2

# NAVAL POSTGRADUATE SCHOOL

## Monterey, California

AD-A247 481

DTIC
ELECTE
MAR 16 1992
S B D

# THESIS

CONVERSION AND RETRIEVABILITY OF
HARD COPY AND DIGITAL DOCUMENTS
ON OPTICAL DISKS

by

Lawrence P. Bittner

March, 1992

Thesis Advisor: Barry A. Frew

Approved for public release; distribution is unlimited

92-06732

Best
Available
Copy

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

# REPORT DOCUMENTATION PAGE

| | |
|---|---|
| 1a REPORT SECURITY CLASSIFICATION<br>UNCLASSIFIED | 1b RESTRICTIVE MARKINGS |
| 2a SECURITY CLASSIFICATION AUTHORITY | 3 DISTRIBUTION/AVAILABILITY OF REPORT<br>Approved for public release; distribution is unlimited |
| 2b DECLASSIFICATION/DOWNGRADING SCHEDULE | |
| 4 PERFORMING ORGANIZATION REPORT NUMBER(S) | 5 MONITORING ORGANIZATION REPORT NUMBER(S) |
| 6a NAME OF PERFORMING ORGANIZATION<br>Naval Postgraduate School<br>6b OFFICE SYMBOL *(If applicable)*<br>37 | 7a NAME OF MONITORING ORGANIZATION<br>Naval Postgraduate School |
| 6c ADDRESS *(City, State, and ZIP Code)*<br>Monterey, CA 93943-5000 | 7b ADDRESS *(City, State, and ZIP Code)*<br>Monterey, CA 93943-5000 |
| 8a NAME OF FUNDING/SPONSORING ORGANIZATION<br>8b OFFICE SYMBOL *(If applicable)* | 9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER |
| 8c ADDRESS *(City, State, and ZIP Code)* | 10 SOURCE OF FUNDING NUMBERS<br>Program Element No. / Project No. / Task No. / Work Unit Accession Number |

11 TITLE *(Include Security Classification)*

CONVERSION AND RETRIEVABILITY OF HARD COPY AND DIGITAL DOCUMENTS ON OPTICAL DISKS

12 PERSONAL AUTHOR(S) Bittner, Lawrence P.

| | | | |
|---|---|---|---|
| 13a TYPE OF REPORT<br>Master's Thesis | 13b TIME COVERED<br>From To | 14 DATE OF REPORT *(year, month, day)*<br>1992, March | 15 PAGE COUNT<br>82 |

16 SUPPLEMENTARY NOTATION

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government

17 COSATI CODES

| FIELD | GROUP | SUBGROUP |
|---|---|---|
| | | |
| | | |

18 SUBJECT TERMS *(continue on reverse if necessary and identify by block number)*

Optical Disks, CD-ROM, Document Retrieval, Information Retrieval, Digital Documents, Document conversion, Image Scanners, Color Scanners, Full Text Document Database, Optical Storage, OCR, ICR, Full Text Retrieval

19 ABSTRACT *(continue on reverse if necessary and identify by block number)*

Paper documents can be converted into digital form, as a collection of images, or a combination of ASCII text and images. Full text and image document databases, display advantages and disadvantages during scanning and conversion processes. Conversion of paper thesis documents could be eliminated, if thesis documents could be submitted in digital form, for storage on optical disks. Utilizing existing paper thesis documents, image and full text databases were developed and evaluated to determine the best digital form for storage of paper documents. Analysis was performed on a thesis document in digital form, to determine the most feasible format for digital document submission. This thesis concludes that conversion of paper documents to digital form should not be pursued. Instead, thesis documents should be submitted in digital form for direct conversion and storage on optical disks. Follow on thesis research is recommended to build an in-house CD-ROM mastering system for this purpose.

| | |
|---|---|
| 20 DISTRIBUTION/AVAILABILITY OF ABSTRACT<br>☒ UNCLASSIFIED/UNLIMITED □ SAME AS REPORT □ DTIC USERS | 21 ABSTRACT SECURITY CLASSIFICATION<br>Unclassified |
| 22a NAME OF RESPONSIBLE INDIVIDUAL<br>Barry A. Frew | 22b TELEPHONE *(Include Area code)*<br>(408) 646-2392<br>22c OFFICE SYMBOL<br>Code 37 |

83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

Unclassified

Approved for public release; distribution is unlimited.

# Conversion and Retrievability of Hard Copy and Digital Documents on Optical Disks

by

Lawrence P. Bittner
Lieutenant, United States Navy
B.S., North Dakota State University, 1983

Submitted in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE IN INFORMATION SYSTEMS MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
March, 1992

Author: ______________________
Lawrence P. Bittner

Approved by: Barry Frew
Barry A. Frew, Thesis Advisor

______________________
C. Thomas Wu, Second Reader

______________________
David Whipple, Chairman
Department of Administrative Sciences

# ABSTRACT

Paper documents can be converted into digital form, as a collection of images, or a combination of ASCII text and images. Full text and image document databases, display advantages and disadvantages during scanning and conversion processes. Conversion of paper thesis documents could be eliminated, if thesis documents could be submitted in digital form, for storage on optical disks.

Utilizing existing paper thesis documents, image and full text databases were developed and e luated to determine the best digital form for storage of paper documents. Analysis was performed on a thesis document in digital form, to determine the most feasible format for digital document submission.

This thesis concludes that conversion of paper documents to digital form should not be pursued. Instead, thesis documents should be submitted in digital form for direct conversion and storage on optical disks. Follow on thesis research is recommended to build an in-house CD-ROM mastering system for this purpose.

Accession For
NTIS GRA&I ☑
□
□
Availability Codes
A-1

## TABLE OF CONTENTS

I. INTRODUCTION . . . . . . . . . . . . . . . . . . . . 1

A. DISCUSSION . . . . . . . . . . . . . . . . . . 1

B. SCOPE . . . . . . . . . . . . . . . . . . . . . 2

C. METHODOLOGY . . . . . . . . . . . . . . . . . . 2

II. OPTICAL TECHNOLOGY . . . . . . . . . . . . . . . . 4

A. INTRODUCTION . . . . . . . . . . . . . . . . . 4

B. OPTICAL SCANNERS . . . . . . . . . . . . . . . 4

1. Scanner Technology . . . . . . . . . . . . . 4

2. Scanning Images . . . . . . . . . . . . . . 6

a. Resolution . . . . . . . . . . . . . . . 6

b. Levels of Greyscale . . . . . . . . . . 7

c. Digitizing Color Images . . . . . . . . 8

3. Optical Character Recognition . . . . . . . 8

a. Matrix Matching . . . . . . . . . . . . 9

b. Feature Matching . . . . . . . . . . . . 9

C. COMPACT DISK READ ONLY MEMORY . . . . . . . . . 10

1. Basic Concepts . . . . . . . . . . . . . . . 10

2. CLV vs CAV . . . . . . . . . . . . . . . . . 11

3. Physical Structure . . . . . . . . . . . . . 11

4. Advantages and Disadvantages of CD-ROM . . . 14

III. INFORMATION RETRIEVAL . . . . . . . . . . . . . . . . . 15

A. INTRODUCTION . . . . . . . . . . . . . . . . . . . . 15

B. DOCUMENT RETRIEVAL VS DATA RETRIEVAL . . . . . 16

1. Response to a Query . . . . . . . . . . . . . 16

2. Question/Answer Relationship . . . . . . . . . 17

3. Utility - The Criterion for Successful Retrieval . . . . . . . . . . . . . . . . . . . 17

4. Influence on Retrieval Speed . . . . . . . . . 18

C. FULL TEXT DOCUMENT RETRIEVAL . . . . . . . . . . 18

1. Index Formulation For Document Retrieval. . 19

2. Difficulties Associated with Document Retrieval . . . . . . . . . . . . . . . . . . . 20

D. MEASURES OF DOCUMENT RETRIEVAL EFFECTIVENESS . 24

E. SEARCHING FEATURES OF FULL TEXT DOCUMENT RETRIEVAL . . . . . . . . . . . . . . . . . . . . 29

1. Word Searching . . . . . . . . . . . . . . . . 29

2. Phrase Searching . . . . . . . . . . . . . . . 30

3. Proximity Searching . . . . . . . . . . . . . 30

4. Boolean Operators . . . . . . . . . . . . . . 31

5. Wildcard Operators . . . . . . . . . . . . . . 31

F. BROWSING'S ROLE IN DOCUMENT RETRIEVAL . . . . . 32

IV. DATABASE DEVELOPMENT . . . . . . . . . . . . . . . . . 34

A. INTRODUCTION . . . . . . . . . . . . . . . . . . . . 34

B. DEVELOPMENT OF TEXT AND IMAGE DATABASES . . . . 35

1. Document Preparation and Scanning . . . . . 36

2. Creation of an Image for each Page. . . . . 36

3. Optical Character Recognition . . . . . . . 38

4. Text Preparation and Markup . . . . . . . . 38

a. Text/Image Database . . . . . . . . . . 38

b. Image/Text Database . . . . . . . . . . 39

5. Automatic Index Generation . . . . . . . . . 40

6. Thesis Conversion Time . . . . . . . . . . . 41

C. PERFORMANCE AND DEVELOPMENT TRADEOFFS . . . . . 44

1. Development Decisions and Tradeoffs. . . . . 44

2. Database Performance . . . . . . . . . . . . 47

a. Level of Document Subdivision . . . . . 48

b. Retrieval Advantages and Disadvantages . 49

(1) Text/Image Advantages and Disadvantages . . . . . . . . . . . 50

(2) Image/Text Advantages and Disadvantages . . . . . . . . . . . 50

D. SUMMARY . . . . . . . . . . . . . . . . . . . . 51

V. SUBMISSION OF THESIS DOCUMENTS IN ELECTRONIC FORM . 53

A. INTRODUCTION . . . . . . . . . . . . . . . . . 53

B. CURRENT THESIS PREPARATION TOOLS . . . . . . . 54

1. Thesis Preparation using G-Thesis . . . . . 55

2. Thesis Preparation using Framemaker . . . . 56

3. Thesis Preparation using WordPerfect 5.1 . . 56

C. FORMAT REQUIREMENTS FOR ELECTRONIC DOCUMENTS . 57

D. CONVERSION OF A DOCUMENT IN WORDPERFECT FORMAT 60

E. CONVERSION OF A DOCUMENT IN ASCII TEXT FORMAT . 63

F. SUMMARY . . . . . . . . . . . . . . . . . . . . 64

VI. CONCLUSIONS, RECOMMENDATIONS . . . . . . . . . . . 66

A. CONCLUSIONS . . . . . . . . . . . . . . . . . . 66

B. RECOMMENDATIONS . . . . . . . . . . . . . . . . 67

LIST OF REFERENCES . . . . . . . . . . . . . . . . . . 70

BIBLIOGRAPHY . . . . . . . . . . . . . . . . . . . . . 72

INITIAL DISTRIBUTION LIST . . . . . . . . . . . . . . 73

## LIST OF TABLES

TABLE 1. BYTE UTILIZATION PER SECTOR . . . . . . . . . 13

TABLE 2. PRECISION & RECALL RATIOS . . . . . . . . . . . 25

TABLE 3. TIME REQUIREMENTS FOR DATABASE DEVELOPMENT . 40

TABLE 4. IMAGE/TEXT DATABASE CALCULATIONS . . . . . . 41

TABLE 5. TEXT/IMAGE DATABASE CALCULATIONS . . . . . . 43

TABLE 6. CONVERSION OF A DOCUMENT IN WORDPERFECT FORMAT . . . . . . . . . . . . . . . . . . . . 62

TABLE 7. CONVERSION OF A DOCUMENT IN ASCII TEXT FORMAT 64

## LIST OF FIGURES

Figure 1 Scanner Operation . . . . . . . . . . . . . 5

Figure 2 Feature Matching . . . . . . . . . . . . . . 10

Figure 3 Microscopic Structure of CD-ROM Disc . . . . 12

Figure 4 CLV vs CAV . . . . . . . . . . . . . . . . . 13

Figure 5 Documents Desired . . . . . . . . . . . . . . 22

Figure 6 Intersection of A & B . . . . . . . . . . . . 23

Figure 7 Terms A AND B AND D . . . . . . . . . . . . . 23

Figure 8 A AND C AND D . . . . . . . . . . . . . . . . 23

Figure 9 Precision - Recall Relationship . . . . . . 26

Figure 10 Variables of Relevance Judgements . . . . 29

Figure 11 Effect of Index Term number on Presicion and Recall . . . . . . . . . . . . . . . . . . . . 46

# I. INTRODUCTION

## A. DISCUSSION

The storage of thesis documents, at the Knox Library aboard Naval Postgraduate School, requires a great deal of space. These documents can be converted to digital form, for storage on optical disks, using optical scanners and optical disk publishing software. Documents can be digitized and stored, using one of two fundamental forms; as bit mapped images, or as full text documents with images of graphics. Each form provides advantages and disadvantages, with respect to document conversion and retrieval. A CD-ROM optical disk will hold approximately 1200 thesis documents (135,000 to 207,500 pages). Since information stored on optical disks, is only retrievable using a computer system, it is vitally important that, different topical areas within this information, be accessible, without spending a lot of time or energy.

The advent of word processors for computers, means that, thesis documents exist in digital form during their formulation and completion. It would be advantageous to receive completed thesis documents in digital form, for direct storage on optical disks.

This thesis analyzes the advantages and disadvantages of storage and retrieval of thesis documents in both image and full text form. It also analyzes the benefits and drawbacks of submitting thesis documents in electronic form.

## B. SCOPE

This thesis includes an in-depth analysis of an image, and full text database. The purpose of this analysis is to provide an accurate determination of the advantages and disadvantages of each digitized form, for conversion and storage of existing documents on optical disks.

Analysis of requirements for submission of completed thesis documents in electronic form is also conducted, and a recommendation for possible implementation is made.

## C. METHODOLOGY

Utilizing thesis documents presently stored at Knox Library, an image and full text database was developed by optically scanning documents, for image generation and optical character recognition. Optical disk publishing software was used to link images of graphics with full text of documents, and perform markup and indexing of documents, for searching and retrieval. Each database was evaluated to determine which provided the greatest benefits to the user (searcher).

Utilizing a digital copy (text with separate images of graphics), of a completed thesis, an experiment was conducted to determine the requirements for turning in thesis documents in electronic form. This experiment focused on the advantages and disadvantages pertaining to a digital document in word processor format; WordPerfect 5.1, versus a document in ASCII text format.

# II. OPTICAL TECHNOLOGY

## A. INTRODUCTION

The generation of a digital database for this thesis, and the retrieval of that information, primarily deals with two areas of optical technology: Optical Scanners and Compact Disk Read Only Memory (CD-ROM). This section will discuss in general, the technologies associated with scanners: their basic operation, optical character recognition (OCR), and image scanning. The basic operation, structure, advantages, and disadvantages of CD-ROM are discussed.

## B. OPTICAL SCANNERS

Optical scanners convert hard copy documents into a digital format. This is accomplished by converting the document into a bit mapped image or into text using OCR. The following sections will describe in general: the basic theory of how a scanner works, the different characteristics of a scanned image, and optical character recognition.

### 1. Scanner Technology

There are three basic types of optical scanners: moving paper scanners, flat bed scanners, and electronic digitizing camera scanners. The primary difference among these scanning technologies is the method of document illumination (or document transport) during the scanning process. (Fowler and Clipper, 1991, pp. 44-46)

The scanning process is basically accomplished by five components: a group of Charged Coupled Devices (CCD); a light source; a set of optics to force the light image onto the CCD; electronics to distinguish between dark (inked) areas, and light (white) areas; and a transport mechanism to move either the paper, or the scanning device depending on the type of scanner.[1] (Wageman, 1989, pp. 3.112.4 -3.112.6) Figure 1 portrays the interaction of components during the scanning process.

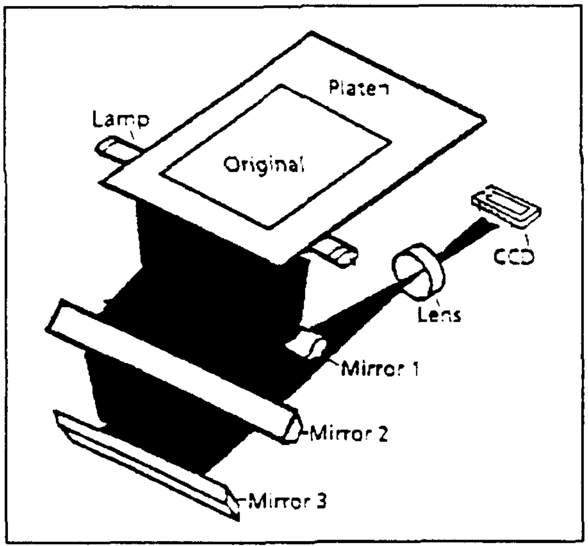


**Figure 1** Scanner Operation
(Taylor, 1987, p. 10)

[1] A Charged Coupled Device is a light sensitive semiconductor which produces an electrical signal which is proportional to the light incident upon it.

During the process of scanning, a document is illuminated, a strip at a time by a light source. The movement of the document relative to the light source is accomplished by the transport mechanism. The transport mechanism either moves the scanning device (light source) across the document, or moves the document across the scanning surface. The reflected light is directed via a set of optics, onto a series of CCD's, which transform the optical signal into an electrical signal. The electrical signal is converted into binary one's and zero's, which represent the scanned image, created by electronics, and the scanner software.

### 2. Scanning Images

Unlike text in ASCII format, an image is not directly accessible. It is a two dimensional bit mapped representation. There are three characteristics used to discuss image scanning: resolution, levels of greyscale, and color. These characteristics of image scanning are discussed in the following sections.

#### a. Resolution

Resolution is a measure of the level of detail represented in an image. Resolution is typically discussed in terms of pixels, or dots per inch (dpi). Pixel, originates from a television industry term; "picture element[2]." Dpi refers to the number of pixel dots used to represent that image. A resolution of 400 dpi implies that, every square inch of that image is represented by 160,000 pixel dots (400 x 400). An image with a resolution of 400 dpi contains more detail and thus better quality, then an image with a resolution of 200 dpi.

Image resolution is often confused with display resolution. Image resolution refers to the resolution of the stored image. Unlike display resolution, which refers to the number of pixels the screen can display. An image with a resolution less than the display will only take up a portion of the display; where as an image with a resolution greater than the display will encompass an area greater then the display. (Apperson and Doherty, 1987, pp. 124-126)

### b. Levels of Greyscale

The number of bits stored per pixel determines the amount of information that can be stored for each pixel. The ability to store this additional information distinguishes black and white images from greyscale images. In a black and white image, each pixel is represented by one

[2] The detail of a television picture image is determined by a combination of picture elements per horizontal scan line, and the number of horizontal scan lines per picture. In the United States a picture contains 525 scan lines with 435 picture elements per scan line. This equates to a resolution of 435 x 525. (Fink, 1984, pp. 83-84)

bit which corresponds to either black or white. In a greyscale system two or more bits are stored for each pixel. The number of grey levels which can be represented by each pixel is determined by two raised to the power of the number of bits stored per pixel ($2^x$). For example in a four bit per pixel system, each pixel could be one of 16 shades of grey.

### c. Digitizing Color Images

The representation of color images is very similar to that of greyscale images. Most color scanners on the market today produce color images by scanning a document three times in succession; once for each of the primary colors: red, green, and blue. Each of the three scans uses as its light source, one of three mercury vapor lamps which produce red, green, or blue light. Each of the three scans produces a color bit map, which is later merged to produce the color image. The number of bits stored for each of the primary colors, determines the displayable colors. Storing four bits per primary would allow 4096 colors ($2^4 \times 2^4 \times 2^4 = 4096$). (Apperson and Doherty, 1987, pp. 143-146)

## 3. Optical Character Recognition

Optical Character Recognition (OCR), or Intelligent Character Recognition (ICR), is the process by which the text (alphanumeric characters) of a scanned image is converted, from a bit mapped representation into ASCII characters. There are two primary methods used in this process; matrix matching and feature matching.

### a. Matrix Matching

The process of matrix matching breaks the image into a matrix of blocks, corresponding to character locations. These blocks are subdivided into a matrix of pixels. The pixel matrix of the character block, is then compared with those of the standard alphanumeric characters, until a match is found that satisfies a preprogrammed level of confidence.

### b. Feature Matching

The process of feature matching is similar to matrix matching, as the character locations are broken down into a matrix of pixels. In feature matching, the comparison is not of a whole pixel pattern, but of the distinct features of the pattern. These features include vertical, horizontal and diagonal lines, as well as loops. An illustration of feature comparison is shown in Figure 2. The determination of a correct match is again based on a preprogrammed level of confidence.

For a more detailed explanation of OCR see Taylor, 1987, pp. 12-17.

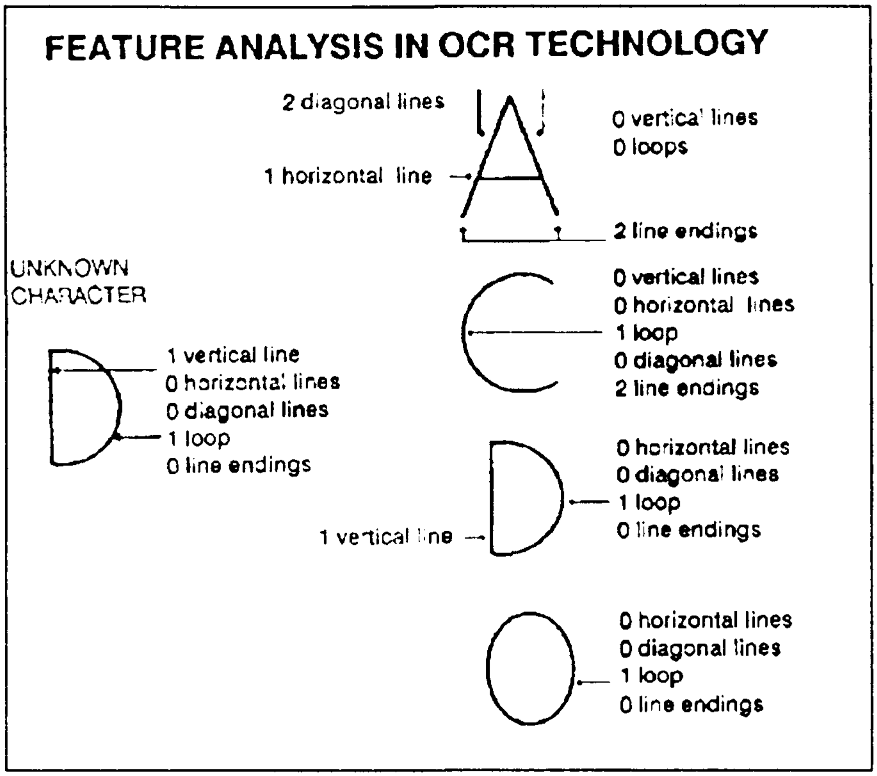


**Figure 2** Feature Matching (Wageman, 1989, p. 2.32.5)

## C. COMPACT DISK READ ONLY MEMORY

### 1. Basic Concepts

Compact Disk Read Only Memory (CD-ROM) is an internationally accepted mass storage media. The "read only" capability of CD-ROM makes it ideally suited for the distribution of large amounts of unchanging data.

The world wide acceptance of CD-ROM as a reliable mass storage media is a result of its standardization. The physical and logical requirements of CD-ROM are established by the specifications in International Standards Organization (ISO) 9660, commonly referred to as "High

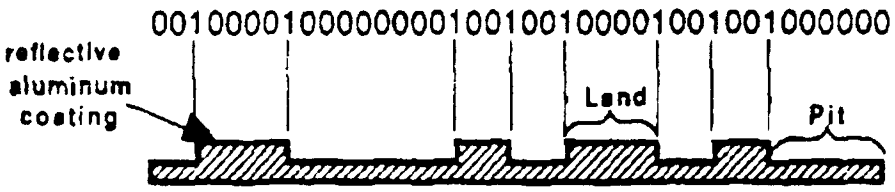
Sierra Group." This standardization ensures that a CD-ROM disc, manufactured in accordance with the ISO 9660 standard can be played on any CD-ROM drive meeting the same standard.

### 2. CLV vs CAV

A CD-ROM drive is a constant linear velocity (CLV) device, rather than a constant angular velocity (CAV) device used for most magnetic drives. This means that the rotation speed of the disc varies, depending on the location of the read head. As the read head moves from the inner to the outer edge of the disc, the speed of rotation would decrease to maintain the disc area under the read head, at a constant velocity relative to the read head. This implies that, the density of the data written on the disk remains constant throughout the entire disk.

### 3. Physical Structure

The physical structure of a CD-ROM disc can be characterized by two distinct characteristics: microscopic pits and lands, and a continuous spiral. As shown in Figure 3, data is represented by a series of microscopic lands and pits. This data is read by measuring the reflected light from a laser beam as it passes over the pattern of microscopic pits and lands. As the laser passes from land to pit, or from pit to land, the amount of reflected lighted light changes; representing a binary one.



**Pit:** microscopic depression in the reflective surface of a CD-ROM

**Land:** raised and reflective surface between 2 adjacent pits.

**Binary 1:** the transition from pit to land and land to pit.

**Figure 3** Microscopic Structure of CD-ROM Disc (Hoge, 1989, p.54)

These microscopic pits are arranged in a continuous spiral, as depicted in Figure 4. The density of this spiral is 16,000 tracks per inch, providing a total storage length of approximately three miles. The continuous spiral of a CD-ROM disc is divided into 270,000 sectors, that are referenced in terms of minutes, seconds, and sector.[3] There are 2352 bytes per sector; 2048 bytes of which are useable. This equates to a useable storage capacity of

[3] Each CD-ROM disc contains 60 minutes of play. Each minute contains 60 seconds. There are 75 sectors per second. 60 minutes/disc x 60 seconds/minute x 75 sectors/second = 270,000 sectors/disk.

540,000 kilobytes per CD-ROM disk (270,000 sectors/disk x 2 kilobytes/sector). The utilization of bytes per sector is shown in Table 1. (Fowler and Clipper, 1991, pp. 36-38)

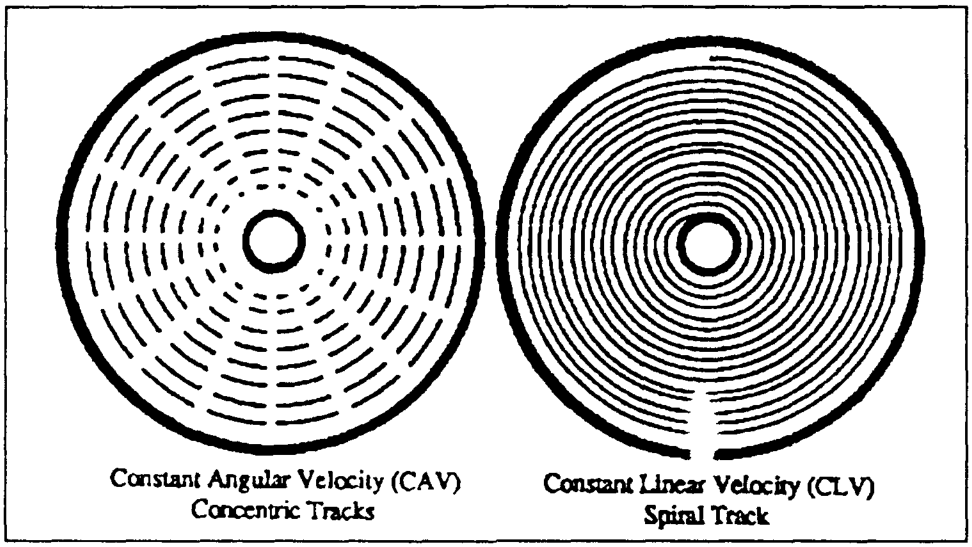

Constant Angular Velocity (CAV)
Concentric Tracks
Constant Linear Velocity (CLV)
Spiral Track

**Figure 4** CLV vs CAV (Fowler and Clipper, 1991, p. 37)

**TABLE 1. BYTE UTILIZATION PER SECTOR**

| | |
|---|---|
| Synchronization data | 12 bytes |
| Header data | 4 bytes |
| User data | 2048 bytes |
| Error Detection data | 4 bytes |
| Unused data | 8 bytes |
| Error Correction data | 276 bytes |
| Total | 2352 bytes |

## 4. Advantages and Disadvantages of CD-ROM

If thesis documents were available in digitized form, what would be the advantages and disadvantages of storing this information on CD-ROM?

1. Compact Storage: No other media provides the storage capacity for such a vast amount of information in such a compact size.

2. Excellent File Integrity: The read only nature of CD-ROM alleviates the problem of inadvertent modification or erasure. The estimated 50 year life of CD-ROM ensures reliable file integrity when used for archival of data. (Arnold, July 1991, p. 40) There is no risk of head crashes resulting in lost data, like those capable on magnetic drives.

3. Extremely Cost Effective: The information storage cost using CD-ROM is 4 cents/MB, compared to 90 cents/MB using floppy disks, 50 cents/MB using WORM and $4.00/MB using winchester drives. The only media which is currently cheaper than CD-ROM is Digital Paper at .006 cents/MB. The draw back of Digital paper is that it is an emerging technology with very expensive drives. (Arnold, July 1991, pp. 40-44)

4. Slow Access Times: The only real disadvantage of CD-ROM, is its relatively slow access times, compared to Winchester and floppy disk drives. The access time for a Winchester or floppy disk drive is more than 10 times faster than a CD-ROM disc drive.

# III. INFORMATION RETRIEVAL

## A. INTRODUCTION

The electronic revolution has enabled the storage of vast amounts of information (in the form of documents), on digital media. Information in electronic form, unlike information on paper, can only be retrieved electronically via a computer system. Consequently, our ability to retrieve information from an electronic document database, is based on our ability to properly index the information, so it can be located effectively and efficiently by the user. A database with documents well indexed, can provide for effective retrieval of information, thus allowing for tremendous storage and time saving benefits, of electronic information, to be fully realized.

The following sections discuss; differences between retrieving documents and retrieving data, the problems associated with document retrieval, some measures of retrieval effectiveness, browsing, and the different methods of searching for documents.

## B. DOCUMENT RETRIEVAL VS DATA RETRIEVAL

Retrieval of data is much different than retrieval of documents. A study conducted in 1984, David C. Blair points out four fundamental ways document retrieval differs from data retrieval:

- "the way queries are answered."
- "the relation between the formal system request, and user satisfaction."
- "the criterion for successful retrieval."
- "the factors that influence speed." (Blair, 1984, pp. 369-370).

These differences are discussed below:

### 1. Response to a Query

Data retrieval systems return the data, which is the direct answer to the question. For example, if asked for Joe Smith's address, the system would return the data that was Joe Smith's address (if it was in the database). Document retrieval systems require that a document(s) be located by matching key words, phrases, or key fields contained within these documents. This implies that, an answer to a document retrieval system query will not be the data directly, but a list of documents likely to contain the answer. The user must then review the documents to determine if they contain the required information.

### 2. Question/Answer Relationship

A document search for key words or phrases provides the user with a list of documents that have a high probability of containing the required information. This is based on the user's perception of what words or phrases are best used to match a subject area to a document search. This highly probabilistic process, depends greatly on the user's prediction of how information is presented in the documents. Conversely, data retrieval systems are based on logical relationships between a question and an answer. The system determines an answer based on a question. If asked for the telephone number of Joe Smith, the system will locate and provide the telephone number of Joe Smith, all other answers would be incorrect.

### 3. Utility - The Criterion for Successful Retrieval

Data retrieval systems provide either, an appropriate answer, or no data at all (i.e. the desired information is not in the database). Success is based on correctness. Document retrieval systems provide a set of documents, that must be reviewed to determine usefulness. The measure of success, is not correctness, but how useful the retrieved documents are to the user; or utility. As you may presume, utility is a subjective measure, that varies from user to user.

### 4. Influence on Retrieval Speed

The specific deterministic nature of data retrieval, results in response times that are closely tied to system speed. Document retrieval is an iterative process, heavily dependent on the user's ability to accurately predict how information is presented. An accurate prediction will induce a query with enough specificity to isolate desired information while excluding most unwanted information. A less accurate prediction will result in a less precise query, with less favorable results. Consequently, the iterations (search/evaluate/query refinement) required to isolate the desired information, are an outcome of query formulation which directly effects overall retrieval time. This results in a total retrieval time that is heavily biased by the user, and less dependent on the system. For this reason, a retrieval system that provides both effective and efficient tools for formulating queries, will out perform one that does not, despite running on a faster system. This is particularly important in CD-ROM databases, whose access times are significantly slower then traditional Winchester Drives.

## C. FULL TEXT DOCUMENT RETRIEVAL

J. D. Fowler and B. Clipper divide electronic document retrieval systems into three classes: database document retrieval systems, full text retrieval systems, and hybrid systems Fowler & Clipper, 1991, p. 61). Our discussion will be limited to full text retrieval systems. For more information on database and hybrid document retrieval systems, see Fowler & Clipper, 1991, pp. 61-67.

### 1. Index Formulation For Document Retrieval.

Full text document retrieval systems save the entire text of the documents. The retrieval system creates an inverted index file of all the unique words (minus stopwords) in the database.[4] The inverted index file is several inter-related levels of indexes. The first level, is a dictionary of all unique words in the document database. Each word in the dictionary has an associated occurrence list of all documents, as well as the location(s) within those documents, of every occurrence of that word. When a search for a word or phrase is conducted, the retrieval software performs a search of the index, in lieu of the entire database, providing faster retrieval times. (Fand, 1987, pp. 84-85)

The storage requirements associated with indexes of this nature range from 30 to 100 percent of the total size of the document database (Fand, 1987, p. 95). Therefore, an actual CD-ROM database is considerably smaller than the 540

[4] Stopwords, are words that are not helpful in the retrieval of a document. Some examples are; an, by, can, do, for, have, if, & when. These words are not indexed, saving space and speeding up the retrieval process.

Megabyte storage space. For example, the size of the database could range from 415 to 270 Megabytes, which is equivalent to 207,500 - 135,000 pages of text, depending on the size of the index.

### 2. Difficulties Associated with Document Retrieval

When searching a document database the user is faced with the task of formulating a query that will satisfy prediction and futility point criteria (Blair, 1984, p. 371).

The prediction criterion, requires the user accurately predict, how the information they are interested in, is represented, in order to formulate a query that will best retrieve that information. The searcher must formulate this query in a manner that will be exhaustive enough to retrieve a satisfactory number of relevant documents, yet specific enough to exclude retrieval of irrelevant documents.

The futility point criterion, requires that in the process of document searching, the user will continue to refine his/her search until the set of documents retrieved, is small enough in number, that the searcher is willing to browse through them. The futility point for most searchers is from 20 to 50 documents (Blair, 1984, p. 372). Research conducted by McCarn and Lewis in 1990, states that the optimum value of information as related to retrieval size, is 35 documents (McCarn and Lewis, 1990, p. 498). The user will therefore maximize the value of the information per retrieval set, if he/she limits the retrieved set to approximately 35 documents.

A user's futility point is not a function of database size. As the size of a database grows, the occurrence list of unique terms in the database also increases. Therefore, in general, a search (using the same terms) of a large database will retrieve more documents than an equivalent search on a small database. Consequently, as the size of a document database increases, so will the number of refinements (or search iterations) necessary to obtain a set of documents, less than or equal to the user's futility point.

Search refinements or iterations usually take the form of search term modification, such as adding or subtracting terms in combination with boolean operators (AND, OR & NOT). This is generally very effective. One should use caution when refining a search query however, as the addition of more search terms will not always increase the probability of finding the information you desire. As the number of search terms increases, the number of different searchable combinations increases. Only if the correct combination(s) of these terms are chosen, will the searcher achieve better results. As an example of this highly probabilistic process, assume that the information desired by the user is the shaded portion of Figure 5. The shaded portion of Figure 5 represents, an intersection of the documents that contain term A, with those that contain term C, or all the documents that contain terms A and C.

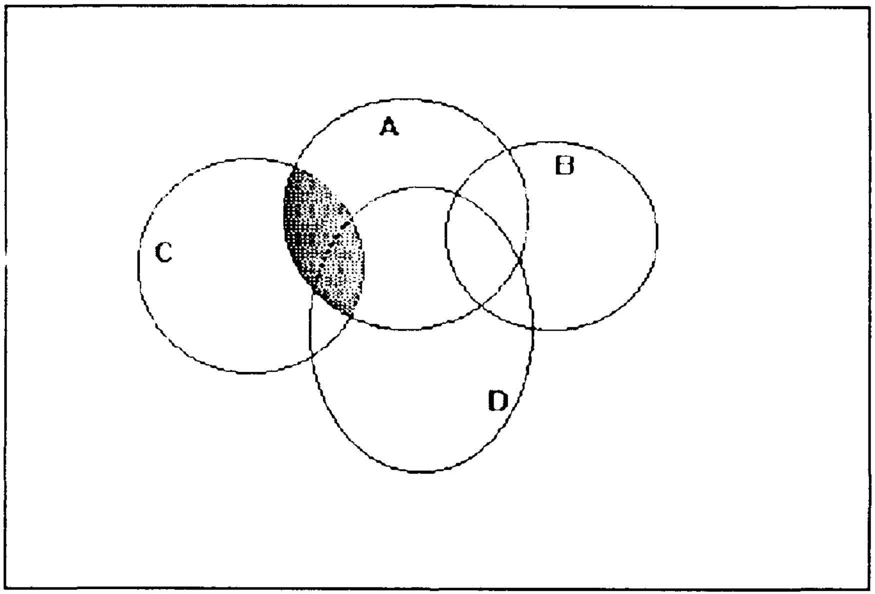


**Figure 5** Documents Desired

**A:** Documents retrieved with term A.
**B:** Documents retrieved with term B.
**C:** Documents retrieved with term C.
**D:** Documents retrieved with term D.

**ASSUMPTIONS:**
A, B, C, D > Futility Point
All set intersections < Futility Point

Suppose the user commenced a search with term A, resulting in set A. Since the number of documents retrieved is greater than the users futility point, the query is modified to include terms A AND B. Results of the query are shown ir. Figure 6.

Since the number of documents resulting from the query (terms A AND B) is below the users futility point, the search could stop here, and the desired documents would not be retrieved. Additional refinement of the query by adding term D (A AND B AND D), will result in a smaller number of retrieved documents, but will not retrieve the documents desired, as illustrated in Figure 7.

Abandoning term B, and picking a new term C will result in a satisfactory query containing a portion of the relevant documents, as shown in Figure 8. Only by combining terms A AND C and abandoning terms B and D can the searcher retrieve all desired documents, as previously shown in Figure 5.

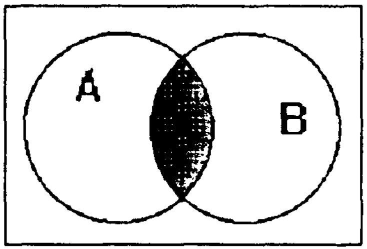


**Figure 6** Intersection of A & B

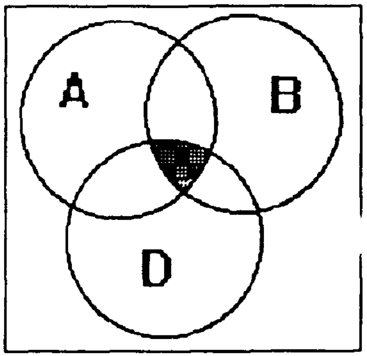


**Figure 7** Terms A AND B AND D

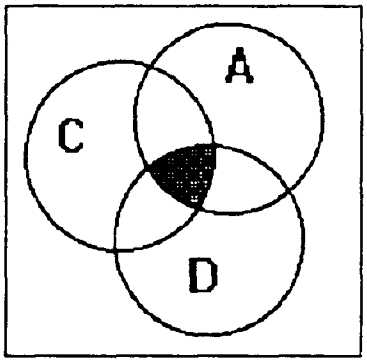


**Figure 8** A AND C AND D

Understanding the capabilities as well as the potential pitfalls of query refinement through term enhancement, is increasingly important as the size of document databases increase.

## D. MEASURES OF DOCUMENT RETRIEVAL EFFECTIVENESS

Document retrieval is indirect, probabilistic, time consuming, and has success measured by utility. These characteristics constrain the measurement of document retrieval effectiveness. Measures have not been found that diminish the effects of subjectivity on document retrieval, however accepted measures are used with these effects considered.

The two most widely used measures of document retrieval effectiveness are, precision and recall (Blair and Maron 1985, p. 290). Precision ratio, the fraction of retrieved documents which are relevant, characterizes the systems capability for retrieving only relevant documents. Recall ratio, the fraction of relevant documents retrieved, describes how well the system is at retrieving all of the documents that are relevant. The mathematical formulas for precision and recall are shown in Table 2. (McCarn and Lewis, 1990, p. 496)

**TABLE 2. PRECISION & RECALL RATIOS**

Precision Ratio: $PR = r/S$

Recall Ratio: $RR = r/R$

R: total number of relevant documents in the database.
S: number of documents retrieved by the search.
r: number of documents relevant, of those retrieved.

(Notation adopted from McCarn & Lewis, 1990, p. 496)

Retrieval systems operate at various levels of recall and precision, by varying the specificity of the submitted query. The average performance for a query will move from low-precision/high-recall to high-precision/low-recall as the query is made more specific (Salton, 1986, p. 651). This relationship is shown in Figure 9.

For example, consider a search for documents pertaining to "micro economics". A general search using the term "economics" will retrieve all documents containing "economics." Many of the documents that contain the term "economics" and not "micro economics", are probably irrelevant (low-precision), however those that are relevant would be found (high-recall). This would place the searcher at the low-precision/high-recall end in Figure 9. Making the search more specific by searching for "micro economics" will increase precision, as many of the irrelevant documents (those containing "economics" and not "micro economics") will not be retrieved. Those documents that were relevant to "micro economics" and only contained the term "economics" will not be retrieved, which will lower recall. This moves the searcher in the direction of high-precision/low-recall.[5]

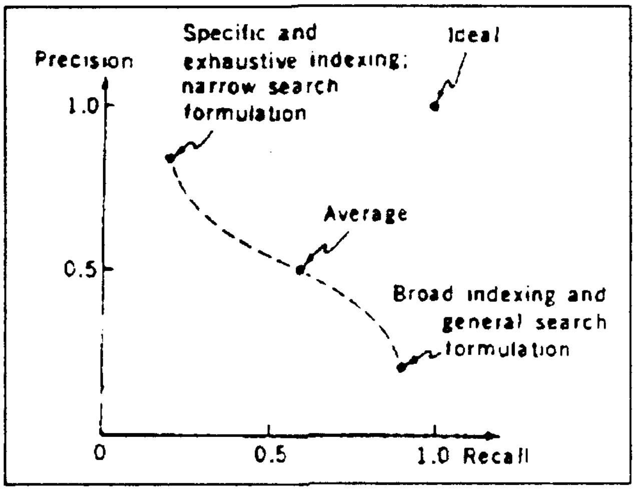


**Figure 9** Precision - Recall Relationship. (Salton, p. 336, 1970)

Traversing from low-precision/high-recall to high-precision/low-recall is the process of satisfying the users futility point. Given that the user's initial query results in a retrieval set greater than his/her futility point, he/she will refine the query (i.e., make it more specific) to reduce the number of documents retrieved. Proper refinement of a general query will result in; a retrieval set of smaller size, with a higher percentage of relevant documents, and non-retrieval of some relevant documents (Blair

[5] Since $S_s << S_g$ and $r_s < r_g$ (as some documents that contain only economics may be relevant to micro economics), $PR_s > PR_g$, and since R is constant, $RR_s < RR_g$.

and Maron, 1985, pp. 296-297). The result is moving from low-precision/high-recall to high-precision/low-recall.

Precision and recall can not be used to compare databases of substantially different size and topical specialty. As the size of a database grows, recall will automatically decrease because of increases in the number of refinements necessary to achieve a retrieval set below the users futility point (Blair and Maron, 1985, pp. 296-297). Precision can not be used, as it is unclear if the changes in size and topical specialty will affect irrelevant retrievals more strongly than relevant retrievals (Doyle, 1975, p. 357).

A point of much controversy in using precision and recall ratios as measures of retrieval effectiveness is their basis on relevance. It is argued that relevance is not really a measure because it is not additive. For example, if documents were weighted on a scale from zero to ten based on their relevance, would two documents with weights of five and five, be equivalent to two documents with weights of eight and two? This controversy can be avoided if one strictly states that a document is either relevant or irrelevant, and dispenses with the determination of the degree of relevance. M. E. Leak and G. Salton define four variables that can affect a relevance judgement (Leak and Salton, 1971, pp. 507-508). These four variables are listed below and are illustrated in Figure 10.

Relevance is affected by:

- "the type of document being judged - including its subject matter, level of difficulty, level of condensation, style."
- "the conditions under which judgements must be rendered, that is the time available, the order of presentation and the size of the document set, the type of task specification."
- "the statement specifying the information requirement which determines relevance."
- "the type of judge used to render the judgements, that is his experience, background, attitude and so on."

A comparison of the two databases developed for this thesis was made by controlling these variables. Each database contained the same documents, although representations were different, the difference did not have any bearing on their relevance. The conditions under which the judgements were made were controlled so as to maintain uniformity, the judge making the determination was the same, and the requirement of relevance was identical for each database.

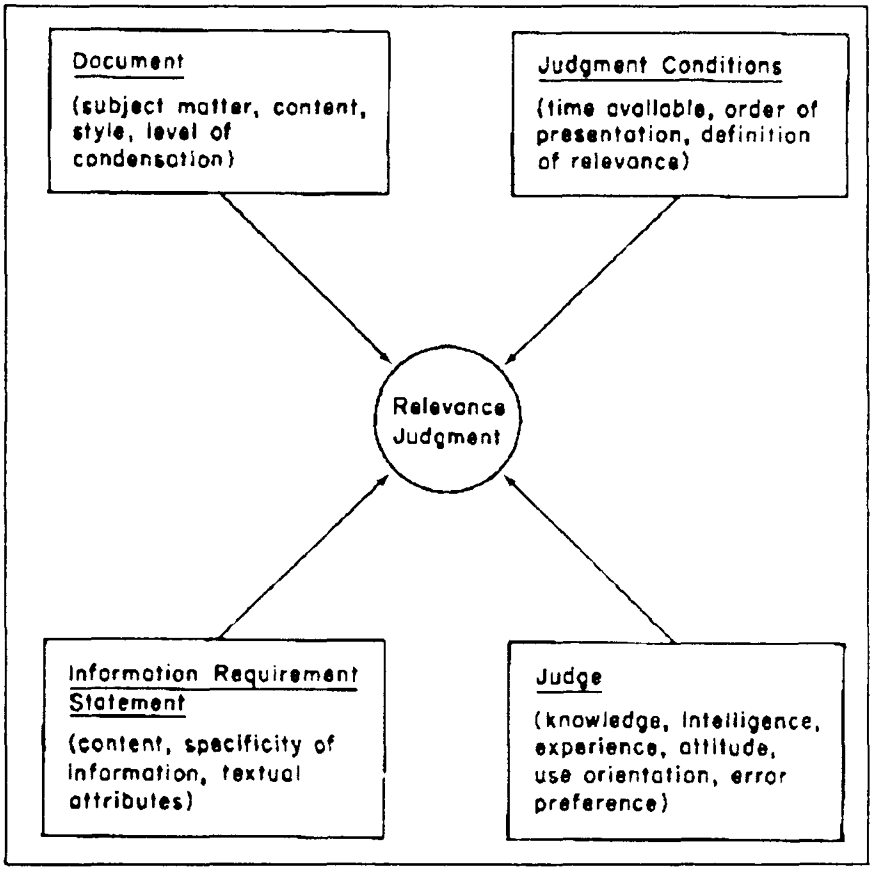

Document
(subject matter, content, style, level of condensation)
Judgment Conditions
(time available, order of presentation, definition of relevance)
Relevance Judgment
Information Requirement Statement
(content, specificity of information, textual attributes)
Judge
(knowledge, intelligence, experience, attitude, use orientation, error preference)

**Figure 10** Variables of Relevance Judgements. (Lesk and Salton, p. 508, 1971)

## E. SEARCHING FEATURES OF FULL TEXT DOCUMENT RETRIEVAL

There are a variety of different methods that may be used when performing a document search, which will provide different results. The searching methods that a software package provides, will significantly effect its usefulness, and should be carefully considered during its evaluation. The most notable searching methods are discussed below:

### 1. Word Searching

Searching for a word, is the most elementary form of text searching. When you perform a word search, the retrieval system will locate all the occurrences of that word in its document database. Single word searching however, is quite general, and will typically result in very large sets of retrieved documents. For example, if you were interested in color scanners, and you performed a word search using "scanner(s)", chances are, you would locate all sorts of documents which discuss scanners, but only a few might discuss color scanners. To locate the documents you desire, would require browsing the retrieved document set, which could be very time consuming.

### 2. Phrase Searching

Phrase searching is the capability of searching for phrases consisting of consecutive words within a document. This capability can reduce the amount of extraneous documents retrieved. Searchable phrase lengths range from approximately two to four consecutive words, depending on the capability of the retrieval software package.

### 3. Proximity Searching

Proximity searching is a mixture of word and phrase searching with greater capabilities. Proximity searching allows the searcher to specify words or phrases that can be searched for within proximity of each other. The proximity can be words, lines, paragraphs, or documents, depending on the capabilities of the software retrieval package. For example, to find information on "how a color scanner works" you might specify a proximity search of "color scanner(s)" in the same paragraph as "operation". As with all full text searching, it is important to think of how the document may have been written, and pick terms accordingly, instead of paraphrasing ideas or topics.

### 4. Boolean Operators

Boolean operators provide the searcher with a powerful tool for formulating and refining queries. Most retrieval programs provide three boolean operators; AND, OR, and AND NOT. These operators can be used in combination with words, phrases and sets.[6] The boolean operators AND, and AND NOT generally shrink the size of retrieval set, as the resulting documents must satisfy both of the search criteria. The boolean operator OR, tends to enlarge the retrieval set, as a document is included in the retrieval set if it satisfies either of the search criteria.

### 5. Wildcard Operators

Wildcard operators can be single or multiple character wildcards. Single character operators, take the place of one character; multiple character operators, take the place of many characters. Wildcard operators are particularly useful due to the varying sophistication of automatic indexing programs. The procedures used by

[6] A set is a group of documents previously retrieved, meeting the criteria of a specified search.

automatic indexing programs for handling word suffixes varies. For example, some automatic indexing software index only the roots of words, where as others do not. This means that a word search with "farm", conducted on a database indexed with a package that indexes only the roots of words, will retrieve documents also containing; farms, farmer, farmed, and farming. For those packages that do not index only word roots, similar results could be obtained using a multiple wildcard operator. A search for "farm?" (where ? is a multiple wildcard operator), would retrieve documents containing terms with "farm" as the first four letters, such as; farm, farms, farmer, farmed, farming, farmyard, farmhouse, farmstead, etc.

### F. BROWSING'S ROLE IN DOCUMENT RETRIEVAL

Browsing is vitally important to a searcher, as it allows them to read through information as one would read a book. Since this is second nature, it provides a familiar way to find information or validate information found through searching. Browsing, can be thought of as a method of document retrieval. Browsing deals with moving from the "where" it is, to the "what" is there. A user may pick a location in the database (a document or point in a document), and then look (browse) to see what is there. Searching, deals with traversing from the "what" they want, to "where" it is, meaning that the system is told "what to look for", and then provides the user with "documents where it is located." (Zoellick, 1987, p. 65)

Browsing is particularly useful with document databases located on CD-ROM. Prior to the availability of CD-ROM document databases, searching was conducted on-line. Slow data transfer rates and high communication costs, meant the searcher did not have the time, or money, to freely browse retrieval sets, or the database. A searcher may now connect to a CD-ROM document database via his/her personal computer. This provides the searcher with fast, easy access for unlimited browsing, as well as searching. Thus, providing the searcher the full utilization of the complimentary search, browse features of full text document retrieval. (Zoellick, 1987, p. 65)

# IV. DATABASE DEVELOPMENT

## A. INTRODUCTION

Documents existing in paper form can be converted into digital form as bit mapped images, or a combination of ASCII text and bit mapped images. Each form provides advantages and disadvantages in the conversion process, as well as the electronic retrieval process. To compare the advantages and disadvantages for each of these processes, and to determine the best overall form, two databases were developed. One database contains primarily text, with few images, the other primarily images, with little text.

The equipment used to develop these databases consisted of a Zenith 286 personal computer operating at 12 MHZ. The computer contains 640 Kilobytes of RAM, two 20 Megabyte Winchester Drives, and a 4 Megabyte Kofax 8202 image memory board, for image compression and decompression when using the attached scanner. Image scanning was performed on a Fujitsu model M3094E flatbed scanner with automatic sheet feeder. All documents were scanned at a 300 DPI resolution with supporting software for image processing and text recognition.[7] A Worm Drive by Laser Drive Limited, model

[7] Calera TrueScan Version 1.071 from Calera Recognition Systems, and Irecognize Profession Plus - Page Image Processing Software for Image Capture and Text

810 with 800 Megabyte cartridges (400 Megabytes per side) provided the mass storage capability necessary for database development. Wordperfect 5.1, by WordPerfect Corporation was used for word processing requirements. The databases were developed using KAware Optical Disk Publishing software, Version 1.0, by KAWARE Corporation Inc.. Any comments made either favorably or unfavorably, about any of the above mentioned software products, are the opinions of this author and are not intended as a software evaluation.

The following sections will discuss development of each of the databases, tradeoffs associated with that development, and advantages and disadvantages of each database.

## B. DEVELOPMENT OF TEXT AND IMAGE DATABASES

Each database contains the same documents, however the representation of the documents is different. The "text/image" database, contains the full text of each document. All figures, graphs, and pages not reliably converted to text (i.e. Cover page, Report Documentation page, Signature page, etc.) are represented as bit mapped images.[8] The "image/text" database contains images of each

Recognition, Version 1.07 from iBase Systems Corporation.

[8] For more information on the reliability and accuracy of optical character recognition, see Taylor pp. 44-51, 1989.

page of each document. The only text in the database, is that which is necessary to provide a description of each image for indexing and retrieval purposes. Since each database contained the same documents, their development required many of the same steps. These steps are discussed below, with their associated time requirements listed in Table 3.

### 1. Document Preparation and Scanning

This process was identical for each database. Documents were prepared by removing staples, smoothing bent pages, and aligning pages properly in the automatic document feeder. During scanning, the iRecognize software creates two files; one containing the image of each page in an image file format, the other a pointer file. The image file size is dependent on the number of pages scanned. Because the automatic sheet feeder was used, storage requirements were an important consideration, as a thesis containing 100 pages would generate an image file of approximately 5.0 Megabytes.

### 2. Creation of an Image for each Page.

This process was required for each database, however only pages not reliably converted to text, and portions of those pages containing figures and graphs were required for the text/image database. The scanned image of each page is in the iRecognize image file format, making it necessary to produce an image of each page in another format. Images of each page were produced using the .PCX format, although the preferred .TIF format was available. The .TIF format was preferred as the disk publishing software will only accept .TIF format. The .TIF format accepted by the disk publishing software however, is not the version exported by most image scanners, making it incompatible. It was therefore necessary to export the images in a .PCX format and then convert them to a useable .TIF format using a commercially available software package called HIJAAK, version 2.01, by Inset Systems Inc.. Once the images were in a useable .TIF format, they were converted to the .CPR disk publisher format (using the disk publisher software) for later retrieval and display. The times required for each of these processes are shown in Table 3. Instances of two different values under the text/image header, are the times required for a figure, (portion of a page) and a full page, respectively.

The storage requirements to accomplish this step were enormous. On average, each page image required 167 Kilobytes in the .PCX format, 65.82 Kilobytes in the .TIF format, and 75.65 Kilobytes in the .CPR (Disk Publisher) format. An average size thesis document of 108 pages, would require; 18 Megabytes of storage space for the .PCX images, 7.1 Megabytes for the .TIF images, and 8.2 Megabytes for the .CPR images, for a total of 33.3 Megabytes.

### 3. Optical Character Recognition

OCR was only necessary for the text/image database. The process of OCR is dependant on such things as, original document quality, font of text used, and proper page alignment. All the documents used for this research, were of good quality, but contained a variety of fonts. On average, a 99 percent or better accuracy rate was obtained, which is about 25 character corrections per page. The times required for OCR and error correction are shown in Table 3.

### 4. Text Preparation and Markup

At this point, the full text of each document, images of all pages of each document, and images of figures, graphs, and charts had been generated. It was then necessary to link the full text of each document with associated images of figures, graphs, and charts to develop the text/image database. To complete development of the image/text database, terms describing each page image (or section of page images) were determined to provide search and retrieval capabilities. This process is discussed below, with applicable times shown in Table 3.

#### a. Text/Image Database

Full text of each document was brought together into one file, where erroneous markings such as file headers, extra spaces, and lines were removed. Markers were placed in the text to divide each original document

(document subdivision will be discussed more in section C, Performance and Development Tradeoffs) into smaller sections.[9] Each section is an individual document from the perspective of the retrieval system. Tags and image calls were placed in the text to provide access to associated images. A tag, is a predetermined unique set of characters that differentiate for the indexing system, text to be indexed, text used for display purposes only, and markings used for image calls. An Image call, is a reference to a file name containing the desired image. Image calls were placed in the text for each of the figures, graphs, diagrams, and pages not converted to text. (KAware Disk Publisher User's Manual, 1990, pp. 6(1)-6(19))

### b. Image/Text Database

Text used to describe images of each page, was initiated using the text from the abstract and table of contents. Key words, phrases, and synonyms for a particular page or section were then added to the table of contents of each document to provide better retrieval. The determination of these key words or phrases, was an extremely difficult, time consuming process. It required reading each page of the document, and then determining words that would best represent the contents. Tags and

[9] A marker is a unique character or string of characters used by the retrieval system, to divide a document into smaller logical documents.

markers were then added for indexing purposes, with image calls inserted for each page of the document. The markers added to this database did not break the original document down as far as the text/image database, due to the limited text available to index.

### 5. Automatic Index Generation

The time required to automatically generate the index, is directly related to the amount of text being indexed. Index generation time varied, due to the differences in the amount of text being indexed. Index generation time for each database is shown in Table 3.

**TABLE 3. TIME REQUIREMENTS FOR DATABASE DEVELOPMENT**

| | Text/Image | Image/Text |
|---|---|---|
| Document Preparation | 5.0 min | 5.0 min |
| Scanning Time/Page | 13.0 sec | 13.0 sec |
| .PCX Image Generation | 30.0/60.0 sec | 60.0 sec |
| .PCX to .TIF Conversion | 45.0/90.0 sec | 90.0 sec |
| .TIF to .CPR Conversion | 35.0 sec | 35.0 sec |
| OCR/Page | 65.0 sec | N/A |
| OCR Error Correction/Page | 4.0 min | N/A |
| Text Correction, Index, Tagging and Image Calls | 90.0 min | 3.0 min/page |
| Automatic Index Generation | 5.0 min | 45.0 sec |

### 6. Thesis Conversion Time

Using the assumption that an average thesis is 108.3 pages long, containing 25 percent figures and graphs, the time to process a hard copy thesis document into an electronically retrievable format can be calculated (Taylor, p. 52, 1989).[10] Tables 4 and 5 illustrate these calculations for a image/text and text/image database.

**TABLE 4. IMAGE/TEXT DATABASE CALCULATIONS**

| | |
|---|---|
| Document Preparation: | 5.0 min |
| Scanning Time:<br>(108.3 pages x 13.0 sec/page) ÷ 60 sec/min = | 23.5 min |
| Image Generation:<br>108.3 pages * (60 + 90 + 35) sec) ÷<br>60 sec/min = | 333.9 min |
| **Text Preparation & Markup:**<br>3.0 min/page * 108.3 pages = | 324.9 min |
| **Automatic Index Generation:** | 30.0 sec |
| | 687.7 min<br>or 11.5 hrs |
| **Note: If the process of .PCX to .TIF image conversion is avoided, 2.7 hrs are saved, yielding a time of 8.8 hrs.** | |

If the conversion of images from the .PCX to .TIF format was not necessary, the total time required to convert

[10] "Selecting 20 thesis documents at random, the average size of a thesis was 108.3 pages in length, of which 27.9 pages were graphs or charts and 7.4 pages were pictures." (Taylor, p. 52, 1989)

a thesis (in paper form) into a retrievable digital form, is approximately the same for either text/image or image/text format. The time gained by not performing OCR and correction for the image/text database, was lost determining proper terms to index the page images on.

This reveals a significant point. When discussing the time requirements for conversion of paper documents to digital form, it is necessary to consider the combined process of, both converting that document to digital form, and proper markup and indexing of that document for later retrieval. As evidenced here, the gains made in electronic conversion, using an imaged based database, were off set by difficulties encountered during markup and indexing, making the overall time for each database method approximately equal.

## TABLE 5. TEXT/IMAGE DATABASE CALCULATIONS

| | |
|---|---|
| **Document Preparation:** | 5.0 min |
| **Scanning Time:**<br>(108.3 pages x 13.0 sec/page) ÷ 60 sec/min = | 23.5 min |
| **Image Generation:**<br>((27.9[†] + 7.4[††]) pages * (30 + 45 + 35) sec) ÷ 60 sec/min = | 64.7 min |
| (3[‡] pages * (60 + 90 + 35) sec) ÷ 60 sec/min = | 9.2 min |
| **OCR:**<br>70[‡‡] pages * (65 sec + 4.0 min) = | 355.8 min |
| **Text Preparation & Markup:** | 90.0 min |
| **Automatic Index Generation:** | 1.5 min |
| | 549.7 min<br>or 9.2 hrs |

**Note: If the process of .PCX to .TIF image conversion is avoided, 0.5 hrs are saved, yielding a time of 8.7 hrs.**

[†] Pages containing Charts or Graphs

[††] Pages containing Pictures

[‡] Full Pages: Cover, Document, and Signature Pages

[‡‡] 108.3 - (27.9 + 7.4 + 3) = 70

## C. PERFORMANCE AND DEVELOPMENT TRADEOFFS

Many of the decisions made during database development affected performance. The following sections will discuss the development decisions made, the impact these decisions will have on performance, and the actual performance of the two databases.

### 1. Development Decisions and Tradeoffs.

One of the biggest decisions during database development was whether or not to divide documents into smaller documents for retrieval, and if so, how small. Subdivision increases search precision, but requires more search time. To ensure that a searcher can always determine the author of a document, an image call for the cover page (or applicable title page) must be placed in each logical document. This means that each level of document subdivision (chapters, sections, subsections etc.) requires additional time to allow for insertion of the proper markers and image calls. When doing a search, the system locates the logical document that contains the word(s) you are searching for. A document 100 pages long not subdivided, is considered one logical document for retrieval. A term or phrase located somewhere in 100 pages, is not as precise as locating that same term(s) in a smaller section of possibly a few pages or less. The effects of subdivision, are most significant when using boolean operators. Without subdivision, the entire document is considered the boolean domain for the search. Proximity searching however, is unaffected by logical document size, as the search is conducted by finding words within a specified range of each other. To determine the proper balance between increased precision, and time required for text markup, queries were conducted on each database under various subdivision conditions. These will be discussed more in "Database Performance".

An additional problem, only affecting the image/text data base was, the association of words, phrases and or synonyms with the page images for proper indexing and retrieval. The under-lying problem with document retrieval, continues to be, that the searcher does not know the "actual words" the author(s) uses when writing about a particular topic, and a particular document may pertain to a number of different topics. It is important to determine the appropriate words to associate with a page image (or group of page images) to best aid the "typical" searcher in finding that information. This author concluded, "there is no magical solution." Different people use different words to describe (index) the images, just as different people use different search terms when looking for information. A related problem, is the best number of words to associate with the page image. As the number of words associated with each image increases; the time required to determine these terms increases, but the likelihood of retrieving that page also increases (recall will increase). Precision, however, will decrease, as more of the image pages will have identical terms associated with them. This is illustrated by Figure 11, which shows how recall and precision ratios change as a function of the number of index terms per document (Doyle, 1975, p. 338). This author compromised between the time requirements necessary to determine the index terms, and the associated recall/precision, by associating approximately five terms with each image page.

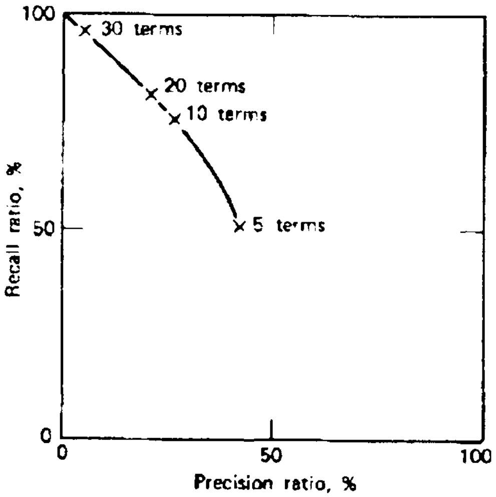


**Figure 11** Effects of Index Term number on Precisicn and Recall. (Doyle, p. 338, 1975)

### 2. Database Performance

To determine the advantages and disadvantages of each database, and properly ascertain the best level of document subdivision, a set of queries was performed on each database. This set of queries was performed on the text/image database, under four different markup conditions (1 - 4), and on the image/text database under three different conditions (2 - 4).[11] These different markup conditions are as follows:

- **Condition One:** Each document was subdivided into the smallest logical section. This ranged from section to subsubsection depending on the document.
- **Condition Two:** Each document was subdivided by sections of a chapter.
- **Condition Three:** Each document was subdivided into chapters.
- **Condition Four:** Documents were not subdivided.

To provide a better understanding of the conclusions, it is important to have a general understanding of the features of the KAware retrieval system. The system provides two modes of retrieval display; "display retrieved documents" or "display all documents." If the system is set to "display all documents", the whole file is available for browsing after a search has located matching documents.

[11] Document subdivision below the section level, was impractical due to the limited amount of text on which to index.

When browsing, the user can jump from term to term (in or among documents), or from document to document throughout the entire file. If the system is set to "display retrieved documents", then only those logical documents, located by the search query, are available for browsing using the same capabilities listed above. All search capabilities discussed in Chapter III are available.

### a. Level of Document Subdivision

Division of documents by section provided the best overall performance considering: ease of browsing; precision; recall; and the time required to insert image calls, and perform text markup. This conclusion is based on the following:

- Document breakdown below the section level did not significantly increase precision, and had no effect on recall.
- Document breakdown below the section level increased significantly the time required to insert image calls, and perform text markup.
- Browsing documents broken down below the section level, was not appreciably easier or less time consuming, then browsing documents broken down at the section level.
- Document breakdown at the chapter level was not conducive to easy browsing of the document. Viewing different sections required scrolling through pages of material, in lieu of simply jumping from section to section, as in section breakdown.
- Boolean searches conducted with documents broken down by chapters, yielded a larger number of irrelevant documents (lower precision) then documents broken down by sections, as associated terms were found pages apart.

- Retaining each document without breakdown, made browsing more time consuming then chapter breakdown, due to the size of each logical document.
- Boolean searches conducted with documents not broken down, resulted in many irrelevant documents, as associated terms were found pages or even chapters apart.

As was expected document breakdown did not affect proximity searches, as retrieval performance for the two databases was consistent over all markup conditions using proximity searching.

### b. Retrieval Advantages and Disadvantages

From the user's view point, it was much easier to obtain data from the text/image database then the image/text database. This was as a result of, being able to read what was located "directly", and an increased capability for finding desired information, by searching the full text of each document. Terms located in the text/image database, are in the actual full text of the document. When browsing for relevance, the user is reviewing the actual text of the document. Terms located in the image/text database are found among terms used to index the images. To determine relevance, the user must scroll through a series of images. Each database had its own advantages and disadvantages, which are listed below.

### (1) Text/Image Advantages and Disadvantages

Advantages:

- Having the full text of each document to index and search on made browsing the document much easier.

- Provided higher precision and recall for the same level of document breakdown than the image/text database.

- Provided direct access to information as text is searched and viewed directly, in lieu of searching an index, and then viewing page image(s) associated with that index, to determine relevance. For example, when a word or phrase is located in a full text document database, the system displays the appropriate section(s) of the document, with the word or phrase highlighted. This immediately draws the user's attention to that section of the document. That same word or phrase would be located in the index of a image based document database. The user would then have to view the page image(s) associated with that index word or phrase to determine relevance.

- Images of figures, charts, or graphs could be printed if desired without added printer memory.

Disadvantage:

- Full page images associated with the database, such as the cover, document, and signature pages, could not be printed without increasing standard printer memory, due to their large size (75.67 Kilobytes average).

### (2) Image/Text Advantages and Disadvantages

Advantages:

- Images of each page provided actual layouts of ea^h page, preserving all bolding, underlines, italics, and line spacing, making the text structurally easier to read.

- The majority of information on the page images, was located using only the terms found in the Table of Contents, implying that much of the time spent developing additional terms on which to index could have been saved.

Disadvantages:

- Access to information was indirect, as images were viewed to determine relevance, in lieu of reading the text directly.
- Enormous storage requirements. Image/text database required 6.7 times more storage space than the text/image database.
- Images of full pages could not be printed with normal printer memory due to their large size (75.67 Kilobytes average).

A problem common to both databases, resulting for the KAware Disk Publisher, is the display of images. Images (pages, figures, charts, or graphs) can be displayed at four different zoom levels, however, only the two greatest zoom levels were effective for reading text from a full page. At that level of zoom, a full line of text could not be read without scrolling, which made reading a page very tedious.

## D. SUMMARY

The paper to digital conversion of a document required approximately equal times for both the full text (text/image) and the image based (image/text) forms. Analysis showed that time during the text preparation and markup may be saved, due to the fact that most of the information was located using the terms from the Table of Contents (image/text database). This would result in a shorter conversion time for the image/text database, then the text/image database. This time savings however, is out weighed by, ease of browsing, direct access to information, and increased precision and recall of the text/image database. Any conversion of paper documents should use the full text approach.

# V. SUBMISSION OF THESIS DOCUMENTS IN ELECTRONIC FORM

## A. INTRODUCTION

Proliferation of the personal computer and word processing software implies that documents exist in electronic form at their completion. Documents in electronic form can be stored on digital media, such as optical disks, with less effort than converting paper documents to electronic form for storage and retrieval. The Naval Postgraduate School receives approximately 230 thesis documents per quarter. On average, approximately 12 copies of each of these documents are reproduced for distribution. Reproduction and mailing costs alone, are estimated to be over $26,000.[12] Since these documents exist in electronic form at their completion, storage and reproduction of an electronic version would eliminate the need for converting future paper thesis documents to digital form.

Turning in completed thesis documents in electronic form, does require some special requirements and decisions. For instance, in what electronic form (WordPerfect 5.1, ASCII text, Microsoft word, etc.) should they be turned in?

[12] Estimate is based on an average size thesis (108 pages) costing approximately $2.90/thesis to mail. Mailing costs are paid from a centralized BUPERS fund, reproductions costs are paid by the NPS.

What special requirements exist for graphics, and pages (Cover, Document and Signature pages) generated on Xerox work stations? What word processing software is currently available on campus for students to produce their thesis?

The following sections will discuss the current thesis preparation tools available on campus, special requirements for turning in thesis documents in electronic form, as well as advantages and disadvantages of documents in WordPerfect 5.1 and ASCII text formats.

## B. CURRENT THESIS PREPARATION TOOLS

Students can produce their thesis one of five ways:

- Use their own personal computer and word processing software.
- Hire a thesis typist to prepare their thesis.
- Use a software package called G-Thesis available on the Naval Postgraduate School mainframe.
- Computer Science department students can use a software package called Framemaker, available on Sun work stations in their laboratories.
- Use a software package called WordPerfect (current version 5.1), which is available in all micro labs with word processing capability, located on the Naval Postgraduate School campus.

A number of different word processing software packages are available on the market, although the most widely used is, WordPerfect, by WordPerfect Corporation. It is the administrative computing standard on the Naval Postgraduate

School campus, and is used by most thesis typists and students. For this reason, the discussion of thesis preparation tools will be limited to; G-thesis, Framemaker and WordPerfect.

### 1. Thesis Preparation using G-Thesis

G-Thesis is a word processing software package, available on the Naval Postgraduate School mainframe. Its capabilities are cumbersome, when compared with software packages available for a personal computer. A formal survey was not done, however, Naval Postgraduate School Computer Center personnel estimate that, G-Thesis is used by few students each quarter. Script, the program used by G-thesis, continues to be offered as it uses little disk storage space and processing time on the NPS mainframe. Script costs very little to buy and maintain as it is one of software packages obtained by the Naval Postgraduate School under the Higher Education Software Consortium.[13] G-thesis can export documents in an ASCII text format, however, there is no way to preserve the graphics, footnotes, endnotes, or page numbers created by the software.

[13] The Higher Education Software Consortium is a program from IBM, where qualifying institutions can obtain as many software packages as they desire (from a predetermined list), for one price.

### 2. Thesis Preparation using Framemaker

Framemaker, is a capable word processing software package available on the Sun Work stations in the Computer Science Department. Documents created with Framemaker are in a Post Script format, but can also be exported in an ASCII text format. Graphics, footnotes, endnotes, and page numbers created by the software are not preserved when the document is exported in ASCII text format. Graphics contained in the document can be preserved by saving them separately, as images, before they are imported into the Framemaker document.

Enrollment in the Computer Science curriculum is approximately 20 to 25 students per class (every other quarter), which is a small number of students, when compared to an entire graduating class, numbering 230.

### 3. Thesis Preparation using WordPerfect 5.1

WordPerfect, by WordPerfect Corporation, is the administrative computing standard for word processing software on the Naval Postgraduate School campus. Every micro computer laboratory at Naval Postgraduate School, that has word processing software installed, uses WordPerfect 5.1.[14] In addition to being widely available, all students in the Administrative Science Department take a course (IS-

[14] WordPerfect 5.0 and Microsoft Word are also available, but only in laboratories where WordPerfect 5.1 is also available.

0123) in WordPerfect 5.1. Thesis styles for WordPerfect 5.1 are available from the Naval Postgraduate School Computer Center. These styles will correctly format a document to meet the requirements for thesis completion. Thesis styles are easy to use, and require little experience with WordPerfect. Instruction is given throughout each quarter to students wishing to use these styles, to provide basic instruction, and answer questions. Beta testing is currently being conducted by Naval Postgraduate School Computer Center personnel, on software by WordPerfect Corporation, that would allow the mainframe's high speed laser printer, to print WordPerfect documents created on micro computers.

WordPerfect 5.1 is a capable word processing software package. Like other word processing software packages, graphics contained within a WordPerfect document are not exportable in an ASCII text format; however, footnotes, endnotes, and page numbers created within the document, can be retained, using WordPerfect's Standard Printer driver, by sending the output to a file, instead of the printer.

## C. FORMAT REQUIREMENTS FOR ELECTRONIC DOCUMENTS

A document submitted in electronic form must contain all aspects of the same document in paper form. This means that all pages of a thesis (from Cover page to Distribution list), must be included in an electronic version. A thesis is composed of: Cover, Document (DD-Form 1473), and Signature pages; List of Tables/Figures, and Table of Contents; Document body containing text and graphics; Appendixes, List of References, Bibliography, and Initial Distribution List. The Cover, Document and Signature pages of each thesis can be produced on Xerox work stations. These pages would be submitted as images to provide author accreditation (Cover Page), preserve Document structure (Document Page), and provide thesis authenticity (signed Signature page). The remaining parts of a thesis are composed of text (and graphics) that can be created with a word processing software package. Graphics contained within a document created by a word processing software package must be saved as separate images, because once imported, the graphics become part of the document, and can not be exported as ASCII text. Regardless of the word processing software package, the final format of an electronic document must be compatible with the Optical Disk Publishing Software being used. The Optical Disk Publishing Software used for this thesis; KAware Optical Disk Publishing Software, Knowledge Access International, was chosen because it provided many of the capabilities available in similar software packages, at a fraction of the cost.

KAware's optical disk publishing software requires that:

- Text of documents being indexed must be in ASCII text format, and only contain ASCII characters between 32 and 126 (decimal).[15]
- Images of graphics must be in .TIF format with one image per file.

Word processing software packages will export documents in ASCII text format; however, many will not preserve items produced using hidden codes, such as page numbers, footnotes, endnotes, and graphics. Since these items are part of the document, they must be preserved. Graphics can be preserved as separate images using the graphics program which created them, or by using an optical scanner.

The requirement of having images of graphics, in addition to, the text of the document, raises a potential problem for many students at Naval Postgraduate School. Most graphics contained in thesis documents, are a result of cutting, pasting, and photocopying of pages for inclusion in the thesis. For most students, the production of images of graphics, will require them to optically scan these graphics, to produce bit mapped images. Most students have never used an optical scanner, thus requiring additional time when producing their thesis. Two optical scanners are

[15] Documents may be in any word processor format while performing markup and inserting image calls. Prior to indexing with the disk publisher software however, the text must be converted to ASCII text.

currently available in the micro computer lab, in Ingersoll I-151. The "Discover" scanner, will produce a .TIF image that is compatible with the KAware disk publishing software. The Xerox scanner, will produce a .TIF image, however it is not directly compatible with the disk publisher software. The Computer Center has purchased, and is currently evaluating (for placement in micro labs), a software package called HIJAAK, that would allow .TIF images produced by the Xerox scanner, to be converted to a .TIF format, acceptable by the disk publishing software.[16] KAware Corporation Inc., is currently working to improve the image compatibility of their software.

## D. CONVERSION OF A DOCUMENT IN WORDPERFECT FORMAT

To ascertain the effort required to transform an electronic document in WordPerfect format, to a retrievable, searchable document, an experiment was conducted, on a document in WordPerfect format. This document was created with WordPerfect 5.1, using NPS thesis styles. This thesis is 245 pages and contains 31 figures or graphs. Each figure or graph in the document, was in a separate image file in .WPG (WordPerfect Graphics) format. A separate analysis of a WordPerfect document, created without thesis styles was

[16] HIJAAK, version 2.02, by Inset Systems Inc., is a commercially available software package that will convert images from one format to another.

not conducted, however, the author does not believe it would take substantially longer to markup such a document due to the search and replace features of WordPerfect 5.1.

Styles used to create a document can be easily modified. Modifications made, automatically reformat affected text. A set of styles can be modified, saved, and then retrieved into a document, replacing the same styles used to create the document. This would preclude making the same changes to styles of each document. Inserting markings for document subdivision, and image calls, for proper author accreditation at section headings, was easily accomplished by modifying the document styles. Style modifications were also used to change document margins. When a document is converted to ASCII text, left margins are filled in as spaces. Top and bottom margins produced significant blank areas in the text. Margin modification alleviated these problems. Search features of WordPerfect, reduced the effort required to insert image calls, for graphics contained in the document. Graphics in the document, were located by searching for hidden codes. Proper markings and image calls were then inserted, by retrieving an existing file, containing this information. Graphics were then deleted from the WordPerfect document, to prevent large blank areas among the text when the document was converted to ASCII text. Upon completion of the markup process, the document was converted to ASCII text, using WordPerfect's standard printer driver. Time required to perform this markup process, convert images to the .CPR disk publisher format, and index the document for subsequent search and retrieval, is shown in Table 6.

**TABLE 6. CONVERSION OF A DOCUMENT IN WORDPERFECT FORMAT**

| | |
|---|---|
| **Image Generation:** | |
| ((31 graphic images) * (45 + 35) sec) ÷ 60 sec/min = | 41.30 min |
| ($3^{\ddagger}$ pages * (90 + 35) sec) ÷ 60 sec/min = | 6.25 min |
| **Text Preparation & Markup:** | 35.00 min |
| **Automatic Index Generation:** | 1.50 min |
| | 84.05 min<br>or 1.4 hrs |
| Note: **If the student performs image conversion to the proper .TIF format, 0.46 hours are saved, yielding a time of 1.04 hrs.** | |
| ‡ Full Pages: Cover, Document, and Signature Pages | |

It should be noted that, the thesis used for this experiment, has over twice as many pages as an average thesis document. The number of graphics however, is approximately equal.

## E. CONVERSION OF A DOCUMENT IN ASCII TEXT FORMAT

To perform this portion of the experiment, the WordPerfect document used above, was converted to an ASCII text document, using the standard printer driver of WordPerfect 5.1.

The process of inserting markings for document subdivision, and image calls for graphic image display, was a tedious process. Without the ability to perform search and replace functions, on hidden codes found at section headings and graphic locations, the entire thesis had to be reviewed page by page. When image calls and markings were inserted in the text, reformatting was required, adding additional time to the process. Time required to perform this markup process, convert images to the .CPR disk publisher format, and index the document for subsequent search and retrieval, is shown in Table 7. Conversion of the ASCII text document, took almost twice as long as the same WordPerfect document.

TABLE 7. CONVERSION OF A DOCUMENT IN ASCII TEXT FORMAT

| | |
|---|---|
| **Image Generation:** | |
| ((31 graphic images) * (45 + 35) sec) ÷ 60 sec/min = | 41.30 min |
| (3‡ pages * (90 + 35) sec) ÷ 60 sec/min = | 6.25 min |
| **Text Preparation & Markup:** | 90.00 min |
| **Automatic Index Generation:** | 1.50 min |
| | 139.05 min<br>or 2.3 hrs |

**Note: If the student performs image conversion to the proper .TIF format, 0.46 hours are saved, yielding a time of 1.84 hrs.**

‡ Full Pages: Cover, Document, and Signature Pages

## F. SUMMARY

WordPerfect 5.1 is a widely accepted word processing software package, available to students in virtually every micro computer laboratory on the Naval Postgraduate School campus. Its ease of use, and many capabilities, save substantial time during the transformation of electronic documents for storage and retrieval on optical disks. These factors make WordPerfect the logical choice for the electronic document standard.

The number of graphics contained in an electronic document, significantly affects the length of time it takes to convert that document to a disc publisher format. Each graphic must be converted to disk publisher image format, and must have an image call placed in the document text to display it; these are both labor intensive processes, as shown in Tables 6 and 7.

Graphics, contained in electronic documents, must accompany the text, as images. This will require additional effort by the students, and support by the school, to provide training in the use of optical scanners and in the graphics capabilities of WordPerfect.

## VI. CONCLUSIONS, RECOMMENDATIONS

### A. CONCLUSIONS

Conversion of paper documents, for storage and retrieval on optical disks, is an expensive, time consuming process. Conversion and storage of electronic documents is more efficient, requiring less than one-eight of the time. Conversion and storage of a document, as a collection of images, is not more efficient than full text conversion and storage. This is due to the large time requirement necessary, to produce an index, which provides search and retrieval capability, for the collection of images.

Full text document databases, provide better search capabilities, than image based document databases. The ability to search on the text contained in a document, provides superior flexibility, to accommodate the wide range of searcher experience levels, and thought processes.

Submission of thesis documents in electronic form, is the best solution for long term future storage and distribution of these documents. The electronic form used for document submission should be WordPerfect (5.1), since it is readily available, and yields the greatest benefits during the conversion process. New versions of WordPerfect (if produced by WordPerfect Corp., and adopted by NPS) would not cause significant problems. A transition period could be established where either version 5.1 or the later version would be accepted. WordPerfect documents created using version 5.1 would be imported and converted to the new version as WordPerfect's software has always been backwards compatible. Submission of thesis documents in electronic form, will require additional work by the student, but will provide efficient use of technology. Electronic documents are easily converted for storage and retrieval on optical disks. Optical disks are inexpensive to mail, and require little storagc space. Conversion and storage of electronic thesis documents directly, will reduce requirements for storage, and digital conversion of paper thesis documents in the future.

Current optical scanning capabilities, will not support the requirement, to provide images of all graphics contained in the document, as well as images of cover, document, and signature pages of a thesis document. Students will require access to software that is capable of converting images to a format compatible with the optical disk publishing software.

## B. RECOMMENDATIONS

Due to the extreme differences between converting paper and digital documents, this researcher does not recommend conversion of paper documents. Instead, a program should be instituted, where students submit one paper, and one electronic copy of their thesis. The electronic copy, would be converted and stored on a CD-ROM optical disk with other theses, for storage and distribution. The paper copy would be mailed to the Defense Technical Information Center for processing and storage. The Defense Technical Information Center would prefer to receive documents electronically; however, no program is currently underway to accomplish this until the Department of Defense establishes a standard for Electronic Data Interchange (EDI), under the Corporate Information Management (CIM) initiative (Viel, 1992).

This researcher recommenas, that the HIJAAK software purchased by the NPS computer center, be made available in the micro computer labs on campus. This will provide students the ability to submit images in a .TIF format that is compatible with the disk publishing software.

The WordPerfect course taught to students in the Administrative Science Department should be broadened, to include training on the graphics capabilities of WordPerfect 5.1. A similiar course should be established in other departments to provide instruction to their students. This will decrease the number of students required to learn WordPerfect on their own.

To provide the optical scanning capabilities necessary for electronic document submission, additional scanners should be purchased for use in the Learning Resource Centers

(LRCs) and micro computer laboratories on campus. Clear, concise instructions should be furnished to students, describing basic operation. Instruction given on WordPerfect thesis styles, should be broadened to include instruction on the optical scanners located in Ingersoll I-151, as well as the use of the HIJAAK software mentioned earlier.

A project for a follow on thesis, would be to establish a system, for mastering CD-ROM optical disks, containing completed thesis documents.

# LIST OF REFERENCES

Apperson, G., and Doherty, R., "Displaying Images", CD\ROM Volume Two Optical Publishing, Ropiequet, S., ed., Microsoft Press, 1987.

Arnold, Stephen E., "Storage Technology: A Review of Options and Their Implications for Electronic Publishing", Online, July 1991.

Blair, David C., "The Data-Document Distinction in Information Retrieval", Communications of the ACM, 27(4), 1984.

Blair, David C., and Maron, M. E., "An Evaluation of Retrieval Effectivene[illegible] For a Full-Text Document-Retrieval System", Communications of the ACM, 28(3), 1985.

Doyle, Lauren B., Information Retrieval and Processing, Melville Publishing Co., 1975.

Fink, Donald G., "Television", Lexicon Universal Encylopedia, 1984 ed.

Fowler, John D., and Clipper, Robert W. Jr., Considerations For Conversion of Microfiche to Optical Storage, Master's Thesis, Naval Postgraduate School, Monterey, California, March 1991.

Hoge, James C., Use of Optical Storage Devices as Shared Resources in Local Area Networks, Master's Thesis, Naval Postgraduate School, Monterey, California, September 1989.

KAware Disk Publisher User's Manual, Knowledge Access International, 1990.

Lesk, M. E. and Salton, G., "Relevance Assesments and Retrieval System Evaluation", The Smart Retrieval System - Experiments in Automatic Document Processing, Salton, G., ed., Prentice-Hall Inc., 1971.

McCarn, Davis B., and Lewis, Craig M., "A Mathematical Model of Retrieval System Performance", Journal of the American Society for Information Science, October 1990.

Ropiequet, S., Einberger, J., and Zoellick, B. (Eds.) Optical Publishing: A Practical Approach to Developing CD-ROM Applications, Microsoft Press, 1987.

Salton, G., "Automatic Text Analysis", Science, Vol. 168, 17 April, 1970.

Taylor, Robert R., Conversion of Hard-Copy Documents to Digital Format Utilizing Optical Scanners and Optical Storage Media, Master's Thesis, Naval Postgraduate School, Monterey, California, March 1989.

Telephone conversation between Ms. Helen W. Viel, Project Officer Electronic Document System (EDS), Defense Technical Information Center, Alexandria, Virginia, and Lawrence P. Bittner, 5 February 1992.

Wageman, C. Peter, The Handbook of Optical Memory Systems, Optical Disk Institute, 1989.

Zoellick, B., "Selecting an Approach to Document Retrieval", CD\ROM Volume Two Optical Publishing, Ropiequet, S., ed., Microsoft Press, 1987.

# BIBLIOGRAPHY

Burkett, T. G., Emrath, P., and Kuck, D. J., "The Use of Vocabulary Files for On-line Information Retrieval", Information Processing and Management, 15(6), 1979.

Gottlick, Richard C., and Victoriano, Edwin A., Optical Storage System for Shipboard Supply Documents, Master's Thesis, Naval Postgraduate School, Monterey, California, December 1989.

Kellett, Daniel A., Hypertext: Another Step Toward a Paperless Ship, Master's Thesis, Naval Postgraduate School, Monterey, California, June 1989.

Lind, D. L., Optical Laser Technology, Specifically CD-ROM, and Its Application to the Storage and Retrieval of Information, Master's Thesis, Naval Postgraduate School, Monterey, California, June 1987.

Rodriguez, Joseph F., Management Concerns for Optical Based Filing Systems, Master's Thesis, Naval Postgraduate School, Monterey, California, March 1990.

Sweitzer, Wayne F., Hypermedia and Digital Optical Media Technologies as applied to a Prototype Geographic and Threat Recognition (GEOTREC) Training and Reference Tool, Master's Thesis, Naval Postgraduate School, Monterey, California, March 1990.

# INITIAL DISTRIBUTION LIST

No. Copies

1. Library, Code 52 — 2
Naval Postgraduate School
Monterey, CA 93943-5002

2. Defense Technical Information Center — 2
Cameron Station
Alexandria, VA 22304-6145

3. Chief of Naval Operations — 1
Director Information Systems (OP-945)
Navy Department
Washington, DC 20350-2000

4. Barry A. Frew — 5
Dean of Computer and Information Services
Code 05
Naval Postgraduate School
Monterey, CA 93943-5000

5. Dr. C. Thomas Wu — 2
Code CSWQ
Naval Postgraduate School
Monterey, CA 93943-5000

6. Dr. Tung Bui — 1
Academic Associate 367
Code AS/BD
Naval Postgraduate School
Monterey, CA 93943-5000

7. LT Lawrence P. Bittner USN — 2
SWOSCOLCOM
Naval Education and Training Center
Newport, Rhode Island 02841-5051