DIGITIZATION OF MARG: CREATING A PROTYPE FOR THE WEB

By

Dr. Bharati Sen*,
Ms. Smita Gubbi**

ABSTRACT

A digital prototype was developed of the journal MARG with the objective of hosting it on the web. The various issues regarding conversion of print to digital format are discussed in this article.

---

* Reader, SHPT School of Lib. Sc. SNDT Women's University, 1, Nathibai Thackersey Road, New Marine Lines, Mumbai 400 020 E-mail : sndtulib@bom3.vsnl.net.in
** MLISc student, SHPT School of Lib. Sc. SNDT Women's University, 1, Nathibai Thackersey Road, New Marine Lines, Mumbai 400 020

0. Introduction

Digitization of Marg as a M.L.I.Sc. Project was planned when Marg Publishers had approached SHPT School of Library Science, SNDT University for indexing journal issues of 50 years. Through the various discussions with the members of the editorial board, it became evident that they were keen to exploit the new opportunities offered by technology, but did not know how to go about it. To get a better idea of what is possible to do using technology, they were willing to allow SHPT School of Lib. Sc. to digitize an issue of Marg. That is how the copyright problem was taken care of. For the project, the hardware and software used were all available within the School.

MARG means pathway. MARG was launched in Mumbai on October 1946. It is a quarterly magazine on Architecture, Arts & Crafts and Culture in India with excellent reproductions in colour as well as in black and white. The journal has a readership all over the world. The contributors are also of various nationalities.

The purpose of the project was to digitize MARG for higher visibility and better access. The specific benefits desired were

- Creating a format for web presence
- Making distribution possible in various electronic formats like CD ROMs, across the Internet etc.
- To facilitate access across space without postal delay.

At the preliminary stage it was decided that the prototype would be designed to create an initial presence on the web. For this a sample issue would be digitized. The end product should be good enough to create interest amongst persons who come across it on the web and motivate them to subscribe to the journal. To put it more succinctly to reach out for potential subscribers. The second stage would be to create a CD-ROM version of a copy of the journal. At this point of time there was no decision regarding the preservation and storage aspect of digitization.

1. Basic Technology

Obtaining a digital image of a physical object can be called digital imaging, image capturing or scanning. Different labels can be attached to this process, but with existing technology the core elements remain the same. A scanning device is used in order to create a digital image from print on paper.

The scanner used for digitization in this project was HP (Hewlett Packard) Scanjet 5100C Scanner which allows dual image scanning, that is, the HP scanner software can capture recognize, and optimize images and text in a single scan. An entire page may be captured with everything on it – images in colour and/or black and white, line drawings-etc. by the press of a button. It also retains the page format, including columns. The correct exposure (resolution and bit depth) is set automatically with colour correction. The black and white line art is vectorised so that clean edges are created. Therefore no "jaggies", are formed even when the image size is increased in application.

For textual material the scanner promises better optical character recognition (OCR) accuracy. When the text is saved as image the font size, italics, bold and underline characteristics; and the page format are retained.

The computer used was a Pentium with a colour monitor. Adobe's PhotoShop and Microsoft FrontPage Editor Ver 98 were the other requirements along with Microsoft Word95.

2. Digital Image

A digital image is composed of a grid of pixels (picture elements) arranged according to a set ratio of rows and columns, similar to the tiny dots that go to make up a newspaper photograph. Each pixel, representing a very small portion of the image, 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. The bits 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. Digital images are also known as bitmapped or raster images, where the graphical information is represented as dots on a grid and image quality is dependent on the initial scanning resolution. The resolution of

an image depends upon the number of pixels in a given area. A computer monitor screen cannot typically display print resolutions.

3. Features of MARG

Marg is a quarterly journal, where each volume consists of four issues. Since this was their 50^{th} year there were 50 volumes to be digitized. The issues were selected using random sampling method i.e., alternate volumes of every 5^{th} year were studied in detail. The study was to find out certain features like the number of coloured images, number of black and white images, number of sketches in the journal. This study would help in calculating average figures for the 50 year volumes.

Certain decisions were taken while studying the issues as follows:

? FULL PAGE BLACK AND WHITE IMAGE- is one where the image is occupying the full page e.g., Volume 50(1), Page 63. Or if there is only one image on that page in the centre, not necessarily covering the whole page, with no other text i.e. part of the article (but the text describing just the image was included) or image present.

? FULL PAGE COLOUR IMAGE- is one where the image is occupying the full page e.g., Volume 5(3), Page 71. Or there is only one image on that page in the centre not necessary covering whole page, with no other text i.e. part of the article (but the text describing just the image was included) or image present.

? FULL TEXT- is a page where the information is in text form only without any kind of images or sketches e.g., Volume 5(2), page 41.

? SKETCH-a drawing in colour or black and white made up of lines and shades. It may be representing the rough structure of the original image on that page e.g., Volume15 (3) Page 10. or trying to show it by drawing what is given in the text e.g., or used for decorating the page.

? BLACK AND WHITE IMAGE-is one which has image in black and white and some shades of gray colour and with no other colour present.

? COLOUR IMAGE-is one with shades of different colours, of paintings, photographs, etc.

? BOTH TEXT AND IMAGES-A page which has text and an image or sketch along with it e.g., Volume 50(1), page 39.

? ADVERTISEMENTS WITH THEME-Every issue of Marg is published on a particular theme and some of the advertisements are based on that particular theme with colour or black and white images with or without references given. E.g., Vol. 35(1) is on Heritage of Karnataka and there are 31 advertisements on this theme. One

of which has a black and white image of “Ram and Sita in a marriage procession”, sponsored by Shri Ambika Mills Ltd.

? OTHER ADVERTISEMENTS-are advertisements with details of their products and no information on the theme e.g., Volume 5(4) Tata Industries Limited.

? SUBSCRIPTION PAGE-Most of the latest issues have a separate flap as subscription form for interested readers to be filled, the slip has perforations and can be easily detached from the journal.

Based on the above decisions, the sample issues were manually scanned and data was collected. The data was then used to calculate the average figures for the 50 volumes as in Table 1.

Table 1

FEATURES OF MARG ISSUES	AVERAGE NUMBERS PER ISSUE	AVERAGE NUMBERS FOR 200 ISSUES*
Number of pages in the issue	80	16,000
Number of full-text pages	25	5000
Number of full-page images	16	3200
Number of full-page colour images	8	1600
Number of full-page black and white images	8	1600
Number of colour images	21	4200
Number of black and white images	64	12,800
Number of sketches	8	1600
Number of articles	6	1200
Number of black and white images per article	10	2000
Number of colour images per article	5	1000
Number of sketches per article	8	1600

***1 VOLUME =4 ISSUES 50 VOLUMES = 200 ISSUES**

4. Digitisation of Single MARG Issue

The single issue of Marg that was digitized to find out about the various aspects of digitization in this feasibility study was Volume 50 (1). It was decided, after discussion, that the issue except the advertisements would be scanned. This included about 5 actual articles and 6 other articles such as book reviews, newsletters etc, and about 82 images including colour and black and white. The development of the product started with the scanning of all the text present in the issue. It is not worth capturing images at higher resolutions if the digitized material will only be looked at on a computer screen. For only screen display, a resolution of 72dpi is sufficient. So after leaving a margin, digitization was done at 100 dpi.

The images were scanned individually and were saved as JPEG file format. This was done in light of the objective of the project. Since the objective was to host the prototype on the web, the files required to be fast loading. At the same time, the images had to retain the looks of the original on the screen. JPEG as a compression algorithm is used to store large colour or gray scale files. JPEG is capable of compressing images with continuous tone or complex colour shadings. Although JPEG is lossy, it has 24 bit colour representation capability.

Each image was stored as a single image file. The reason for saving images in different files is that it is easy to link the text in the file to the image by giving file name. Therefore merging images in one file serves no purpose and it becomes rather complicated for the machine to identify a single image as an image, but it is easier for it to identify a particular file, which in turn has that particular image.

The digitized textual matter was first saved in one single file. However since the file size became too large, it was felt that it would not be appealing to the readers, as users would have to scroll down endlessly to read the articles. Thus the single file was divided into smaller files containing one article each. The text was then edited for spelling mistakes, missed lines, font size, etc in Microsoft Word. The text files were finally saved as html documents so that it would be easier to put html tags for external and internal linking.

The text files were linked with the image files through html tags using Microsoft FrontPage editor 98. Two types of linkages were provided, internal linking i.e., within the file and external linking, i.e., from one file to another file.

Every image had a caption describing the image. These captions were enclosed within tags and linked to the pertinent text. E.g., </I></FONT><A HREF="50_1F1.JPG"><I>Image-1 "Young Raja Raj Singh Watching a Dance Performance". Guler painting. Chamba by Ranjha (Ram Sahai). 1772 Opaque water colour on paper</I></FONT></A><I><FONT FACE="Bookman Old Style">; </P>

Saving the text as a word document changed the original format of the page i.e. originally the text was arranged in the form of two columns. After converting to word document it was changed to separate paragraphs. As the size of the pages was large, many a times last few lines could not be scanned and that part had to be manually typed. The same was the case for the text very close to the binding region. Some of the alphabets scanned appeared as follows: I=1; 15=is; a=@; l=I; u=n; n=u; etc. and had to be corrected manually. Sometimes the text became Italics or bold, superscripts were not recognized. The numbers had to be manually added.

File Structure

The final arrangements of the files were in a folder named V_50I_1, the main folder or a directory for the full issue. The folder has about 82 image files i.e., single file for each images, eleven text files, (named as A1.htm, A2.htm etc.), and 4 small image files, contents.htm page acting as a home page, 1 small image file for cover image, 1 small

image for the title image created using PhotoShop. As mentioned above, a single folder for an issue with number of text files was created for the issue i.e. one folder with two sub-folders for a single issue, for Text and Image respectively. The sub-folder for text was further divided into various single text files for each article.

Since the files had to be compact and quick in downloading, the textual matter had been separated from the images and stored in separate files. This obviously detracted from the visual impact of the articles. To create a balance between the two requirements viz. fast downloading and visually attractive, thumbnail images were inserted with the text. These images being small, took up little space (between 4 and 7 KB) and at the same time indicated to the reader what to expect. Each of these thumbnail images was linked to its larger version. Anyone who wanted to see the large image with all the details and colours could click on the thumbnail image.

A content page was created as in the original issues. Each item on the contents page was linked to the appropriate text file. Finally a title page was created with the project title and names of the guide and the student.

Buttons for navigation were added to move from one part of the text to the contents page or to the top as required. A Home Page was created with the title of the project and names of the student and her guide. The final product was saved on a CD-ROM (RW).

5. Conclusion

This project did not use digitization for preservation and archiving purpose. These objectives would have required a different set of decisions resulting in file formats that would not be lossy and maintain absolute image fidelity.

This development project explored various steps involved in digitization. This pilot study was helpful in finding out various file formats in which files can be saved, various software options available for creating desired output of the text and images. During the project it became very evident that clarity of objectives and thorough planning is required as otherwise a lot of time, effort and skills are wasted in producing outputs which do not measure up to the expectations or requirements.

6. References

101 TWEAKS FOR YOUR COMPUTER. 2000. Available at WWW: http://www.chip-india.com

ADVANTAGES OF HP SCANJET 5100C SCANNER. Available at WWW: http://www.scanjet.hp.com/products/classics/5100C /advantages.htm

HAMPSON, ANDREW. 1999. An Introduction to Digital Imaging. Library Technology. Available at WWW: http://www.sbu.ac.uk/litc/lt/ltcover.html

LESK, MICHAEL. 1997. Practical digital libraries: books, bytes and bucks: Morgan Kaufmann Publishera; 1997. 297p. ISBN: 155860-459-6.