DIGITIZATION: AN OVERVIEW OF ISSUES

By

Prof. Harsha Parekh*

ABSTRACT

Digital resources are a comparatively new category of information materials in Indian libraries. Although in many ways managing digital resources is similar to handling other resources, there are significant differences. One major difference lies in the fact that libraries are also increasingly getting involved in the very creation of digital resources.

Digital resources can broadly be grouped into two categories – those that are originally created and distributed in digital format and those that are originally created in another format and are later converted into a digital format through a process generally referred to as digitization. Several organizations – libraries, governments, research institutions, and commercial organizations - at local, regional, national and international levels are involved on digitization activities.

This paper begins with an understanding of digitization - the meaning and the processes – and then seeks to look at the issues involved in the digitization process from the perspective of libraries. As such, the focus is greater on the digitization of printed materials, rather than objects, or sound, film or video recordings.

---

* Prof. of Library Science & University Librarian, SNDT Women's University, 1, Nathibai Thackersey Road, Mumbai - 400 020. E-mail : hsparekh@bom3.vsnl.net.in

0. Introduction

Digitization refers to the conversion of an item – be it printed text, manuscript, image, or sound, film and video recording - from one format (usually print or analogue) into digital. The process basically involves taking a physical object and essentially making an "electronic photograph" of it. An image of the physical object is captured – using a scanner or digital camera – and converted to digital format that can be stored electronically and accessed via a computer **(1)**. It is rarely, however, that the process stops at this stage. To optimize and exploit the use of digital documents, a considerable amount of value addition may be undertaken to the "electronic photograph".

One great advantage of electronic materials is the ability to search the entire contents of textual material for any word. For those documents that are searched rather than read (many reference books, compilations, etc.), electronics offer a tremendous advantage. Similarly, classical texts that are studied, examined and commented upon for the use of particular words and phrases is another category of materials which lend themselves to this format. In order to make the scanned "electronic photographs" searchable, it is necessary to convert the image into ASCII text files. Thus, the first level of post-processing value addition to the digitized document is its conversion into text using, most often optical character recognition software.

A second advantage of digital documents is ubiquity. A single electronic copy can be accessed from great many locations, and to many simultaneous users (assuming copyright permission is available). This feature, in conjunction with the development of the information superhighway and particularly the World Wide Web, has meant access over different kinds of machines and multiple platforms. To take advantage of the situation and enable global access to local text based documents, these need to be structured. Usually this implies tagging and the use of mark up language such as HTML, XML, etc.

Putting up a digital document on the Internet is of limited use unless it is retrievable through search engines and directories. This further requires the addition of metadata.

Sometimes, these value additions and post-scanning processes are implicitly assumed in the meaning of "digitization" at other times the word is used in a restricted sense to include only scanning. The narrow or broad interpretation generally depends on the context in which digitization is undertaken and the expected use of the digital materials.

1. Technology

The basic process of digitization is fairly simple though a wide range of sophisticated techniques and tools may be used. Essentially, a digital image is composed of a grid of pixels (picture elements) arranged according to a set ratio of rows and columns. Each pixel, represents a very small portion of the image, and is allocated a tonal value; namely, black, white or a particular colour or shade of gray. These tonal values are digitally represented in binary code (zeros and/or ones). So a digital image is actually a grid made up of zeros and ones. The binary digits for each pixel are called bits and are stored in a sequence. When the digital image is displayed on a computer screen or sent to a printer, the bits are interpreted and read by the computer to produce a physical representation of the original material.

1.1 Scanning

Capturing a digital image is known as scanning. Image resolution i.e. the number of pixels in a row and colour depths determine the quality of the scanning. Digital cameras and scanners may both be used to capture the image. Both have photo-sensors, which consist of a charge-coupled device or CCD array. This is an array of electronic components, which converts light into electrical signals. The camera or the scanner image processing unit converts the resulting electrical out put into digital bit patterns.

As technology currently stands, scanning is the most cost-effective way to create a digital file. Creating a digital image of the original source material is the only way of accurately reproducing its information content, layout and presentation. In the case of printed documents, this means that the typefaces of the original text can be retained in the electronic copy as well as diagrams, photographs, and even hand-written annotations that have been added in the page margins. There are various types of scanners available. They include flat-bed scanners which can have sheet-feeders attached, overhead scanners and drum scanners.

An alternative to scanning is to photograph a document using a digital camera. Digital cameras may be hand-held or fixed. Hand held digital cameras are not suitable for archival scanning, excepting, the high-end digital cameras. They have no scanning limitations when it comes to size and shape, and can scan at an extremely high resolution (up to 15,000 pixels

across the long dimension). They however have certain lighting requirements and need a high-level of operator skill. Overhead fixed digital cameras present great potential for scanning oversize materials, media in all formats, bound material with the aid of book cradle and present a lower risk to fragile materials by allowing face up

1.2 File Format

A related issue with reference to images is the file format for storing image data. Images are represented by a set of numerical values specifying the colours of individual pixels. The number of possible values that may be assigned to a pixel varies with the format selected for image representation and data storage. In a two-bit (or binary) file, each pixel is designated as being either black or white. In the case of an eight-bit gray-scale image, each pixel may be assigned a different level of 256 shades of grey with gradations from white to black. In a twenty-four bit color image, each pixel may be any one of several million (16,777,216) possible colors. Images of greater depth require more disk space to accommodate the increasing number of possible values that may be assigned to each pixel. Colours are defined by specifying three values. RGB (or Red, Green, Blue). These three colours are considered to be fundamental and "un-decomposable".

In addition to the number of bits used to represent colours and their shades, since image files are very large, techniques of compression become critical. Compression techniques used affect the quality of the image. Although this may not be visible to the normal eye, some compression techniques result in data loss and are referred to as "lossy" file formats. There are hundreds of image file formats, many of which are proprietary. GIF, JPEG and TIFF are some common examples of image file formats. Table 1 summarizes the qualities of the common formats, which are portable across various platforms.

Format	Encoding	Compression	Quality	Portability	Origin
GIF Graphic Inter change Format	Binary	LZW	8 bits	Mac/PC/ UNIX	Compuserve
JPEG/ JFIF Joint Photographic Expert Group	Binary	RLE & JPEG	24 bits	Mac/PC/ UNIX	C-Cube Microsystems
TIFF	Binary	CCITT Gr.3 & 4 LZW, RLE, JPEG	24 bits	Mac/PC/ UNIX	ALDUS & Microsoft
PDF	ASCII & Binary	None Recently added JPEG	32 bits	Platform Independent	Adobe System

Table 1: Image File Formats

1.3 Optical Character Recognition

Another technology involved in digitization is Optical Character Recognition or OCR. Scanning results in creating images of the pages of a document. These pages may consist of text comprising of letters and words and sentences as well as line drawings, half tone pictures and symbols. When a page is stored as an image, manipulation of the text is not possible as the image file only contains a digital representation of the ``look" of a printed page but lacks understanding of any of its contents. Thus editing, cut-and-paste, correction, retrieval etc. are

not possible. This restricts the use of the scanned document and limits the advantages of digital documents until a way is found to extract the contents of the digital image into text.

The usual process by which a page image is transformed into a text file is Optical Character Recognition (OCR). The purpose of the whole OCR process is to recognize the letters, words, and symbols printed on a page. Presently, there is a wide range of commercial OCR software in use.

OCR systems usually first receive a page image as input, then they segment out characters, and finally they recognize these characters. Additionally, OCR systems may use spell checkers or other lexical analyzers that make use of context information to correct recognition errors and resolve ambiguities in the generated text. The output of the OCR process is a text file, corresponding to the printed text in the image file.

No OCR software is able to give a 100% error-proof results. If the OCR software gives up to 95% correct conversion it can be considered good. Less than 80% is of no practical use, since the correction time and effort required will be equivalent to full keying in. Thus all OCR will need a considerable manual editing, adding to the cost and time involved.

There is no proven OCR software to handle Indian language texts. Today, if Indian language materials have to be digitized there are two options – maintain the files as digital images or manually key in the material.

1.4 Markup

To make it possible to send and receive digital documents across various networks, independent of any special hardware or software platform, and to take full advantage of the format, conformance to some standards is required.

An electronic document has no inherent structure other than that of linear character/byte string. Therefore if parts of the document have to be made identifiable, conventions must be established. For example, tagging may be used to designate special parts of the text. Tagging consists of inserting into electronic documents short character strings called tags, which indicate the start or end of a part of the document. The tags found in an electronic document are collectively referred to as *markup*.

The three most commonly known markup languages are Standard Generalised Markup Language (SGML), Hypertext Markup Language (HTML) and Extensible Markup Language (XML). SGML is considered to be the mother of all markup languages, while HTML and XML are subsets of SGML. The defacto markup language on the Web is HTML and several editors - such as EditPlus, FrontPage, etc. - are available which will automatically insert the appropriate tags.

1.5 Metadata

A digitized product that is to put up on the Web needs information that makes it possible to be located. One of the principal challenges is to determine what information is essential in describing an electronic product. The “Dublin Core” (see http://purl.oclc.org/metadata /dublin_core/) and other special initiatives for structuring and standardizing descriptive data propose to combine information about the technical characteristics of digital files

(how they were created), their location, and a summary of their contents. The resulting information is known as "metadata" and is located in the header of a tagged document. Their function is to provide users with a standardized means for intellectual access to digitized materials.

1.6 PDF

Another alternative to tagging is the use of a proprietary format such as Adobe Portable Document Format (PDF) which is the open de-facto standard for electronic document distribution worldwide. Consisting of a package of software, PDF can handle scanning, OCR conversion and structuring both of text and images. Adobe PDF is a universal file format that preserves all of the fonts, formatting, colors, and graphics of any source document, regardless of the application and platform used to create it. PDF files are compact and can be shared, viewed, navigated, and printed exactly as intended by anyone with a free Adobe Acrobat Reader.

2. Library Digitization Projects

Libraries approach the digitization process from different perspectives. They may undertake digitization projects for a number of reasons e.g. they wish to share their unique and valuable resources with a larger and dispersed groups of readers, they may want to preserve rare documents they possess or they may want to save valuable shelf space by converting paper based volumes into digital documents. Individual libraries or groups of libraries working in tandem may undertake digitization projects. Collaborative projects may work under a national or regional policy. Any initiative to digitize documents needs to be carefully thought out and has the following phases:

1. Setting objectives/Clarifying purpose
2. Selecting Materials
3. Digitization Assessment and Benchmarking
4. Implementing the project – preparation of materials, image capture
5. Preserving the digitized documents

2.1 Setting Objectives/Clarifying Purpose

While there may be different immediate concerns for digitization, the underlying purpose of digitization is generally to improve access to materials. This need to improve access can occur under different circumstances. Some documents need to be made accessible over a wide geographical and cultural region. These could include government policy documents (e.g. the IT 2000 policy of the Government of India), historical documents which constitute a national heritage (e.g. the documents in the American Memory Project) or even textbooks which are part of the national curriculum (e.g. the national curriculum in UK). In situations where physical access is limited – either because of remoteness of location (e.g. accessing a rare book at the Bhandarkar Oriental Institute Library from all over the world) or inconvenience of timings, digital surrogates may serve the purpose.

Sometimes the concern is preservation and digital reformatting is seen as a means of keeping the world's heritage alive for future generations. However, as has been pointed out, the greatest collections in the world would have diminished scholarly value if access were inhibited. Preservation, therefore, is, also, access [(2)].

Occasionally, particularly in collaborative or commercial projects, a third concern is manifested – one of electronic document delivery. The desire here is to facilitate access to materials, without necessary ownership and storage costs. For example, the digitization of back volumes in the social sciences is being undertaken by the JSTOR project, with the intention that libraries may access back runs of journals smoothly (without any breaks in the collection) and without the need to lock up valuable storage space. Similarly, the ADONIS project digitizes several hundreds of current scholarly journals in the field of medicine and health in order that individual libraries can acquire them.

The desire to improve intellectual access and promote scholarship in a particular area may also motivate research and academic organizations to create digital documents. Ease of search (e.g. the Constitution of India), ability to compare different versions or editions (e.g. of Shakespeare's plays), collocating different documents and developing a virtual collection (e.g. the Raagmala Paintings) may be reasons for digitization. Statistical data from different sources may be collated in a digitized form so that future processing may be easier (e.g. selected Census data useful for Women's Studies may be compiled to enable and encourage scholars to do further statistical analysis, trend identification and forecasting).

Whatever the primary driving force for the initial digitization project, all benefits – and limitations – of digital documents accrue. It is frequently impossible to distinguish which benefit is greater. However, when national and government bodies undertake digitization projects, they are generally concerned with universal access. On the other hand when individual libraries take up digitization work, they are primarily looking at digitization to improve physical access to their resources. Universities and academic institutions are more concerned with intellectual access. (Policies are discussed at a later session).

Although the purposes are frequently indistinguishable, the decisions regarding digitization – what to digitize and how to digitize – depend on which purpose predominates – access or preservation.

2.2 Selecting Materials

In selecting individual materials for digitization, it is important to consider how closely the document fits into the purpose. Presuming the document is relevant to the purpose, several other questions need to be asked to determine its suitability.

Do you have the right to digitize?

If the document is in the public domain, or if the period of copyright is over or if you own the copyright to the document, you have the right to digitize it; if not, it may be necessary to get copyright permission. Government policy statements, reports, budgets, are some examples of public domain documents. Old materials, which are no longer under copyright restrictions, such as publications of the nineteenth century, can also be digitized. The copyright of reports and other internally generated documents rests with the institution and no permission is required to digitize them. University and academic libraries, the world over, have been involved in digitizing theses and question papers [(3)].

For other materials, permission from copyright holders will be necessary. Getting this permission may be time-consuming, difficult and involve the negotiation and payment of copyright fees. However, even when copyright is involved, if the purpose is not commercial but academic, copyright permission is not necessarily difficult or expensive. A recent

experience at SNDT Women's University indicates that when developing an electronic library of research papers of women's reproductive health and human rights, (http://www.hsph.harvardedu/grhf/) most publishers and authors were willing to give permission freely; only 3 of the 300+ authors contacted asked for copyright fees.

In this connection, it is noteworthy that the concepts of 'copyright' and 'fair use' are also undergoing a change. The Digital Millenium Copyright Act (1998), Section 404 of USA – in line with the WIPO Treaty - permits libraries and archives to make three digital backup copies of print information, but these copies may not be used by patrons outside the library premises [(4)].

Is it possible to digitize?

The nature of the source documents (the material that is to be digitized) should be considered next. The source documents can be viewed in terms of the original medium they are stored on and their physical attributes. For example a list of the most common physical attributes that need to be considered would be:

- *Physical constituency*: Paper (matt and gloss), Microform, book-bindings, Vinyl Records, Audio Cassettes, Audio CDs, Audio Tape Spools, Film, Video, etc.
- *Physical size*: The actual dimensions of the object are extremely important, i.e. it is difficult to digitize large maps or posters using conventional scanning equipment, and this may require creating a surrogate (e.g. a photograph) and scanning from that.
- *Physical robustness*: Can the document be unbound, for example? Or is it so valuable or delicate that it needs to be digitized under certain conditions? Automatic sheet feeders are fast and efficient, but they may destroy brittle paper. Digital cameras can minimize the manipulation of source materials, but subjecting certain media —watercolors, for example —to prolonged lighting is problematic.

A detailed decision making matrix for selecting materials for digitization has been developed by the Council of Library and Information Resources [(5)].

2.3 Digitization Assessment and Benchmarking

Having selected the items to digitize, the next step is to make a digital assessment to decide on goal qualities of the digital product. Since digitization encompasses a range of procedures and technologies with widely varying implications and costs, it is necessary to determine the most suitable goal quality requirements for each project. Goal qualities may be based on a number of factors - particularly the purpose of digitization and an idea of how the digital product is going to be used. A balance between complete and comprehensive details and convenience of use may need to be decided and this depends on the purpose.

For example, if the goal is to provide an image-based finding aid that helps users identify original materials of interest, slow-loading high-resolution images would not serve the purpose. If, on the other hand, the intention is to reduce or eliminate handling of original materials, an image must convey all critical information embodied in the original. If the plan is to use the matter in print i.e. desktop publishing then one needs to send the images as TIFF. If the images are going to be looked at, or used online then they should be converted to GIF (if the images are small and less than 256 colours) or JPEG if they are large and/or have more than 256 colours. If there was a need to bind a group of image into a single file and then view them, a PDF file may be more suitable.

To determine appropriate quality of a digitized output, since there are no absolute standards each project needs to develop its own benchmarks. At this preliminary benchmarking exercise, the resolution and depth of the images and the image file format must be established. Thus a digitization project for preserving rare photographs may opt for full details with the associated large size of files (say a TIFF loss-less file), whereas a national history project aimed at wide dissemination of photographs may opt for a more standard but "lossy" JPEG files.

Frequently, when preservation is the main objective, access to the digitized product is also required. In such cases, it is common to develop both a faithful master copy and other "downsized" derivatives for convenient access. It may also make economic sense, as Michael Lesk has noted, to "turn the pages once" and produce a sufficiently high level image so as to avoid the expense of reconverting at a later date when technological advances require or can effectively utilize a richer digital file [(6)]. Once captured, the archival master can be used to create derivatives to meet current, but varied user needs: high resolution may be required for printed facsimiles, moderate resolution for OCRing, and lower resolution for on-screen display and browsing. The quality of all these derivatives may be directly affected by the quality of the initial scan. Frequently, therefore, a digitization project makes several images of the same pages.

3. Implementing the Project – Preparation of Materials, Image Capture

Having selected the material and established the benchmarks and goal qualities of the digitized product, the actual implementation of the project must begin. This phase involves decisions regarding outsourcing or in-house allocation of work, preparation of materials, actual image capture and file management.

3.1 Outsourcing or In-house

The decision to undertake the digital image capture in-house or to outsource the process to an external bureau or agency will depend upon the value and condition of the source material, the scanning equipment and expertise available in-house and time and cost parameters. Andrew Hampson summarizes the advantages of outsourcing digitization projects in the following table [(7)].

Advantages	Disadvantages
? Quick Delivery Time	? Copyright of digital images needs to be assigned to client in the contract, and not retained by the bureau
? Costs can be favourable compared to in-house costs	? Lack of control over scanning environment
? Range of scanning equipment available	? Need to transport materials
? Bureau absorbs equipment depreciation and obsolescence costs	? Degree of trust involved in Quality Assurance
	? Service level agreements needs to be right

3.2 Preparation of Materials

Assembling materials for digitization, disbinding and cleaning them may be necessary, before actual image capture begins. Establishing safe handling procedures is an important aspect

when rare materials are being digitized and a balance may needs to be struck between the potential for damage and acceptable risk.

3.3 Actual Image Capture

Figure 2, which represents the key stages in the process, indicates how the actual scanning compromises only a small part of the entire process. As discussed earlier, more than one digital image may be required and if value-addition is to be made, OCR, tagging and addition of metadata are also to be undertaken.

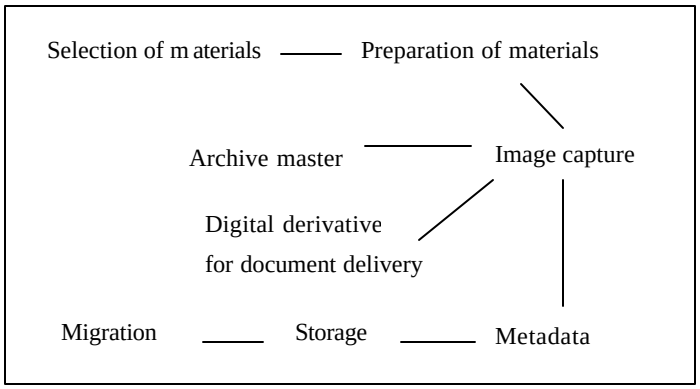

Figure 2: Key Stages in the Digitisation Chain [8]

3.4 File Management

A robust file naming convention should be set up with a view to efficiently manage the digital masters and their derivatives. The file directory structure should help in identifying the individual units of information.

4. Preserving the Digitized Documents

Rapid developments are taking place in both the hardware and software involved in digitization. This means that the present technology will soon be supplemented by newer technology. The stability of current systems and the digitized products is thus questioned. Systematic efforts will be needed to ensure that what we digitize today is not slide into obsolescence tomorrow. Migration to newer systems and media and regular refreshment are two possible solutions. However, they are both costly and time consuming; they also carry a risk of data loss.

5. Conclusion

This paper has identified a variety of issues relating to digitization. It has not examined the financial issues and costs of digitization, since they vary significantly depending on the technology used. Digitization efforts in a library require a good assessment of user needs, a clear understanding of the value of individual information resources and strong project management skills. Several libraries in India are at present engaged in digitization projects. Sharing the lessons learned in this area will be a positive step in the transformation of print-based libraries to digital libraries.

6. References

1. Hampson, Andrew: Scanning in the Right Direction. Library Technology 4 (5) November 1999. p.79.
2. Shoaf, Eric C: Preservation and Digitization: Trends and Implications. IN Advances in Librarianship. Edited by Irene Godden. V.20 New York: Academic Press, 1996. p.224.
3. Dugdale, David & Dugdale, Christine: Growing an Electronic Library: Resources, Utility, Marketing and Policies. Journal of Documentation 56 (6) November 2000. p. 644-659; Hampson, Andrew, Pinfield, Stephen & Upton, Ian: Digitisation of Exam Papers The Electronic Library 17 (4) August 1999. p.239-246.
4. Levy, Neill A: The Long Arm of Copyright Law: Problems in the Electronic Age. Part 2: Libraries, Fair Use and Document Delivery. CINAHL News 19 (1) Spring 2000 p. 4.
5. Hazen, Dan, Horrell, Jeffrey & Merrill-Oldham, Jan: Selecting Research Collections for Digitization. New York: Council for Library and Information Resources, 1998.
6. Kenney, Anne R: Benchmarking Image Quality: From Conversion to Presentation at http://www.uky.edu/~kiernan/DL/kenney.html (visited February 10, 2001
7. Hampson, Andrew: Managing a Digitisation Project Managing Information 5(10) December 1998. p.31
8. ibid.