Facilitating Enhanced User Access Through Palm-Leaf Manuscript Digitization – Challenges and Solutions

[1]P SRINIVAS, [2]P.RAGHAVENDRA PRASAD, [3]G.NAGALAKSHMI
[1]Dept of IT, MRIET, Hyderabad
[2]Dept. of CSE, Brilliant Engineering College, Hyderabad
[3]Dept. of CSE, Tirumala Engineering College, Hyderabad

***Abstract*— Palm-leaf manuscripts are a unique medium that serve as a repository of knowledge from across a few centuries. Digitization is an essential process to preserve this knowledge and provide access to users. Due to their unique nature, palm-leaf manuscripts pose some specific challenges. This paper highlights such challenges identified during the course of implementation at SCSVMV University Palm Leaf Library and provides suggestions for possible solutions to these challenges.**

Keywords—Palm Leaf Manuscripts; Digitization; Palm Leaf Library; Challenges;Solutions

I.INTRODUCTION

India is home to a vast treasure trove of knowledge, inscribed, stored and passed on from one generation to the next through a unique medium – palm leaves. The palm-leaf manuscripts have proven to be a very good medium for its single-most important property – its ability to remain intact for a few hundred years when maintained well. However, they are very susceptible to degradation by nature if the conditions are not ideal. The Manuscript Library at Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya (SCSVMV University), is home to several ancient manuscripts, and served as the source for the work carried out in this study and implementation. More than 5 lakh pages have been cataloged and are in different stages of digitization.

Palm-leaf Manuscripts are found across the country, inherited by families from their preceding generations. However, with the passage of time, their importance has greatly decreased, and few people understand the need for preserving and maintaining them. It is now left to individual scholars, institutions like libraries and universities, and the governments to take the onus to preserve the palm-leaf manuscripts in their original form, as well as to take up their processing especially through digitization.

Palm leaf manuscripts contain information from a variety of domains including [1]:

Indigenous medicine, including:

- Siddha,
- Ayurveda and
- Yunani Systems

Human anatomy (Varmam, surgery)
Veterinary science

1. Agriculture
2. Traditional art and architecture:
   - Temple art
   - Temple architecture
   - Shipbuilding
   - Carpentry
   - Metalworking
   - Sculpture
3. Traditional musicology
4. Techniques of writing
5. Astrology & astronomy
6. Yoga
7. Animal husbandry
8. Martial arts
9. Physiognomy (Samudrika Laksanam)

Fig.1: A set of digitized palm leaf manuscripts

Keeping in mind the rich content and knowledge available in the manuscripts, the goal is to transform the manuscript content into an easily accessible form through the application of Information and Communication Technologies. The National Mission for Manuscripts [2] is the pioneering effort in our country for the digitization of manuscripts. As part of the NMM, various Manuscript Resource Centres, Conservation Centres and Manuscript Partner Centres have been identified for survey and documentation of manuscripts as well as conservation. Kritisampada [3] is an outcome of the project wherein a national database of manuscripts is made available on the internet, with some essential metadata. Guidelines for digitization have been prepared and formalized. More than a crore pages of manuscripts have been digitized across various centres.

II. PALM-LEAF MANUSCRIPTS – SOME UNIQUECHALLENGES

Palm-leaf Manuscripts have some unique challenges, which are different from other media like paper.

Volume

The palm-leaf manuscripts have survived for a phenomenal time period of several hundred years. What this means is that the number of palm-leaf manuscripts available are quite huge. While it is a good sign that such a lot of content over a few centuries are available with us now, it also means that the volume of this content poses a new challenge for its preservation and access. The volume is also in terms of the geographical spread – the palm-leaves are found in various geographies, and with a lot of individuals or families who have inherited them from their forefathers.

Manuscripts are Fragile, and prone to damage, Large number of manuscripts are already damaged. Hence, they are to be handled carefully, and cannot be put through some automated scanners or other machines. Further, there is also reluctance on the part of the owners (both individuals and institutions) to share the original manuscripts even for the digitization process. Multiple copies of the same manuscript exist, possibly with differences in text.

Variety

The palm-leaf manuscripts are quite divergent in terms of

- Scripts – Grantha, Modi, Manipravalam, Nandi Nagari, Tamil, Telugu, Malayalam etc. Many bundles have more than one script – Grantha + Telugu, grantha + Malayalam.
- Languages – A number of languages including Sanskrit, Tamil, Telugu, Pali,
- Medium – Different kinds and shapes of palm – leaves, other types like Birch Bark etc.
- Time-period- The manuscripts are across different time-periods, thereby adding to the variety in all aspects.

Fig.2: Sample image of Damaged Manuscripts

Numbering

The leaves in the manuscript are not numbered. They are tied together with a thread, and if this thread comes off, the bundle sequence is lost. They can be re-arranged only with the help of the contents. In some cases, the library staffs have inscribed numbers directly on the original palm-leaves using a pen, thereby permanently altering its original state.

Resource Persons

Very few persons are familiar with the different aspects of manuscripts – right from how to handle manuscripts, how to digitize, understanding the various scripts and languages to catalog the manuscripts, etc.

III. DIGITIZATION

Objectives of Digitization – The ultimate objective of digitization is to provide complete content access to the end-user and to other systems.

1) Preservation
2) Dissemination (Enhanced access)
3) Reduction in handling

Digitization Process

Digitization involves four primary steps – Meta-data generation, digital image capture, linking the images with meta-data and finally, end-user access facilitation. Each of these steps involves very specific challenges. These are described, followed by a discussion on possible solutions.

Metadata Generation- Catalog Creation

Meta-data is very critical in any system to facilitate efficient retrieval and other operations. Quite often, most operations involve sifting through large parts of the meta-data and subsequent access to some selective portions of the actual data. Hence, the meta-data has to be defined with care. The National Mission for Manuscripts has defined a list of 24 properties as part of the Subject Metadata of Manuscripts. The various Manuscript centres

with financial support from the National Mission for Manuscripts has collated this information after identifying and surveying manuscripts that are with individuals and in institutions.

Fig.3: Subject Metadata of Manuscripts defined by the NMM

Based on this data collected, NMM has created a publicly accessible database which can be searched on Title, Author, Script, Subject, Language or Material. For every manuscript, details are made available in the following form:

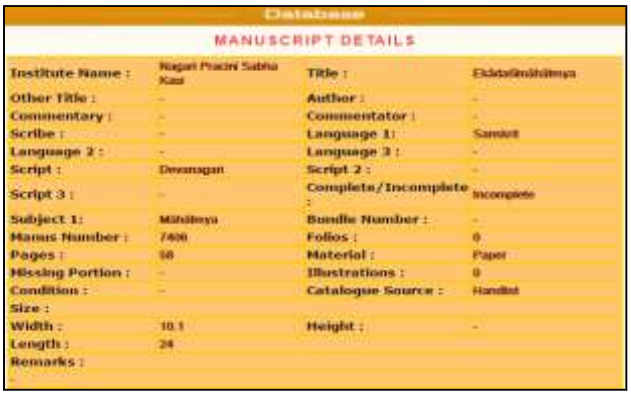

Fig.4: Manuscript Details in Kritisampada

A cursory glance at the above record highlights the fact that 14 values are recorded, some fields are empty and some have a hyphen value. There is no clarity on whether the fields are not applicable in the particular instance (for example, the Manuscript has only 1 script, hence Script 2 is marked with a „-„) or whether the particular value is not known (for example, Author is marked as „-„as the author name is not known), or whether the value has not been recorded.

Along with the above set of attributes, a few more attributes are required to describe the palm-leaf completely and then, they can be grouped into several categories of meta-data so that based on the need, only a subset of the metadata will only be used leading to faster and more accurate access–

1. Content Metadata pertaining to the actual text itself
2. Physical Characteristics Metadata – that describe the height, width, color etc.
3. Physical Condition Metadata- Condition of the palm leaves (Good, Bad, Worm-eaten, text not visible etc.), pages missing, pages in wrong bundle etc.

Further, the structural aspects of a manuscript are to be documented at two levels - the bundle level as well as individual leaf-level, as some of the meta-data attributes (like title and author) are uniform for an entire bundle of palm-leaves whereas attributes like page number, condition of the palm-leaves etc. vary from one leaf to another. Having the meta-data at different granularities will facilitate retrieval also at different levels. For example, the following is a top-level graph depicting the subject-wise count of palm-leaves available at SCSVMV University Palm Leaf Library:

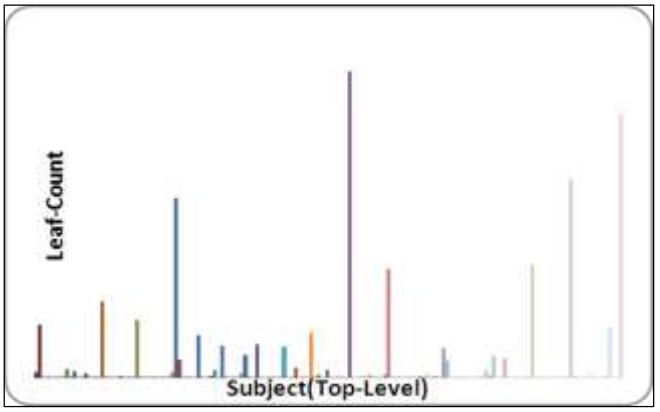

Fig.5: Subject-wise leaf count of Manuscripts

Such visualization will help an end-user understand the nature of collections available in the palm-leaf library with little effort, thereby aiding the retrieval process.

There is a strong scope for automatic extraction of structural aspects from the digitized copies through image processing, and pattern recognition etc. This will enhance our understanding of the manuscript collections, help automate the meta-data generation, and prevent erroneous entries in the catalog.

The output of this step is meta-data repository. As discussed earlier, the quality of the meta-data repository is very vital. Hence, to ensure the quality of meta-data, two steps are suggested:

- Several attributes in the meta-data (both structural and content) can be generated automatically by using digital image processing techniques and machine learning. This will result in massive meta-data generation with minimum human intervention. For example, Neethu S Kumar [4] has suggested an

automatic approach to detect and identify the characters from a manuscript.

- Massive digitization without associated quality control may be a futile exercise. Hence, there is a strong need for an automatic quality evaluation tool that can analyze the quality of metadata vis-a-vis the digital images to identify any inconsistency.

Image acquisition and OCR

Step two of digitization is the Capture of actual images of manuscripts followed by conversion of digitized images into editable text form. The National Mission for Manuscripts provides detailed guidelines on the resolution and other aspects for image capture. If the images are not of the desired quality, the digitization itself is futile. The challenge here is that the quality evaluation of digitized images is at present restricted only to physical verification, mostly by the same person who digitizes the manuscript. Hence, there is a need for automatic evaluation of the digitized images to ensure that the image is of the required resolution, etc.

Some other challenges include:

1. Location of palm-leaf manuscripts, especially in few and far places.
2. Unwillingness of owners to provide manuscripts even for image acquisition purposes.
3. Due to the fragility of manuscript leaves, automated scanners etc. cannot be used and only manual scanning is in vogue.

Once the scanned images are found to be of acceptable quality, the next step is to retrieve text from the images. The palm-leaf manuscripts contain content in a variety of scripts. Most of the scripts are ancient, and further, the scripts have changed over time. Hence, it is not common to find Optical Character Recognition systems (OCRs) for the scripts and therefore, development of OCRs for a variety of scripts is an essential part of digitization.

Even in this process, some unique aspects peculiar to manuscripts are to be borne in mind including–

1. absence of punctuation marks in the texts,
2. non-uniform layout even within a single manuscript bundle(for example, varying number of columns)
3. binding holes on the manuscript leaves in different positions
4. Damaged leaves (leading to loss of portions of text)
5. The specific way in which edits by the manuscript scribes are made in the original documents.

Hence, the development of OCR is to be preceded by an intensive pre-processing activity. Researchers have proposed the use of common image processing techniques like normalization [5] and adaptive binarization method [6] to enhance the quality of the scanned document images. An intelligent approach to noise reduction is also implemented. Character segmentation for particular languages is also suggested. Approaches to text classification for ancient scripts have been implemented in the concept of Thai Manuscripts [7]. Recent works are exploring the use of bio-inspired algorithms for handwriting recognition.

In the above processes, the challenge is also in terms of variety of Scripts and Languages making it impossible to have a single common solution. Therefore, the approach would be to develop script-agnostic systems for the preprocessing and OCR processes.

Image Storage and Meta-data linkage

The catalogues and scanned images are usually in independent silos. The linkage between the meta-data and the actual palm-leaf images should be maintained at all times, thereby providing easy indexed access. The storage of the digital images of the manuscripts should be well-defined such that any access is transparent to the physical location of the files. With a view to ensuring the reliability of the storage systems, replication is also to be considered and in the event of any loss of data, the redundant copies will be used for restoration.

End User Access through Search and Retreival Systems

The current systems including the Kritisampada by the National Manuscripts Mission provide basic search and retrieval system where one can access the catalog entries based on some attributes like title, language, script etc. There is a need for enhanced, user-friendly search mechanisms. Further, support is to be provided for access to the Palm-leafs of different scripts & Languages primarily through transliteration & translation systems. In addition to such direct access to the content, systems must be developed to form linkages with other repositories to make the manuscripts more relevant to everyday needs. For example, the contents of manuscripts could be mapped to Wikipedia pages.

Text-mining algorithms can be developed in order to automatically classify and cluster manuscripts into similar groups. Visualization is also required to provided inituitive access to end-users.

IV. CONCLUSION

Palm-leaf manuscripts are a unique repository of historic knowledge. To uncover the knowledge available in the manuscripts and make them easily accessible to domain experts and common-man, a comprehensive digitization process is required incorporating the unique aspects of palm-leaf manuscripts. Various tools and technologies are to be developed for the same including the challenges with respect to palm leaf manuscripts

REFERENCES

[1]. "Institute Of Asian Studies". Instituteofasianstudies.com. N.p., 2016. Web. 18 Oct. 2016.

[2]. "National Mission For Manuscripts". Namami.org. N.p., 2016. Web. 11 Oct. 2016.

[3]. "Kritisampada: The National Database Of Manuscripts". Namami.org. N.p., 2016. Web. 11 Oct. 2016.

[4]. Kumar, Neethu S., Dinesh S. Kumar, S. Swathikiran, and Alex Pappachen James. "Ancient indian document analysis using cognitive memory network." Presented In IEEE International Conference on Advances in Computing, Communications and Informatics, ICACCI 2014, pp. 2665-2668.

[5]. Shi, Zhixin, and Venu Govindaraju. "Historical document image segmentation using background light intensity normalization." In Proc. SPIE, vol. 5676, 2005, pp. 167-174.

[6]. Cherala, Sridhar, and Priti Rege. "Palm leaf manuscript/color document image enhancement by using improved adaptive binarization method."presented in Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP , 2008.

[7]. Chamchong, Rapeeporn, and Chun Che Fung. "Character segmentation from ancient palm leaf manuscripts in Thailand In "Proceedings of the Workshop on Historical Document Imaging and Processing", ACM, 2011.