MEMORANDUM FOR: Robert E. Browning, Director
Division of High-Level Waste Management

FROM: Avi Bender, Senior Project Manager
Engineering Section
Systems Engineering and Evaluation Branch

SUBJECT: TRIP REPORT, PRESENTATION OF PAPER AT OPTICA 87' AMSTERDAM, HOLLAND

The attached paper was presented at OPTICA 87', Holland. About 100 people from European industry and government attended my session and great interest was expressed at the technical approach NRC is using to address text and image records management to potentially streamline its licensing process. It was apparent that although our concept and approach are unique, there are a growing number of system vendors and integrators who will be capable of delivering similar products in the future. Of particular interest to the attendees was the way we have integrated document source capture from our 5520's and optical character recognition devices directly into a searchable full text retrieval system.

I was proud to represent the NRC at this international conference and the positive feedback I received was most encouraging. Given the leading role that we, as an agency, have established in the government for optical disk based systems, I believe that we should continue to share our experierces with DOE, other government agencies and industry groups in the future. Thank you for your support.

Avi Bender, Senior Project Manager
Engineering Section
Systems Engineering and Evaluation Branch

Enclosure:
As stated

cc: Thompson
Bernero
Amenta

FULL TEXT SEARCH AND IMAGE RETRIEVAL USING OPTICAL DISC TECHNOLOGY

A. Bender, U.S. Nuclear Regulatory Commission, USA

Keywords: Full Text Search, Optical Disc, Information Retrieval.

Abstract: The United States Nuclear Regulatory Commission is conducting a pilot project to demonstrate the application of optical disc technology and full text storage search and image retrieval for licensing records management. An open architecture system has been developed with off-the-shelf hardware and integrated into an existing office environment. The combination of document source capture and optical character recognition provides an effective means for building a searchable database.

1 INTRODUCTION

The United States Nuclear Waste Policy Act (NWPA) of 1982 established into law a Federal policy to manage the disposal of high level commercial nuclear waste (HLW). The Department of Energy (DOE) is required to identify potential sites for HLW disposal, recommend a site and submit an application to the Nuclear Regulatory Commission (NRC) for construction authorization.

The NWPA requires the NRC to review DOE's application within a period of three to four years and to deny or approve the application. The time frame allotted for license review, adjudicatory proceedings, and for reaching a final decision, is shorter than has traditionally been necessary for licensing nuclear power reactors. One of the most significant contributors to the length of past reviews has been the time associated with finding and processing relevant documents. The license application is expected to be submitted by DOE to the NRC in 1991. Millions of documents, which provided information for major site selection decisions, will need to be readily accessible for review by all the parties to the licensing process which include affected Indian tribal organizations and States. The complexity of the document management process is evident and is of most concern to the NRC which is mandated to license a HLW repository and protect public health and safety.

Given the adjudicatory nature of the licensing proceedings, lawyers and technical staff will need to quickly gain access to individual names, places, issues and other information within documents. Traditional indexing and document surrogate search systems do not provide an acceptable level of recall and precision which will be needed during the licensing proceedings.

instead, these systems restrict the inquirer to a few predetermined fields and key words. With full text search the ability to access relevant documents is less dependent on the required congruity between the vocabulary of the indexer and that of the user. The additional use of a thesaurus, clustered file arrangements and other techniques, make full text search an extremely powerful retrieval tool.

There have been a number of recent studies, however, which provide conflicting viewpoints on the merits of full text search. The often quoted Blair and Maron study (Ref. 1) concluded that full text provided very poor recall and precision for a large database in a major legal proceeding. A more recent study by Carol Tenopir (Ref. 2) concluded that full text search in combination with other search methods is superior to either surrogate searches or just full text searches. Tenopir concluded that "The presence of full text often allows articles to be retrieved that could not be found by searching on titles, controlled vocabulary descriptors or abstracts". Our experience during the pilot project supports this conclusion (Ref. 3).

Most integrated optical disc systems on the market today rely on simplistic indexing of documents for future retrieval and can be best characterized as optical disc <u>filing</u> systems. Our requirement for enhanced information retrieval has necessitated the creation of a searchable full text/image system. The concurrent search of both full text and a surrogate of the document coupled with a thesaurus is expected to provide a much more powerful content search capability than one could achieve using abstract, subject or key word searches. New innovative approaches, using full text storage, search and retrieval coupled with image capture, will be necessary to address the complex document management task which lies ahead.

2 NRC OPTICAL DISC PILOT PROJECT

A pilot project was initiated by the Division of Waste Management (DWM), in 1986, to demonstrate to DOE and NRC the potential application of an optical disc based full text search storage and retrieval system. The objective of the pilot is to demonstrate emerging document handling technologies as they may apply to facilitating NRC's future licensing review requirements and to integrate such technologies with the existing operational office environment.

Previous studies by the DWM had established some minimum system requirements. One of the most important requirement included the ability to quickly and accurately store, search for and retrieve the image of the document. The original image is of great importance since it is the true representation of the document and may have signatures and marginal notations which could have legal significance during the licensing proceedings. A further requirement was identified for enhanced document retrieval techniques to improve recall and precision and reduce potential future uncertainty associated with the system's reliability.

During the first phase of the pilot a full text search and retrieval system was developed using a mainframe based timesharing system. With IBM's STAIRS as the search software, users were able to query a correspondence database. Response time to a search query was between one to three seconds. Although the system represented a vast improvement over traditional manual search and retrieval, it was expensive and had an important limitation.

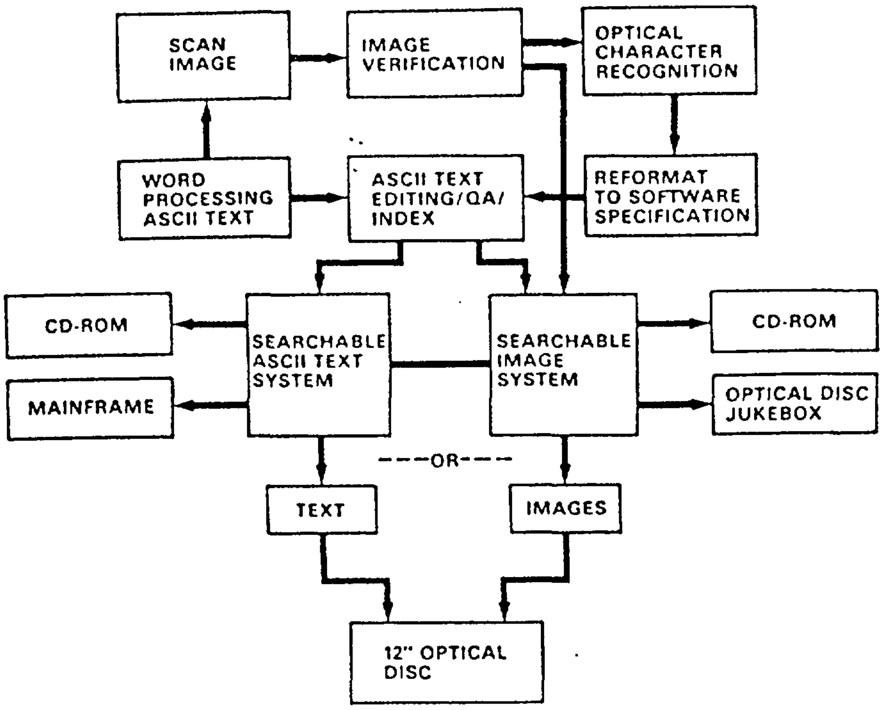
A user was still required to go to a paper filing system to obtain the original image of the document and all the associated attachments.

With a future need to mass distribute the document database in a cost effective and comprehensive manner, optical disc technology offers several opportunities. The distribution of a self contained full text search/image database on optical discs, including the search software and utilities for downloading at the host system, offers a significant cost reduction over on-line timeshare access systems. In addition, the potential storage densities and capacities of optical discs far exceed both magnetic disk and microfilm storage mediums. Full text search and retrieval coupled with optical discs as either strictly storage mediums or as a combination of storage and database management medium, appears to hold great promise for optimizing future text and image management requirements.

3 CREATING THE TEXT/IMAGE DATABASE

The most labor intensive and time consuming aspect of creating a searchable full text and image system is the process of converting text to ASCII. Although optical character recognition (OCR) systems are becoming increasingly competitive with rekeying, they cannot be relied upon completely for text to ASCII conversion. The DWM has, therefore, implemented procedures for source capture of documents to streamline the database building process. Over sixty percent of the documents produced by the DWM are created on an IBM 5520 word processing system. These documents can be saved as ASCII files, thus eliminating the need to rekey. The OCR is used for those documents received over which there is little control. The DWM has evaluated several OCR's and is presently using the PALANTIR, an omnifont recognition OCR and digitizer. Future procedures will be written to require contractors to submit their reports in both hard copy and in ASCII to further reduce the reliance on either character recognition or rekeying. Capturing the image is a simpler task. With an average scan time of three seconds per page, documents can be quickly captured and continuously stored on a 12" WORM optical disc. Each document that is scanned for image capture is tagged with an accession number which becomes the link with the full text retrieval of the ASCII file. The text and image capture process is illustrated in Figure 1.

Whether a document is received from an outside source or from our word processing system, it will be scanned and digitized in its entirety. Portions of the resulting image file will be sent to the OCR for character recognition (except for what is already available as ASCII from the word processing system). Not all documents can be readily converted to ASCII text and some presorting must take place. For example, a letter with several attachments such as photographs, maps, or poorly duplicated originals will be digitized



SCAN IMAGE
IMAGE VERIFICATION
OPTICAL CHARACTER RECOGNITION
WORD PROCESSING ASCII TEXT
ASCII TEXT EDITING/QA/ INDEX
REFORMAT TO SOFTWARE SPECIFICATION
CD-ROM
SEARCHABLE ASCII TEXT SYSTEM
SEARCHABLE IMAGE SYSTEM
CD-ROM
MAINFRAME
OPTICAL DISC JUKEBOX
---OR---
TEXT
IMAGES
12" OPTICAL DISC

Fig. 1. Text and Image Capture Process

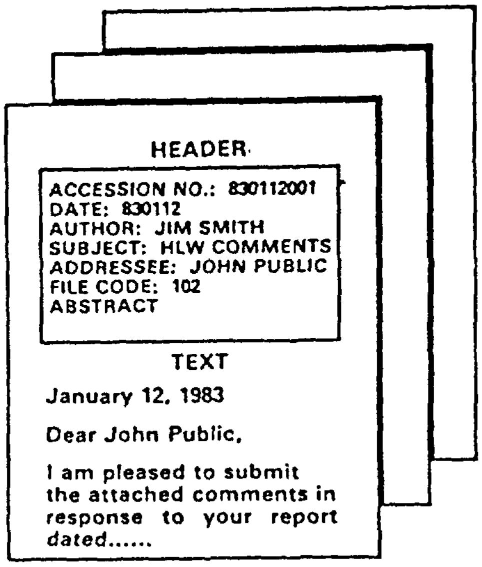
and stored on the optical disc, but only the cover letter will be available for ASCII conversion and subsequent full text search.

Once the documents are converted to ASCII the next step is preprocessing the text into a format suitable for the full text retrieval software. Text files sent to STAIRS must conform to a width not to exceed 69 characters, formatted and unformatted fields need to be defined and each paragraph must be coded using a three character alphanumeric on the left margin. This process has been automated. As the raw ASCII file is indexed by an operator, to input any or all of the information listed in Table 1, the coding is accomplished automatically. In addition, all text which exceeds the required width is reformatted by the same program to 69 characters.

Each document receives what we call a header. The header is a surrogate of the document containing some or all of the fields listed in Table 1. which is created during the coding process. The header plays an important role in future retrieval for several reasons. The header contains formatted and unformatted fields to facilitate Boolean type searches to narrow the resulting search documents to only those pertaining to a range of dates, authors, and subjects. Often times a search results in a large population of documents and a quick review of only a few fields may be desirable. Since the surrogate is physically attached as the front portion of the total document, a search of the system allows us to

TABLE 1. HEADER FIELDS

ACCESSION NO.
DOCUMENT DATE
AUTHOR
SUBJECT
ADDRESSEE
REPORT NO.
TITLE
SUBTITLE
FILE CODE
ABSTRACT



HEADER

ACCESSION NO.: 830112001
DATE: 830112
AUTHOR: JIM SMITH
SUBJECT: HLW COMMENTS
ADDRESSEE: JOHN PUBLIC
FILE CODE: 102
ABSTRACT

TEXT

January 12, 1983

Dear John Public,

I am pleased to submit the attached comments in response to your report dated......

Fig. 2. Sample Header (Surrogate) and Full Text

simultaneously browse either the full text or the surrogate. A sample of the header and the full text is illustrated in Figure 2.

The process described above for text and image capture has enabled us to integrate our office record management with full text storage search and retrieval capability. At this writing, the system's potential future configuration is being evaluated in light of the myriad of possible options now available to meet our requirements.

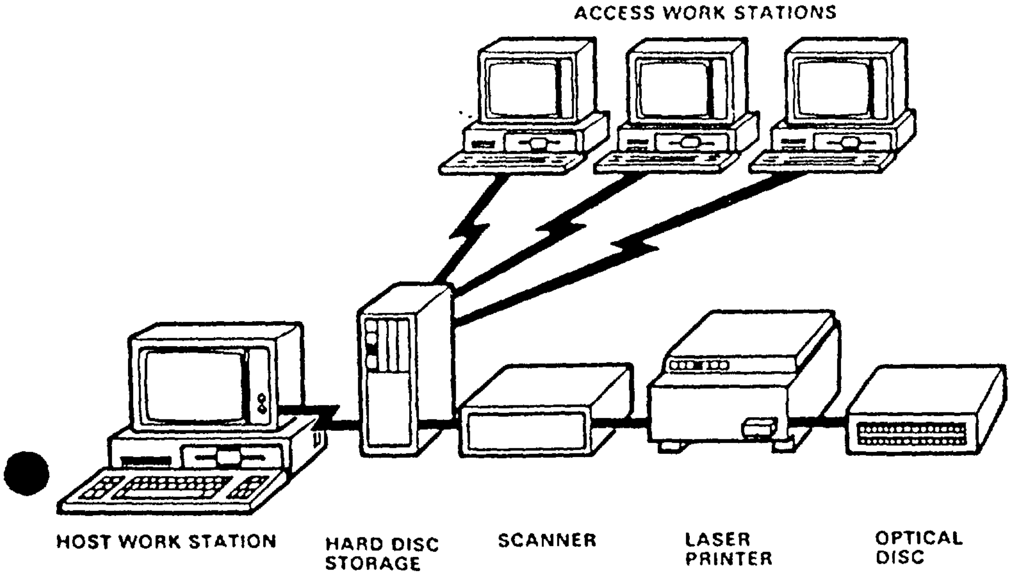


Fig. 3. Single Workstation Configuration

4 SYSTEM CONFIGURATION

The present pilot project consists of a single MS-DOS workstation configuration which is illustrated in Figure 3. Given the open architecture system approach and the integration of off-the-shelf hardware, a number of possible options will be available in the future (Ref. 4). The use of an Ethernet type communication network described in Figure 4 would provide the maximum flexibility for growth of multiple work stations and access to other systems. Given the rapid pace with which this new technology is evolving, the open architecture system will enable us to add or replace outdated devices.

The hardware consists of the host microcomputer, a scanner, an omnifont OCR, high resolution monitor,the optical disc drive and controller and ancillary boards for compression and decompression of the image.

Hardware integration assistance is being provided by the Smithsonian Institution. The major system components, at this writing, are as follows:

IBM PC AT Controller

The controller is a standard IBM PC AT Model 339. The PC AT utilizes the 80286 microprocessor operating at 8 Khz. 512 Kbytes of RAM memory is provided on the system board with an additional 2.0 Mbytes of above board RAM provided by a J-Ram AT3 add-on board. The system has one 1.2 Mbyte floppy disk drive one 360 Kbyte floppy disk drive and one 30 Mbyte high performance voice-coil activated winchester drive with an average access time of 35 ms. In addition, the controller contains one parallel and two serial ports.

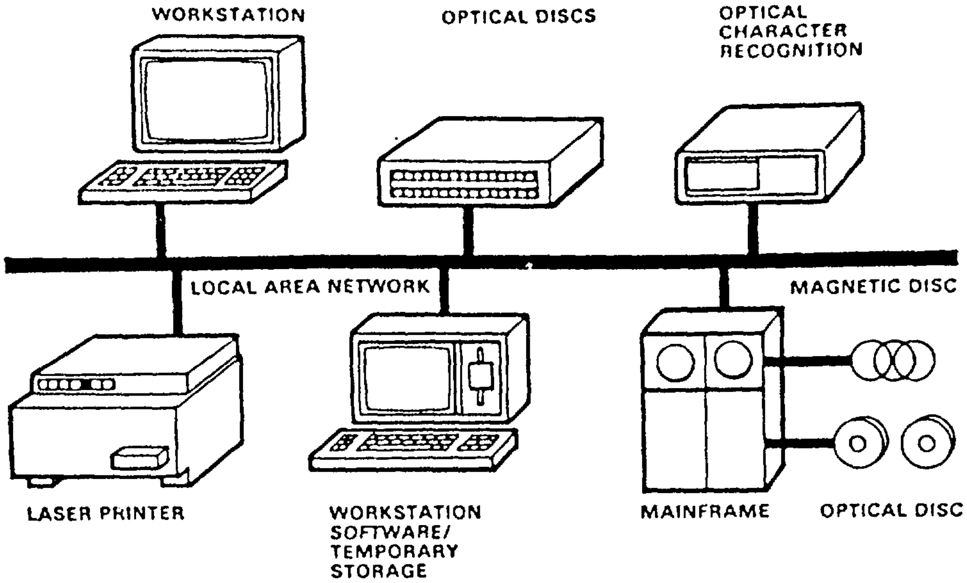


Fig. 4. Potential Future Communication Type Network

Laser Printer

The printer is a Ricoh Laser Printer. which is capable of printing up to eight pages per minute, depending on the type of output. The printer has horizontal and vertical resolutions of 300 DPI, and can therefore produce extremely accurate copies of original documents.

Image Scanner

The image scanner is the Ricoh IS-400. The IS-400 is a desktop sized device that is capable of scanning flat sheets as well as books at a resolution of up to 400 dots per inch. Scanning speed is rated at 2.5 seconds per page. Scanning density control is accomplished automatically by the system, however manual override capability is also available. The time required to scan 500 pages of 8 1/2" x 11" documents, exclusive of operator intervention, is approximately 21 minutes. However, operator intervention is required for document presorting, collating and sequencing.

Magnetic Mass Storage Device

The magnetic mass storage device used for temporary storage of image and ASCII files, as well as all applications software is the Priam CT02-06A Shared Space. The Priam tower is a 190 Mbyte high speed floor standing Winchester disk drive with an eleven head configuration and an average seek time of 25 milliseconds. The Priam device is equipped with a 60 megabyte capacity tape backup subsystem, and an average transfer rate of 86.7 Kbytes per second. The tape subsystem allows backup of all data as required. The Priam can also be configured for shared user access that would allow multiple use of the storage space for both data acquisition, data preparation and header(document surrogate)addition, as well as database search and retrieval.

High Resolution Display Subsystem

The high resolution display subsystem is provided by Laserdata. The display subsystem consists of a scanner controller/compression and decompression processor boards, processor/monitor controller and laser printer controller, and a high resolution monitor.

Laser/optical Storage Devices

The Laser/Optical Disc Drive selected for the NRC pilot System is the Sony WDD-3000 write once read many (WORM) drive with removable disc platters. Each disc has an active storage capacity of 3.2 gigabytes of information. Data transfer rate is 5.18 Mbits per second with an average head seek time of .35 seconds.

The CD-ROM Drive that will be used for text storage is the Sony CDU series with a storage capacity of 540 megabytes. This CD-ROM Disc system uses 5 1/4" compact discs, which would enable mass distribution of the text data.

Laser/Optical Disc Drive Interface

The Instar Optical Disc interface has been chosen as the interconnect between the Sony WDD-3000 and the IBM PC AT Controller. The interface employs the standard small computer systems interface (SCSI) board and protocol. The Instar software provides the directory management routines required for management of data being written to the laser disc.

Character Recognition Engine

The character recognition engine is the PALANTIR Compound Document Processor (CDP). The CDP is a tabletop OCR unit capable of recognizing over 15,000 different font styles and character sizes from 6 to 26 point. At this writing system enhancements were being announced for recognition of dot matrix print and improved recognition speed.

Full Text Search Software

Several microcomputer based full text search software packages are being evaluated at this writing. Each package will be judged on the basis of how it meets the needs of the NRC pilot Project. In the current mainframe system configuration, IBM,s STAIRS is the search engine. For the single microcomputer based workstation, we are evaluating Bluefish by Computer Access Corporation and will be looking at other systems such as BRS.

OCR and Indexing Utility Software

The NRC has developed its own unique software utilities to streamline the database creation process. The program allows the user to take a file from an OCR or a word processing system and convert it to a workable ASCII file which can be recognized by the host full text storage and retrieval software package. The text can then be processed by a menu driven program which prompts the indexer to assign codes. The program automatically creates the necessary coding for the host software.

5 SUMMARY

Optical disc based systems have created new opportunities for information management. To make full use of their power industry standards will need to be established. In addition, greater attention should be focused on information retrieval, particularly content value of information within documents. Current indexing and document surrogation approaches may be outdated techniques which need to be reassessed in light of the new technological advances which have changed the basic construct of our approach to information management.

The DWM pilot project has demonstrated that the technology is available to develop cost effective document management systems. Further work will be required to streamline the database creation process and to complete comprehensive operating and record management procedures.

6. ACKNOWLEDGMENT

Special thanks to Joseph Bunting and Philip Altomare for having the vision and leadership which was necessary to bring this effort to fruition, and to Ms. Sharon Root for her invaluable support as the system administrator.

7 REFERENCES

1. Blair, C.D. and Maron, M.E. "An Evaluation of Retrieval Effectiveness for a Full Text Document Retrieval System." Communications of the ACM (28) 3, March 1985.

2. Tenopir, C. "Full Text Database Retrieval Performance". Online Review Vol.9, No.2, April 1985.

3. Bender, A. "Application of a Full Text Storage and Retrieval System for Records Management". Journal of Information and Image Management, Vol. 19, No. 4, April 1986.

4. Bender, A. "Optical Disc Document Management System: Design Considerations". Submitted for February 1987 publication, Journal of Information and Image Management.