IASLIC Bulletin 52 (2). 2007. p.105-114

Online Digital Asset Creation : Technology and Standards

Devika P. Madalli

Lecturer, Documentation Research and Training Centre, Indian Statistical Institute, Bangalore

Manju P.U.

Project Assistant, Documentation Research and Training Centre, Indian Statistical Institute, Bangalore

Digital library initiatives in India consist of either born digital resources or material that is retro-converted from hard copies while some are hybrid having both the types. Majority of the digital libraries have some kind of digitization activity like the partner projects of the Universal Digital Library project or the mammoth National Manuscript Mission digitization project. Often before embarking on digitization projects, implementers are faced with various steps in decision making regarding components such as hardware and software, storage devices and network options. In addition, they also need to make the right decision on what formats to store the digitized collection in, how to manipulate the data, organizing techniques and adhering to world standards that are applicable at different stages in the project. This paper leverages the experiences of the DRTC digital library of seminar and conference proceedings. The paper discusses about the digitization project and the issues and the steps involved in the process of digitization. The paper also outlines some of the standards followed in the process, and the formats used for the digital file storage.

Keywords : *Digitization, Retroconversion, File formats, OCR, Institutional Repository, Standards.*

1. Introduction

For almost four decades now, Documentation Research and Training Centre, Indian Statistical Institute (DRTC/ISI) has been regularly conducting annual seminars and conferences in the area of Library and Information Science (LIS). The proceedings are published by DRTC, Indian Statistical Institute. These volumes, especially the earlier ones, are in hard copy only and many of the volumes are no longer available. However, owing to the seminal contributions of eminent LIS researchers in the seminars and conferences, there is demand to this day for the articles and discussions documented in the seminar and conference proceedings. These volumes are deemed a treasure house of the ideas of Library and Informatio Science (LIS) especially in evolving period c important areas of research such as facete classification, scientometrics, subject indexing and artificial Intelligence applications to LIS among several others. DRTC/ISI has initiated a project aiming to make available the conten of these seminars and conference proceedings through digital repository, so tha it provides an online organized collectior along with useful search features to retrieve the articles.

However, there are various issues to be addressed in the process of establishing digital repositories. In this paper, we have documented the issues and research experiences in the process of digitization, tro-conversion of the data, file format ındling and repository implementation. Also scussed are digital repository solutions and e standards that apply to this work.

Digitization

The process of converting printed format to digital format is called digitization. The ·ocess involves creating a digital image, text ɔnversion and then presenting it on a ɔmputer, local area network or the Internet. ı other words, it can be said that digitization the process of converting information into digital format. In this format, information is rganized into discrete units of data (called its) that can be separately addressed.

Digitization is the act of making something igital, expressing a physical object "in umerical form, especially for use by a omputer."[1] By digitizing an original and lacing a digital copy of it on a computer, the le can be manipulated, transferred and stored ·ith ease.[2] Being able to access digital copies f material across the web allows patrons reater access to the content without ıcreased wear and tear on the original.

·. Issues in Digitization

?.1 *Digitization equipment*

·he hardware for a digitization project epends on a variety of factors such as the ype of material to be digitized, amount of lata on disk expected, projected the growth ıf data and such. Corresponding to the ·arious types of material that are to be ligitized there are many types of scanning quipment. The variety in scanning equipment ıvailable not only represents the multitudinous ıature of the types of source documents dealt vith by digitization services, but also the increasing commercial competition in this area.[3] In general, the types of digitization equipment used are[4] :

- **Flatbed scanners**

These are most common scanners. Flat bed scanners are more like the Xerox machines or copying machines. The page or book has to be kept flatly on the glass surface and the scanning heads moves underneath it. The advantage of this type of scanner is that it is commonplace and hence support and spares are easily available. Also they come in low end models and hence costs are lower. However, flatbed scanners are always "manned" as there should be an operator to turn the page of the book and place face down on the glass plate for every scan. Also with thicker books there would be a skew near the binding hence it would result in poor output.[5]

- **Sheet feed scanners**

The flatbed scanners were complemented with sheet feed mechanism or the Automatic Document Feeder (ADF). In the earlier models the ADF cost almost as much as the scanning unit. But the later scanner models have integrated feed mechanism where a sheaf of papers may be fed and scanner automatically feeds and scans the pages. The advantage is of course that the scanner need not be manned for every scan. Constraint is that the sheets should be "sheets" - that is we cannot do this with books unless we could cut one and separate the sheets. However, for high volume scanning these are good solutions.

- **Drum scanners**

These are used for special applications where the scanner scans images that are mounted on a rotating drum. Hence these types of scanners are known as rotary scanner. The advantage is that this gives very good resolution of scanned output. However, they require very proficient operators and also the material should be amenable to be wrapped around the drum. They are also slow.[5]

- **Slide scanners**

As the name suggests, these scanners are used to scan slides and also documents that are sensitive and not amenable to direct scanning. The use of transparent media generally delivers an image with good dynamic range, but depending on the size of the original, the resolution may be insufficient for some needs.[6]

- **Microfilm scanners**

Microfilm scanners are popularly used in libraries for retro-converting microfilm based bibliographic records. This type of scanners cost more as there are only a few products in the market. The scanner output also depends on the film and data quality.

- **Handheld scanners**

The handheld scanners are especially popular for selective scanning of parts of a page or any document. But they require real precision when passing the scanner over the material it must have with coordinated speed and consistency. Any shake or varying of distance from the material will result in poor output. Hence, these are not so popular. However, handheld scanners found a special application in libraries in building bibliographic databases by selectively and progressively scanning title pages of books for different bibliographic elements and their respective tags from standards such as MARC21.

- **Digital cameras**

Scanners with digital cameras ar increasingly becoming popular. Usually digit cameras are mounted on a movable an adjustable arm of scanner. These also requir proficient handling and tend to be slow an manned operations. But the advantage is th these scanners can handle a wider variety objects.[7]

Advanced models of scanners today a mounted with high resolution digital camera balancing device that balances the halves an open book so that scanning surface even in the first page or the last page. Som of these scanners come with bundled OC software for text recognition, skew fact correction, error flagging and such advanc features.[8]

In the DRTC digital library, we a digitizing the seminar volumes that are abo 200-250 pages and mostly of A4 size pap We use a low-end flatbed scanner as th volumes are not thick. An ADF is attach and where there were spare copies of semin proceedings available, the binding has be cut off to automatically feed 50 sheets at time and perform scanning also. However, material that is very old and cyclostyl material, this is not very effective and the volumes have to be literally typed.

4. Issues in the Scanning Process

Though scanning is a routine task the are several factors to be considered. First t scanning process, albeit with sophisticat technology and tools, is still arduous. Henc it is important to make the right decisions what form we store the scanned documen The most often considered factors are t solution, bit depth and file format on disk. A w of these issues and terms often mentioned the context are discussed here :

.1 Pixels

"Pixel" is short for picture elements, which nake up an image, similar to grains in a hotograph or dots in a half tone. Each pixel an represent a number of different shades r colors, depending on how much storage pace is allocated for it. Pixels per inch (ppi) r "dots per inch"(dpi) are preferred term, as more accurately describes the digital image.

.2 Resolution

Resolution refers to picture clarity and harpness. Higher the resolution we chose, arger the file size. Hence, utmost caution has o be exercised when making image scans. Big file size on networks clog bandwidth and often are timed out when a user on low andwidth tries to access. Also, with images we often make conversion from one format to other and these conversions are often lossy ransformations. Hence, it is advisable that we should use a high format master and then make conversion to other formats as required.

The digital library federation lists in its recommendations: 300dpi, 24-bit, color TIFF for images and 600dpi, 1-bit, bitonal TIFF for pages of text.[9] First part of the recommendation - 300 dpi, 600 dpi - addresses the resolution at which the image is segmented. The number equals the number of pixels (picture elements) captured within an inch, abbreviated as dpi (dots per inch) or ppi (pixels per inch). The higher the resolution the finer the grid used to segment the image. Resolution has a proportional effect on file size. i.e. higher the resolution used it increases the file size.

Image Resolution is of two types i.e. Optical resolution and interpolated resolution. Optical resolution is the maximum number of pixels a scanner is physically capable of capturing. Interpolated resolution is artificially generated: software takes pixels captured by the scanner, expands the grid pattern, and estimates pixels that lie in between the pixels that are captured by the scanner.[10] The table given below shows how the file size (for a particular file) varies with the change in resolution and color.

Table 1 : Resolution chart

Document type	600ppi Resolution	400ppi Resolution	300ppi Resolution
Grayscale (Tiff)	27.8 MB	12.2 MB	6.87 MB
Color(Tiff)	85.8 MB	35.5 MB	20.8 MB
Grayscale (PDF)	497KB	493 KB	484 KB
Color(PDF)	680KB	663 KB	657KB

4.3 Bit depth

The second part of the recommendation is regarding Bits i.e.-1-bit, 24-bit indicating bit depth. Images may be either pure black and white or grayscale. In black and white image, the shades are described using bits. A bitonal image is represented by pixels consisting of 1 bit/inch, which can represent two tones (typically black and white), using the values 0 for black and 1 for white or vice versa. Black and white images are called 1-bit because there are only two shades, black and white. With grayscale images the numbers of bits has been increased to describe additional shades from black to white, so 8-bit grayscale means there are exactly 256 (i.e. 2^8) gradations.[13]

4.4 Color Images

Color images are created through the RGB combination of colors: red, green and blue. When objects are scanned, the reflected light is separated into red, green and blue channels. Each color is then described in bits. 8 bits are assigned to each of the colors by the specification above - thus for RGB it amounts to 24 bits. When described, each color - red, green, and blue- in 8 bits, in total it generates 16.7 million colors. If bit depth is increased file size also increases.

5. Scanner and Scanner Software

Physically scanner can only scan the document. It is the associated software or 'drivers' that helps manipulate, store and process the scanned output.

Like any device driver the scanner diver software acts as a bridge between hardware and application software. These drivers operate the scanner and transfer the digitized file to the hard drive. The scan driver may be a standalone driver provided by the scanner hardware company. The stand alone driver runs the scanner without involving other software and saves the file to the hard drive. In other cases it may be included as utility, as in Photoshop software or commercial OCR software.

Scanner software provided by the manufacturer is desirable as they give the driver specific to the model. The third party scanner software bundled with image processing packages and OCR software is also good but sometimes they may exclude the newer, older and specialized models of scanner.

Digital Imaging projects mainly use the two types of software discussed above. Some software, such as Adobe Photoshop, can serve as both the scanning software and the image editing software. The scanning software is usually limited in its functionality. The important functionality to check is the ability of the software to be able to store images in different image standard formats such as TIFF, JPG, GIF, etc. Software should also be able to convert image files from one format to another.

6. Editing of Images

While the images are scanned using a scanner, only a few source documents are in perfect condition. They might have go discolored, stained or otherwise difficult to read as with many older seminar volumes that we scan in our project. In such instances to clean and edit the original, image editing facility is needed. Adobe PhotoShop and Corel photo/paint are two widely used professional editing tools for images. In our present study 'e are using Adobe Photoshop to edit and lean the images. We first convert the scanned nages into TIFF. Then using Photoshop tool ıe noise and dust marks are removed from ıe images. In the next stage, the image is djusted into correct position then cropped s necessary.

. File Formats

There are various file formats for storing nages on disk. TIFF is the widely used file ormat. TIFF stands for Tagged Image File Format and is accepted as a standard for ırchival files. When scanning images, the ecommended practice is to save the file lirectly after scanning as a TIFF before any mage editing or compression is performed. The main reason is that TIFF format store the mage details in tagged style and literally oreserves all information of the image. The TIFF files are huge compared to other space efficient formats. Hence, they are advisable for offline storage and archival purposes only. Also TIFF file may be stored and used as Master copy to produce other formats as required by production systems like Digital Libraries.

Other common file types are GIF, JPEG and JPEG-2000, all of which are display file formats, meaning the emphasis is quick image display rather than data fidelity.[11] GIF stands for Graphics Interchange Format, a low resolution file format used primarily for web graphics and icons. GIFs are suited for image display because the color palette of GIFs is limited to 256 colors, whether that's shades of grey or individual colors. GIF is also suited for animated images such as a "flying flag" that we see on many websites. It is fairly easy to achieve such animation using GIF.

JPEG is another well-known display file format. JPEG stands for Joint Photographic Experts Group, an international coalition of photographic experts who created an image compression routine to make photographs small enough for displaying on the internet, while retaining image quality. The actual name for the file format is JFIF, JPEG File Interchange Format.[11] One draw- back of the JPEG is that the compression is lossy. Hence it is not advisable with images such as line drawings. However JPEG 2000 specification has improved and is known to give less lossy transformations.

One of the major concern in hosting images is the file size especially when we are planning image intensive websites or digital libraries. Bandwidth is often limited and expensive and hence it should be judicicusly decided what resolution, bit depth, file format will be used for a production system without bringing down the access speed or taxing the bandwidth.

8. OCR Technology

After the scanning stage we only get an image of the page. Though the letters may be "readable" they are not represented as letters to the computer. To convert the letters or digits from image to computer processible letters

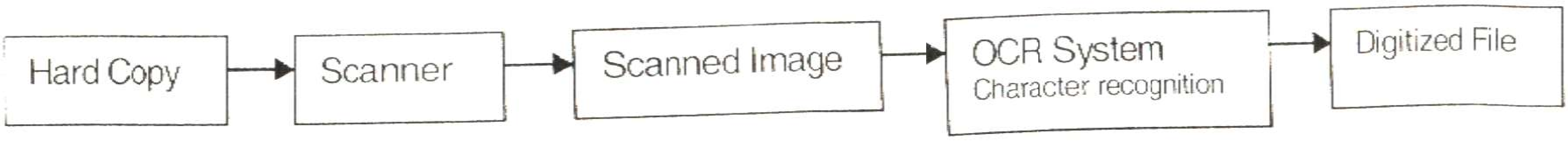

Fig. 1: OCR Process [13]

or data we have to convert it into text. OCR (Optical Character Recognition) technology converts printed characters into electroni ones that can be processed by a compute

9. Work flow of Digitization and Digital Repository

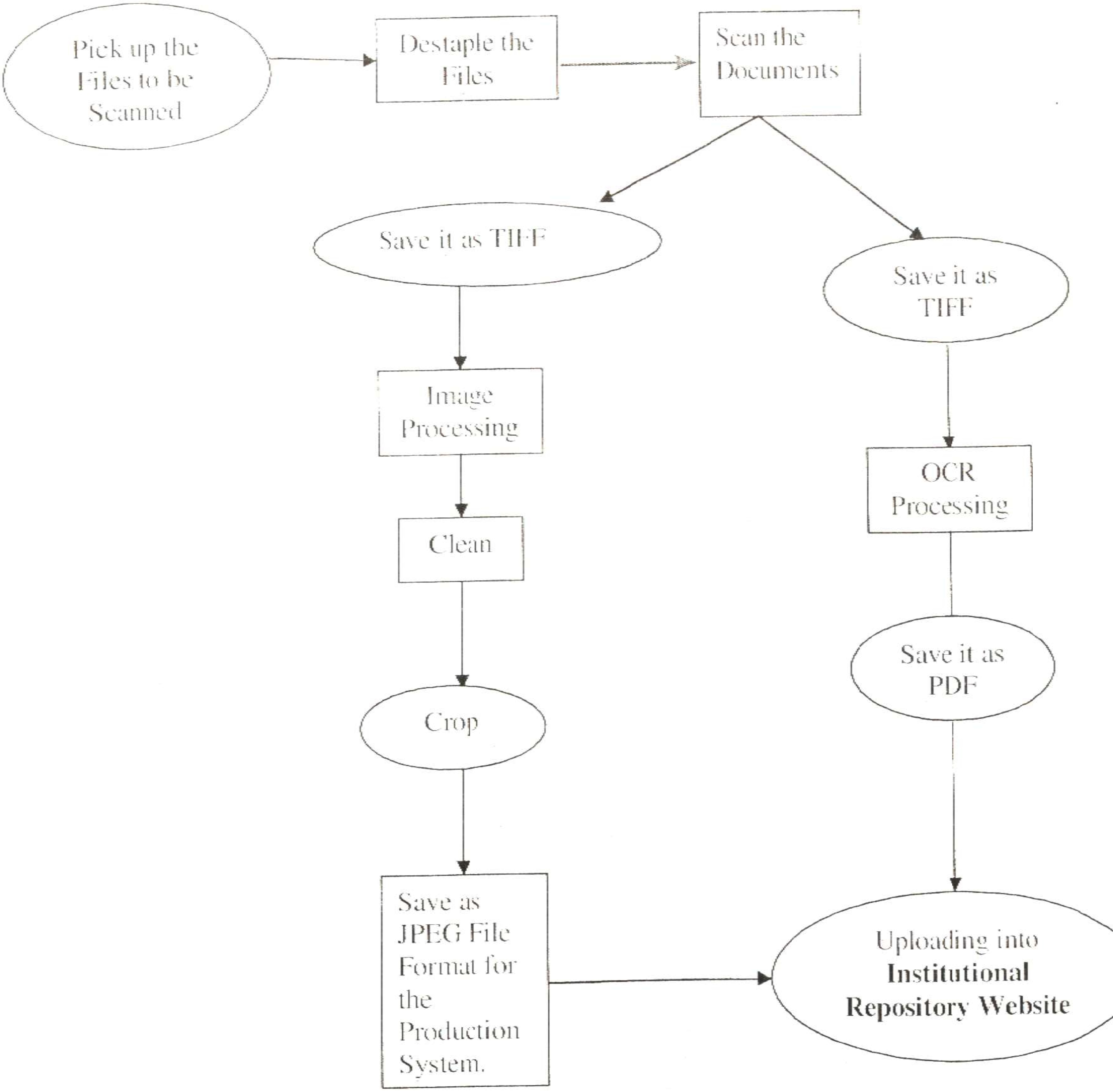
Pick up the Files to be Scanned
Destaple the Files
Scan the Documents
Save it as TIFF
Save it as TIFF
Image Processing
OCR Processing
Clean
Save it as PDF
Crop
Save as JPEG File Format for the Production System.
Uploading into Institutional Repository Website

Fig. 2: Workflow of digitization process

- *Identification of resources*

The seminars and conference roceedings are in various formats including ype-written and cyclostyled (an archaic eplicating style) forms. The volumes starting om years around 1990s are in some digital rmat produced using software like WordStar, hi-writer, TeX etc. The various seminar olumes and forms in which they are available re recorded in this stage.

- *Scanning*

HP Scanjet 4070 Photosmart Scanner Flat bed) is used for the digitization work. he documents are scanned in 300dpi esolution. As the resolution increases the file ze also increases, hence we maintain 300 oi. After the scanning the file is saved as DF image format. In the process of OCR Optical Character Recognition), it is found at 300 dpi impressions of the proceedings re resolved to text without many pographical mistakes.

Cleaning

As the printed pages of the volumes are ometimes old and stained, the scanned ersion is not very clear. Much noise in the rm of dark marks is found in the scanned utput. The file is saved as TIFF. This image cleaned using Adobe Photoshop. TIFF is ored offline in DVDs as master copies. For nages, another copy in JPEG format is saved r production System.

Storing and Retro-conversion of files

The files have to be retro-converted from eir original formats to PDF. At present PDF nage format is used for first-gen files. Retro-onversion will be undertaken on PDF text rmat in the next phase.

❖ *OCR (Optical Character Recognition)*

For indexing and searching purposes we have to convert the document into searchable text form. Here, at first we tried Scan soft Omni page Pro12.0 software for optical character recognition. About 77% accuracy is achieved on scanned volumes. After the recognition, it needs lot of editing and formatting. At present, we are using Abby fine reader 8.0. Professional Edition, compared to Omni page, it gives the highest level of recognition accuracy and format retention After necessary editing the documents are stored as PDF.

❖ *Digital Repository*

Repositories are important and helpful in managing and capturing intellectual assets as a part of the institutional and organizational information strategy. A digital repository can host a wide range of materials for a variety of purposes and users. It can support research, learning, and administrative processes. The great advantage of repositories is that they help institutions to develop coherent and coordinated approaches to the capture, identification, storage and retrieval of their Intellectual Assets.[12]

There are various commercial as well as open source solutions to host digital repositories. The decision of which system to use chiefly is based on the following criteria:[13]

- What type of resource the repository would host?
- How much data (size)?
- How many users and what type of interaction?
- What file formats are to be hosted?

* Whether streaming video/audio?
* Whether multilingual collections?
* Whether interoperable with similar domain DLs online?

There are many such criteria that should be carefully considered before the decision. One good booster to building digital repositories is the availability of very good open source software that is highly customizable offering a variety of features.

We have decided to use DSpace Digital Library software. DSpace is 100% open source software from the HP-MIT joint project. It has maximum number of installations among the digital libraries in India and abroad. It is customizable, scalable, and multilingual and has a very active developer as well as user community[14].

10. Conclusion

Building a digital repository encompasses a very serious study into the components, standard practices and norms. It also includes decisions on hardware and software that endures the test of time as the collections grow and the user community expands and diversifies in terms of the services expected. It is a good balance of sophisticated technology with very simple and customizable interfaces that works in most academic environments. However, it is sound policies that contribute for the success of any digital repository. All research and academic digital repositories should strive to work towards making available, openly and freely accessible information in order to foster better knowledge dissemination and scholarly interaction without the encumbrances of costly and time consuming buying and subscribing methods followed hitherto.

References

1. THE AMERICAN heritage dictionary of the english language, 4th ed. 2000. Houghton Mifflin, Boston, MA. http://dictionary.reference.com/search?q=digital (2006)
2. WENTZEL (Larry). Scanning for Digitization projects. *Computers in Libraries*. 27, 3; 2007. p. 6- 8.
3. LEE (Stuart D). Digitization methods. http://www.bodley.ox.ac.uk/scoping/digitization.html (Mar.1999)
4. DIGITAL IMAGING tutorial: image creation - scanner types. Cornel University Library/Research Department. http://www.library.cornell.edu/preservation/tutorial/technical/technicalB-03.html
5. DIGITAL IMAGING tutorial: image creation - scanner types. Cornel University Library/Research Department. http://www.library.cornell.edu/preservation/tutorial/technical/technicalB-03.html
6. SCANNER GUIDE: first time users. Nuance Communications Inc. http://www.nuance.com/scannerguide/firsttimeusers/types/ (2007)
7. DIGITAL REPOSITORIES: helping universities and colleges, JISC Briefing Paper. http://www.jisc.ac.uk/uploaded_documents/HE_repositories_briefing_paper_2005.pdf (Aug 2005)
8. SULLIVAN (Michael J). Types of Scanners. http://cui.unige.ch/db-research/pedsi/scanners/www.hsdesign.com/scanning/types/types.html (1996).
9. THE DIGITAL library federation benchmark working group (2001-2002). Benchmark for Faithful Digital Reproductions of Monographs and Serials. Version 1. http://www.diglib.org/standards/bmarkfin.htm#benchmark (Dec 2002)
10. WENTZEL (Larry). Scanning for digitization projects, *Computers in Libraries*. 27, 3; 2007. p. 6- 8.
11. WENTZEL (Larry). Scanning for digitization projects, *Computers in Libraries*. 27, 3; 2007. p. 46-48.

2. MADALLI (D P) and SETH (Renu). Open source software for building digital repositories, Proceedings. NACLIN, Aug 2005 p 394 - 405.

3. MADALLI (D P). DRTC digital library of seminar and conference proceedings: a proposed model, DRTC Workshop on Digital Libraries: Managing Convergence, Continuity and Change, Bangalore. Mar 12-16 2001. Paper H.

14. NOERR (Peter). Digital library tool kit, 3rd ed. 2003. Sun Microsystems.

Appeal to the Authors

Authors are earnestly requested to follow meticulously 'Information for Authors' brought out in the second cover of every issue of this Bulletin, before sending their papers to us. We find that instructions regarding some of the following issues are usually not being taken into consideration by the authors in several occasions –

1. *Two copies* of articles typed in double space on one side of A4 size paper with wide margins (at least 3 cm.) should be submitted.

2. Each article must be accompanied with an abstract preferably not exceeding 150 words.

3. The name of the author(s) should appear *only* in the *title page.*

4. Each article must be accompanied with at most eight key words just after the abstract.

5. Illustrations should be submitted in original, as they serially cited within the text, preferably on tracing paper drawn with black Indian ink.

6. Reference to literature collected from hardcopies or websites should be mentioned consecutively as they are serially cited within the text. For a web document date of search should be incorporated as given in item no. 4.13 of the Publication Policy and Guidelines for Authors. Accuracy and completeness of literature should be ensured. Authors are requested to follow the "Publication Policy and Guidelines for Authors" published generally in the March issue of each year.

7. Total number of words in the paper should not exceed 5000.

8. Each and every contributor is requested to send a forwarding letter duly signed along with the article. Contributors are also requested to give their official designation along with e-mail address.