VTT TIEDOTTEITA - MEDDELANDEN - RESEARCH NOTES 1592

Optical character recognition in microfilmed newspaper library collections

A feasibility study

Riitta Alkula & Kari Pieskä

VTT Information Service

TECHNICAL RESEARCH CENTRE OF FINLAND
ESPOO 1994

ISBN 951-38-4707-1
ISSN 1235-0605

JULKAISIJA – UTGIVARE – PUBLISHER

Valtion teknillinen tutkimuskeskus (VTT), Vuorimiehentie 5, PL 2000, 02044 VTT
puh. vaihde (90) 4561, telekopio (90) 456 4374

Statens tekniska forskningscentral (VTT), Bergsmansvägen 5, PB 2000, 02044 VTT
tel. växel (90) 4561, telefax (90) 456 4374

Technical Research Centre of Finland (VTT), Vuorimiehentie 5, P.O.Box 2000, FIN–02044 VTT, Finland
phone internat. + 358 0 4561, telefax + 358 0 456 4374

VTT Tietopalvelu, Vuorimiehentie 5, PL 2000, 02044 VTT
puh. vaihde (90) 4561, telekopio (90) 456 4374

VTT Informationstjänst, Bergsmansvägen 5, PB 2000, 02044 VTT
tel. växel (90) 4561, telefax (90) 456 4374

VTT Information Service, Vuorimiehentie 5, P.O.Box 2000, FIN-02044 VTT, Finland
phone internat. + 358 0 4561, telefax + 358 0 456 4374

Technical editing Maini Manninen

VTT OFFSETPAINO, ESPOO 1994

Alkula, Riitta & Pieskä, Kari. Optical character recognition in microfilmed newspaper library collections. A feasibility study. Espoo 1994, Technical Research Centre of Finland, VTT Tiedotteita - Meddelanden - Research Notes 1592. 53 p. + app. 1 p.

UDC 025.17
651.5
681.3.02

Keywords optical character recognition, information systems, scanners, archiving, information retrieval, electronic archives, optical archiving, microfilm

ABSTRACT

The aim of the OCR Index project was to investigate the feasibility of optical character recognition (OCR) in generating full-text indexes for newspaper collections. The project comprised a literature survey and a controlled experiment with 35 mm microfilm frames and original newspaper pages. The test material was scanned with a microfilm scanner and a A4 size page scanner. The resulting image files were processed by OCR software to produce editable text files. The purpose was to determine, whether OCR is accurate enough for producing indexes automatically.

A major problem with microfilm scanning is the relative newness of the technique. Scanners suitable for volume conversion of roll film have only recently become available and require extra accessories for handling 35 mm roll film, the commonest format used in libraries. Another problem is incompatibility of image files and OCR software, requiring special conversion programs.

In the OCR Index project, text files produced by the OCR software were analysed and recognition errors grouped into main categories. These were: unrecognised, substituted, split, joined, inserted, and deleted characters. The first two appeared to be the commonest error types. The accuracy of recognition was poorer than that of ordinary office documents. This is partly due to the problematic nature of newspaper text, which has tight character spacing, and contains multiple columns and various typefaces.

Information retrieval demands good accuracy, not only of character accuracy but also of words; a misspelled word will not match the search term. The amount of correct words appeared to be much lower than that of correct characters. A character accuracy below 98 per cent gives such poor word recognition that it no longer seems feasible to produce indexes from text files obtained via OCR.

To improve recognition results, OCR software dedicated to newspaper text should be produced. Also, automatic spelling correction methods for text produced by OCR should be improved. Text retrieval methods that can cope with incorrect words should be developed.

FOREWORD

To gain some insight to the feasibility of optical character recognition in generating full-text indexes for newspaper collections, NORDINFO financed the OCR Index project. The goal of the project was to obtain information about the technical feasibility of current microfilm digitisation and OCR techniques in the described application. The work was carried out by the Information Service of the Technical Research Centre of Finland (VTT). The tests were performed with commercial off-the-shelf scanners and OCR software.

The authors would like to take this opportunity of thanking NORDINFO for their financial support of this project. Thanks are also due to Director Sauli Laitinen and Senior Research Officer Eero Sormunen of the VTT Information Service for their useful comments in producing this report.

Espoo and Oulu, June 1994

Riitta Alkula Kari Pieskä

TABLE OF CONTENTS

1 INTRODUCTION

Many libraries, especially national ones, have extensive newspaper collections as original printed volumes or microfilm copies, or both. Microfilming is a well-established technology that offers a tried-and-true, low cost and easily-managed method of archiving different kinds of documents.

There is, however, a major problem in using newspaper archives on microfilm. Reading a microfilm is uncomfortable and can only be done by one person at a time. Processing and distributing microfilms is cumbersome and time-consuming. Microfilm records cannot be accessed with content-based retrieval. Newspaper archives are an essential information source for research in various fields of history as well as social and political sciences, but if they were more searchable they could be used more actively in these and other areas such as education, business, and industry.

In theory, the access problem can be solved in various ways. A conventional answer is to produce a bibliographic database of newspaper articles using controlled or uncontrolled index terms, but in practice libraries cannot afford the high cost of manual indexing. In some limited special collections, cataloguing and indexing of articles could be reasonable. Very few bibliographic databases covering daily newspapers are currently available in Scandinavia.

Full-text databases provide another solution. Some Scandinavian newspapers are publicly available as full-text databases, but most of these are business oriented like *Dagens Affärer* or *Kauppalehti*. The availability of general purpose, daily newspapers in electronic form is very poor.

Thirdly, major newspapers are moving from manual archives of press cuttings to in-house full-text online databases. One could imagine these opening up a wider audience, but unsolved copyright and data protection problems, as well as the economic interests of newspaper companies, promise to keep most newspapers off the online and CD-ROM market in the nearest future (Kantola 1991). How long newspapers are willing to maintain their old files, which they do not need in their daily work, is an open question.

The problem of access to newspaper collections is similar in all Nordic countries, and it is not likely to improve in near future. Copyright and data protection are more problematic here than in most countries. The market for electronic information is narrow and does little to stimulate commercial interests. The access problem with the old newspaper volumes and microfilm collections of libraries will remain.

Digital imagery is an interesting alternative to microfilming. In document image processing (DIP) systems, the pages of documents are scanned and stored as bitmap images. There are no restrictions to the content of pages and graphics are not a problem. Accessibility and browsability of the material are improved as it can be used by many people simultaneously and remotely via networks. Also, information from different sources can be accessed at a single point. Copying or transmission is possible without deterioration of the material. Copies of a database can be distributed using optical discs. The productivity of paper copy delivery services can also be improved.

The major problem with first generation DIP systems was that they worked like computerised microfilm systems and did not support content based searching (Locke 1991). They were seen as records archiving systems rather than information retrieval systems. Recently, free-text management software packages have been integrated into the major DIP systems (Anon. 1992). Although bitmapped document images require much more storage space than ASCII files, this problem is becoming less important due to new compression techniques, larger (optical) discs, and lower memory prices. DIP systems are currently used for managing technical and office documents, press cuttings, etc.

Optical character recognition (OCR) and full-text indexing seem to offer a potential solution to the access problem. OCR has been applied successfully in the retrospective conversions of library card catalogues into electronic form. Errors have been corrected by manual editing, but this is too expensive for full-text indexing of newspaper collections. On the other hand, a small number of recognition errors in the latter is less dramatic than in library catalogues as full-text contains numerous redundant term entries. In addition, the effect of errors on the searching reliability of an index can be minimised by applying morphological analysis and pattern based searching techniques. There are some examples of OCR having been used to transfer texts from printed newspapers to ASCII format (e.g. Ougham & Williams 1992b, p. 13).

The OCR Index project was carried out To gain some insight into the feasibility of current microfilm digitisation and OCR techniques for generating full-text indexes for newspaper collections. The project comprised a literature survey and a controlled experiment. Newspaper pages from *Hufvudstadsbladet* (language: Swedish), *Vasabladet* (Swedish), and *Politiken* (Danish) were used as test material. Microfilm and paper samples were digitised using a microfilm scanner and A4 size scanner. The resulting image files were processed by OCR software to produce editable text files. Recognition errors were analysed and grouped into main categories. The purpose was to determine whether OCR is accurate enough for producing indexes automatically. Assessment was also made of the possibilities and problems of OCR-based full-text searching in newspaper microfilm collections.

2 DOCUMENT MANAGEMENT

Document management encompasses the management of all forms of documents held in one of four types of media: paper, microfilm, magnetic or optical. It is difficult to understand what document management is really about as so many different terms are employed to describe a similar mixture of technology and systems. *Document image processing*, *document management*, *document imaging*, *image processing*, *imaging*, and *engineering document control systems* are some of the names used by suppliers and consultants to describe one major area of document management that relies on the technology of image processing and optical discs.

2.1 MICROFILM ARCHIVES

Microfilming has its origins in the last century. Its suitability for archival storage has been well proved, and it is accepted by courts as a legally acceptable alternative to paper. Microfilm records are durable and able to withstand even rigorous daily use, allthough heavy wear and tear on an old film may obscure some of the information. Microfilm offers several advantages over other archiving media:

- excellent longevity (over 500 years when stored in a standardised environment)
- low price
- standardised and approved media
- no risk of obsolescent technology.

Microfilms exist in three different forms: rolls, fiche and aperture cards (Figure 1). For reading and making paper copies there are several kinds of reader/printer devices available. Nowadays, as digital document management proliferates, special purpose scanners for all types of microfilm (16 and 35 mm roll films, 35 mm aperture cards, microfiche and jacketed microfilm etc.) have arrived on the market. The equipment may be intended for general use or production scanning. Mekel is the market leader in bureaux scale roll film scanners and Wicks & Wilson and Photomatrix in aperture card scanning. Although TDC offers a fairly limited system for fiche scanning, virtually nothing is available to cope with jacketed microfilm. SunRise Imaging and Bell & Howell are marketing modular scanners, which with a range of changeable heads can scan all types of microfilm from roll film to jackets. In addition to these large scale microfilm scanners is the MS-100 microfilm scanner/printer by Canon, designed for scanning individual frames of any kind of microfilm.

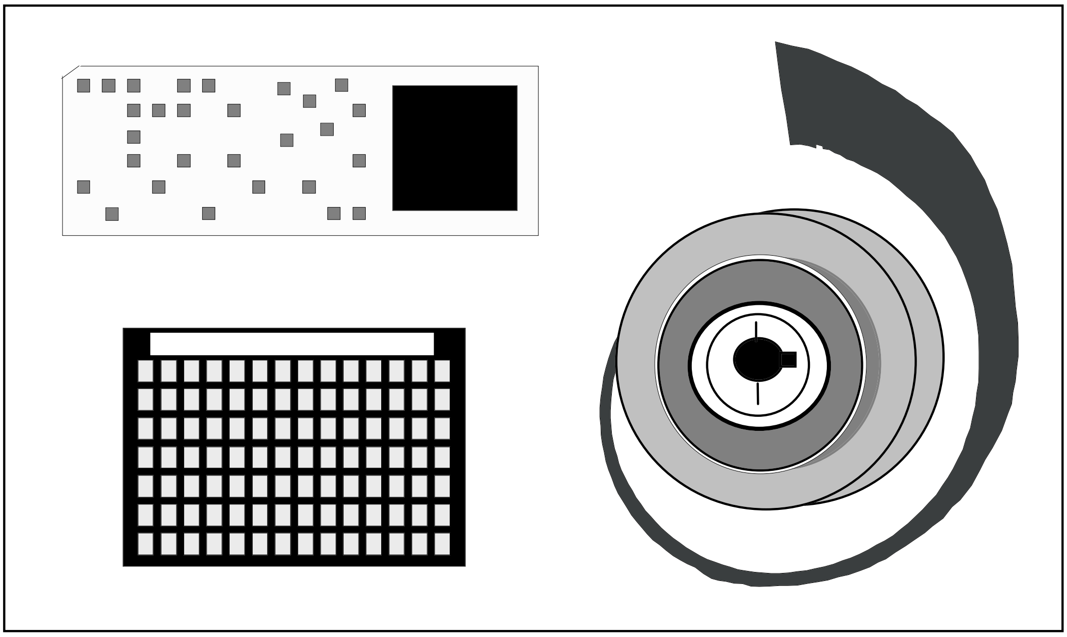

Figure 1. Different types of microfilm: aperture card, microfiche, and rollfilm.

2.2 ELECTRONIC DOCUMENT MANAGEMENT

A document image processing system (DIP) can be defined as follows (Saffady 1993):

> *DIP is an integrated configuration of hardware and software components that produces pictorial copies (images) of office files, reports, publications, and other source documents for storage, retrieval, dissemination, or other purposes and converts the documents to electronic images for storage on optical or magnetic media.*

The electronic images typically produced from paper originals by scanners are usually stored as bitmapped images on optical discs and therefore require plenty of memory space. For retrieval purposes DIP systems maintain a computer database which serves as an index to stored images. This index usually resides on magnetic disc, which guarantees fast retrieval. The use of bitmapped images differentiates DIP systems from Text Data Management systems which handle information in computer readable (usually ASCII or EBCDIC) character form. However, more sophisticated document management systems offer an option to convert bitmapped images into editable and searchable text form through OCR (see section 2.2.3).

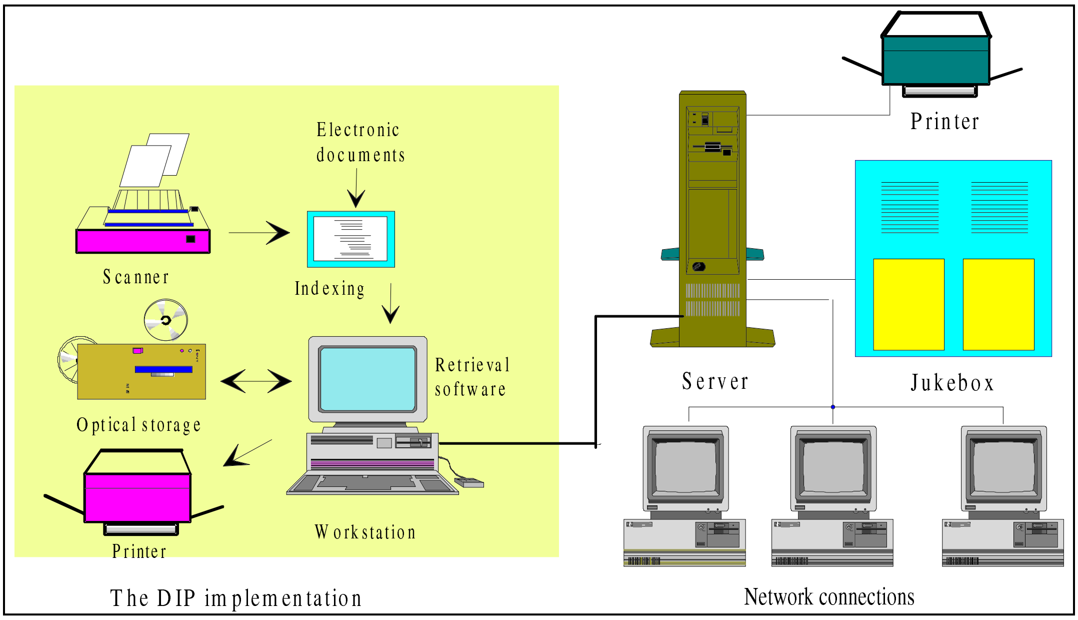

Figure 2. Typical implementation of a DIP system with connection to local area network (LAN).

2.2.1 Configuration

The first document image processing systems in the early to mid 1980s were minicomputer based turnkey systems (e.g. Canon, FileNet, Kodak, Panasonic, Philips, Toshiba, Wang). The first microcomputer based optical filing system, the Discus 1000, was introduced in 1985 (Advanced Graphic Applications). A new approach to DIP was offered by Canon in 1988 when they introduced their inexpensive personal optical filing system, Canofile 250. However, IBM's Image Plus system, also introduced in 1988, is said to have legitimised electronic document management technology in the minds of many prospective users. Since 1990 several open (or half-open) systems based on existing PC networks have been introduced (e.g. Daisy, Olivetti, Plexus). Software based implementations, sometimes called "shrink-wrapped" imaging solutions, represent the latest development positioning as an add-on application for customer-owned computer systems like word processing. A typical configuration comprises a workstation, scanner, printer, and a storage device which may be further connected to a local network (Figure 2).

At the lowest level a DIP system employs scanning technology to store and index simple documents such as invoice and billing information. At higher levels are integrated networked systems that manage fax, scanner input, OCR, full-text indexing, electronic data interchange (EDI), E-mail, and data collection from the field. This information can be made available to anyone on the network from within a variety of applications (Harvey 1991).

The ultimate scope of the electronic document management system or DIP is the complete management of documents throughout their life cycle from creation to destruction. The contribution of document management software to the capture, storage, indexing, retrieval, processing and communication of documents is therefore crucial to the operation of DIP. There are different approaches in the way of arranging the components of DIP. The system can be put together from several specialised programs controlled by the DIP program. On the other hand, the DIP software itself may handle all operations necessary for processing documents. Examples of the latter are recently introduced document management software packages such as Lotus Notes:DI, Page Keeper and Watermark.

2.2.2 Input

The input task of a document management system is to convert information to a form suitable for computer processing and storage. Documents can be complex mixtures of text, graphics, photographs, logos, signatures and tables. For these documents, often called compound documents, not only the information contents but also the format, layout, and appearance of the information are of vital importance.

Input to DIP can be done in different ways depending on the type of source documents. Information already in electronic form is accepted as such by most DIP systems. Miscellaneous low quality textual information may be transferred most efficiently by re-keyboarding. Photographs and other pictorial material can be captured by video camera, but the commonest equipment for inputting any kind of document is a scanner. The capture stage is the most critical, since what can be achieved during any subsequent processing depends on the quality and type of image information initially secured (Saffady 1993; Wiggins 1992a).

Scanners

Optical scanners provide the means to capture the image of original documents (paper, microfilm or slide). Typically a light source illuminates the document page. Page scanning is the action of converting the light reflected from the page into electrical impulses. Black areas on the page absorb most of the light, while white areas reflect it. From grey areas, light is reflected in proportion to shades of grey. With most scanners a row of tiny sensors is placed over a narrow strip of the page. The charged-coupled device array (CCD) comprising a linear arrangement of individual photo-sensitive cells generates a charge proportional to the amount of light reflected from the small illuminated area beneath it. The sensor row moves to the adjacent strip of the image and the process is repeated until the whole page has been scanned strip by strip. This electrical representation of the image may be stored, processed, or transmitted, ultimately to be converted back into visual form by electronic means (see Figure 3) (Cawkell 1989).

The image of the document page is captured as a two-dimensional grid of individual picture elements (pixels). This process is called bitmapping. For high volume scanning there is also equipment available which provides an electrical representation of the whole page thus allowing direct feed-back for adjusting brightness, contrast, etc.

For bitonal scanning the binary digital signals (bits) produced by the scanner need only represent black (bit 1) or white (bit 2) to create the required bitmap. Thus one bit per pixel is sufficient information to record bitonal images. Once scanned, the image can be automatically compressed and transferred to a storage device.

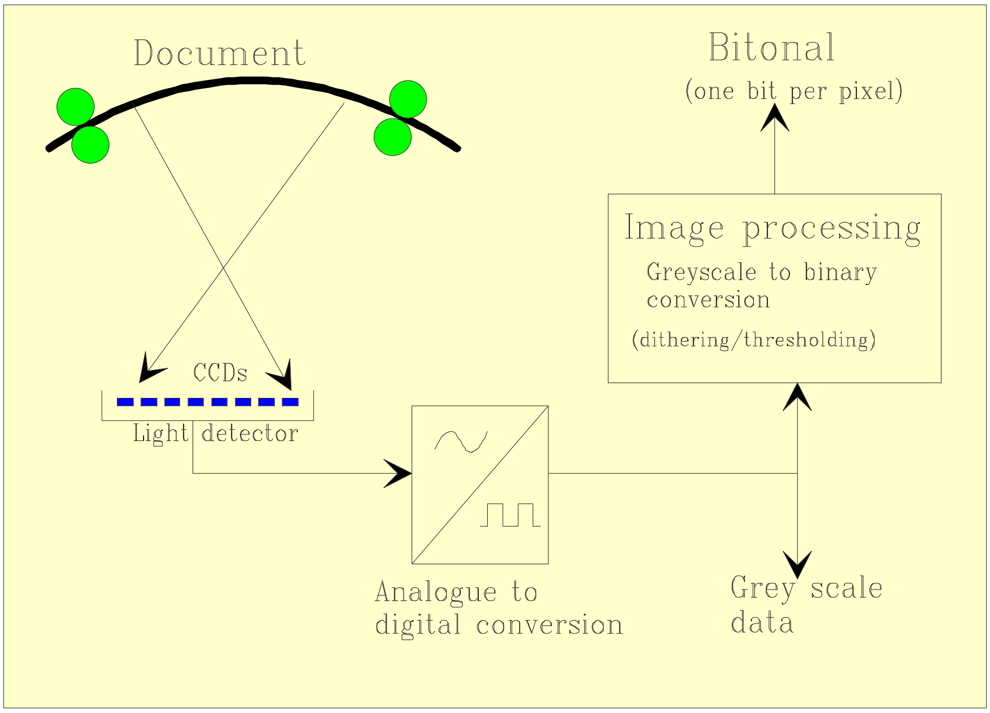

Figure 3. Document scanning (Wiggins 1992a).

Commercially available scanners include small hand-held devices, conventional flat bed and roller fed scanners, and more expensive high speed machines. Special purpose devices such as microfilm and slide scanners are also available. Well known scanner manufacturers are Agfa, Hewlett Packard, Microtek and Xerox. Microfilm scanners are available from e.g. Mekel, SunRise, and Canon.

Resolution

The number of pixels in a line and the number of lines used to create the bitmap indicate the resolution employed. Scanner resolution is normally expressed in terms of linear dots (i.e. pixels) per inch (dpi). The resolution chosen must at least be equal to that required for any subsequent processing and output. The optimal scanner resolution depends on the particular application. For most office applications 200 to 300 dpi is sufficient, since it is well matched to the 300 dpi output of laser printers. For applications requiring OCR, 400 dpi provides improved detail.

Thresholding and dithering

Most scanners capture information in multi-bit pixel or greyscale format, making the scanned image appropriate for photographic or colour applications. However, to reproduce line art or to enable the greyscale of photographic material to be simulated on bitonal laser printers, the greyscale information must be processed to produce single-bit-per-pixel images. Thresholding a greyscale image means setting a threshold, or contrast level, to distinguish between black and white. Dynamic thresholding automatically adjusts the threshold level in relation to the quality of the document being scanned. Dithering in scanning technology is often based on distributing and overlapping pixels within an area normally containing a 4 x 4 pixel matrix. Thus the reproduced image is divided into small areas filled with from zero to 16 pixels, giving anything from white, through shades of grey, to black when full (Wiggins 1992a).

Special requirements for microfilm scanning

Film scanning for back file conversion is not yet been widely used as film scanners are significantly higher in cost than paper scanners and so are best suited to service bureaux operation. Although much of the hard copy to be scanned will be from paper, there are strong arguments for microfilming documents prior to scanning them into a DIP system (Sturt 1992):

- Microfilming is a low tech operation and provides a hard copy back-up for security and possibly legal reasons (allowing for destruction of original paper).
- Micrographics bureaux can handle large volumes of paper efficiently and cheaply. Microfilming prior to large scale scanning may well turn out to be the most cost effective method for digitising paper documents.
- Film scanning is a high tech, efficient process and is best divorced from paper handling and preparation. (According to Sturt, three operators could

run more than five film scanners twenty four hours per day, which he estimates to total over 150 000 scanned A4 pages per day.)

- After microfilming, key indexing from a cheap film copy can be carried out simultaneously with scanning of the film. The index and image information need only be brought together when the data is up-loaded to the retrieval system.
- Pre-index information for automatic indexing can easily be added to the film.

The microfilm digitisation process comprises several phases. First, the microfilm is scanned with appropriate equipment. If necessary, the image is further processed, i.e. corrected with de-skew, enhance, and rotate options. The image can then be temporarily stored for viewing on a high-resolution display for quality control. If the image is of acceptable quality, it is indexed and saved in permanent storage (image database), where it can be retrieved and displayed, printed, or transmitted for use.

A scanner purposed for high volume scanning, such as needed in a micrographics bureaux, should allow high levels of flexibility and throughput. It should offer maximum automation and robustness, e.g. automatic frame detection, automatic image location, dynamic thresholding, automatic exposure and intensity control, and fast image processing and enhancement. The more automatic the scanning process is, the less the scanning costs per frame will be. Flexibility means that the scanner should support various frame formats, variable resolution (about 200 - 600 dpi), and output in generally acceptable formats (e.g. TIFF, PCX). A support for scanning compound documents would be useful, as library material includes different types of documents containing variable amounts of text, pictures, photographs, tables etc. (Shiel & Broadhurst 1992; Broadhurst 1992).

In the actual film scanning, attention should be paid to ensure the absolute integrity between the film frame number and scanner count, or the indexing will be non-sensical. Potentially problematic frames should be identified beforehand to prevent a false frame count. Problematic frames are those that overlap, have rogue shape, contain images that are out of position, or in any way differ substantially from the norm of the film. (Shiel & Broadhurst 1992, pp. 51 - 52)

As virtually all microfilm was produced before digitisation was envisaged, the specific requirements of the scanning process were not known. If it is possible right from the beginning to produce the microfilm specially for scanning, then different conditions should apply to such microfilm in comparison to that produced to be used and accessed conventionally. To ease the problems of scanning microfilm, the microfilming procedures should aim toward (Shiel & Broadhurst 1992):

- consistency of image contrast, density, position etc. from frame to frame
- low amount of skew
- only one image per frame. Two or more images per frame should be avoided if possible. If there must be more than one image, then images should be separated to give a clear edge.
- low distortion of image (e.g. when filming an open book, pages should be square and flat)
- ordering the frames to give long runs of similar frames. Occasional dissimilar frames (e.g. photographs among predominantly text) should be separated out and grouped together.
- indexing aids - because it is extremely important to keep the frame count in phase with the index of the scanned image file, consideration should be given to adding check marks at specified intervals.
- microfilm produced for scanning - microfilm standards should be revised so that new microfilm will facilitate easier digitisation. E.g. Kodak's Imagelink High Quality microfilm is claimed to have improved image sharpness and lower levels of minimum density compared with earlier products (Shiel & Broadhurst 1992). For office documents on microforms a good resolution value is 140 line pairs per millimeter, when images are captured at a reduction ratio of 24. This corresponds numerically to an approximate scanning density of 300 dpi (D'Alleyrand 1993).

2.2.3 Optical character recognition

The term optical character recognition (OCR) was launched over 30 years ago with a piece of equipment that could read just one font in one specific point size. In time, multifont systems were developed capable of recognising ten or more common typewriter fonts. However, in order to be recognised characters had to match precisely the matrix stored in the memory.

In 1978, Kurzweil introduced a radically new approach: a system that could in principle be trainable to read any font. Thereafter, several aiding components were taken into use: artificial intelligence, dictionaries, contextual rules, etc. In 1986, Calera introduced a system mimicking the learning power of the human brain, based on a trainable algorithm which enabled the computer to learn from arbitrary data input. The acronym ICR was initially used as a marketing term for intelligent character recognition, to differentiate new products of the 1980s from the original

OCR products. Lately, ICR has stood for image character recognition or "read from image" as opposed to "read from paper" technology (Schneider 1993).

The terms OCR and ICR have sometimes been used indiscriminately. In this report the term OCR is used to describe the process of character recognition, regardless of the recognition device, the types of characters recognised or the technology used to recognise them.

OCR technology has made agile progress from an approach devoid of variation to the ability to train, and then further to learning without explicit training. The First installations relied on matrix matching of scanned characters with stored sets of bitmaps. This approach proved to be very fast but eventually problematic as laser printers became more common and the variety of fonts increased. Multiple character sets required many matrices, which took a lot of memory space and thus was expensive. Then came systems based on feature extraction or topological analysis, by means of which characters could be identified by their features, irrespective of font type or size. These feature extraction systems replaced the matrix matching systems; although they were slower, they did not devour memory space or require careful character training like matrices. Most recently, neural network techniques have been developed to complement these approaches, making the system learn as it goes along (e.g. adaptive recognition of Calera).

OCR implementations may be hardware- or software-based. In the former, document digitisation and character recognition are performed by the same device in a single scanning operation. In a variant approach, recognition may be performed by a coprocessor board to which a conventional scanner is attached. Software-based implementations divide OCR into two worksteps: document digitisation and character recognition. Relatively inexpensive OCR programs are designed to operate with customer-supplied scanners on any common PC platform. As a potential limitation in high-volume applications, software-based implementations are typically slower than hardware-based OCR readers (Saffady 1993).

The most broadly used feature extraction algorithm identifies characters on the basis of their distinctive features. As an example, the lowercase "k" is recognised as a character with one vertical line joined in the middle to two diagonal lines, while an uppercase "D" is recognized as one vertical line joined at both ends by a loop (Figure 4). The best OCR products supplement feature extraction with contextual analysis, dictionary searches, frequency counts, or other techniques that increase the likelihood of accurate character identification. Although the trainability of OCR products is normally linked to matrix matching algorithms, for maximum versatility some OCR products combine built-in, omnifont recognition with trainability for unusual fonts or special characters.

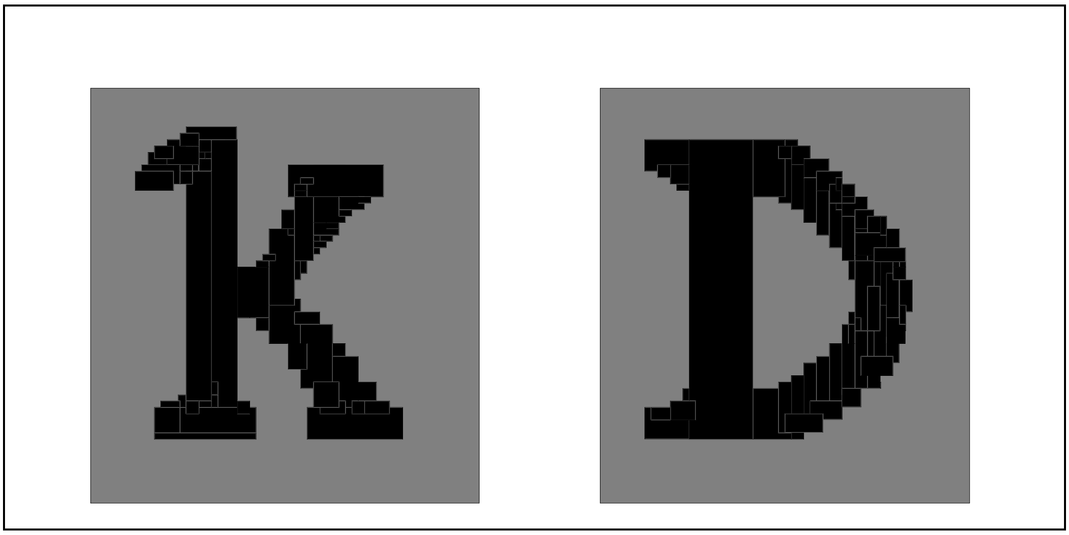

Figure 4. Bit-mapped images of characters.

The best recognition results are obtained with high-contrast documents that contain sharply defined, dark characters on a white background. Although faded printer ribbons and speckled photocopies impair recognition, scanning resolution selected for document digitisation will have the greatest impact on recognition accuracy. Most OCR products recommend or require a scanning resolution of 300 dpi. Some OCR devices and programs can accommodate document images scanned at 400 dpi but increasing the resolution will significantly retard the recognition speed. As with key-entered data, the recognised characters must be proof-read to detect and correct errors. Unrecognisable characters are marked with a specified symbol, e.g. circumflex or tilde.

The accuracy of OCR has improved greatly during the past two years. Recognition accuracies from 99.9 to 99.5 per cent have been reported in tests with printed material containing columns, different typefaces, font sizes and styles, pictures, boxes and non-alphabetic symbols (e.g. bullet points). This equals to roughly 4 - 18 errors on an A4 page of 3 500 characters. (Ougham & Williams 1992a). These results were achieved using inexpensive, general purpose OCR software. Many OCR packages support a wide range of national character sets.

Despite the rapid development of OCR technology, it is still rather difficult to proceed from image to character representation. The problem of any OCR engine is to decipher shapes of characters formed by black pixels in a sea of white pixels. Background noises such as smudges, colours, carbon stains, etc. encumber the recognition, although e.g. image enhancement offers a solution for cleaner images. There is a direct correlation between image quality and OCR accuracy. OCR programs now on the market are adequately fast for a job that is relatively easy to read (e.g. clear, uniform text), but are not accurate or versatile enough to handle non-standard type newspapers. Re-keying is still the preferred solution for large scale text conversion projects, being as economical as OCR followed by enough proof-reading to maintain accuracy.

Re-keying

If the objective is to create a 100 per cent correct text, OCR accuracy levels must be very high to avoid unacceptable error correction workloads. An A4 page with 2 000 marks recognised with 90 per cent accuracy will contain 200 errors. At five seconds per error, correction will require approximately 17 minutes, compared to the fact that a professional typist will write the entire page in less than 12 minutes! (Saffady 1993).

Consequently, the condition of the material must be controlled before scanning. If the text is not crisp and clear, or there are stains and dirt, these are likely to cause trouble in scanning. E.g. pages produced by dot matrix printers are likely to be problematic. Too much space between dots produces a broken character, which results in false recognition. Other material likely to cause trouble are old photocopies, as copying smears text edges, and tinted and coloured paper, as colour reduces the contrast between background and text and makes it difficult to separate the text from the noise (Warner 1993). With problematic material, the most cost-effective way is to retype it. This takes less time than correcting a text file full of errors.

2.2.4 Indexing and retrieval

Index planning involves careful analysis of application requirements, with particular emphasis on the ways in which document images will be retrieved. To find and retrieve a required document, it must be indexed and its location recorded. For output purposes, also details of its format must be available. Indexing provides means for fulfilling these requirements.

As for DIP systems handling large amounts of documents, indexing may cause a major problem as it may be the most labour-intensive, time consuming and expensive phase of document input. Automatic indexing is usually supported in text retrieval tools of the best DIP systems, which extract index entries (tokens) automatically from character-coded source documents. These in turn may be created from bitmapped images by an OCR engine. Just like OCR processing, automatic indexing can require considerable computer time but is still invariably faster than manual indexing. The index entries contain pointers to the digitised images from which the text files were produced.

For retrieval purposes, many DIP systems include database management (DBMS) or text retrieval systems. DBMSs (like Oracle, Ingres and dBase) normally support standard query language (SQL) to retrieve information. Most text retrieval systems rely on non-standard text search languages (Boolean search). Although DBMSs and text retrieval systems have different applications there is a trend to integrate these in DIP systems. An example of this approach is Total Recall which integrates BRS/SEARCH text retrieval software with the leading relational database management systems such as Oracle and Ingres. Other such integrated systems are Pagenet, Invesdoc 2000, and Dorodoc (Ramsden et al. 1993).

Traditional Boolean search is based on an exact match between search terms and index terms. Text retrieval systems using extended capabilities, like concept-based retrieval and fuzzy retrieval, have lately appeared on the market. The former provide retrieval using text structures and topics, the latter allow partial match of search terms. In some systems, search is made easier by allowing input of the search statement in natural language. Some retrieval systems automatically rank the search results according to their relevance to the query, which helps the user to identify quickly which documents should be read. The pattern recognition approach is capable of retrieving information using not only words but also images and other multimedia objects. Some examples of more advanced text retrieval systems are Ful/Text, Memex, Naturell, Personal Librarian, PixTex, and Topic.

When building an index from the text produced by OCR software, the accuracy of recognition is important. Although today's OCR systems may achieve an accuracy of over 99 per cent with good quality printed material, there will still be a few words with errors on each page. As most (text) retrieval systems are based on an exact match between search and index terms, recall can be impaired since a misspelled word will not match the search string. Also, the size of the index increases, or at least the index becomes garbled, as it contains a variety of misrecognised words. The traditional Boolean retrieval or even extended search methods like concept-based retrieval or relevance ranking cannot cope with misspelled words. Only search methods allowing detection of forms not identical to the word in query may be of help in retrieving incorrect words. Such pattern matching looks for similarities rather than exact matches. The more closely the letters in the index match those in the query terms, the higher the score of documents containing those terms. Missing, incorrect or extra letters will reduce the score, but will not eliminate the document as a hit until the score falls below a given threshold (Glassco 1993).

2.2.5 Storage

Potential storage media for holding documents are based on microfilm, magnetic, or optical technology. Microfilm as a storage medium has been overshadowed by enthusiasm for optical discs. Microfilm can still offer remarkable benefits, particularly on economic grounds for smaller document collections. Magnetic media such as magnetic tapes, hard discs and floppy discs are typically used for short term storage during image processing and transmission, since they offer better data access speeds and transfer rates than do optical systems. Weak long term stability, lack of portability, proportionally high costs, and low capacity are some disadvantages of magnetic media. Memory capacities of different magnetic and optical storage media are presented in Table 1.

Table 1. Typical capacities for current magnetic and optical storage media (Wiggins 1992a).

MEDIUM	Size [inches]	CAPACITY [Mbytes]
Floppy disc	3.5	.72 - 1.44
	5.25	.36 - 1.2
Hard disc		20 - 1000
CD-ROM	4.75	640
WORM	5.25	600 - 1280
	12	4400 - 9000
	14	6800 - 10200
Optical card		2 - 4
Rewritable	3.5	128 - 256
	5.25	512 - 1024

Optical storage technology involves writing data onto a sensitive recording medium using a laser beam, and reading data back by shining a laser beam onto the surface of the recording medium and detecting the level of light reflected back. As with magnetic recording, optical recording supports a number of formats, including optical discs, optical cards and optical tapes. The capacity of an optical card (2 - 4 MB) is too small for serious document management. Optical tapes manufactured by e.g. ICI offer a flexible alternative to rigid optical media for large scale document storage (1 terabyte per reel).

Optical discs fall into three groups: Read Only, Write Once, and Rewritable (Anon. 1993b). Optical discs offer high storage capacity, portability, durability, and easy access to the data. With Compact Disc Read Only Memory (CD-ROM) the user cannot store data or change what is already on the disc, whereas WORM storage devices permit the user to write data once, which then cannot be changed or rewritten. Erasable or rewritable disc technology combines the erasability of magnetic storage with the high capacity and permanence of optical storage. An erasable disc 5.25 inches in diameter, holding 600 MB, can store up to 1000 times more information than a floppy disc of the same size (Heimbürger 1990).

2.2.6 Output

The output device should be capable of handling the required resolution. Comfortable reading of an A4 page of text requires a resolution some three times greater than the 80 dpi available on a regular television set. If it is only necessary to identify the required image on the screen, and request a copy of it via some other route, standard PC displays (EGA or VGA) may prove adequate. Complex or large size images, as exemplified by engineering drawings, may need dedicated workstations with 19" or even larger screens. The resolution of a given display device may be expressed in two ways: (1) as the number of pixels along the display area's horizontal and vertical dimensions, or (2) as the number of pixels or dots per inch that are displayable within a digitised document image.

The familiar resolution of 14" VGA monitors can thus be expressed as 640 x 480 lines per screen or 60 dpi. Graphic display devices, sometimes described as bit-mapped video monitors, intended specifically for electronic document imaging applications can offer much higher resolution displays. A 15-inch portrait-mode display with a resolution of 2200 lines with 1700 pixels per line is a standard grade for modern document image processing systems. While this equals 200 dpi, the commonest resolution used by scanners, the digitised document images can be displayed at the same resolution at which the source documents were scanned.

19-inch landscape-mode bitmapped video monitors are increasingly supplanting the 15-inch monitors in electronic document imaging installations. Most of them operate at resolutions of 110 to 150 dpi, which permit the legible display of 6-point type. The higher resolution displays (200 to 300 dpi) have never become commonplace due to their high cost and because they exceed many application requirements and customer expectations. At 150 dpi, the display area of a 19" monitor contains 1538 lines with 2048 pixels per line for a total of about 3.15 million pixels per image. As e.g. a digitised A4 sized page scanned at 200 dpi typically contains over 3.8 million pixels, the digitised image must be scaled for display at lower resolution. Special circuit boards for such scaling are available, but it may also be performed by the image processing software (Saffady 1993).

Electronic document imaging systems always include one or more laser printers for paper output, which is especially important where retrieval workstations operate with low-resolution or partial-page displays. In most installations, laser printers support black-and-white printing resolutions up to 300 dpi, which is quite acceptable for most electronic document imaging applications. Higher resolution printers (600 dpi or more) and colour printers are increasingly available but rarely required since scanning resolutions seldom exceed 300 dpi and colour documents play a niche role in today's DIP systems.

As with display screens, the choice of a printer for hard copy output should be matched to the application. For normal text and bitonal line work scanned at 300 dpi or less, output to a standard 300 dpi laser printer is acceptable. For colour or greyscale images, using a 600 dpi printer could yield better results. If a typeset quality output of 1200 or even 2400 dpi is required, the printer must be capable of operating in the equivalent range (Wiggins 1992b).

Typical single-workstation laser printers are slow, producing usually only 4 to 6 pages per minute. For multi-workstation electronic document imaging configurations there are much faster, more expensive printers available. Connected to a LAN-based system a high-speed laser printer can operate at a rate exceeding 50 pages per minute, giving a duty cycle of about 4 million pages per year.

As an alternative output option, electronic document images can be recorded on microfilm through a computer-output microfilm (COM) process. As output media for electronic document images, COM-generated microforms offer several advantages (Saffady 1993):

- compact dissemination media for large quantities of printed images
- human-readable media independent of the DIP system used
- stable backup and legally accepted storage media.

The transmission of electronic document images is still another way of output. The ability to transmit digital images between system components is one of the principal advantages of electronic document imaging technology. Unlike paper documents or microfilm images, digitised document images are transmission-ready. As a standard or optional feature, many electronic document imaging systems support a facsimile gateway consisting of a fax modem and its associated software (Saffady 1993).

3 EXPERIMENTS ON DIGITISING MICROFILM

Microfilm plays an important role in library environments. It has been a medium of choice for preserving large newspaper collections. The New York Public Library first experimented with microfilming newspapers in 1933. Helsinki University Library started microfilming newspapers in 1951 and today has a collection exceeding 30 000 rolls of 35 mm microfilm. Because microfilm has limited physical and intellectual access, recent experiments have focused on converting microfilmed library materials to digital form to improve access to collections.

3.1 CIMTECH MICROFILM DIGITISATION PROJECT

35 mm roll film is by far the commonest format used in libraries. Unfortunately, scanners suitable for the volume conversion of roll films have only recently become available. Most of the experience with digitising microfilm has been gained with 35 mm aperture cards, which are widely used in industry for technical drawings, maps and plans.

A very extensive test programme with 35 mm roll film digitisation was performed at Cimtech (National Centre for Information Management and Technology) and funded by the British Library Research and Development Centre. Cimtech undertook an eighteen month project designed to investigate the feasibility of digitising both printed library material and microfilm (Shiel & Broadhurst 1992; Broadhurst 1993).

At the start of the Cimtech project, no suitable 35 mm roll film scanners for high volume digitisation were available. Later, Mekel M400XL and SunRise DMS 50i-R were introduced and a controlled test programme was performed with these. Both of the machines are designed for volume production in bureaux and the like, and can operate in automatic mode to allow unattended scanning. They cover all film formats and have outputs in user selectable file formats (e.g. TIFF).

The trials revealed a number of factors which must be taken into account to capture high quality digital images from 35 mm roll microfilm. Good digital images require good quality film; from poor film they are of variable quality. Unfortunately, the quality of library microfilm can vary considerably and not only from one roll to the next, but also from frame to frame within a roll.

A poor microfilm is characterised by low photographic quality (poor contrast and resolution), skewed images, inconsistent placement of the image within the frame, shadows and image distortion caused by the pages of the document not being flat during filming, and "show-through" of print from the reverse side of the page. Image orientation may be horizontal or vertical. If the film contrast is particularly low it will be difficult for the scanner to distinguish between the image and the background.

One of the difficulties with 35 mm film is that it does not include image retrieval marks of the sort common with 16 mm commercial microfilm. Such marks (blibs) would facilitate the accuracy of automatic scanning as the machine could advance the film the required amount between frames. Without blibs the scanner should be able to sense the edge of the frame or even the edge of the page within the frame. If the images are located squarely and centrally within the frame, scanner parameters can be set for the first frame on the roll and all subsequent frames can be scanned without further adjustment.

Digitising microfilm such that the whole frame is captured causes the fewest problems. However, many 35 mm roll films used in the Cimtech tests contained two images (e.g. the full spread of a tabloid newspaper) per frame. Splitting double-page images into individual frames can be done but the economics of so doing will depend on the quality and consistency of the image. If the division between the pages is unclear or the image moves within the frame, a great deal of manual adjustment is necessary.

The scanning of paper-based library documents was also tested at Cimtech. The main problem with this material was that most document scanners have been designed for business applications where the documents tend to fall within a limited range of sizes and are often single sheets that allow automatic handling. Library material introduces a number of additional factors such as bound, fragile or oversized material. Few document scanners will accommodate material larger than size A3. As there are no suitable scanners for bound material, the spine has to be removed and the book reduced to single sheets. This entails slow manual operation and is impossible with unique, archival books.

As the preparation of paper documents for scanning can cost as much as the scanning itself, it may be more cost-effective to microfilm the paper documents first and scan the microfilm. In some cases the results have been found to be indistinguishable from those obtained by scanning the original material on a document scanner and, equally important, could be achieved more easily as the film is designed to be handled mechanically (Broadhurst 1992).

The scale of reduction of a film image must be known in order to determine the scanning resolution (Broadhurst 1992). Reductions of some eight to 10 times on 35 mm film (typically books and tabloid newspapers) have been shown to give results comparable to those obtained from a paper document scanner at the equivalent resolution. For broadsheet newspapers scanned at one page per frame in cine format the reduction factor is roughly 16. In the Cimtech tests, each page of the sample film (*The Scotsman* newspaper) contained a large amount of small print and yielded a compressed image file nearing one megabyte. In this case it was concluded that for this type of material good quality scanned images are possible, with detailed attention to setting scanner and enhancement parameters. This was, however, demonstrated by work on partial images only.

In the Cimtech tests, image enhancement was used to clean up the digital image and thus give better prospects for OCR. Contrast and sharpness of the image were sharpened, noise-like stains and speckle removed, and broken lines and solid areas filled. A useful spin-off of image enhancement was that it usually produced a smaller image file.

In most cases image enhancement appeared to give better recognition results than no enhancement, but occasionally resulted in loss of information when pixels were removed from degraded characters, making them more broken. Enhancement could not be applied automatically, the operator having to determine the optimum setting either by experience or trial and error. Thus the technique cannot yet be considered as an anytime solution, but should be used with care and only with appropriate materials.

Attempts to apply intelligent character recognition to raster images from library microfilm achieved only limited success in the Cimtech tests. Recognition of image files is highly sensitive to the quality of digital images. OCR works best with simple type faces and layouts. Unfortunately, library material is not such and recognition appeared to be slow and unpredictable with test images. This was partly due the nature of the material used in the tests, most of which was of historical nature and not of particularly good quality.

Short extracts from four different sheets were selected to see whether OCR errors follow any particular pattern. This was not found regarding either words or characters, except for a few of each which anyway were predictably difficult. An unexpected result was that the error rate did not quite follow the same track as quality. Originals processed directly were best, followed by the paper scan. Generally 35 mm was better than 16 mm microfilm, which is predictable as 35 mm film has higher resolution due to the larger size of film frames (Shiel & Broadhurst 1992, p. 45).

The Mekel and SunRise scanners used in the Cimtech tests can output image files in TIFF format, which can be imported by most OCR systems. However, as there was no interaction with the scanner to optimise the image for recognition, the OCR system was dependable on the quality of the TIFF file. It was found that the best images for viewing and printing are not necessarily the best for OCR. To ensure that the OCR software is compatible with the image file, it should be integrated with the scanner. A trainable OCR system is probably necessary to cope with the variety of library material. Cimtech recommends that at present, with other than high quality material it may be safest to concentrate OCR on tables of contents, indexes, notes, introductions, synopses, and other material used to assist with indexing etc.

3.2 YALE UNIVERSITY LIBRARY

Another large scale project involving the conversion of library materials from microfilm to digital images was launched at Yale University Library in 1991. The purpose of Project Open Book is to test the feasibility of digital imaging as a preservation tool. Some 10 000 volumes of books will be digitised from 35 mm microfilm. Project goals and scope, with a preference list for system developers, have been defined by Waters (1991).

Project Open Book arose from the realisation that microfilm, although the medium of choice for preserving library materials, makes access to information more difficult for readers. It is highly durable and able to save the entire contents of a collection in compact form using relatively simple and well established technology, but is harder to read than a book. The key hypothesis of the project is that digital imagery can improve access through printing and network distribution at a modest incremental cost over microfilm. Assuming that the material is already microfilmed, the simplest and most straightforward imaging system would be one that produces from the microfilm a high quality printed copy of the original document for the library to return to its shelves. Naturally, capturing and storing documents in digital image form is necessary to gain further improvements in access, e.g. through the application of OCR (Figure 5).

Technically, the implementation of Project Open Book will have four main components. The first is a conversion subsystem for capturing microfilm in digital image form, for structuring individual image components into a document, and for storing the results on a magnetic or optical medium. The conversion process includes scanning, quality control, structure composition and indexing. The documents are converted at the highest possible digital resolution (i.e. 600 dpi) to avoid repetition of the task. The image data is stored in TIFF format, with the CCITT Group IV standard for compressing files.

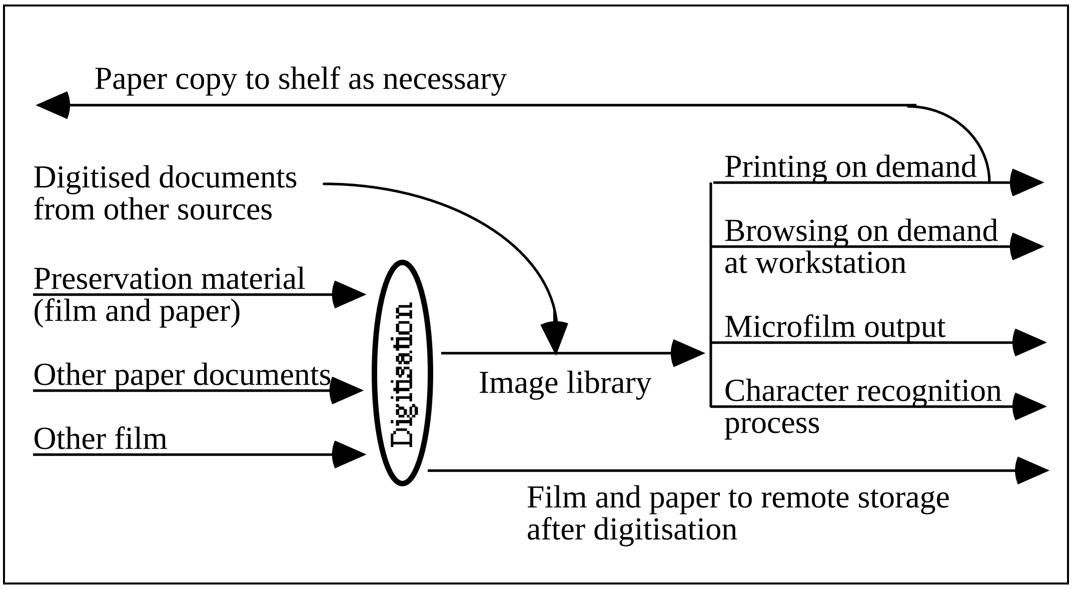
Paper copy to shelf as necessary
Digitised documents from other sources
Preservation material (film and paper)
Other paper documents
Other film
Digitisation
Image library
Printing on demand
Browsing on demand at workstation
Microfilm output
Character recognition process
Film and paper to remote storage after digitisation

Figure 5. Digital imagery at Yale University Library (Waters & Weaver 1992).

The second component is a storage subsystem for managing document images and associated structure files. It is planned to be accessible through the network and contains the digital image documents in two formats: low resolution form at 200 dpi primarily for browsing, and a relatively high resolution form at 600 dpi primarily for printing. The images will be stored an optical, and the document files on magnetic, media.

Bibliographical information referring to a document in image form will be entered into the library's online catalogue in standard MARC format. Browsing through an image document requires description of the contents and structure of image documents, which is planned to be done in conformance with ISO 8613, i.e. Office Document Architecture (ODA) and Interchange Format (ODIF). These structure files need to be stored in and retrieved from a database file linked referentially to the document image files. This database is expected to be relational and conform to Structured Query Language (SQL).

Third, the Open Book system supports browsing stations distributed on the campus network within Yale University Library. For network transmission, the system supports TCP/IP ethernet network protocols. The personal computer workstations will allow users to zoom and rotate images, place bookmarks, and advance to specified pages. The Xerox CLASS software will be used for this purpose.

Finally, the imaging system will provide network access to high quality Xerox Docutech image printers for reproducing the image documents on paper upon demand. The CLASS software is also planned to support page printing by printers attached to the user's workstation.

Project Open Book comprises six phases over a period of three years. By the end of 1992, the project had been established and a formal bid process conducted. As a result of the vendor selection process, Yale University Library chose the Xerox Corporation to serve as its principal partner in the project. The outcome of the first phase is described in greater detail by Waters and Weaver (1992).

3.3 THE FRENCH NATIONAL LIBRARY

A project to scan and index the literary works of the French National Library Service was launched in 1993. The French government decided to change to electronic archiving methods for its library service in order to increase accessibility of information and turn it into a form which can be networked and supplied to a number of display locations from a single source. (Anon. 1993a)

The aim of the project is to archive part of France's collected literary works in a form which will give readers, researchers and academics much faster access than traditional printed books. Some 300 000 documents are to be digitised. Image Scanning Services (ISS) are currently converting traditional microfilm and micro–fiche literature into more accessible electronic form, involving the scanning of over nine million frames. Pindar Infotek is responsible for scanning over 50 000 books. The pages will be scanned to create a database featuring pages of text and images exactly as they appear in the original book. A full index will be compiled simultaneously. Pindar Infotek will also create an index for books currently stored on microfilm and microfiche (Anon. 1993a; Feretti 1994).

4 VTT EXPERIMENT ON NEWSPAPER DIGITISATION

4.1 GENERAL

The aim of the OCR Index project was to investigate the possibility of producing indexes from scanned newspaper pages, primarily by scanning 35 mm microfilm frames. The purpose of this study was to find the correlation between character recognition accuracy and quality of search. The idea was to scan a full film frame once and feed the TIFF file into OCR software to test the feasibility of producing indexes from such an "automatically" produced text file.

Although some earlier projects have been carried out on microfilm scanning (chapter 3) these have focused primarily on scanning for preservation, and have dealt mainly with books. Only the Cimtech project had newspapers as part of its material. Other experiments have looked at specific problems and possible systematic recognition errors in performing OCR, but the test material has been paper documents, not microfilm (Sun et al. 1992; Nartker et al. 1994; Taghva et al. 1993, 1994). To the best of our knowledge, examination of the whole process from microfilm digitisation to OCR performance has not been systematically done elsewhere.

The scanning tests in this project were formulated to follow the programme performed by Cimtech (Shiel & Broadhurst 1992), although in more limited scope. 35 mm microfilm rolls from two Finnish-Swedish newspapers, *Vasabladet* and *Hufvudstadsbladet*, were used as the test material. The film samples (duplicates) were obtained from Helsinki University Library. The selected newspapers were of normal size, no tabloids. For scanning from paper originals we used the above two Finnish newspapers and a Danish newspaper, *Politiken*.

The project workstation consisted of a 486 PC with 16 megabytes of RAM and a one gigabyte hard disk, running under DOS 5.0 and Windows 3.1, with a VGA display. The OCR software mainly used was Caere´s OmniPage 2.11, with additional tests performed with XIS K5200, Recognita 2.0 and GigaRead. OmniPage was chosen as it was felt to be state-of-the-art and effective in handling typescript, and offered a wide range of options and settings.

4.2 MICROFILM SCANNING

In the OCR Index project, microfilm scanning was intended to be a relatively routine task, with the main emphasis on OCR. As it turned out, problems with the scanning forced us to change the project plan and pay more attention to the scanning phase. Due to these problems most of the test material was finally scanned from paper originals.

The original plan was to use a service bureaux to scan the microfilm, but it appeared that no bureaux in Finland could handle 35 mm roll film. A Swedish bureaux serving the whole of Scandinavia offered a test scan, although the actual scan would have to be done in a Dutch conversion centre. The price of the test scan, however, was prohibitive and the approach was dropped.

A suggested alternative was to cut the microfilm into individual frames and set each frame into a slide scanner. The scanned images would then be saved on Kodak Photo CD, from which editable TIFF images could be captured. Although this type of procedure might be applicable in specific circumstances it is not viable for large quantities of microfilm. As one Photo CD contains about 100 pictures, converting a full 35 mm microfilm containing 600 - 700 frames would require 6 - 7 discs. This procedure is highly labour intensive and consequently very expensive. In addition, as the microfilm would be mutilated, it would first have to be copied to preserve the original.

The third option was to obtain a microfilm scanner and scan the samples in-house. At the time, the only off-the-shelf product available in Finland was the Canon MS-100 microfilm scanner. Equipped with an ordinary lens it could not handle a full 35 mm frame, which meant that we could only scan half an A2 page at a time. To get an adequate image file, the test frame from *Hufvudstadsbladet* was scanned in four pieces, each corresponding to one A4 sheet. The resulting image files were of high quality as the film was a brand new copy.

In preliminary tests the SunRise scanner was also tried but gave such a poor image that OmniPage was unable to read the image file. One reason could be insufficient luminous intensity due to the SunRise scanner not being equipped with a wide-angle lens for 35 mm frames. The maximum scanning area was roughly half a frame, which gave insufficient luminous intensity for a clear image. Similar obstacles were also encountered with this equipment in the Cimtech tests. Some tests have been performed by scanning only part of a page, which has been reported to produce image files of acceptable quality (Shiel & Broadhurst 1992). Another possible reason for the poor quality is that the film used in OCR Index project was from 1985 and had already been used in the library, causing scratches.

Scanning only parts of a newspaper page is not as fast and cannot be performed as automatically as for an entire frame. Another and perhaps more severe problem is that lengths and layouts of articles vary, making it difficult to focus on the scanning area without splitting words or columns. If the image file contains parts of words, broken and incorrect words will appear in the text file and reduce accuracy. If these are then saved in the index its size will increase needlessly, and false index terms may create noise in the retrieval results. Splitting may result in totally meaningless character strings or in words with a totally different meaning; for example the word *talade*, which means 'spoke', might give *ta*, meaning 'take', and *lade*, meaning 'laid' or 'put'. This could be avoided with software that automatically stitches together partial scans to form a single image. One example is the CatchWord Pro software for Logitech ScanMan hand-held scanners.

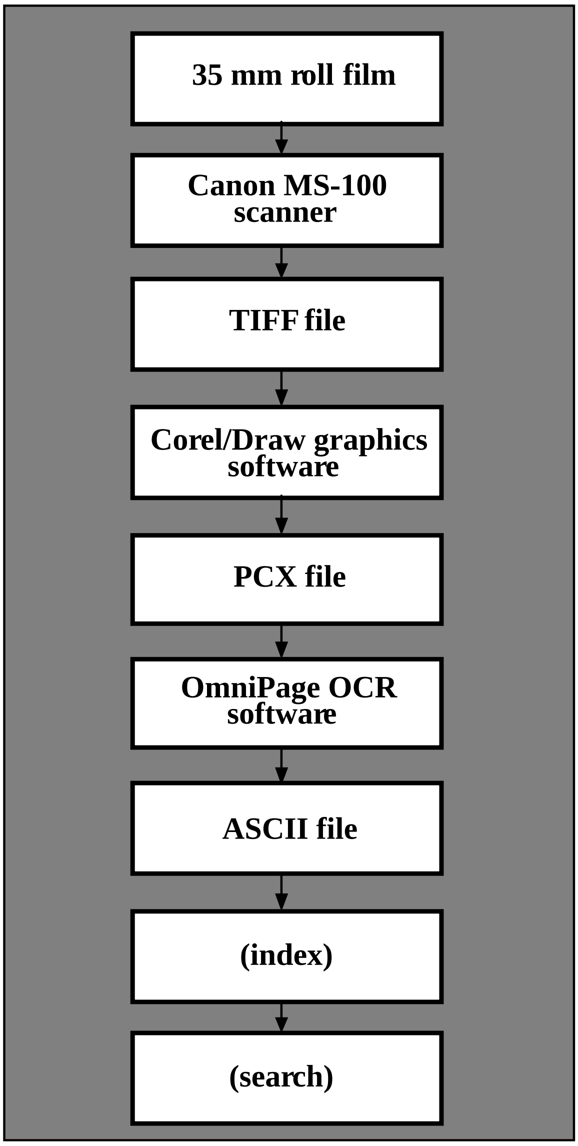

Figure 6. Microfilm scanning process in the OCR Index project.

One set-back with the Canon MS-100 scanner is that it produces TIFF files unreadable to the software we tried in the preliminary tests (OmniPage, Recognita, GigaRead, proLector). Incompatibility of TIFF formats is by no means uncommon, and was dealt with in the Cimtech tests by using a special image conversion programme to make TIFF files comprehensible to OCR software (Shiel & Broadhurst 1992). In OCR Index project the unsuitable TIFF code was converted to PCX format by CorelDraw 3.0, a PC illustration package. The resulting PCX files could be read quite easily by the OmniPage OCR software. The final microfilm scanning process is illustrated in Figure 6.

4.3 PAPER SCANNING

As scanning from microfilm did not prove technically feasible, additional tests were performed by scanning directly from newspaper originals with an HP Scanjet+ scanner. The A2 test pages from *Hufvudstadsbladet* and *Vasabladet* were scanned in four pieces, each corresponding an A4 sheet. In some tests an A4 sample was also taken from *Politiken*. The scanning process was controlled by the OCR software itself (OmniPage 2.11), thereby producing no incompatibility problems. The test pages were scanned at 400 dpi resolution and saved in TIFF format, which was compressed according to CCITT group IV format.

Scanning newspapers with an A4-sized scanner is troublesome, as an A2 page must be placed with extreme care to produce an unskewed image. The resulting A4 images were of good quality, but the columns were often split into more than one image, creating false characters and words in the resulting text file.

For test purposes the scanning of an A2 page in four (or two) pieces may be feasible, but in operational use takes far too much time. In operational use the newspaper page should be scanned in one pass with an A2-sized scanner. Some of these are now available. In our project a German Ratio-A2 scanner was tested with a couple of sample pages. The resulting image file seemed to be quite clear, but unfortunately could not be read by our OCR software as the compressed TIFF format was non-standard.

4.4 OPTICAL CHARACTER RECOGNITION

In earlier microfilm scanning projects very little attention has been paid to the problems of OCR technology. Emphasis has been mainly on conservation, fulltext indexing being only a secondary issue. In these projects a catalogue entry has usually been regarded as sufficient for retrieval tests.

Conservation projects usually assert that once scanned, image files are apt for OCR, although this option has not been applied in the project. From the experience gained in the OCR Index project it can be said that recognition of foreign image files is not a straightforward task. OCR programs normally incorporate scanner drivers of their own and allow adjustment of scanning parameters such as resolution. However, the amount of adjustable parameters is restricted when dealing with foreign image files. OmniPage 2.11, for example, supports scanning in 200, 300, and 400 dpi resolution, but only when scanning with OmniPage itself. With importation of foreign files only 300 dpi resolution is possible. Similarly, it is normally possible to select only part of the image to be recognised, which is convenient when cropping pictures or unnecessary text. With foreign files this option is disabled in OmniPage. To get optimum results, OCR software should be integrated with the microfilm scanner.

The total number of errors and the distribution of error types in the text generated by OCR has been claimed to depend on the recognition environment in which the OCR process is executed. Such an environment usually consists of a scanner, recognition software, and the characteristics of text to be recognised. The major characteristics of text are the fonts, sizes, and styles of characters, and the printing quality of the text to be recognised (Sun et al. 1992). In our project the variables of the recognition environment were resolution (300 or 400 dpi), the scanner (Canon MS-100 film scanner or HP Scanjet+ paper scanner), and the newspaper (*Hufvudstadsbladet*, *Vasabladet* or *Politiken*), while the OCR software remained constant, i.e. OmniPage.

All OCR software can cope neatly with individual, clearly separated characters. With newspaper text this is not the case, as newspaper pages contain tightly spaced small type. As letterspacing is very tight, characters are easily adjoined. The paper quality is not very high, and the pages contain multiple columns, narrow gutters, photos, and special symbols. Consequently, newspaper text is one of the most difficult text types to be recognised, as many hard-to-read elements are combined in it.

A special problem when dealing with the Nordic languages is their special character sets. Although the major OCR programs support several character sets, the test results achieved with texts in English cannot be directly applied. An interesting problem with OmniPage trials was that although the Swedish character set was selected, OmniPage 2.11 did not do well with Scandinavian characters "å", "ä", and "ö", but often recognised them as "a", "a" and "o", respectively.

The error rate of recognition is defined as the number of misrecognised or unrecognised characters divided by the total number of characters in the text (Sun et al. 1992). An unrecognisable character cannot be recognised by the OCR software and is usually marked with a special non-recognition mark, say tilde (~). This kind of error is relatively easy to correct, as the OCR software itself sees the character as problematic, and marks it clearly in the text. More troublesome are misrecognised characters, where one or more characters are mistakenly recognised as one or more different characters than in the original text. These are more difficult to detect, as there is no indication of false recognition. In this case the errors can be found only by comparing the recognised text to the original.

According to Sun et al. (1992), some earlier researchers have stated that errors introduced by OCR input devices are only the substitution of one letter for another, i.e. characters are never inserted or deleted by the OCR reader. Sun and others, however, noted that this is not the case: in fact, it is rather common for two adjacent characters to be misrecognised as one character. For example, "ll" may be recognised as "ü", "rn" as "m", "in" as "m", etc. Sun and others did not encounter the opposite type of error, where a character is misrecognised as two or more characters. In the OCR Index project, however, did; e.g. "ä" was recognised as "ft" or "n" as "ri".

In the OCR Index project the following classification for errors was produced: Erroneous characters can be:

- unrecognised: OCR does not recognise the character and replaces it with a non-recognition mark (e.g. "~")
- substituted: OCR gives a false interpretation of a character (e.g. recognises "t" as "l" or "e", "ä" as "a", "g" as "&")
- split: OCR interprets a character as two (or more) characters (e.g. sees "B" as "l3", "m" as "rn").
- joined: OCR joins two (or more) characters together, which reduces the amount of characters (e.g. joins "rn" to give "m" and "ll" to give "D")
- inserted: dirt or other stains are recognised as characters not in the original text (e.g. ".", "_")
- deleted: a character is e.g. too obscure for OCR and is deleted altogether (e.g. ".", "-", "i", ",").

A special case of character deletion or insertion is where a space is inserted into a word or a space is removed from between words. If a word is split by a space (e.g. "talade" to "ta lade") or several words are joined together (e.g. "väcks mot Sami" becomes "väcksmotSami"), they are not the same words as in original text and are not retrieved correctly.

An additional group termed "unclassified" was used in cases where OCR produced a result that could not be analysed in detail (e.g. "ddf~rtifr,i~tna;t" for "därifrån, trygg och glad. Jag hade träffat"). Characters present in the text file due to splitting of columns via an A4 scanner were not included in the calculations, as they resulted from the test arrangements, and not from the OCR technology.

Although OmniPage is an omnifont program, and should recognise any kind of text, the recognition accuracy proved to be relatively poor. With images from the microfilm scanner, the commonest OCR error was substituting a character with another. Most problems were caused by Scandinavian characters (å, ä, ö). Other types of errors were less frequent. The amount of characters that OmniPage considered problematic, i.e. unrecognised, was small compared to the total amount of false characters. Even the amount of joined and split characters seemed slightly higher than that of unrecognised characters. Character insertion and deletion seemed to be the rarest error types (Table 2, Figures 7 and 8).

Table 2. Results of character recognition with OmniPage. HB = Hufvudstadsbladet, VB = Vasabladet; Canon = Canon MS-100 film scanner, HP = HP Scanjet+ A4 scanner; 300, 400 = resolutions of 300 dpi or 400 dpi.

Characters	HB Canon 300	%	HB HP 400	%	VB HP 400	%
Unrecognised	119	4.8	195	46.3	162	41.4
Substituted	1 276	51.3	121	28.7	112	28.6
Joined	243	9.8	13	3.1	14	3.6
Split	167	6.7	19	4.5	30	7.7
Inserted	63	2.5	8	1.9	36	9.2
Deleted	293	11.8	19	4.5	33	8.4
Added space	62	2.5	45	10.7	1	0.3
Removed space	191	7.7	1	0.2	3	0.8
Not classified	71	2.9	0	0.0	0	0.0
Incorrect characters total	2 485		421		391	
Characters in original text	14 058		15 755		19198	
Accuracy %	82.3		97.3		98.0	

With images from the A4 scanner the commonest error was leaving a character unrecognised. Character substitution was another very common error. As seen in Table 2, the error profiles and amounts of errors in microfilm samples were clearly different from those in paper samples. Accuracy was roughly 82 per cent for recognition from a microfilm scanner and 97 or 98 per cent, from an A4 scanner.

One reason for the inaccuracy of OCR results was the newspaper text itself, which has tightly spaced characters regardless of whether it is scanned from a microfilm or from a paper original. The recognition of a scanned newspaper page typically does not reach the accuracy level of a scanned office document (Warner 1993). In our tests, apparent sources of errors were large sized characters in headlines and italicised text. The problem with a poor recognition level is that when the accuracy falls below about 90 per cent, the user spends more time proof-reading and correcting the text than if he or she were to retype it.

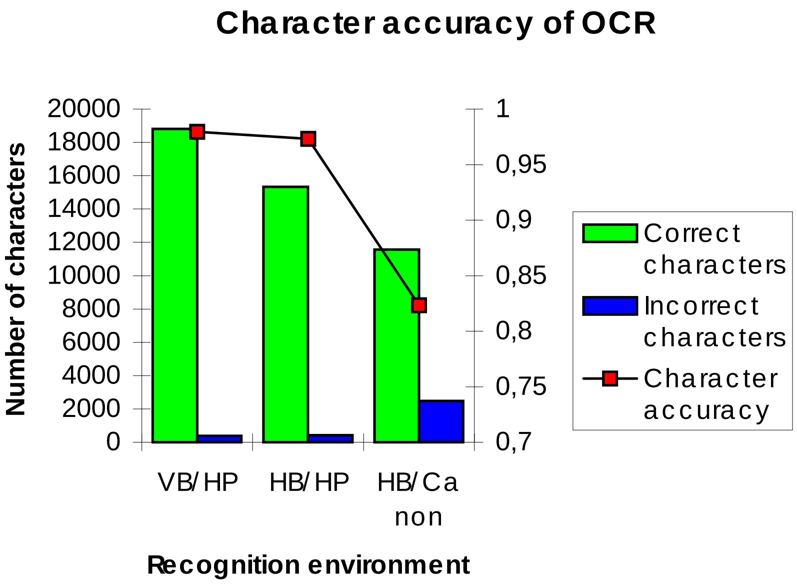

Figure 7. Results of character recognition with OmniPage. HB = Hufvudstadsbladet, VB = Vasabladet; Canon = Canon MS-100 film scanner, HP = HP Scanjet+ A4 scanner; 300, 400 = resolutions of 300 dpi or 400 dpi.

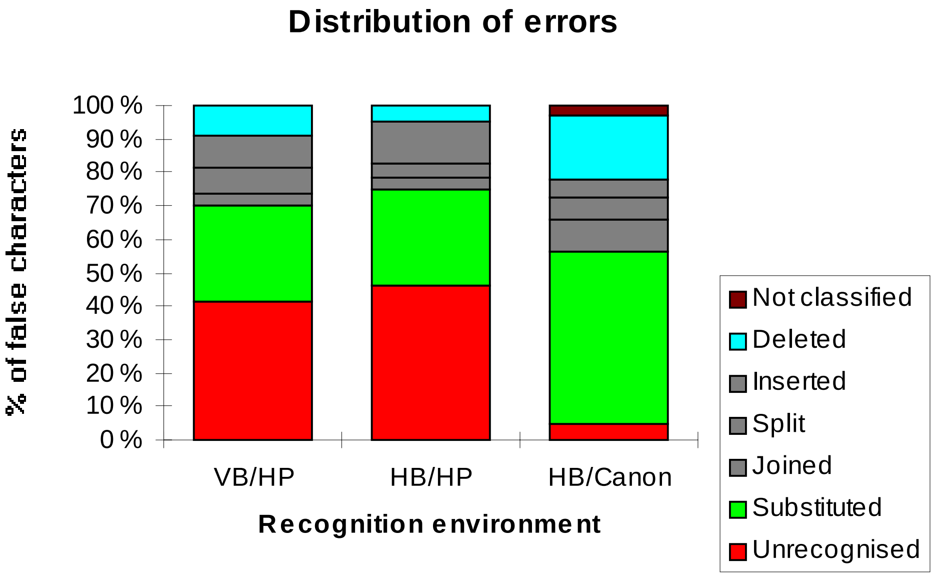

Figure 8. Distribution of recognition errors. HB = Hufvudstadsbladet, VB = Vasabladet; Canon = Canon MS-100 film scanner, HP = HP Scanjet+ A4 scanner; 300, 400 = resolutions of 300 dpi or 400 dpi.

As the test material used in the OCR Index project was not very exhaustive, nothing definitive can be said about the differences between microfilm and paper scanning. Possibly the differences were not actually due to the microfilm itself (besides getting it scanned in the first place). Microfilm and paper samples were not commensurable, as conversion from Canon's TIFF format to standard PCX format decreased resolution from 400 to 300 dpi. The conversion was performed because OCR software misinterprets Canon's TIFF as corrupted and fails recognition. Another limitation of OmniPage software was that it used a fixed value of 300 dpi for foreign image files. Apparently, conversion to PCX format somehow decreased the quality of the image file more than direct scanning at 300 dpi would have done. The conclusion drawn from the results is that scanning and OCR software should be integrated, as recognition of foreign image files is problematic.

To test the effect of scanning resolution on recognition, minor samples from newspapers were scanned at both 300 and 400 dpi. Tests were performed with an HP Scanjet+ by scanning A4-sized samples from *Hufvudstadsbladet, Vasabladet* and *Politiken.* A corner of the newspaper page was scanned at both resolutions. The page was not removed from the scanner between scans. Table 3 and Figure 9 show that when using the same scanner and scanning software, changing the resolution level does not affect recognition accuracy as much as the results in Table 2 suggest. In one case, namely with *Vasabladet,* 300 dpi even seems to give more accurate results. This seems peculiar, but is partly due to the fact that the test page contained an advertisement that was of different text type than the actual body face. This advertisement text was the main source of error in the 400 dpi scan test. One clear difference between scans at 300 and 400 dpi is that in the latter the proportion of unrecognised characters is greater. This gives better possibilities for proof-reading and error correction, as problematic characters are clearly marked in the text.

One reason for the unexpected variety in scanning results is so-called borderline pixels. Even when the paper is not moved, OCR results vary between scans because the scan head is not in exactly the same position from one scan to the next. Thus a borderline pixel may be black on one scan and white on another (Zimmermann 1993). If the borderline pixel is in a critical position it may cause an "e" to be recognised as a "c" and so on.

Table 3. Results of character recognition at different resolution levels. HB = Hufvudstadsbladet, VB = Vasabladet, Pol = Politiken; 300/400 = scanning resolution 300/400 dpi.

	HB 300	HB 400	VB 300	VB 400	Pol 300	Pol 400
File size in kilobytes	160	213	133	192	107	147
Unrecognised	16	51	8	59	52	42
Misrecognised	117	83	69	75	107	39
Incorrect characters total	133	134	77	134	159	81
Characters in original text	6591	6591	5528	5528	3757	3757
Accuracy %	98.0	98.0	98.6	97.6	95.8	97.8

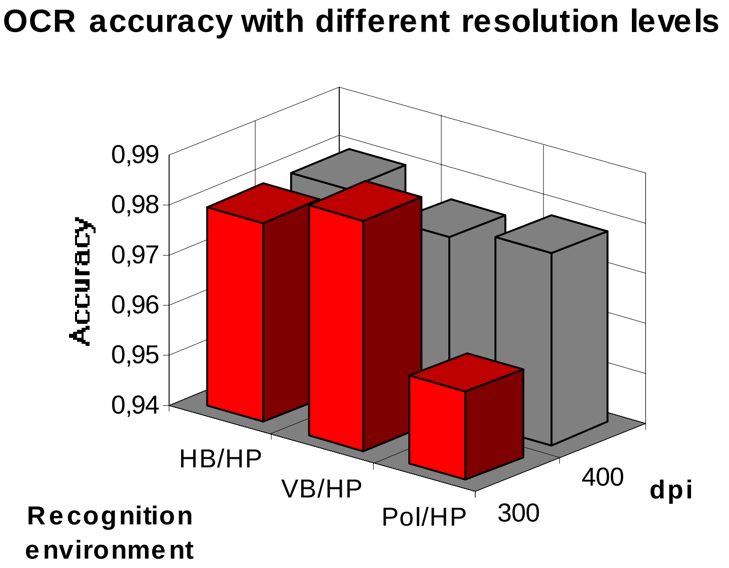

Figure 9. Results of character recognition at different resolution levels. HB = Hufvudstadsbladet, VB = Vasabladet, Pol = Politiken; 300/400 = scanning resolution 300/400 dpi.

When dealing with information retrieval, not only the accuracy of characters but also the correctness of words is important. If the word is misspelled, it does not match with the search term - unless the false character happens to be after the truncation point at the end of the word. In the OCR Index project a word was regarded as correct only when it did not contain any unrecognised or misrecognised characters and was not split or joined. If a word was divided between two lines, it was counted as one word if there was a hyphenation mark at the end of the first part. To enter words correctly into the index, software should be able to recognise a hyphenation mark and an end-of-line mark and join the pieces on different lines together. In other words, software must be capable of logical word connection instead of recognising the parts as (physically) different words.

The amount of correct words appeared to be much lower than the amount of correct characters. The accuracy of recognised words was only 44 per cent with scanned microfilm images. Images from an A4 page scanner gave a far better accuracy of 97 or 98 per cent (Table 4, Figure 10). The correct words were mostly short ones, prepositions and the like, which for retrieval purposes are trivial or non-meaningful. In retrieval systems many of these words would be likely candidates for a stop-word list, i.e. words that are excluded from the index. These results are in accordance with those of Nartker et al. (1994), who reported that a fall in character accuracy affects the word accuracy even more dramatically. Removing stop-words further exacerbates the results.

There is a certain statistical basis for this phenomenon. If the probability that a character is recognised correctly is p $(0\ ^2\ p\ ^2\ 1)$, the probability of recogning a two character word correctly is $p*p$, presuming that every character has the same probability of being recognised correctly. (In reality the probability is not the same, but here it is presumed that p represents the average of individual probabilities.) The probability of recognising an n character word correctly is p^n. The longer the word, the smaller is the probability that the whole word will be

Table 4. The accuracy of recognised words with OmniPage. HB = Hufvudstadsbladet, VB = Vasabladet; Canon = Canon MS-100 film scanner, HP = HP Scanjet+ A4 scanner.

Words	HB	HB	VB
	Canon	HP	HP
Incorrect	1 397	359	256
Originally	2 511	2 773	3535
Word accuracy %	44.4	87.1	92.8
Character accuracy %	82.3	97.3	98.0

recognised correctly, although the probability of each individual character is the same in different words. If accuracy is low, such as 80 per cent, the probability of recognising a five-character word correctly is only 33 per cent. If the recognition accuracy is higher, say 98 per cent, the probability is much higher at 90 per cent (Appendix 1).

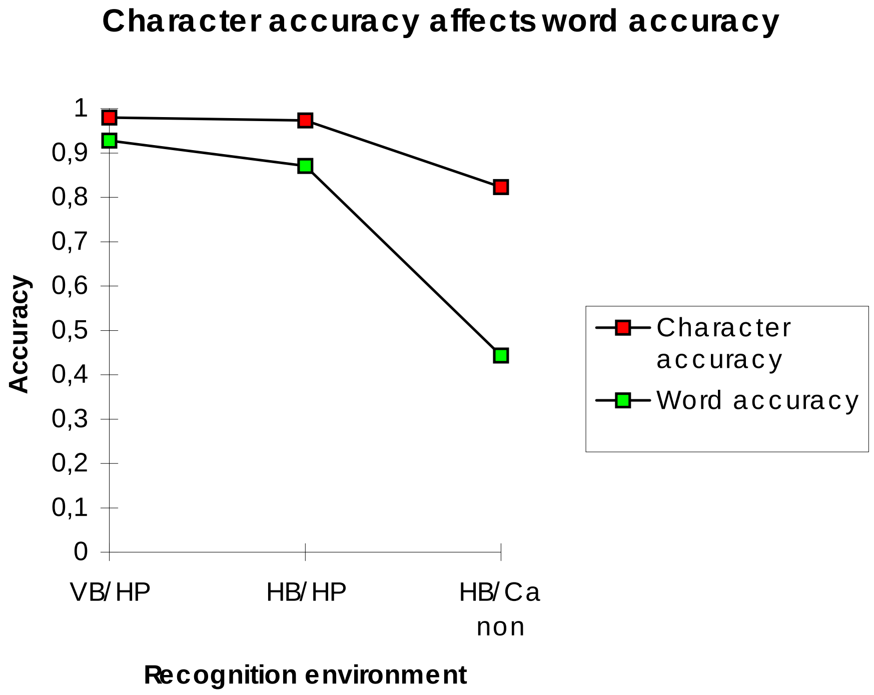

Figure 10. Accuracy of recognised words with OmniPage. HB = Hufvudstadsbladet, VB = Vasabladet; Canon = Canon MS-100 film scanner, HP = HP Scanjet+ A4 scanner.

This means that if OCR software is not accurate enough there is little use in producing indexes automatically. From the Figure 11 it is clear that with poor accuracy, such as below 90 per cent, the word accuracy soon drops below an acceptable level. To get relatively good results even with longer words, the character accuracy must be at least 98 per cent to make the word accurate enough for indexing. Judging from the results of our tests (98 per cent accuracy at best), and assuming that methods for automatic correction of recognition errors will be further developed (e.g. Sun et al. 1992), this accuracy level seems to be achievable. To a certain extent, retrieval methods allowing inexact matching of search terms make it possible to retrieve even misspelled words (Robertson & Willett 1992). In an environment where 100 per cent accuracy is not imperative, even traditional Boolean retrieval can be applied to text produced by OCR software. In retrieval tests performed by Taghva et al. (1993 and 1994), only a few documents were missed due to OCR errors. They suggest that with additional automatic correction procedures even missed documents should be retrievable.

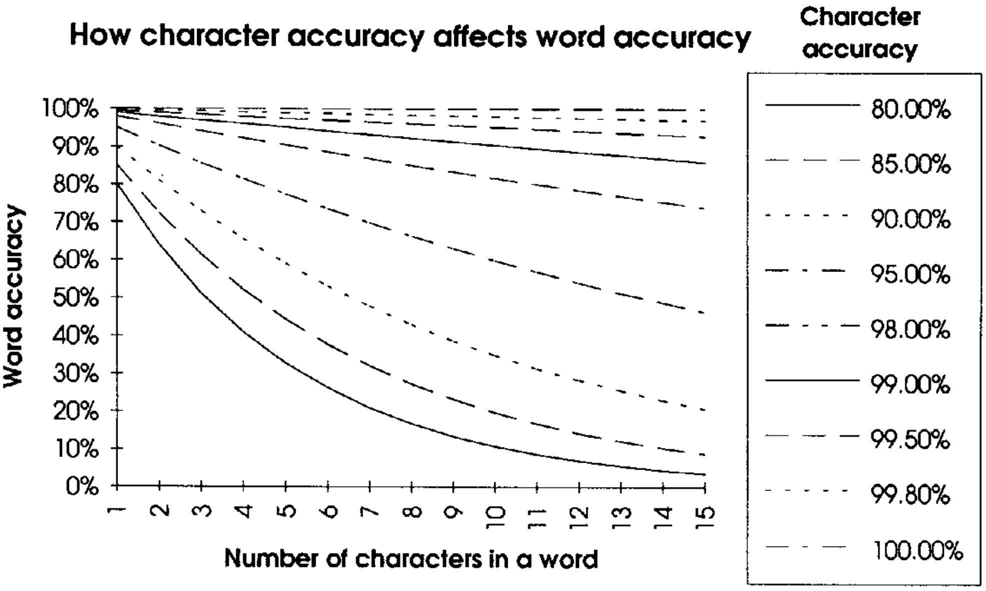

Figure 11. Effect of word length on the probability of recognition accuracy at fixed precision levels of character recognition (see Appendix 1).

5 COST FACTORS

As microfilm is designed to be handled technically, it may prove to be cost-effective to scan from the microfilm and not from paper originals. Turning pages is one of the most expensive parts of microfilming and digitisation and should be minimised (Lesk 1992). However, the immature stage of the microfilm scanning technique makes it difficult to predict the total costs of scanning large volumes. The costs depend greatly on the quality of the microfilm, but also on the steps taken after scanning.

Digitising paper documents (e.g. newspapers) creates costs in several phases:

- preparation of documents: the pages must be opened and placed on the scanner in the correct position (if the documents are available on microfilm this can be omitted). Scanning parameters must be adjusted according to background colour, fonts, and other properties of each document
- scanning
- indexing
- quality control
- storing.

5.1 SCANNING COSTS

The changeover from microfilm to digital image currently takes a theoretical 2 seconds per image with a Mekel M400 scanner, which costs about FIM 300 000 ($ 50 000), providing the film quality remains constant and the frame movement is smooth. Operator intervention is needed only to change the roll, roughly twice an hour. Assuming that the machine should pay itself off in three years (about 5 000 working hours), it would cost below FIM 120 per hour to run. Since in an hour it can do from 1 000 to 2 000 frames, the cost per frame to convert from microfilm to digital should be about 0.12 marks. Compared to the 0.8 - 1.7 marks per page for scanning, using microfilm is a reasonable intermediate step to digital imagery (Lesk 1992).

The newspaper archive of Helsinki University Library comprises some 20 million microfilm frames on 30 000 rolls of about 600 - 700 frames each. On each frame is an image of an A2-sized page or full spread of a tabloid newspaper. Assuming an average working time of one hour per microfilm roll, the digitisation process of 30 000 rolls would take about 20 man years. The corresponding labour costs would be around FIM 3.6 million.

The cost of a large scale microfilm scanning station (e.g. SunRise DMS 50i-R) is about FIM 500 000, including a rollfilm adapter, a special lens for 35 mm rollfilm, and image enhancement software. Assuming that a qualified operator is able to run two systems simultaneously, the total cost could be decreased by investing in two scanner stations. The labour costs could be lowered to about FIM 2.0 million, yielding total costs of about FIM 3.0 million, instead of 4.1 million with only one scanning station.

The above figures give some idea as to the productive capacity of current scanning technology. Without practical experience, however, it is difficult to estimate the actual time needed to scan a microfilm roll. Digital technology has not yet settled down and the volume conversion of microfilm is very rare. Only a few service bureaux in Europe have invested in expensive microfilm scanning equipment.

5.2 STORAGE COSTS FOR A DOCUMENT IMAGE ARCHIVE

In addition to scanning, there would be some costs for storing and handling the data on optical discs. For retrieval purposes the texts should be fed into OCR software and index files created. The system could be automated so that user intervention would not be needed in the indexing process. The ideal solution, from the viewpoint of retrieval and availability, would be to convert all documents to electronic form regardless of the original. However, newspaper material presents one major problem: the huge amount of documents.

When digitised, each A2-sized page scanned at 400 dpi fills about 8 MB of uncompressed space. Before storing the information, a generally accepted compression such as the CCITT G4 should be used to give a mean compression of 1:20. One frame would still occupy about 400 KB, which with 20 million frames gives an estimated

$$20\,000\,000 * 400\,000 \text{ bytes} = 8 \text{ TB (terabytes)}$$

for the backlog archive. In addition, a yearly cumulation of 1 000 roll films gives 280 GB (gigabytes) per year.

Few solutions are available for storing such vast amounts of information for relatively fast access. A 1 TB jukebox with 14" optical discs from Kodak or a 1 TB CREO optical tape recorder from ICI are possible candidates, but around nine of these would be needeed in parallel. Considering that the price of one jukebox is about FIM 4.5 million, the total purchasing cost for storage equipment would be over FIM 40 million. Of course, the costs could be drastically cut by organising only 1 TB for active (online) storage and moving the rest to a collection of discs

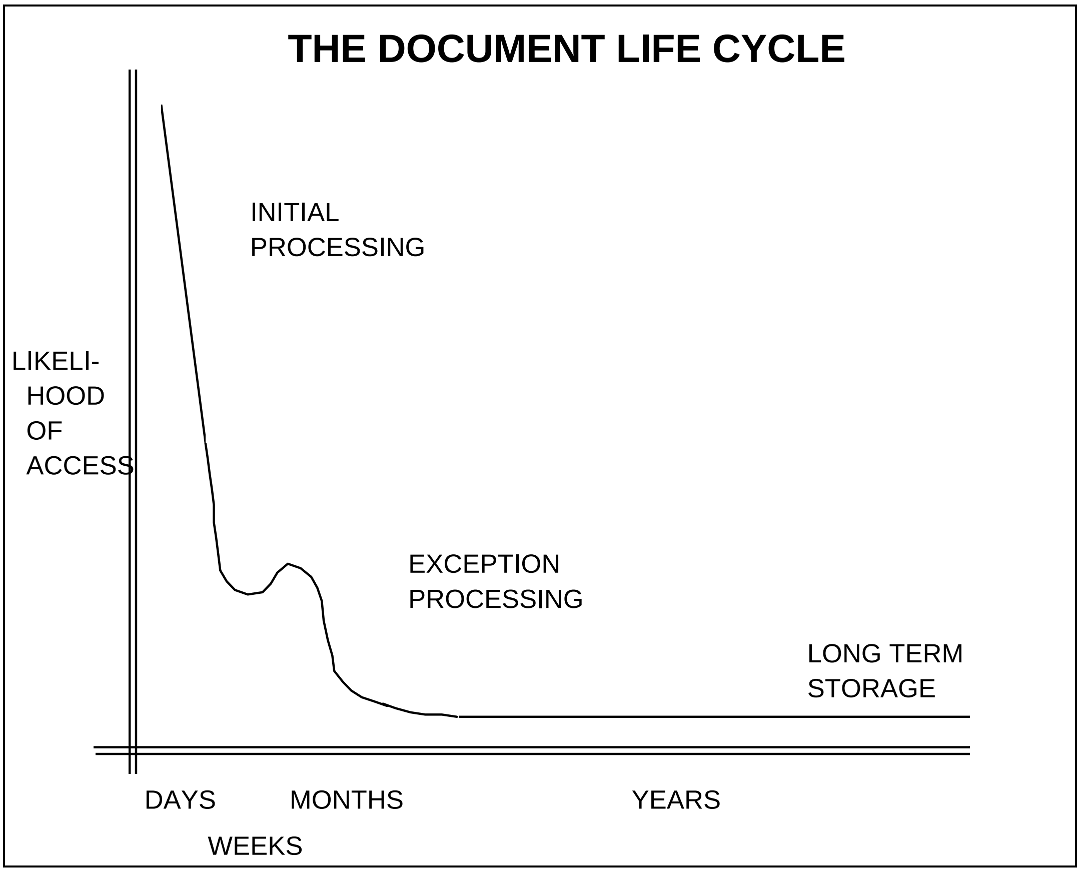

Figure 12. Life cycle of an office document.

on a shelf. There could be retrieved by asking an operator to fetch the appropriate disc and load it into the jukebox.

Another argument against a full image archive for newspapers is that documents are needed only occasionally. When an office document is in active use, it would be useful to scan it first and store it onto optical discs to allow fast access. When a document ages and there is only occasional, if any, need to retrieve it, the most cost effective solution would be to save it on a microfilm by using microform printers, i.e. film recorders (D'Alleyrand 1993). This solution, however, is pertinent only to office documents used actively at the beginning of their life cycle (Figure 12). In this case it is usually relatively easy to distinguish between active and passive documents. With newspapers in library microfilm archives this distinction is not so clear. A specific microfilm is used only occasionally, if ever. In proportion to the size of the collection, only a minor part of it will be in active use. The problem, however, is that any of the films may suddenly be needed at any time.

5.3 STORAGE COSTS FOR A MICROFILM ARCHIVE WITH FULL TEXT INDEXES

A recommendable solution is to keep microfilm as the main archival medium and assure efficient retrieval by creating full text indexes. In this case only the index files would be stored in machine-readable, character-coded form for retrieval purposes. The indexes would contain all or nearly all words captured from digitised pages with an OCR engine.

Although the scanning process would be expensive, the memory size required for index files remains quite small and inexpensive. This alternative would assure retrieval of the whole archive, although the originals would only be kept on microfilm and output on demand with microfilm scanners. Maintaining microfilm is cheaper than maintaining current electronic media, which needs updating on a regular schedule.

One newspaper page image is reduced to about 20 KB of ASCII text by OCR. Assuming that the size of the index, i.e. inverted file, is of the same size as the document file, an index of 20 million pages would need a storage of "only" 400 GB with a yearly cumulation of 14 GB. This amount of information could be stored in three jukeboxes of 150 GB each or in one 1 TB piece of equipment. The storage purchasing costs would be about FIM 4.5 million in each case.

5.4 STORAGE COSTS FOR A CD-ROM ARCHIVE WITH FULL TEXT INDEXES

With a microfilm-based archive as presented in the previous chapter, the problem is that heavy use may deteriorate the microfilm. Again, if the film is copied, the result is poorer than the original. The copying problem can be solved by saving the images on a cheap digital media like CD-ROM with a CD recorder. CD-ROM is a standard technique already widely used in libraries. A copy of a digital disc is of the same quality as the original and the distribution of discs is easy to arrange. As one CD-ROM disc can contain 600 MB of data, an image archive of 8 TB would take about 14 000 CD ROM discs (instead of 30 000 microfilm rolls). The discs would form an off-line archive, where a specific disc is fetched on demand like the microfilms in the previous example. The problem with digital technology is that it may become obsolete, even when well established like the CD-ROM technique. Thus the regular updating costs of a digital system should be taken into account right from the start.

With current technology, even a relatively cheap digital device like CD-ROM is much more expensive than microfilm. Nevertheless, the development of OCR technology may improve the possibilities of CD-ROM as a newspaper storage medium. If the text in pages can be saved in ASCII format and only the pictures in image format, the memory size of files will be reduced. E.g. the most recent version of OmniPage software, Professional 5.0, contains a TruePage feature that is claimed to preserve the original outlook of the scanned page. Consequently, if the layout information of the newspaper page can be coded and saved with ASCII text and images, it would be possible to produce a practically identical copy of the original newspaper page.

6 CONCLUSIONS

Scanning

Microfilm scanning is a promising technology, but currently not adequately established for the purposes of this project. 35 mm microfilm is used mostly for archival purposes, as the large frame enables capture of the finest details of the text and pictures. Unfortunately, few scanners are capable of handling 35 mm roll film. As the 35 mm film does not include image retrieval marks (blibs) of the sort common to 16 mm commercial film, it is fairly difficult to advance the film automatically the required amount between frames. No scanner yet exists that is capable of real automatic, unattended scanning of 35 mm microfilms.

Another problem with digitisation is the incompatibility of files. Although the scanners are claimed to support standard file formats such as TIFF or PCX, the files may actually be incomprehensible to OCR software. As the whole scanning and character recognition process is not integrated, additional file conversion software has to be used.

The Third problem with scanners is that a full 35 mm film frame cannot be scanned without a special wide-angle lens, which is available only for special microfilm scanners. The frames must be scanned in several pieces. Currently, there is no software support for scanning large (over A3) sized pages, e.g. so that scanned A4 images could be automatically stitched together to form an A2 page.

Optical character recognition

As scanners produce image files in various, non-standard formats, some kind of image conversion software is needed in the current microfilm digitisation process. Adding an extra phase naturally increases costs. OCR software should better tolerate image formats and have better conversion filters. The Best solution, of course, would be to integrate the whole process from the start, the scanner and OCR software being designed to work together. As film scanners represent a relatively new technique, such integrated systems have not yet been produced, probably because scanning projects have focused primarily on conservation, with only minor interest in OCR.

The peculiar problem of newspaper text is that it is printed on relatively poor paper and letters tend to smudge making the text of inherently poor quality. This kind of a problematic text type may need an OCR software tailored to the needs of the text. At least current OCR cannot cope very well with newspaper text. Another problem is the size of the page: automatic operation would necessitate that a whole A2 page is scanned and recognised in one pass. Currently the software seems to interpret an A2 page as an A4, which means that the characters may fall below the recognisable point size (the minimum font size manageable to OCR software is about 6 points).

Recommendations

As more and more newspapers produce their texts in electronic form, the most cost- effective solution to the access problem would be that the newspaper indexes are compiled directly from these editing systems. As plain ASCII text would be sufficient for this purpose, the process would be relatively straightforward and technical problems mediocre compared to the anticipated benefits. The problem with this approach is that it may not be applicable due to copyright restrictions (Kantola 1991).

Currently, for example in Finland, many newspapers have their own in-house databases, which cannot be accessed publicly due to copyright regulations. Only some reference or abstract databases are available, but do not provide comprehensive coverage. One solution for copyright problems might be that the index of a newspaper database is produced from the full text of articles, but the articles themselves are not stored in the newspaper database. Instead, a surrogate could be saved containing article title and publication information etc. needed for identification.

The above solutions are not, however, applicable to retrospective newspaper archives and newspapers produced in traditional manner. These materials are not in digital form and need to be converted. Another problem is that the appearance of the newspaper is lost. Microfilming as a photographic process makes a faithful copy of original printed material.

There is no readily designed solution for digitising newspaper archives on a large scale. For smaller archives a document image processing (DIP) approach with an OCR option would be the most appropriate. Several successful experiments have been reported also in Scandinavia. For the large newspaper archive of a national library, storing the entire material as a full-image archive allowing fast access would be very expensive. To build up and maintain such a massive (terabyte class) system for relatively passive use makes the costs per retrieved item very high.

Another solution would be to retain the base archive on microfilm and digitise only the indexes for content based searching, i.e. to produce a hybrid imaging system that makes use of both microfilm and electronic media. By committing only indexes to digital form and using a film scanner to access microfilm frames on demand, the user can take full advantage of both types of medium, and the storage costs are smaller than with a full-image archive. The strengths of each have been recognised: digital systems for user aspects such as ease of access, multi-access, transmission etc.; microfilm for long term archival storage, low risk of obsolescence of equipment, legal acceptability and cost effectiveness (Shiel & Broadhurst 1992). A new range of products that enable a document to be scanned and microfilmed in a single pass have also appeared on the market, although not yet for 35 mm roll film.

The third solution would be similar to the second one, except that images would be stored on a cheap digital medium, like CD-ROM. The discs would form an off-line archive, where a specific disc is fetched on demand as the microfilms in the previous example. It is, however, easier to access, copy and deliver CD-ROM discs than microfilms. Maintenance costs should include regular updating. With current prices the CD-ROM approach is more expensive that a microfilm based archive, but the development of OCR technology may improve its possibilities if OCR software can produce both ASCII text and layout information (possibly also SGML tags) from the image file. ASCII text would be more compact than images, and with layout codes it would be possible to produce a virtually identical copy of the original newspaper page.

Further research

The tests in the OCR Index project were not performed with optimal equipment and software. Further tests with enhanced professional equipment may show that the problems with microfilm scanning can be solved and optical character recognition accuracy increased. Such a test environment for automatic microfilm scanning and conversion would require more resources than were used in the OCR Index project.

If the scanning and OCR technique can be sufficiently improved, the possibility of establishing a centralised Nordic digitisation centre should be investigated. Such a centre would control the entire process from microfilm to text retrieval program. The centre could be a subsidiary of a national library, or a service bureaux that would purchase the equipment in co-operation with the libraries concerned.

An extended research project would contain e.g. following phases:

1. Finding and evaluating appropriate image enhancement, OCR, and text retrieval software that can be used with microfilm scanners and/or A2-sized scanners and can cope with newspaper text.
2. Developing OCR software that can retain the appearance of the original page in addition to ASCII, and possibly also produce SGML tags.
3. Developing automatic spelling correction methods for texts produced by OCR software, e.g. by using language-specific spelling checkers.
4. Developing text retrieval methods that can cope with incorrect words, e.g. by using pattern matching techniques.

REFERENCES

Anon. 1992. Text retrieval technology and electronic document imaging. Document Image Automation Update, vol. 11, no. 11, pp. 1 - 4.

Anon. 1993a. The French Connection. Information Management & Technology, vol. 26, no. 5, pp. 196 - 197.

Anon. 1993b. Optical storage: an overview of the technology and its use within the United Nations system. Geneva: United Nations. 114 p. ISBN 92-1-100390-3

Broadhurst, R. 1992. Microfilm scanning and digitisation. Information Management & Technology, vol. 25, no. 3, pp. 132 - 134.

Broadhurst, R. 1993. The digitisation of library material. Information Management & Technology, vol. 26, no. 3, pp. 128 - 132.

Cawkell, A. E. 1989. Image processing and page presentation: part 2. The Electronic Library, vol. 7, no. 2, pp. 106 - 110.

D'Alleyrand, M. R. 1993. Producing microforms with your imaging system. IMC Journal, vol. 29, no. 6, pp. 16 - 22.

Feretti, M. 1994. Bibliothèque Nationale de France: Digitalisierung für die Recherche und zur Archivierung von Büchern, Bildern und Tönen. Nachrichten für Dokumentation, vol. 45, no. 3, pp. 167 - 169.

Glassco, R. A. 1993. Evaluating commercial text search-and-retrieval packages. Information Technology and Libraries, vol. 12, no. 4, pp. 413 - 421.

Harvey, D. 1991. Catch the wave of DIP. Byte, vol. 16, no. 4, pp. 173 - 182.

Heimbürger, A. 1990. Electronic images. Espoo: Technical Research Center of Finland. 73 p. + app. 50 p. (VTT Research Notes 1083).

Kantola, A. 1991. Mahdollisuudet perustaa yleinen sanomalehtiaineiston tietopankki [On the possibilities to establish a national newspaper databank]. Helsinki: Helsingin yliopiston kirjasto. 40 p. + app. 1 p. (Helsingin yliopiston kirjaston monistesarja - Helsingfors Universitetsbiblioteks stencilserie 16). ISBN 951-45-5750-6 (in Finnish)

Lesk, M. 1992. Image formats for preservation and access: A report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access. Microform Review, vol. 21, no. 1, pp. 17 - 24.

Nartker, T. A., Rice, S. V. & Kanai, J. 1994. OCR accuracy: UNLV's second annual test. Inform, vol. 8, no. 1, pp. 40 - 45.

Ougham, H. & Williams, B. 1992a. Reading typefaces. Electronic Documents, vol. 1, no. 5, pp. 1 - 32.

Ougham, H. & Williams, B. 1992b. Recognising characters. Electronic Documents, vol. 1, no. 9, p. 1 - 32.

Ramsden, A., Wu, Z. & Zhao, D. 1993. Selection criteria for a document image processing system for the ELINOR Electronic Library Project. Program, vol. 27, no. 4, pp. 371 - 387.

Robertson, A. M. & Willett, P. 1992. Searching for historical word-forms in a database of 17th-century English text using spelling-correction methods. In: Belkin, N., Ingwersen, P. & Mark Pejtersen, A. (eds.) Proceedings of the fifteenth annual international ACM SIGIR conference on research and development in information retrieval, Copenhagen 21 - 24 June 1992. New York: ACM Press. Pp. 256 - 265.

Saffady, W. 1993. Electronic document imaging systems: Design, evaluation, and implementation. Westport, CT: Meckler. 182 p. ISBN 0-88736-840-9

Schneider, J. 1993. The future of OCR in document management. Inform, vol. 7, July, pp. 18 - 24.

Shiel, A. & Broadhurst, R. 1992. Library material digitisation demonstrator project: Project report. Hatfield, Herts: Cimtech. 88 p.

Sturt, J. P. 1992. Interfacing microform scanners for Document Image Processing applications. Headway Computer Products info. 6 p.

Sun, W., Liu, L-M., Zhang, W. & Comfort, J. C. 1992. Intelligent OCR processing. Journal of the American Society for Information Science, vol. 43, no. 6, pp. 422 - 431.

Taghva, K., Borsack, J. & Condit, A. 1993. Automatic error correction and query evaluation of OCR generated text. In: Raitt, D. & Jeapes, B. (eds.) Online information 93. 17th International Online Information Meeting Proceedings, London 7 - 9 December 1993. Oxford & New Jersey: Learned Information (Europe). P. 115 - 128. ISBN 0-904933-85-7

Taghva, K., Borsack, J., Condit, A. & Erva, S. 1994. The effects of noisy data on text retrieval. Journal of the American Society for Information Science, vol. 45, no. 1, pp. 50 - 58.

Warner, T. 1993. Which documents are good candidates for OCR? Macworld, vol. 10, no. 11, p. 95.

Waters, D. J. 1991. From microfilm to digital imagery: On the feasibility of a project to study the means, costs and benefits of converting large quantities of preserved library material from microfilm to digital images. A report of the Yale University Library to the Commission on Preservation and Access. The LIBER Quarterly, vol. 1, no. 3, pp. 239 - 280.

Waters, D. J. & Weaver, S. 1992. The organizational phase of Project Open Book. A report to the Commission on Preservation and Access. Washington, DC: The Commission on Preservation and Access. 11 p.

Wiggins, B. 1992a. Document image processing - An overview, Part 1. Document Image Automation, vol. 12, no. 3, pp. 3 - 9.

Wiggins, B. 1992b. Document image processing - An overview, Part 2. Document Image Automation, vol. 12, no. 4, pp. 12 - 20.

Zimmermann, K. A. 1993. Why no two scans are alike. Imaging Magazine, vol. 2, no. 8, p. 14.

Effect of word length on the probability of word recognition correctness on fixed precision levels of character recognition. APPENDIX 1

OCR	Number of characters in a word														
precision	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
80,00%	80%	64%	51%	41%	33%	26%	21%	17%	13%	11%	9%	7%	5%	4%	4%
85,00%	85%	72%	61%	52%	44%	38%	32%	27%	23%	20%	17%	14%	12%	10%	9%
90,00%	90%	81%	73%	66%	59%	53%	48%	43%	39%	35%	31%	28%	25%	23%	21%
95,00%	95%	90%	86%	81%	77%	74%	70%	66%	63%	60%	57%	54%	51%	49%	46%
98,00%	98%	96%	94%	92%	90%	89%	87%	85%	83%	82%	80%	78%	77%	75%	74%
99,00%	99%	98%	97%	96%	95%	94%	93%	92%	91%	90%	90%	89%	88%	87%	86%
99,50%	100%	99%	99%	98%	98%	97%	97%	96%	96%	95%	95%	94%	94%	93%	93%
99,80%	100%	100%	99%	99%	99%	99%	99%	98%	98%	98%	98%	98%	97%	97%	97%
100,00%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%