SRELS Journal of Information Management
Vol. 41, No. 1, March 2004, Paper D. p29-42

DIGITAL IMAGING SOLUTIONS FOR STORING INDEXING AND RETRIEVAL OF FULL TEXT LITERATURE: A SPECIAL REFERENCE TO UNPUBLISHED DOCUMENTS

*Wathmanel Seneviratne**

The paper discusses the digital imaging solutions for storing, indexing and retrieving full text documents with special reference to grey literature. The various digitization processes in practice described

KEYWORDS/DESCRIPTORS: Digital Iaging; Grey literature; Full text storage, Indexing and retrieval

1 INTRODUCTION

Modern libraries and information centers, despite their origin, type, size etc. receive plethora of information in different types and formats. These include not only books and periodicals etc. of published type, but many a kind of literature such as manuals, reports, reprints and also loose-leaf documents in unpublished type. The formats differ from familiar printed type to micro media, A/V media, electronic media and other modern storing methods and systems. Whatever the format, published forms are easy in access and control but the literature in unpublished form are considered to be difficult for bibliographic control. The librarians think about the bibliographic control very much in system terms perceiving that if they know where the material is then the user accessibility is achieved. Hence the traditional librarianship bothers immensely about preparing bibliographic entries. When talking in user genera speedy full text access is important as well as the information on materials available. So that the demand for full text forms specially in electronic form is being ever increasing.

This paper discusses alternative solutions needed by libraries in providing full text solutions to the readers. The paper tries to bring forth some solutions that

* *The Senior Assistant Librarian*, Library University of Colombo, P.O. Box 1698, Colombo 03, Sri Lanka. E-mail: salweb@cmb.ac.lk

could be adopted from the products and systems available in the market. The author tries to accommodate needs of different library types for storing, indexing and retrieving full text. It is hoped that the solutions proposed in the paper diver widely used digitization practices (keying-in process and well known flat bed scanning) followed in libraries to a potential that can be applied in libraries.

2 DEFINITIONS FOR THE MATERIALS

Digital imaging systems proposed in the paper specially is being discussed in relation with the unpublished, grey literature. Therefore it is necessary to identify the materials marked for digitization.

D.N. Wood in his report to the British Library, in defining role of the British Library Lending Division, states grey literature as 'literature, which is not readily available through normal book selling channels and therefore difficult to identify and obtain'. (Wood, 1982) The Chambers' Dictionary states 'materials published non- commercially' as grey literature.

Examples of grey literature include reports, technical notes and specifications, conference/seminar proceedings (when not published), supplementary publications, data compilations, manual, codex, trade literature etc. (Auger 1998). Above mentioned definitions are valid for all sorts of grey materials including loose-leaf forms, that carries greyish qualities as mentioned in the definitions. Most of these materials are organization oriented and available in different paper formats. Van der Heij (1985) has also pointed out about these greyish aspects of this literature. The greyishness of the literature itself makes acquiring, processing, storing and maintaining of the same tedious and difficult.

Types of loose-leaf literature consists of different formats depending on the nature of the purpose and printed in different paper sizes (size of the documents is important for the electronic solution proposed). The main formats identified are; Brochure format, Handbill format, Card format and Report format (Seneviratne, 1994).

Brochure format is used in disseminating trade and technical literature(Evans & Child, 1963), advertising institutions, services and products. In many cases brochures are produced for temporary dissemination especially in advertising or as a tool of creating current awareness among the desired community, for example to be distributed at exhibitions, conventions etc. and some times for continuous distribution but not on permanent basis. There are standard paper sizes

suitable for brochures ranging from B6 to A3. Some brochures are clipped to form a booklet and some of them come as a sheet of paper folded in to 2 or 4 and type set in a desired way. The brochures come in different paper thickness too.

Hand bills, often are of B5 size. There are bigger and smaller sizes too. Cards are also comes in different sizes (business card size to B5 size). The report format is used for standard applications like printing out reports of commissions, seminars, round table discussions, etc. In most cases the report comes in A4 paper or in A3 folded in to two. One can understand the range of paper sizes used in loose-leaf forms and the difficulty in maintaining those in vertical files in cabinets or in boxes made in different sizes.

3 IMPORTANCE OF DIGITIZATION

The documents of this category is normally kept in vertical files or in pamphlet boxes ordered and/or indexed by broad subject headings or special areas of interests of that particular library. In general this type of literature may have been collected by all kinds of libraries or in separate sections that collect special materials. However, it is a widely heard complaint about the difficulty of maintaining the documents in discussion.

Difficulty and inefficiency of the vertical file management is worth mentioning here when discussing importance of digitization.. The processing staff has to spend a considerable time in numbering the document, taking out relevant file/pamphlet box from the drawer/shelf each and every time, and insert the document in the correct order and place back the file in the drawer. In some libraries an index card is typed and filed in an information index. But this process is found to be very difficult to carry on when there is a heavy load of loose-leaf literature in one mail. Even if the users are guided by an information index (card based) or by a computer database which contains and retrieves bibliographic references, the users are still have to follow the same cumbersome, tedious process of pulling out the necessary document from the storage.

In special libraries this kind of literature is collected in good numbers. In most of the cases maintenance is difficult due to accumulation of dust and infestation of pests in the vertical files over the years. Fungus and pests tend to threaten the control of the tightly packed filing cabinets and drawers. The problem is specially visible in humid regions. This is not because of lack of proper care, but due to the

difficulty of cleaning and maintaining the papers in unbound form and in varied paper sizes and also due to the traditional storage systems. The temperature and humidity level inside the drawers is comparatively high. All these problematic situations count for looking into an alternative electronic solution.

When it comes to digitization of loose-leaf forms, the information officer holds another responsibility of ensuring the effectiveness in storing, maintaining and retrieving the full format as it was. New trend in libraries is to use a database management system to provide wide range of facilities with ready made or customised integrated solutions. The solutions are proven best in providing bibliographic information. When it comes to the full text and graphical or image oriented information the librarians mostly count on CD-ROM or online sources. But these standard sources do not cover, as mentioned above, the variety of unpublished literature collected over the time in remote by many institutions.

With the advancement of desktop publishing, loose-leaf literature stays multiplied in production. The users are often disgusted or almost lost in searching the required piece of information among hundreds and thousands of literature collected in drawers and cabinets. Even though we are in a society of high IT usage, more than 95% of the grey documents are still on paper. The growth of it is so alarming so that it has now became a critical need to control and to retrieve the full formats instantly.

It is not only the informal and loose-leaf forms that can be benefited from the kinds of solutions proposed. Reprints and journal articles also can be converted into digitized form using these methods. Particularly the special libraries have collections of reprints in thousands, which cause immense difficulty in control and retrieval. In getting rid of this kind of in-house document management problems discussed above, digitizing solutions available in the market could be readily used as a remedy. The systems available are far more efficient in producing and providing the full text solutions for the user and takes off the loose document management problems in libraries.

4 SOLUTIONS AVAILABLE

There are three options available for librarians in obtaining digitized products.

- Ready-made digitized products which can be bought off the shelf (e.g. Full text CDs available with publishers and online full text services).

- Commercial digitization solutions. (conversion of printed information to digitized products by a commercial vendor of solutions)
- Digitization solutions adopted in-house.(e.g. using modern digitization solutions).

Many libraries with limited budgets might not be able to subscribe to the first category, or may not be able to go for commercial services in the second category. But if a library is in a position to think about the third category the paper tries to bring forth some workable and experimented solutions for them.

The systems discussed here, obtained their popularization mainly in managing office documents and in science and technology laboratories. But the feasibility of using the system in libraries is proven after the usage of the system in two special and one academic libraries in Sri Lanka. (Library- Peoples Bank, Information Services Centre, Industrial Technology Institute former CISIR, University of Moratuwa) At the Peoples' Bank the system is mainly used to store hundreds of internal circulars and other information which comes in loose-leaf form. The ITI library (former CISIR) use the system to store several thousands of reprints, pamphlet literature, data sheets, etc.

5 EXPECTED FEATURE OF THE SOLUTIONS

There are three main facilities expected by the librarians from digitization of full text documents, namely;

1. Archival solution;
2. Indexing solution; and
3. Retrieval solution

Digital archival solution is expected to achieve under mentioned objectives.

a. Transform the printed media into reliable, retrievable electronic media. (may be to a hard disk or to an external storage). The solutions should be reliable and standard. Should be usable in long run. Data stored should be able to retrieve under many access points.

b. Should be readable using ordinary PC and re-writable into another digital media. The librarians specifically expect data to be stored in a computer accessible format. It is highly appropriate because a necessity may arise to

make available the digitized literature in networked/online environment and to deliver the document electronically and also for many other applications in an electronic environment.

c. Conversion to digital format should be fast. These solutions are supposed to get rid of the over flown cabinets and stacks and thousands of documents await to go through the process. Hence the speed matters in the conversion process.

d. The solutions should not be much sophisticated, and should be easy to handle by staff members with varying technical literacy levels. In libraries specialists are not always there to handle a specific job. So that easy to handle processes will create fun in performing the job.

e. Should be affordable.

f. Indexing ability (discussed separately below).

Features expected by a librarian from an Indexing Solution can be discussed as follows.

a. Facility to index and categorise the documents simultaneously or before or immediately after once a document is converted.

b. Ability to categorise the documents into files, folders or into some sort of order, thereby facilitating the staff select the document type as and when needed. This specially because of the varied nature of the documents available to be converted.

c. Facility to index the documents by the access points desired by the librarian. There should be a flexibility for the librarian to select and index the access points for later retrieval.

Retrieval solution is envisaged to produce following facilities.

a. An IR interface should be provided as in bibliographic packages.

b. Should be able to browse the entire database.

c. Should be able to search by the fields intended by the user.

d. Ability to conduct simple and complex searches.

e. Should be flexible and user friendly for the user.

6 SOLUTIONS PROPOSED

The solutions are in different models. These solutions can be discussed on the line of facilities expected by the librarians as mentioned above and extra facilities provided by the systems it self.

Features available in the systems are discussed in the paper with a general approach as the facilities available in each and every product are slightly varied according to the brand. The solutions allow scanning, searching and retrieving of PC-archived documents.

7 METHOD OF DIGITIZATION

High Speed Image Scanning is the method employed in these solutions. The system is named as *Digital Filing Systems* or *Digital Cabinet* or *E-Cabinet* or *Electronic Filing System* etc.

Depending on the type of majority of the documents available, a Scanner Model should be selected. For loose-leaf documents like data sheets, reprints etc., Vertical Scanning Model can be adopted. It is also important to consider about the scanning of both sides of a document in one scanning run. A facility called Duplex mode is available in the systems that offer high quality single and double sided scanning of text, photos and graphics. For books and other bound volumes Flat Bed Model can be considered.

The Text Mode available in these systems is used for character-based documents, while its Photo Mode delivers optimal scans of materials incorporating photos or graphics. Images captured in either mode can be made at resolutions ranging from 300dpi x 300 dpi (the resolution varies according to the brand). The systems adjusts all these requirements to its operations automatically. Image output of the scanner is in TIFF/DOS format. The documents scanned can be viewed in enlarged form, rotatable and can be edited through normal word processing functions.

In some products (ScanDoc 1998-2002) each electronic file confines to not more than 100 images because of the limits of most computer's RAM memory, which prevents viewing of larger files. If a particular physical file is more than 100 pages, it will span more than one electronic file. This will lead to assign file names manually, and requires to define extra directory levels.

Some products offer (ScanDoc1998-2002) conversion of paper documents to editable file formats such as word processing files, using commonly called OCR or Optical Character Recognition software. Accuracy of conversion is entirely dependent on the quality of the original. Accuracy with a high-quality original comes to average of 95-99% based on industry experience. A document produced by a laser printer or printing press with standard fonts on white paper is considered a high-quality original. Photocopies, low-resolution printouts, documents with non-standard fonts such as script, and documents with poor contrast due to similar colors of paper and ink, are all considered low-quality for the purposes of OCR and will probably require extensive manual processing to improve accuracy.

The systems support manual and automatic document feeding, automatic detection of paper size and paper thickness. The image can be viewed while scanning and scans documents of the size from business cards right up to the size of B4. It also provides a choice as to text or half-tone resolution.

A very admirable feature of the systems is the scanning speed. The approximate scanning speed is 40 – 50 A4 pages per minute. The scanning speed varies depending on the resolution of the print, e.g. 40 A4 pages per minute 40 ppm. 300x 150 dpi – simplex, 30 A4 pages per minute – 30 ppm. 200x200 dpi-simplex etc. When considering the scanning speed the scanner is faster than a normal flat bed scanner as the equipment is specially designed for the purpose. To make the process easy the documents to be scanned should be prepared as mentioned below. The speed is the prominent feature and main advantage of this equipment (Canon elec.Inc. 2000).

The loose-leaf leaf forms, if in one size can be fed into the scanner in stack mode, (automatic) or it has to be fed manually. If the articles of periodicals or books or any bound form to be scanned, flat bed models can be considered.

These systems allow us to scan documents, organize them into digital files, drawers and folders. The location of documents can be password protected. The systems are provided with a PC interface. These file cabinets can be transferred into any drive or any other external storage solutions like Magnetic Optical Discs (MO) and CD-ROM. If the user prefers to use MO discs an external MO drive is to be bought. If the user prefers to use CDs as storage, a CD writer is essential. The MO disc specifies the space to store 13000 up to 15000 A4 size papers depending on the text and image resolution (Mitsubishi 1999).

It is the user's preference to select a storage media. The user may do away with the external storage devices if high capacity hard disk is at hand. The specialty with the storage in the system is not the hardware, but the pre-coordinate indexing method available in storing the documents. Indexing and retrieval functions will be discussed under software section.

An easy -to-use touch-panel display lets the documents scan, index, and record documents for later retrieval in Windows 95 environment. Documents are fed and received from the front of the unit. This means that the device require limited installation place. In addition, for easy compatibility, most of the products supports both TWAIN and ISIS drivers for use with an extensive range of application software.

8 INDEXING METHOD

Before starting scanning sessions, an indexing officer should plan the indexing method. Such as broad subject categories should be decided and which subject fields are going to which drawer, etc.

As librarians we are very much interested in the indexing part and the retrieval part of any system proposed. Once the broad subject categories are defined it is very easy to scan the documents to the desired cabinet and index box after highlighting the same from the index cell panel. Then the system opens the highlighted drawer and store the document under the indexed field. The system offers the user a pigeon hole like lots for Index Cells or Boxes(Canofile 1999). Annotations, marking the additional information like text, keywords, markers, on top of the scanned image/text are also available. Hiding of information with black-out and white-out boxes is possible.

Indexing system varies depending on the make of the product. For example in some systems like Canofile a limited space is available for the title (title can be typed in only up to 35 digits). In this system the disadvantage may be over come by breaking down the index cells in to specific and narrower subject areas.

Keyword indexing is possible in the systems using broad terms and narrow terms. Another facility is ability to incorporate class numbers into the index cells and drawers. The documents can be cross referred too.

If the materials in question are already organised in vertical files or in pamphlet boxes, same broad subject arrangement can be used in defining the

Cabinet and Index Cells. If this literature is being indexed for the first time using Digital Cabinet, a technical thought has to be given prior to organising the literature collected using some sort of subject headings such as Library of Congress Subject headings, Macro thesaurus, OECD thesaurus, Root thesaurus etc.

The librarians who are conversant with the bibliographic software packages may find the indexing function available in the systems are not comprehensive enough. Nevertheless systems experimented have their own characteristics and can not be compared with the indexing methods available in the text retrieval software, electronic databases and in the Internet as these solutions take the user direct to the full text documents.

9 INFORMATION RETRIEVAL FACILITY

Information retrieval facilities provided cannot be compared with the standard information retrieval software hence the guidelines of software evaluation can not be applied with this kind of system. The system has a user interface for document retrieval with different user levels and security levels (user levels has to be already specified) and provides visual access to all stored documents (Browsing function).

The user can search for documents by drawer/cabinet and by indexed subject fields. The index boxes may contain broad subjects or class numbers or both. The system displays all the documents in that particular cabinet/drawer or in the index cell. The user can also search through keywords. It should be noted that unlike complex search facilities provided in the commercial IR systems, search facility provided here is very simple and straight forward. Simple truncation methods are available in the systems.

After executing a search, a listing of relevant records will be displayed first. The user may select required records and then the full document will be displayed page by page. As mentioned above the pages displayed can be enlarged and rotated. Rotation is very important when viewing diagrams or maps etc., as the facility allows the user to turn the page and browse as if the user turn printed pages of a document length wise.

There is an author/title search available for the user. As mentioned above under the indexing section the field offer limited number of digits. A very narrow or specific search is not possible through title or author search. If a user remember

a title, words may be entered up to a certain limit. For example, in the Canofile system search criteria may be entered up to 35 digits. The system will display all the titles starting with the words entered in the search criteria.

10 THE PRODUCTS

The products are of desk top feed on models with a facility to interface with a normal PC. Dimensions approximately covers around 36.2(W) x33.6(D) x19.6(H) cm. The SCSI interface ensures connectivity of the computer system and the scanner. A built-in SCSI interface makes connection to your system as effortless as possible.

The products of the kind allows to stack as many as 100 sheets onto the feeder tray, and detect the paper size automatically and starts the scan. This equipment offers manual and automatic document feeding, automatic detection of paper size and paper thickness. The image can be viewed while scanning.

11 ADDITIONAL FEATURES

Optional Imprinter prints specified numerals, characters and symbols at any position on scanned documents. Optional Endorser allows users to print up to 6-digit character strings on documents as they are scanned for verification. Optional Bar Code Decoder may be added to some models for automatic interpretation of bar code symbols.

A "Count Only" mode is also available for simple document counting and quantity verification without scanning. The skew correction function automatically detects if the document is fed at an angle, and straightens them during scanning.

12 TRAINING THE STAFF

No special training is required to scan, view or print the documents when using the system. Most computer users were able to get results in just a few minutes.

13 PREPARATION OF THE DOCUMENTS FOR THE PROCESS

The documents to be scanned should be made "Scan Ready". Scan ready, means all the staples and other paper fasteners have to be removed from the files and the pages. The pages are expected not excessively wrinkled, torn or otherwise

damaged. The pages are all oriented the same way (i.e., all tops up) and facing the same way (if single-sided). If the pages are of odd or mixed size, manual feeding is recommended.

The files should also be separated and labeled in a way that is logical to how you would store and access them normally. We will assign a file name based on the label on the physical folder, unless some other naming scheme has been planned.

To make the process easier the documents to be scanned should be made loose from binders. Folded papers should be torn or cut into one page leaf.

Another important aspect to consider is, the documents should be dust free as the mirrors in the machine would get scratched by the dust particles. It is advised that all the old documents should be cleaned before inserting in to the scanner.

14 SYSTEM SOFTWARE AND USER INTERFACE

The software comes with the system functions to control the scanning mode, store the scanned documents in the selected storage media, index the information as assigned by the system user (library staff) and retrieval of information for different user levels.

15 DOCUMENT SHARING AND COMMUNICATION

Once documents are scanned and stored into a normal PC hard disk document sharing and delivery can be performed through available network system. The user may e-mail a scanned page by using an option provided. An e-mail address may also attach to a folder (Runningman Software 2001)

16 ALLOCATION OF STAFF

Unlike in the manual organizing, staff involved in storing/indexing documents with the system should be trained for the purpose. But it is not difficult to train the staff for this kind of system as the process is simple. Yet the indexing officer should have a subject knowledge about the document indexing as well as information retrieval aspects.

17 ADVANTAGES OF THE SYSTEM

- Affordable cost of the system.

- Other departments in the organization can use the system for other office documents by using separate external storages.
- High speed scanning assist to digitize loose paper collections quickly.
- Different user environments are entertained- single user and multi user systems.
- Get rid of cumbersome vertical filing process in cabinets.
- Document security. Data stored in MODs or CDs can be locked away if important documents like board documents; technical literature or trade literature is stored. If the documents are on MODs, information can not be retrieved through a normal computer without the MO drive.
- Data security. Ability to define user levels ensures barricading users to view confidential documents and through pass word protection.
- Windows compatible.
- No need of worrying about data entry to the data base.
- The full text and format is preserved as in the original.

18 DISADVANTAGES

- Indexing system and retrieval system offered is different from the systems available in text retrieval software. The user has to get familiar with the new system.
- The document should be dust free. Hence extra labour is to be employed to clean the documents.
- All the documents to be scanned should be in single sheet form in case of vertical models.

19 CONCLUSION

In this paper it has been tried to discuss alternative thoughts for electronic document management system that can be adoptable in libraries. There are adoptable solutions in the market with varying facilities. Please see the table in the appendix for product brands and facilities available.

Computerization and automation aspects, communication models, software systems, DBMSs, TXT packages, are highly discussed in many library and information technology seminars, forums, conferences etc., which involves highly academic and technical knowledge for systems analysis and design. For this

particular solution, it is only necessary to have an indexing librarian than a software specialist.

When making a decision to divert into a new system a library manager may conduct a cost benefit analysis between two aspects. In one side informal literature stocks available, frequency of receiving, demand for this kind of information, number of staff and staff time involved in manual organizing/ processing, number of staff and time involved in retrieving the information, space occupied in the cabinets and other points that involve with manual system. On the other side, organizing and processing methods with the new system, storing time, labour involved, retrieval efficiency etc. can be evaluated. The analysis could be conducted selecting a system user and end user samples using above mentioned parameters.

REFERENCES

1. Auger (C P). *Information sources in grey literature.* Bowker Saur, London. 1998.
2. *Canofile software (an advertising brochure).* Saitama,Canon Electronics Ltd. 2000.
3. *Canofile software: a Digital Document Management System.* Metropolitan Computers (pvt) Ltd. Colombo, 1999.
4. Evans (Agard B); Child (P). *Trade literature. In Special materials in the library.* The Library Association, London. 1963.
5. *Magneto-Optical Discs (an advertising brochure).* Mitsubishi Elec. Co. Ltd.Tokyo.1999.
6. *Raunniman Software Digital file cabinet, duplicate image finder.* 2001. *http://www. runningmansoftware.com*
7. *ScanDoc: Document Scanning and Imaginf System by Panasonic.*1998-2002. http://www. scandoc.com
8. Seneviratne (W). *Report on feasibility of using Canofile system for storing reprints and pamphlet collections of the CISIR library.* (written for the old version of Canofile) Colombo.1995.
9. Van der Heij (D G). Synopsis publishing for improving accessibility of grey scholarly information. *Journal of Information Science.* Vol. 11; 1985; p.95-107.
10. Wood (D N). Grey literature: role of the British Library Lending Division. *ASLIB Proceedings.* Vol. 34 (11/12); p.459-465.