Digitization Perspective of Medieval Manuscripts

S M Shafi

Abstract

The paper discusses importance of digital interface of manuscripts as a prototype for evolving a strategy for manuscript digital library. Focuses on different issues regarding various technical and planning issues of digitization process, organization, role of metadata and delivery and access. Enumerates key projects taken over in western countries and also initiatives in India. Highlights main issues regarding manuscript digitization and archiving.

Keywords : Digitization, Digital Library, Manuscripts.

0. Introduction

Medieval Manuscripts are rich sources of tradition, history and culture. These ought to be preserved, organized and disseminated to make them available to the world at large. The western countries have taken a lead in starting digital initiatives to preserve the manuscripts but such initiatives in India are either poorly organized or in primitive stage of development. These rich treasures of our country are distributed in private and public institutions in variety of media (parchment, vellum, palm leaves, paper etc), languages, scripts, collection sizes and in different conditions. Hence there is a scope to address the problem and focus it from different dimensions to suggest solutions. The information technology has come with a promise to develop digital initiatives for manuscripts for preservation, dissemination and delivery. It is further desirable to initiate digital libraries for medieval manuscripts (DLMM) to achieve following goals:

a. A cost effective bibliographic control of manuscripts.

b. An unhindered access to contents over computer and communication network without any barrier (political, geographical, social etc).

c. The effective retrieval is possible in digital format because surveying and cataloguing them manually has not been carried out in a scientific, systematic and uniform way and that has resulted in their poor recall and precision etc.

d. The medium, generally paper, used in these manuscripts is old and not of superior quality. Most of the manuscripts are moth eaten, torn out, water stained etc. Thus digitization is a viable alternative to enhance the longevity of its life and preserve them from different hazards.

e. The digital format can serve as an electronic archival copy accessible in the network from where users can make multiple copies at different points.

f. Modern times witness new weapons of warfare and destruction coupled with escalating terrorism having inbuilt tendency to devastate treasures of culture and scholarship. Therefore, it is significant to address the problem and use image technology advances to preserve the manuscripts from such hazards.

g. Manuscripts are not fully exploited by the scholars and researchers and to make them available in digital form around the globe via internet and fully searchable through search engines will make their diffusion much wider and give more visibility to cultural heritage of ancient and medieval scholarship.

h. The revival of culture, moral values and traditions eroded by technological and industrial developments is desired these days. Hence, MMDL may help to rediscover it by making manuscripts (containing rich culture & traditions) more transparent, accessible to one and all. The attention of world towards rich past will help to have better understanding of each others values and traditions and will generate a fresh quest for knowledge among scholars who are already well set for making increasing use of computers and communication networks.

1. Digitization Process

The main focus of digitization is preservation and access of manuscripts. The digitization confronts many problems and requires utmost caution and research and exploration. The project chain normally (Beagrie Greenstein (1998) [1] consists of:

I. Instigate the project
II. To assess and select the material
III. Prepare for digitization
IV. To digitize
V. Edit
VI. Delivery
VII. Support

In context of Indian manuscripts it can be summarized in the following paras:

Most archival programme begin as an idea which may stem from a new funding opportunity or it may be a part of institutions overall strategy in managing its rare collections. Thus the initiative can be either reactive digitization in response to external request or proactive digitization when authorities target an archive of manuscripts. In reality, many institutions in India operate in reactive manner when catching a new call or opportunity for sponsorship. This immediately follows to perform an initial assessment of the manuscripts under consideration and select material from it which could be digitized. The medieval manuscripts are normally copyright free but need adequate information about the manuscript for its retrieval which involves expertise of orientalis, scholars, codocoligists, linguists etc at one or the other stage of the project. However, ‘cherry-picking’ is acceptable proposition (i.e. focusing a sized item within chosen archive when funds are limited and extended when funds are available) by using a matrix covering all varying categories by which an item can be assessed. This matrix needs to devise to meet diverse situation when many such examples can be located. (Ayris, 1999) [2]

Digitization is the core of an archival project which need source item (here manuscript or its surrogate), capture device, computer with required storage to house the digital image and necessary software interface. The picture is clear when one focuses on three components i.e. anatomy of image, image formats and digitization hardware and software. The anatomy of dots, pixels, resolution, bit depth, shade etc effect the quality of image. For example one major problem arise when resolution concerning the size of the image scanned compared with the original dimension of the manuscript but the resolution at which one chooses to scan is not always the “effective” resolution. This creates more problems especially in surrogate like microfilm but technology seems to remedy such problem in future. Bit depth and dpi are probably important factors which influence quality and dynamic range (Dmax-Dmin) all depending upon the hardware used. However it will be everyone’s interest to opt for high quality though affected by many constraints and pressures. The most frequently cited reason for not choosing maximum resolution is resulting file size. Simple formula for calculating the file size is:

Resulting file size= dpi x dpi x bit depth x dimension/8

Let us assume one page of 10 inch square item scanned at 24 –bit colour at a resolution of 600 dpi Applying formula it comes to:

600 x 600 x 24 (10 x10)/ 8

= 864,000, 00/8

=108, 00,000 bytes

108,000 Kb or 108 Mb

It clearly shows the file size extremely large— equivalent to 75 floppy disks and nine such images would fill 1 GB hard drive. There could be other reasons as well for sacrificing higher quality for better speed. However, one solution to the balancing of quality versus size is to employ some form of compression using encoding /decoding algorithm to make it much smaller. Ideally this would mean no loss of information but many file formats provide 'lossy' compressions although not necessarily obvious to human eye, do involve degradation in quality. Therefore TIFF is generally used for master images of bi-tonal documents for being 'lossless'. Interpolation is also used in connection with compression but it is best to avoid it in capturing or resizing to maintain quality true to its original.

Good number of file formats (.TIFF; .JPG; .GIF; .PCD; .PNG; .PIC; .BMP; .PDF; .DJVU; AND PYRAMID File) have their relative qualities and limitations but still TIFF is widely used as cross –platform and archiving format for high quality images to be saved without any loss in original capture but alone not tied to any particular scanner or displayed and conversion from TIFF to other formats is relatively straight forward. However .JPG and .GIF prove ideal for displaying via web in all browsers. The others may have advantage for compression, printing etc but remains to be seen how far these may go.

The digitization process has complexity in deciding about many problems of utmost importance which influences are curatorial concerns on how source material is to be digitized. For example, handling of a rare and delicate item like manuscript or painting may be restricted regardless of digitization, one could not be able to run ones' fingers across an original copy and therefore certain types of scanning will be prohibited and this contact digitization (using flatbed scanner with bright light from CCD) is often prohibited for handling fragile manuscripts, as against non – contact digitization (like camera) at a distance with employing facilities (ultraviolet or infrared) illumination for controlling light to reduce heat. So a survey of digitizing hardware (flatbed scanner; sheet feeders; drum scanners, slide scanners; microfilm scanners; different variety of digital cameras; oversized document scanners and scan backs) need to be undertaken so to arrive at right type of software for different types of manuscripts. Outsourcing to the reputed and well established firms can be another alternative for bigger archives, of course, with extreme care and vigilance. The digital software (covering capturing, processing and delivery of images) is required though much hardware comes bundled with a range of imaging software but as referred to earlier, many decisions are to be undertaken regarding file formats etc.

Processing may involve many things like cropping, fine tuning, color schemes, applying filters etc and creating of derivatives

After understanding hardware and software requirements, the archival body responsible for digitization should be able to perform digitization assessment as under:

- Suggest digitization procedure to meet objectives of the archive/library and satisfy needs and recommendations of curators/ conservation experts
- Establish digitization from original material / or surrogate

- ✍ Place of digitization : In-house or out sourcing
- ✍ Establish cost of conversion (i.e. unit cost for digitizing each unit)

  To redefine to make it more viable

The next step is in the practical issue of actual digitization and steps involved therein before capturing image are:

I) material is to be checked and prepared

II) Sample shots taken to allow for accurate <u>benchmarking</u> (process at the beginning of digitization project to set levels in the capturing process to ensure that the most significant is captured. It is a dual process involving not only the study of the source document itself but also sample shots and the bench marker should have full knowledge of present / future needs of the user)

III) Cameras calibrated accordingly when digitalization is under way.

IV) Quality Assurance (Different from Benchmarking being post process check to verify decisions made earlier were right ones) checks to be undertaken to assure digital files match all applications.

2. Organisation

Without a brows-able or searchable catalogue, end users will struggle through to find manuscripts in the collection they may seek. This descriptive information attached to digital archiving may be commonly referred to as metadata. This needs to be a thorough understanding and debate as one third of overall cost of a project of digitization is involved in it (Piglia) (1999) [3]

Text based cataloguing provides most straightforward approach to recording of information about digital manuscripts and the director/manager has to decide the structure of its catalogues and its resultant format – an appropriate metadata system which should meet the current and future requirements of the library and its users.

Metadata as term crept into standard Vocabulary of IT related projects has invited much discussion, confusion and problems that a single accepted standard have not evolved and yet user requirements are eluded out of sight. However a successful metadata system should satisfy needs of cataloguers, users, technical experts and administrators. Thus it should be flexible, extensible and forward looking which allow easy searching and browsing by the user at different points of access to collection preferably as a part of online catalogue.

Many large digital projects looked elsewhere leaving MARC or Dublin Core and RDF systems out strapped to cope up with the complexities. The solution that has been adopted is that of SGML /XML. However RDF is also used in combination and has proven more logical and practical policy. The findings of the CEDARS (The CURL Exemplars for Digital archives) (4) are extremely influential on future digital initiatives. Like many other projects, CEDFARS has recognized the need for a comprehensive metadata system and currently looking at open archival system. JIDI (JISC image Digitization initiative) project has set out a lengthy set of definitions for metadata categories and has segmented cataloguing into five levels via Collection, Work, Visual document, Person and Organization. But RLGs[5] working group on preservation issues on metadata(1998) has taken 16 categories under which digital image should be recorded. The CEDARS/ JIDI/RLG is extensive, complex and overlapping and XML approach seems to offer best possible solution for the present. Assuming that one accepts XML as the metadata system of choice, the next step is to decide the Flavour of XML i. e. which DTD is used. One may write ones own DTD, though most digital imaging project select an off the shelf DTD and at present two DTD that are greatly used are TEI and EAD. The Text Encoding Initiativer (TEI) was established in 1987 and reviewed in 1998 as TEI-lite. Besides its complexity, people have looked to other DTDs for their projects on account of the literary and linguistic background of TEI. The notable example being of EAD (Encoded Archival Description) seems to be attractive to both curator and librarians as it emanated from background of archival description. Most digital archival projects chose TEI flavor (DTD) of XML; EAD occasionally or a combination of both. Some are started to embed Dublin Core metadata or others with XML and write their own DTD.

The other aspect is describing content, file naming and creation of metadata. The knowledge about a manuscript, its relation with traditional knowledge, history and scholarship, lack of standard terms in vocabulary tools are problems which crop up very significantly and may affect the retrieval system especially at delivery stage. The uniformity, consistency and expertise of the cataloguers under the supervision of orientalists well versed in different languages will be of utmost importance. However, secondary tools (descriptive catalogues/ Directories etc) will be of immense help at this stage especially in varied cultural and linguistic diversity of Asian subcontinent. One may have to follow different extremes—simply dealing with broad description of the manuscript and other to decipher the depth matter about the nature, significance and importance of the document. The file naming of manuscripts is very useful especially if Meta get lost and thus should be meaningful and /or mnemonic conveying information about file and its relation to other files besides persistent and consistent. The first one need to convey is some form of meaning to a file name. The most obvious example would be using a file naming system that reflects the shelf mark/ accession numbers of the volume under digitization e. g. The filename “ms25fir.jpg” would suggest that it is an image of folio one, recto from manuscript number 25. Similarly .jpg suffix may indicate that this is derivative of master image (.TIF). Some users are frequently frustrated by the fluid state of WebPages and their sudden disappearance, rendering URL obsolete. It is therefore important that the identifiers one uses for digital file are persistent-hence the notion of persistent URLs (PURLS) (http:// purl.org).

3. Delivery and Dissemination Systems

Flat – file Browsing (series of Linked files) can be a quick and easy solution to many programmes but it is also extremely unsophisticated from both cataloguers and users perspective. Complicated searches are not available besides updating the site is always cumbersome. The structured approach is seen as desirable and database programme can be used to store the metadata and link it to image files. The simplest way is to use a standard desktop database system such as Microsoft access link aging images and can interface relatively easily with the web (using Microsoft ASP). The second option is to choose a system designed specifically for image database—an option that has been chosen for many projects [(6)]. These have limitations due to high cost, lack of interoperability and focused on museums and galleries. A third option is SGML/XML route to treat the catalogue as a text base. It will allow one to catalogue all the material in XML, place XML file or files on a server and allow user search the texts. The user would probably do this in web form. Then, through CGI scripts the system could convert the users input (i. e search or browse) to syntax usable by XML search engine on the server. This search engine in turn would analyze the catalogue files and return the hits; with further set of scripts converting material back to HTML for web display (although with XML this intermediate conversion process will eventually will not be necessary). Experience shows that the solution works well and is customizable to such an extent as to satisfy the most demands of users. This, however, needs knowledgeable and experienced staff.

The first digital image the user sees is the thumbnails image-i.e. a small derivative of around 100 pixels high created from a larger image. The thumbnail is usually small enough (PFG file of up to 40:1 compression) to allow to download quickly and convey at the same time meaningful graphical information about full image. From Thumbnail, the user could be allowed to download an image of sufficient resolution (say 72 dpi – a screen resolution) to fit a low end monitor of 640x480 resolutions. Some difficulty arises when one considers the next level of image for advanced analysis or study (e. g. 150- 200 dpi). The files would be larger in terms of file size (and slower to network) and one may not be keen perhaps to allow such images to be freely available. Besides creating, delivery of the full text has invited decisions to choose between transcriptions or by OCR. Having done so, it is surrounded with some form of structural make–up using a recognized SGML/XML DTD, with level of encoding to be chosen by the project management. Future copyright decisions, if any and image protection decisions are to be made at an institutional level. It will vary from collection to collection and importance and policy of the parent institution. However employing banners and watermarks may cut right across the image and always be visible and defeat the entire purpose of the programme of digitations.

The dissemination programme can be offline using CD_ROM technology and need to save as a backup for future use and transfer to other media keeping the life span of the medium in consideration. However to make the cultural and historical assets more visible on-line access will be the viable solution but proper organization and management is to be taken into account in keeping large size of the collection in view in each country of Asia. It is reported by National Commission of Manuscripts in India that there are 50 million manuscripts about India available in many media, languages and scripts. Thus such diversity in a particular country can't be managed by a single organization but needs to be organized according to the region, language, script and available expertise in the institution etc. However coordinating bodies in particular country can coordinate and monitor in order to achieve possible uniformity, consistency and maintaining of accepted standards so that overlapping is avoided and interoperability is achieved in a cost effective manner. This will also solve the web-hosting problems to a greater extent and a Portal can be maintained by the coordinating agency so to serve a Gateway for important resources of Asia with links to global digital archives in the coming future.

4. Digitization Initiatives

a. Western countries

Western countries have taken a lead in undertaking digital library projects for medieval manuscripts resulting in development of various lists, union catalogues of manuscripts in various libraries in the region or country. These have evolved into OPAC's of the manuscripts with many fields for their retrieval and displaying images of manuscripts with different resolutions. The prominent among them is the prestigious programme of Laurentius digital library (http://laurentius.lub.lu.se/) of the Lund university, MASTER (Manuscript Access Standards for electronic Records), "Medieval manuscripts of Bodleian Library", The Digital scriptorium (http://sunsite.berkeley.edu/scriptorium/), Oxford university manuscripts, (http://image.ox.ac.uk), MEDIEVAL MANUSCRIPTS OF SYRACUSE UNIVERSITY LIBRARY (http://libww.syr.edu/digital/collections/m/Medieva), The BIBLIOTHEQUE NATIONAL DE FRANCE (http://www.bnf.fr/enluminures/), "Medical Manuscripts in NLM", "European Manuscript server Initiative" (EMSI), "Unesco Memory of World", "The Humanities Text Initiative (HTI)" and "Manuscript digitization Demonstration Project of LC". These projects have used photographic and digital methods to develop a full or partial archival copy of the manuscripts available on-line using different software with JPEG format for image processing in association with different metadata initiatives. These attempts could be very useful in evolving a common strategy for digitization of vast manuscripts resources.

b. Indian initiative

The following paragraphs summarize progress on the Indian front :

i) Khuda Baksh Oriental Public Library (www.kblibrary.org)

It is one of oriental libraries having rich collection of Persian, Arabic, Urdu and other language manuscripts. The descriptive catalogue is in 30 vols which appeared in 1923 and reprinted in seventies, besides other catalogues and publications compiled by different authorities. Now the institution has launched a website and a link is given to the whole catalogue of the library in a tabular form. The table has hyperlinks to different volumes of the catalogue which can be browsed as JPEG files and browsed from as a file from top to bottom. The whole project has been undertaken by National Informatics Centre. It has not introduced any retrieval mechanism as the document is treated as collection of image files in JPEG format. The catalogue, ironically, can't be accessed under metadata or any useful descriptor.

ii) Raza Library ,Rampur, Uttar Pradesh (www.razalibrary.com)

The library developed under the patronage of "nawabs" contains a collection of about 10, 5000 vols. has appeared recently on the web with an interface for the manuscripts. It displays information of two manuscripts on a single screen, giving brief bibliographic information in Romanized form along with a folio of the manuscript scanned. However the collection is limited to few manuscripts only and further work is in progress. The retrieval engine for the data is missing as they have not structured the data nor made use of any digital library software.

iii) Kashmir University Project (www.makhtootat.org)

The Department of Library & Information Science, University of Kashmir, Srinagar, J&K has undertaken a project sponsored by the UGC for designing a database of medieval manuscripts available in Kashmir. The work is in progress and is made accessible on web after proper test and assessment.

5. Issues :

✍ High band communication with computer networks which support efficient image document transfer

✍ Data storage

✍ Data compression and interpolation

✍ Scanning and conversion technologies

✍ Planning and execution (in house, outsourcing, financial implications, funding etc)

✍ Data image file formats (.TIF. .JPG, .GIF,..PCD,.PNG,and others like.PIC.,.PDF,.BM)

✍ Text encoding initiative (TEI)

✍ Single interchange standard for both First level (abbreviated)and full description of the manuscript

✍ Development of the software for simple descriptive records and provide means for fully detailed records

✍ Link manuscript images and full text transcript

✍ Adaptation of Unicode for different scripts and languages and research thereof for developing multilingual databases.

6. Conclusion

Besides many technical issues the magnitude of medieval manuscripts and their interface may appear quite a big challenge to Indian librarians in terms of funding. It is however, quite a vast and attractive field to exploit resources and take various initiatives and provide global access to indigenous knowledge. Equally important is to plan and design them in such a way so as to learn from developments around the world in general and from the western world in particular which has given us the lead in such endeavors.

7. References

1. Beagrie, N and Greenstein, D (1996). A strategic policy frame work for in creating and preserving digital collections: a report in Digital archiving working Group. British Library and Innovation Report 107. British Library and Information centre (http:// ahds.ac.uk / image/ framework.html)
2. Ayris, P (1998). Guidelines for selecting materials for digitization. Joint RLG and NPO preservation, 21-7. (available at Http :// www.rlg.org/preservr / joint/)
3. Puglia, N (1999). The costs of digital imaging projects. RLG Dgi News 3 (5)
4. http:// www.curl.ac.uk.cedarsinfo.shtml.
5. RLG working group on preservation issues of Metadata (1998) Final Report. Mountain View (http://reg/preserve/ presmeta.html)
6. http://www.willo.com

About Author

Sh. S. M. Shafi is working in the Department of Library of Information Science in University of Kashmir, Srinagar, J&K. He has presented number of papers in seminar and conferences. E-mail : shafi_sm @rediffmail.com.