, possible refinements become obvious and a wish list grows rather quickly - more work stations for patrons and staff, more information in the retrieval data base, faster scanners, faster printers, remote access and all the new products that this rapidly advancing technology is going to offer.



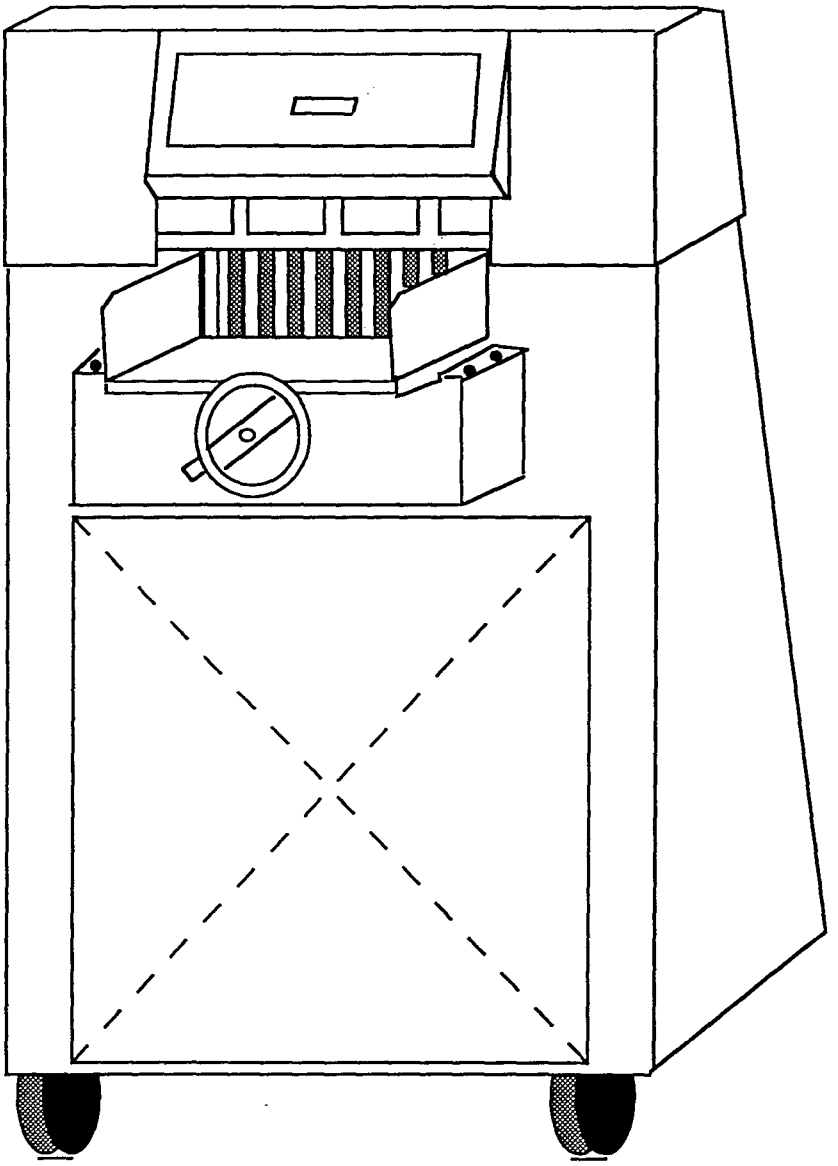
NRL Document Retrieval System

Figure 1



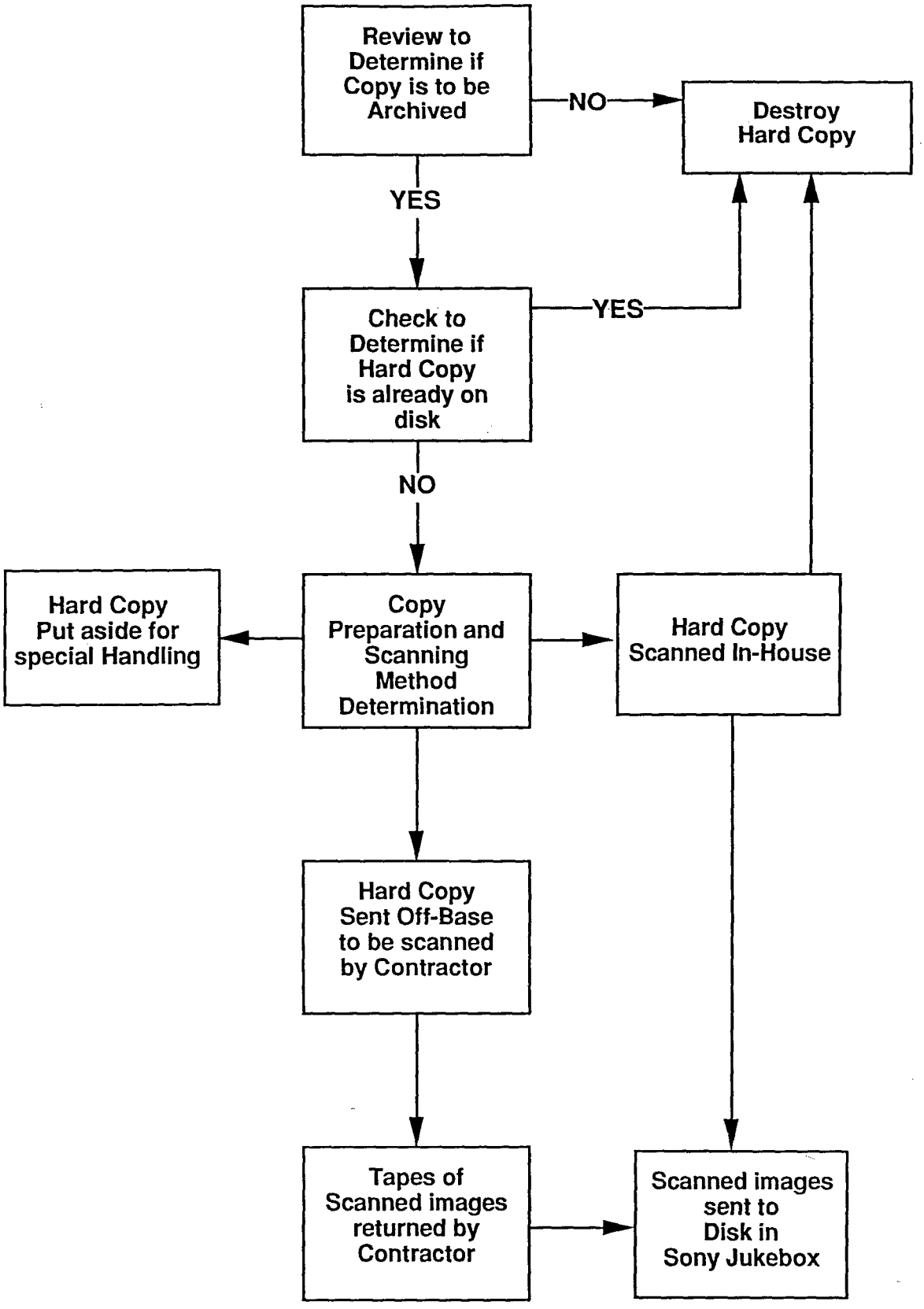
SONY WDA -610 AUTOCHANGER

Figure 2



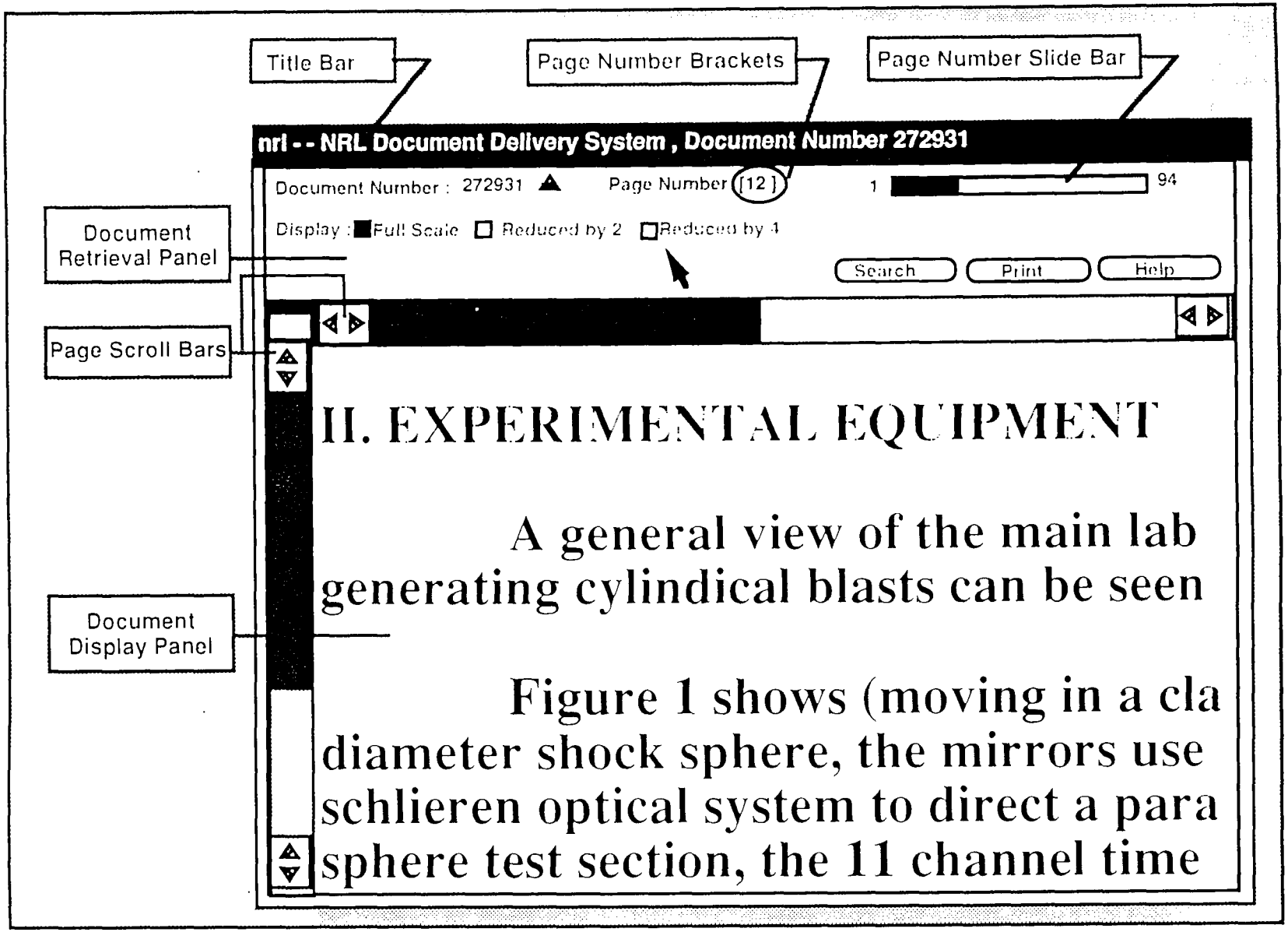
Challenge Automatic Paper Cutter
Model 20 Power

Figure 3



WORK FLOW FOR PROCESSING
HARD COPY FOR OPTICAL DISK

Figure 4



The NRL Document Delivery System Window

Figure 5

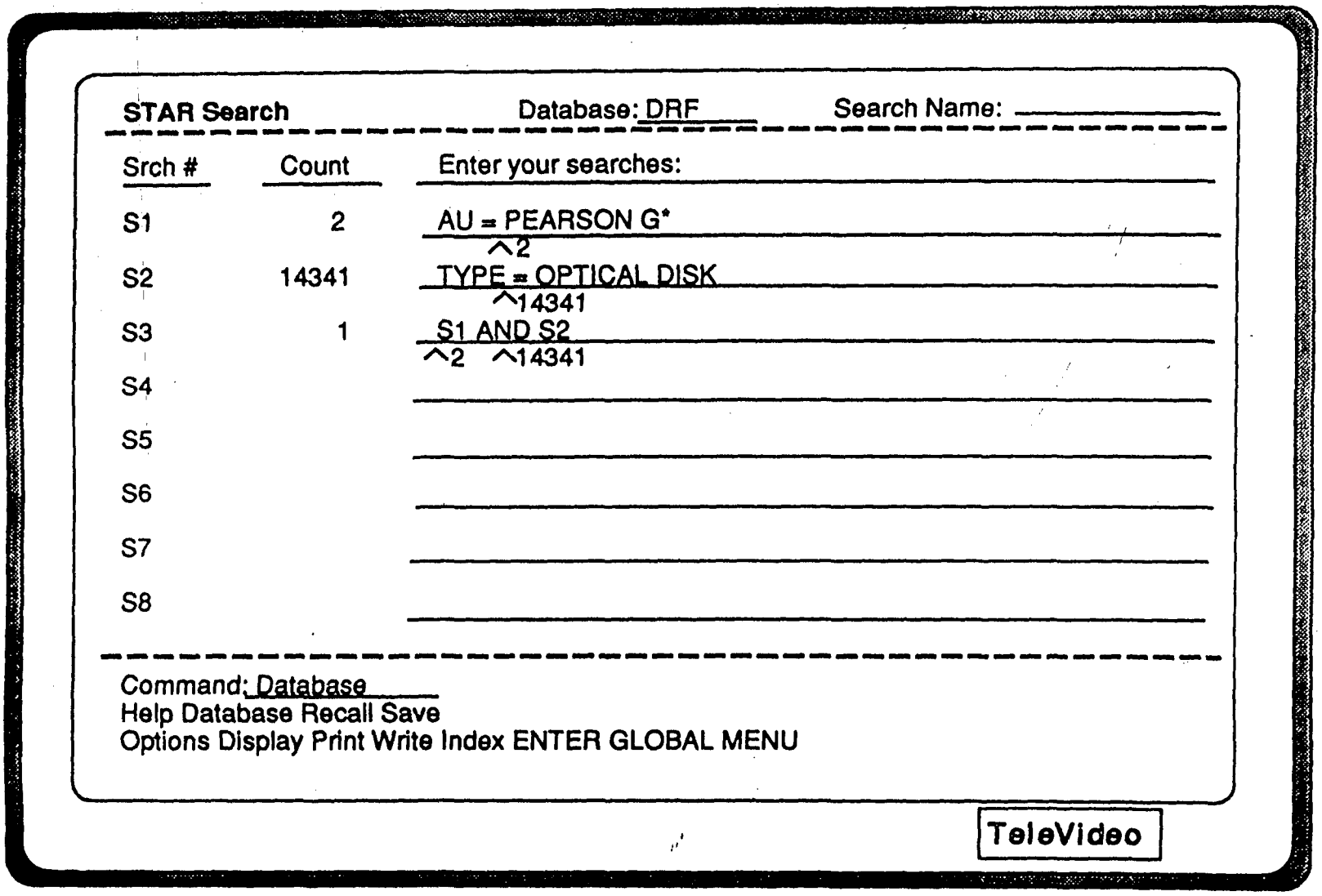

STAR Search
Database: DRF
Search Name:
Srch #
Count
Enter your searches:
S1
2
AU = PEARSON G*
^2
S2
14341
TYPE = OPTICAL DISK
^14341
S3
1
S1 AND S2
^2 ^14341
S4
S5
S6
S7
S8
Command: Database
Help Database Recall Save
Options Display Print Write Index ENTER GLOBAL MENU
TeleVideo

FIGURE 6

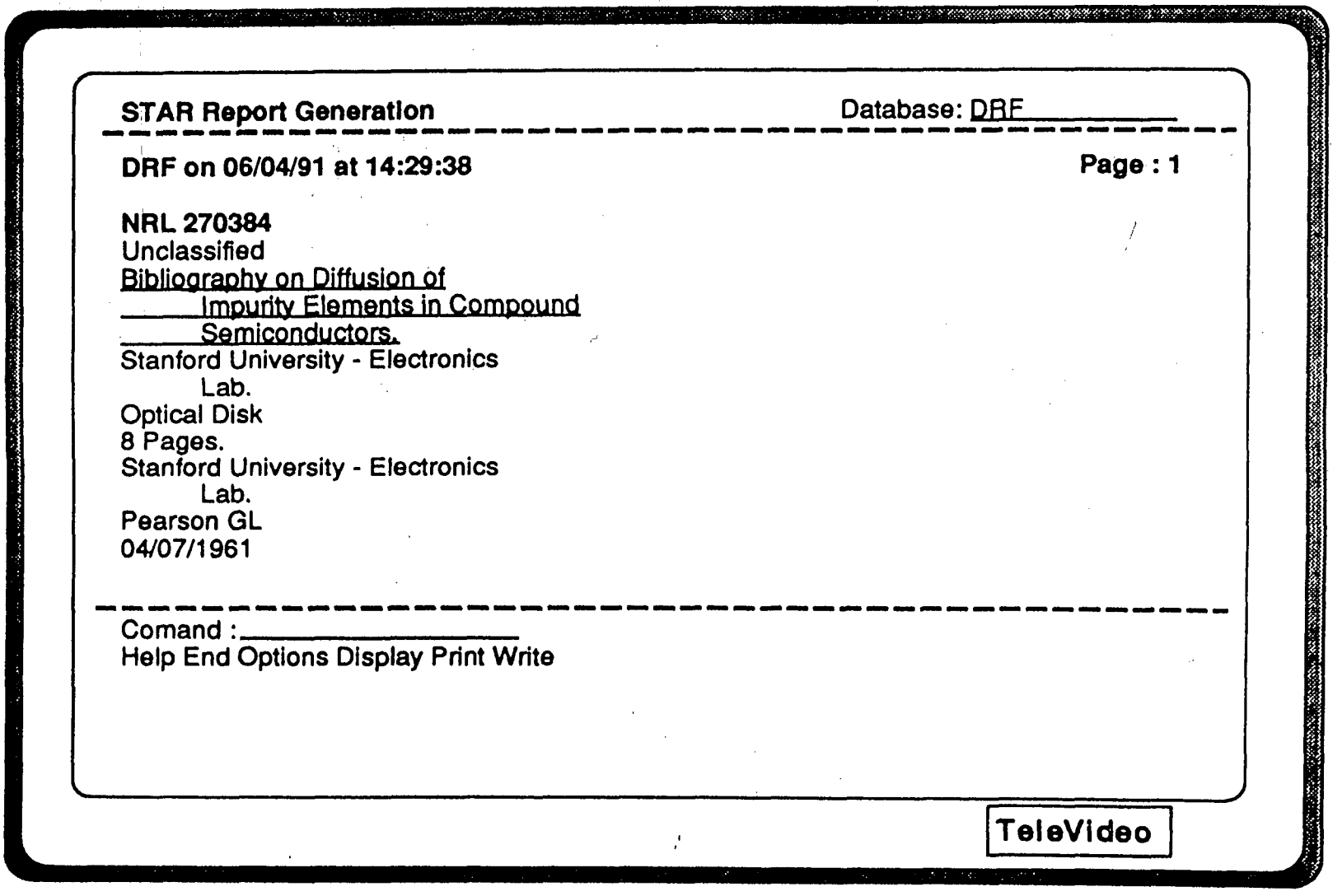

STAR Report Generation
Database: DRF
DRF on 06/04/91 at 14:29:38
Page : 1
NRL 270384
Unclassified
Bibliography on Diffusion of
Impurity Elements in Compound
Semiconductors.
Stanford University - Electronics
Lab.
Optical Disk
8 Pages.
Stanford University - Electronics
Lab.
Pearson GL
04/07/1961
Comand :
Help End Options Display Print Write
TeleVideo

FIGURE 7