ROLE of SCANNING and OCR in DIGITIZATION

▪ DIGITIZATION :

Digitization is the process of converting information into a digital format . In this format, information is organized into discrete units of data (called bits) that can be separately addressed (usually in multiple-bit groups called bytes). This is the binary data that computers and many devices with computing capacity (such as digital cameras and digital hearing aids) can process.

▪ NEED ARISES :

- **Information Explosion -** The information explosion is the rapid increase in the amount of published information or data and the effects of this abundance. As the amount of available data grows, the problem of managing the information becomes more difficult, which can lead to information overload.
- **Limited coverage of Print Document -** As the information increases day by day and the users demand changes form document to specific information; it is impossible to cover all aspect of knowledge with the print documents. Library space is limited and it can't face this overgrowing collection of printed documents.
- **Inadequate space in Libraries -** Library space is limited. Now a days libraries are suffer from inadequate space problem due to overgrowing collection of printed documents day-by-day.
- **Budget and Staff Problem -** Libraries always suffer due to lack of sufficient budget. Value of paper and charges of printed document increases day by day. At the same time the users number and need increases day-by-day.

Again due to lack of fund, libraries can't provide sufficient staff for handling the huge physical collection of the library.

- **Remote Access -** Library time is fixed for any library. Geographically dispersed users can't effectively use the library collection everyday. Remote access or of-campus access is mandatory for these users. It is quite impossible with the printed document to do so. Again many sources like, reference books can't be borrowed and user face problem in such situation.
- **Multiple Access -** In case of printed document multiple accessibility is possible only when multiple copies of the document available. That means for every needy user one copy must be allotted. This is quite impossible due to lack of funds and physical space.
- **Physical Harm -** Printed material always face some physical harm due to irresponsible user and staffs, also by dirt, sunlight, moisture etc. Also every paper upon which the document is printed has a limited life-time. Slow rate of degradation start at the time the paper made and after its life span it become fragile.

Also many old and rare manuscript can't be given to the user for open access .

- **Changing nature of publishing house -** Due to easy retrieval process many Publishers and Government agencies now published document in digital form. For this reason

libraries also co-ordinate with those agencies and start acquiring digital documents in its collection.

For the upper mention reasons there is a need arises for the use of digital document in every library collection.

▪ BENEFITS :

- Convenient to store securely
- Better protection from physical harm
- Easy and Multiple Accessibility
- Easily Cloned into multiple copies
- Quickly Retrieved and Updated
- Shared or Transferred in no time
- Shared Securely − either as an encrypted URL or as an attachment
- Increased Staff productivity
- Environmentally friendly, Paperless document storage
- Save money on storage space

▪ TOOLS OF DIGITIZATION :

- COMPUTER
- FLATABED SCANNER
- DIGITAL CAMERA
- MICROPHONE
- AUDIO RECORDER
- VIDEO RECORDER
- OCR(Optical Character Recognition) WEBSITES and SOFTWARES
- ICR (Intelligent Character Recognition) TECHNOLOGIES
- DIGITAL STORAGE

▪ DOCUMENT SCANNING :

- Document scanning is the practice of using scanners to convert paper documents into digital images. Scanning of a document provide a way to cut costs of manipulating, increase productivity and improve access to the information.

- Document scanning converts ordinary paper documents into useful and accessible digital files. Paper documents create a barrier to productivity, accessibility, and scalability for any organization(libraries) due to their inefficient nature. Paper documents inherently are hard to manage, secure, and protect. Converting these same documents into electronic documents solves these issues.

▪ OCR (Optical Character Recognition) :

- Optical character recognition (also optical character reader, OCR) is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded editable text, whether from a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image .

- With the help of OCR, people no longer need to manually retype important documents when entering them into electronic databases. Instead, OCR extracts relevant information and enters it automatically. The result is accurate, efficient information processing in less time.

▪ Digitization of Document With the help of SCANNING and OCR :

Digitization of a Prited document can be done with the help of Scanner and OCR through the following steps :-

- **Document Scanning** - Document scanning results in an imaged version of a document, or an electronic copy of the paper document. It is just a picture of the document without any intelligence, ie. the ability to search or cut and paste into other documents.

- **Document Conversion** - Document Conversion goes beyond basic scanning and includes the OCR (Optical Character Recognition) version of the text within the document. In this step scanning image converts into searchable, editable text with the help of OCR technology.

- **Document Remediation** - Document Remediation takes document conversion another step further by including meta tags to images and signatures and correcting the read order of tables and columns for use by assistive technologies such as screen readers.

It is the process of converting a normal digital publication into a resource that is accessible to individuals with visual and auditory disabilities.

▪ THE OCR PROCESS :

The OCR process differs from one OCR program to another, and each one requires a considerable amount of learning. The program's manual will explain this process in detail. Four points deserve particular attention: quality control, tables, images, and specialized material such as formulas, foreign characters etc.

A. QUALITY CHECK -

Normally there are four quality checks.

The first is performed at the same time as OCR. Every OCR program has a built-in spell-checker that highlights every suspect letter. At the same time the image of the word appears too, making it easy to check and correct the error.

The second is a general check of the text once the OCR process is finished. Common errors are to miss a page, a paragraph, chapter titles, and so on. A general overview is necessary to check if pages are missing. It is essential to check titles, chapter headings, paragraphs, and tables.

The third is a spelling check using Microsoft Word. This program has a dictionary that is often more sophisticated than the one embedded in OCR programs. By importing the book into Word and performing a spelling check there, more errors can be found and corrected. Be sure to add to the spell-checker any particularly difficult or error-prone words, or scientific and technical terms common in that type of publication.

Finally, the completed document should be checked by an independent person who samples the complete book and checks for errors, problems with tables and images, tagging, and the general look of the resulting text. Only after this final check can a book be considered ready for digital dissemination.

B. TABLES :

OCR programs do not cope well with tables. Moreover, tables are hard to check. They contain many digits, sometimes with points and commas, and entries are easily misplaced into the wrong row or column. They require concentrated effort, dedicated work, intensive proof-reading, careful checking, and good quality control. They can be handled in three basically different ways.

First, tables can be treated as images. This involves scanning them as black-and-white images and placing them in this form at the appropriate point in the document. This is the easiest solution. There are no errors, and the only time taken is that involved in creating the image. However, this solution consumes more memory than others. Also, the resolution is not always sufficient when large tables are displayed on a computer screen. If you make the complete table fit, the resolution is too small. If you make the table over-wide, the user must scroll to see all columns and rows, and cannot get an overview of the contents.

Second, tables can be recreated manually by making a table with the same number of rows and columns and filling the entries by typing them in, character by character.

Third, the table can be OCR'd. This saves time compared to the manual process, but has a potential for more errors. Columns sometimes get merged, and commas and points are not recognized.

C. IMAGES :

Publications contain three different general types of image :

- black and white line art;
- black and white photographs;
- color photographs.

Black and white line art should be scanned in line art mode and saved as GIF or PNG files. Black and white photographs should be scanned in greyscale mode and saved as GIF or JPEG files. Color photographs should be scanned in color mode and saved as JPEG files. Generally speaking, medium-quality JPEG provides adequate resolution.

For most collections, images consume the bulk of the space required on a hard-disk or CD-ROM. This makes it important to optimize each image for clarity and visibility, while minimizing its size. To save space you might drop some or all of the images if they are not relevant to the text.

Once the images have been scanned, you can put batch-processing programs to work to resize or enhance all the images at once.

D. SPECIALIZED MATERIAL :

Many documents contain specialized material such as special characters, formulas, and difficult pages. Special characters generally relate to different languages and diacritical marks. The language option for the OCR program should be set for the specific language being read. Formulas will have to be recreated manually. Sometimes this is not possible in the OCR program, but only in a word processor like MICROSOFT Word. Difficult pages that contain complex material or are damaged so that a clear image cannot be obtained might have to be retyped manually.

▪ EXAMPLES :

Digitization of a print or hand-written document can be done by the following three process :

1. DIGITIZATION THROUGH MOBILE APP (SCANNING + OCR)

"PDF Scanner: Document Scan+ OCR (Android Users/Free)"
One of the most popular OCR apps, which continues to receive rave reviews for its easy to use functionality is the 'PDF Scanner : Document Scan+ OCR'.

2. DIGITIZATION THROUGH PC SOFWARE (SCANNER + PC S/W)

"FreeOCR SOFTWARE"
The basic version is free to download and use and this can be used in home system.

3. DIGITIZATION THROUGH ONLINE WEBSITES

"i2OCR (http://www.i2ocr.com/)"
Digitization not only done through Free OCR Software but also there are many free websites which provide the same function and with less complexity of installation as in free OCR software .

▪ CONCLUSION :

- Thus we see that with the help of Scanning and OCR we can digitized any print document even image of a document or handwritten text into computerized editable text. This process is very much useful to make any rare manuscripts from physical harm and also make it user accessible so that user can get knowledge from it.
- In case of digitization of any pics or handwritten text through Scanning and OCR technology, one thing must be kept in mind that the rate of digitization and extraction of text is depend on the clarity of the pics and the writing. 100 % extraction from a handwritten document is quite impossible.
- However, Digitization of printed document via Scanning and OCR make a new era of dissimination of knowledge and preservation of document through Web.