01

राष्ट्रीय पाण्डुलिपि मिशन

॥ विज्ञानमुपास्व ॥

National Mission for Manuscripts

National Mission for Manuscripts

Guidelines for Digitization of Manuscripts

सत्यमेव जयते

A Mission about Memory

The vast manuscript wealth of India contains the 'memory of the world'. Featuring hundreds of themes, India's manuscripts represent sophisticated ideas and the most timeless of pursuits – of capturing ideas through language. It is said that a people's spoken and written language is their most important cultural attribute. In India, over thousands of years, manuscripts have been written in a vast number of languages and each in itself embodies her history.

Today, thousands of manuscripts lie neglected in institutions and homes around the country, in urgent need of conservation. India possesses more than an estimated five million manuscripts, making her the largest repository of manuscript wealth in the world. But this tremendous pool of knowledge is under threat and manuscripts are disappearing at an alarming rate. They are found on materials such as birch, palm leaf, handmade paper and cloth that require specialized care and conservation.

The National Mission for Manuscripts was launched in February 2003 by the Department of Culture, Ministry of Tourism and Culture, Government of India, to save this most valuable but less visible, of our cultural inheritances. An ambitious five-year project, the Mission seeks not merely to locate, catalogue and preserve India's manuscripts but also to enhance access, spread awareness and encourage their use for educational and research purposes.

Introduction

This document addresses the standards for creating archival quality digital still images of manuscripts. These guidelines are meant for any organisation planning to digitise these materials. The guidelines specify factor affecting image quality, file formats, storage and access standards for images.

Guidelines are prescribed to maintain consistency, high quality scan and meeting the global standards. It is advisable to use these guidelines to create images of long-term usage and reduce the need to rescan material.

Out of scope of these guidelines are:

- digitization standards for audio and video recording
- born digital material
- conservation processes for preparation of material for digitization
- scholarly work required deciphering manuscripts like folio numbers etc.

Background

India has the largest collection of manuscripts in the world. They are spread all over the country and also abroad in different libraries, academic institutions, museums, temples and monasteries and in private collections. The rich manuscript wealth of India today faces a threat of survival. The invaluable heritage of India in the form of manuscripts has to be documented, preserved and made accessible to us and to succeeding generations.

National Mission of Manuscripts (NMM) has the primary objective of using digital technology to preserve the manuscripts for the posterity. There are no digitization standards thus far available those the mission in its massive digitization initiative can adopt. NMM and National Informatics Centre – a premier Govt. of India IT organisation – has studied the best practices being adopted in several digitization projects by world libraries, museums etc. In depth study of the digitization processes of organisation such as Library of Congress, USA, National Library of Australia, National Library of New Zealand helped our efforts to develop the workable standards. National Informatics Centre has conducted prototype digitization of manuscripts of Orissa and Chennai using these specifications. Guidelines are indicative and evolving in nature. As the technology advances these need updating regularly.

Digitization

Digital technology opens up a totally new perspective. The World Wide Web holds millions of websites and the Internet is market place for research, teaching, expression, publication and communication of information. Libraries and Archives are society's primary information providers with respect to cataloguing and processing management. Besides preserving and providing access to 'born digital material' a great number of libraries nowadays have also turned to creating digital surrogates from their existing resources. **Digitization** means acquiring, converting, storing and providing information in a computer format that is standardized, organized and available from demand from common system. With specialized scanner's manuscripts are converted into compressed digital formats and stored systematically for future reference.

Target Audience

These standards are aimed at decision makers, library managers, and curatorial and technical staff members.

Why Digitize?

The reasons for implementing a digitization project, or more precisely for digital conversion of non-digital source material, are varied. The decision to digitize may be in order to:

- To increase access: this is the most obvious and primary reason, where there is thought to be a high demand from users and the library or source has the desire to improve access to a specific collection.

- To improve services to an expanding user's group by providing enhanced access to the institution's resources with respect to education, long life learning.
- To reduce the handling and use of fragile or heavily used original material and create a "back up" copy for endangered material.

Components

- Selection Policy
- Conversion
- Quality control programme
- Collections management

Selection

Background

It is important to see digitization as a series of choices where competing requirements and demands have to be balanced. When selecting source material for digitization it comes down to three basic questions: whether the source material

Needs to be converted?
Should be converted?
Can be converted?

The selection therefore has to be converted in such a way that it will assure that not only issues like the value of the selected material and interest in its content are considered but also demands concerning technical feasibility and institutional requirements.

Issues involved in the selection of material will be examined from two perspectives:

- Principal reasons for digitization
- Criteria for selection

Principal reasons for digitization

For enhanced access

Mainly for research purposes.

To facilitate new forms of access and use

The main purpose in this case is to enable the use of original manuscripts that cannot be consulted in its original form other than by visiting its specific repository, and for manuscripts that has been damaged and where technology is needed to reveal its content or shape.

For preservation

The purpose is, in the first place, to create accurate reproductions of the original manuscripts on a long lasting medium. These reproductions need to satisfy both users of today and future potential users, and must therefore both be of high quality and possess a physical stability that can be maintained over time.

Criteria for selection

Content

Regardless of the purpose for implementing a digitization project, the selection of source material will always be more or less content driven.

Demand

The level of demand is of course of great interest when selecting source material for digitization. Involving scholars and other researchers in the original decision is therefore a traditional selection methodology.

Condition

Selection of material for digitization will be affected both by its physical condition and by the existing quality. Material, which is fragile, damaged and in poor condition may present too many risks of further damage being caused by handling to allow it to be scanned without special care, or some basic conservation techniques. Similarly, if the material being considered as a candidate for digitization lacks detailed cataloguing or descriptive data, it is essential for future access to such material to create such data, and it will therefore need to be considered whether the necessary costs of doing this can be included in the overall budget of the digitization project.

Technical Requirements and Implementation

Conversion

A digital image is an "electronic photograph" mapped as a set of picture elements (pixels) and arrange according to a predefined ratio of columns and rows. The number of pixel in a given array defines the resolution of the image. Each pixel has a tonal value depending on the level of light reflecting from the source document to a charged-coupled device (CCD) with light-sensitive diodes. When exposed to light they create a proportional electric charge, which through an analog/digital conversion generates a series of digital signals represented in binary code. The smallest unit of data stored in a computer is called a bit (binary digit). The number of bits used to represent each pixel in an image determines the number of colours or shades of grey that can be represented in a digital image. This is called bit-depth.

Digital images are also known as raster images to separate them from other type of electronic files such as vector files in which graphic information is encoded as mathematics formulas representing lines and curves.

Source documents are transformed to bit-mapped images by scanner or digital camera. During image capture these documents are "read" or scanned at a predefined resolution and bit-depth. The resulting digital files, containing the binary digits (bits) for each pixel, are then formatted and tagged in a way that makes it easy for a computer to store and retrieve them. From these files the computer can produce analog representations for on-screen display or printing. Because files with high-resolution images are very large it may be necessary to reduce the file size (compression) to make them more manageable both for computer and the user.

When a source document has been scanned, all data is converted to a particular file format for storage. There are a number of widely used image formats on the market. Some of them are meant both for storage and compression. Image files also include technical information stored in an area of the file called the image "header".

The goal of any digitization programme should be to capture and present in digital formats the significant informational content contained in a single source document or in a collection of such documents. To capture the significant parts, the quality assessments of the digital images have to be based on a comparison between those digital images and original source documents that are to be converted, not on some vaguely defined concept of what is good enough to serve the immediate needs.

Input Specification

- The input documents are manuscripts of generally A2 – A4 size.

- Manuscripts are available at various repositories located at different parts of the country and need to be digitised at the Digitization Centre to be located at site.

- Manuscripts are primarily available on Paper (various types), Palm leaves, Bhoj Patra, Bamboo leaves, Pamera Leaves, Cloth, Clay Tables, Perchments, Tamra Patra, Wooden Covers, Ivory Covers/Sheets, Wooden Beads, Scrolls, Dear Skin, Micro film etc.

- They are generally very old and brittle and need special and sophisticated handling techniques.

- Some manuscripts are having illustrations/charts created using ancient inks, vegetable dyes, metals such as silver, gold etc. They are very likely to get oxidised with the effect of bright light and heat.

- All pages of the manuscripts shall be numbered before scanning, if not already numbered.

Handling and Preparation of Manuscripts

- Manuscripts taken up for digitization should have undergone conservation process as necessary.

- Standard conservation techniques are to be used for cleaning and to increase the legibility of manuscripts as and where necessary.

- Manuscripts are handled in the best way as told by scholars. Therefore, while digitization the placing the manuscripts on the scanning, filming platform shall be done by the custodian scholar or strictly under their guidance.

- In general, binding is not allowed to be taken out as it may damage the manuscripts, however in the some cases where it is absolutely necessary, due care shall be taken to remove the binding and rebind them using sophisticated methods.

- Scanner/Camera operators should wear surgical gloves so as not to damage any of the manuscripts.

- Soft bristled paint brushes to be used to wipe away the years of accumulated dust and dirt as necessary.

- Long, horizontal format requires special handling considerations.

- To maintain the sequence of loose leaf manuscripts local scholar are to be engaged to remove the thread, enumerate the pages and record the missing folios and rethread the manuscripts after digitization.

Image Capture

1. Selection of Imaging Equipment

It has an important impact on the quality of the image. Equipment from different manufacturers can perform differently, even if it offers the same technical capability. Face up scanners/ Digital Camera or any other non-touch device shall be used to capture images of the manuscripts. Flat bed and other touch devices shall not be allowed to be used as they might harm the original state of manuscripts. While selecting the imaging equipments the cost play a crucial role. The total cost of imaging not only involves the cost of the capture device but also associated peripheral devices, lighting equipments, labor cost, processing equipments, storage cost etc.

The workflow of image capturing, processing, and storing should be automated to reduce the cost. Face up scanners with low processing time can be better choice. While in digital photography flexible lighting arrangements can make the object better lit.

2. Image quality

Image quality at capture can be defined as the cumulative result of the scanning resolution, the bit depth of the scanned image, the enhancement processes and the compression applied, the scanning device or technique used, and the skill of the scanning operator.

Resolution

It is determined by the number of pixels used to present the image, expressed in dots per inch (dpi) or pixels per inch (ppi). Increasing the number of pixels used to capture the image will result in a higher resolution and a greater ability to delineate fine details, but just continuing to increase resolution will not result in better quality, only in a larger file size. The scanning of images hence will take place at 300 dpi.

Bit depth

It is a measurement of the number of bits used to define each pixel. The greater the bit depth used, the greater the number of grey and colour tones that can be expressed.

The mission follows two kinds of scanning:

- *Bitonal scanning* using one bit per pixel to represent black and white.
- *Greyscale scanning* using multiple bits per pixel to represent shades of grey, the preferred level of grey scale is 8 bits per pixel, and at this level the image displayed can select from 256 different levels of grey.
- *Colour scanning* uses multiple bits per pixel to represent colour, 24 bits per pixel is called true colour level, and it makes possible a selection from 16.7 million colours.

Illustrated manuscripts

Illustrations and charts shall be scanned separately and merged with the text at the appropriate location or any other damage. While capturing images of illustrations especially when they are created using metal such as silver, special care should be taken to avoid oxidation.

- **Image enhancement process**

  Original raw image shall be saved as per Master Image Specification.

  The raw image shall be processed to remove dirt, worm marks, water marks, noise, shadow, scratch marks, skew etc.

  Adjustment of brightness and contrast, gamma correction, sharpening and blurring, removing patterns and adjusting colors will also be the part of Image Processing.

  Cleaned image shall be saved as per Clean Master Specification.

  Two derivative images namely Access Image and Thumbnail image will be derived from the cleaned image.

- **Compression**

  It is normally used to reduce file size for processing, storage and transmission of digital images. Methods used are for example to abbreviate repeated information or eliminate information that the human eye has difficulty in seeing. The quality of an image can therefore be affected by the compression techniques that are used and the level of compression applied. Compression techniques can be either "loss less", which means that a decompressed image will be identical to its earlier state because no information is thrown away when the file size is reduced, or "lossy" when the least significant information is averaged or discarded in this process. In general "loss less" compression is used for master files and "lossy" compression techniques for access files.

Quality control

It is an important component in every stage of a digital imaging project. Without this activity it will not be possible to guarantee the integrity and consistency of the files.

3. Methods

The automated image evaluation tools that are available today are normally not sufficient for materials that are required for cultural and scientific purposes. Therefore, visual quality evaluation has to be done:

On-screen or
Print-outs

Technical limitations that can affect the evaluation must be considered, beginning with the possibilities of getting good quality printed hard copies of grey scale and colour images.

Recommended methods for:

On-screen evaluation

- View scanned images at 1:1 (100% enlargement).
- Use of target to evaluate greyscale and colour reproduction.
- Use resolution targets and histograms to evaluate spatial resolution and tonal reproduction.
- See Image Quality Control Checklist at Annex I

Print-out evaluation

- Examine by human eye hard copies created from the images to see if they fit the quality requirements.
- Compare the printouts with the source documents.

4. Scanner quality control

Before a scanner is bought, vendors should be required to deliver measurable digital results from relevant digital image quality evaluation tests. When a digital image project is running, scanning quality control measures must be set to enable operators to ensure that the scanning device is operating within anticipated tolerances. Issues of main concern in performance are: spatial resolution, tonal reproduction, colour reproduction and noise.

Collection Management

The possibility of being able to use a collection of digital images in the way it was intended depends not only on conversion standards and quality controls, but also on how the collection is managed. If the purpose is to meet not only short term needs but also to provide accessibility over time, steps have to be taken to satisfy both current use and the expectations of future users.
Plans must be made for example to:

- Make scanned images appropriate to the ultimate intended use.
- Up grade distribution of images and user interface functionality.
- Transfer images to new technical platforms to meet increasing capacity for processing and handling of digital information.
- Migrate digital images to new file formats or physical media to ensure long-term accessibility.

To make scanned images usable, great concern should be taken relating to their storage. All image files that are produced by a digital image project must be organized, named and described in a way that fits the purpose of the project.

5. Organization of images

Before a name and a description of an image file is considered is has to be decided how it should be stored. Normally, the source documents being scanned are physically organized according to principles of library management. Collections of manuscripts generally have numbers given by the library or repository where they are stored. So the organization is in such a way that just by looking at the name one could tell about the manuscript digitized.

6. Naming of images

There are two approaches for this: (1) to use a numbering scheme that reflects numbers already in use in an existing cataloguing system, or (2) to use meaningful names. Both approaches are valid, and what fits a certain collection or source documents should be chosen.

The images will be named according to Manus Id generated by the cataloguing software of the mission called **Manus e-Granthavali.** Meta Data information for each manuscript scanned will be stored in the database and is identified by its Manus Id. So the Manus Id and the Accession Number given by the institute where the digitization is taking place forms the basis of naming the digitized images of each manuscript page.

7. Description of images

To describe digital images there is a need for metadata that is structured data about data. Metadata can also be defined as data that facilitates the management and use of other data. Metadata information according to Dublin Core standards is stored through the Manus Granthavali software for each manuscript.

The highest quality file produces is referred ad Digital Master. These file are created as a result of direct result of image capture. Master file represents the original manuscripts as close and correctly as possible. Derivative images are generated from the master image using photo editing software like Photoshop.

Output Specification

1. Master Image (Original Uncleaned and Uncom-pressed)

2. Clean Master (Cleaned loss less compressed image)

3. Access Image (Derivative lossy image)

4. Thumbnails

The detail specifications of these images are as follows:

Master Image (Original Digitized Image)

File Format:	Tiff latest version
Compression:	Uncompressed
Spatial Resolution:	300 dpi
Subject Metadata:	As per standards fixed by NMM
File Naming:	As Specified

Clean Master (Cleaned Image)

File Format:	Tiff latest version
Compression:	Group 4 CCITT compression
Spatial Resolution:	8" X 10" at 300 dpi
Subject Metadata:	As per standards fixed by NMM
File Naming:	As Specified

Access Image (Derivative Image)

File Format:	JPEG latest version JBIG (in case of Black and White)
Compression:	Group 4 CCITT lossy compression
Spatial Resolution:	1024 x 768 pixels
Subject Metadata:	As per standards fixed by NMM
File Naming:	As Specified

Thumbnail

File Format:	JPEG latest version
Compression:	Group 4 CCITT lossy compression
Spatial Resolution:	1" x 1"
Subject Metadata:	Nil
File Naming:	As Specified

Metadata Creation

Subject Metadata

8. Manuscript number
9. Title
10. Other title
11. Author
12. Organization
13. Commentary
14. Commentator
15. Scriber
16. Language
17. Script
18. Complete/Incomplete
19. Subject
20. Bundle number
21. Folio number
22. Pages
23. Material
24. Missing portion
25. Illustrations
26. Condition
27. Catalogue source
28. Remarks
29. Manuscript date
30. Manuscript length (in inches)
31. Manuscript width (in inches)

Subject metadata is created according to the Manus Data Record using Manus Granthawali software of NIC already in use with in the National Mission for Manuscripts.

Technical metadata

Technical Metadata is that which describes the features of the digital file. Technical Metadata is automatically generated and assigned to the image file at the time of creation.

File Name	Assigned at the time of scanning
Date Created	
Date Modified	
Image Format	
Width	Pixels
Height	Pixels
Color Mode	RGB, grayscale
Resolution	Pixel per inch
File Size	
Color Profile	ICC color profile
Make	Scanner Software
X Resolution	Scanning Resolution in the X axis
Y Resolution	Scanning Resolution in the Y axis
Resolution Units	
Software	Imaging Software

Using Photoshop these technical metadata can be viewed.

Technical Expertise

National Informatics Center,
Cultural Informatics Division, New Delhi

Indira Gandhi National Centre for the Arts,
Cultural Informatics Lab, New Delhi

Mahabharata Samshodhana Pratishthanam,
Bangalore, Karnataka

References

Library of Congress:
Output Specifications

National Library of New Zealand:
Image quality checklist

National Library of Australia:
Output Specifications

UNESCO:
Digitization: selection criteria etc.

Annex I

Image Quality Check List

This is a indicative quality checklist for assessing digitised images. Some of the assessments will need to be made with direct comparison with the original.

1. Image is correct size/resolution in the long dimension
    a. Digital Master = 3000-5000 pixels
    b. Clean Master = 3000-5000 pixels
    c. Access File = 600-800 pixels
    d. Thumbnail = 150 pixels
2. File Name is correct
    a. Digital Master
    b. Clean Master
    c. Access File
    d. Thumbnail
3. File Format is correct
    a. Digital Master - Tiff
    b. Clean Master - Tiff
    c. Access File - Jpeg

Compare Digital copies with the original for:

4. Image is correct color made (8 bit gray scale etc.)
5. Cropped Correctly
6. Not rotated /flipped
7. Not skewed
8. Lack of sharpness/ excessive sharpness
9. No moiré patterns
10. Not Pixilated
11. Not color cast
12. Histogram (Not clipped – No loss of detail in highlights shadows – 256 shades of gray represented – tonal variation)
13. Over all Too Dark/ Too Light
14. Un even tonal values/flare
15. Excessive noise
16. Remarks