Medicon Engineering Themes
Volume 3 Issue 2 August 2022
Editorial Note

Post Digitization Challenges and Solutions for India Palm Leaf Manuscripts

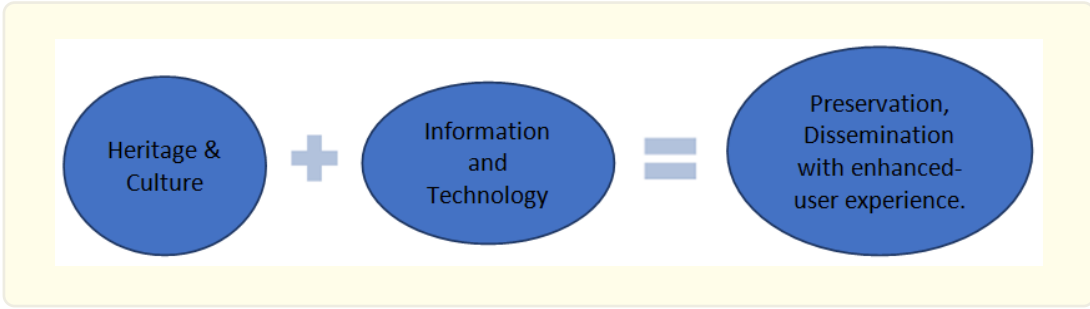
Nagendra Panini Challa*

Assistant Professor Senior Grade-2, School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India

***Corresponding Author:** Nagendra Panini Challa, Assistant Professor Senior Grade-2, School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India.

Received: August 01, 2022; **Published:** August 02, 2022

India is a country with a very deep and significant history and a grand heritage. Its culture and traditions are very unique, varied and celebrated the world over. Each and every district in the country has its own set of heritage structures, traditions and cultural practices. One of the important facets of our heritage is the palm-leaf manuscripts. Our country is home to more than five million palm-leaf manuscripts – the largest collection in the world. The manuscripts cover a variety of subjects - from music, to yoga, languages, art, architecture etc. Their fragile nature makes palm-leaf manuscripts susceptible to damage. It is the duty of our society to safeguard its culture and heritage and pass it on to the next generation. Information and Communication Technologies (ICT) provides us with an array of solutions for data collection, processing and presentation. The use of ICT in the preservation and dissemination of our Culture and Heritage is an area with high potential, and is to be explored in depth, in a domain-specific manner.



The National Mission for Manuscripts (NMM) 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 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 page of manuscripts has been digitized across various centres.

Researchers have proposed the use of common image processing techniques like Image Segmentation, normalization and adaptive binarization method to enhance the quality of the scanned document images. An intelligent approach to noise reduction is also implemented. Methods for Character segmentation for particular languages are also suggested. Approaches to text classification for specific scripts have been implemented. Recent works are exploring the use of bio-inspired algorithms for handwriting recognition. The digital journey of a manuscript commences from image capture (or scanning stage) to pre-processing, segmentation and text extraction, translation/transliteration and visualization, integration. The research carried out in the domain of Manuscripts images is primarily oriented towards enhancing a single step of the process (eg. Reducing noise, improving OCR Accuracy etc.), and mostly single-script oriented. There is a lot of scope for Research and Development in each and every one of these stages, along with a strong underlying, comprehensive framework for the entire process.

Research Challenges

Cataloging (Meta-data)

A well-built catalog (or meta-data repository) is a vital requirement to facilitate effective and efficient information retrieval, especially when pertinent information is to be retrieved from among several lakhs of manuscripts. There are several catalogs of manuscripts that are available in different organizations. A catalog of catalogs is also available, which serves as a top-level index that facilitates the search for locating a manuscript. The existing catalogs suffer from various drawbacks including the following:

- Variations - Different types of cataloging with varied levels of details.
- In most cases, there is no direct link between the catalog and the actual manuscript described in the catalog.
- The catalogs do not facilitate search on different parameters, and some are not available online.
- Text-mining algorithms can be used to identify similar manuscripts, interlink related manuscripts, as well as to group manuscripts into categories, thereby facilitating easy access.
- A centralized –access mechanism will facilitate generation of statistics on the utilization of the manuscripts. Analytics may be performed to learn about most popular manuscripts, evolution of scripts etc.
- Cross-linking with other repositories and overlaying information from other sources.

Digitization

The guidelines for digitizing the manuscripts have been formulated in the form of output image specification as well as naming standards. However, there is no quality control or evaluation beyond physical verification through on-screen evaluation or print-out evaluation. 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 the digital images, as well as the metadata and classify the quality of digitized images as acceptable or otherwise. Further, digital enhancement of the images can be undertaken to ensure that the images are of acceptable quality.Some domain experts can annotate the manuscript, and provide commentaries and interpretations, that will help people from different domains to understand the manuscript content. Various text-mining algorithms can be applied to ensure easy and varied access to the full-text.

Digital restoration

The manuscripts (especially palm-leaf) are fragile and prone to damage. A large number of manuscripts have already been affected and are in different states of damage. While a few meta-data items give an indication of the amount of damage, there is a need to assess the damage objectively. Thus, a tool for automatic measurement of damage (in terms of the actual leaf and specifically in terms of the content loss) is very essential. The damaged images can be restored through application of both machine learning and image processing, thereby making the digitized images more amenable for further processing like OCR etc. Thus, the utility value of the scanned images will increase.

Access

The catalogs and scanned images are in silos in various institutions and organizations in the country. It is essential to network all the centres and facilitate any-time, any-where access to the manuscript content, for experts and the common-man. Hence, in addition to enhancing the search engine and access solutions at the software level, it is also required to optimize the storage and access mechanism for improving the performance. Technologies and topologies are to be evolved to give a seamless access to a large number of users who are likely to access the content. Issues of IPR could arise, and hence, a controlled access to the content will be enabled to address the same. Easy access will facilitate the participation of more users, and this could also be utilized for initiating crowd-sourcing efforts to interpret manuscripts where users can interpret and describe the manuscripts.

Solutions:

The key objective of the project is to uncover the knowledge available in the manuscripts and make them easily accessible to domain experts and common-man. It is proposed to achieve the same through the development of the following resources:

- Establishment of a model for a Manuscript Digital Library – This model will consist of the hardware, software and best practices to be followed by any agency which has a sizeable collect of manuscripts.
- Hardware – Identification of optimal Storage and network technologies and topologies to provide fast access to a large user base which will serve as a model for replication at the various Manuscript Digital Libraries.
- Software – covering a broad spectrum including the following:
  i. Tool for effective data cataloging and for automatic measurement of meta-data quality.
  ii. Tool for automatic evaluation of digitized image quality.
  iii. Tool for automatic restoration of damaged manuscript images.
  iv. Multi-lingual Search-engine.
- Data Sets – Digitization of different varieties of manuscripts to produce data sets that will be made available for use by the research community.

In conclusion, the current efforts at manuscript digitization in the country are progressing at various centres. However, a quality measurement and control system need to be incorporated. Efforts to process the scanned data and transform it into readily usable form are required. The current research in the area of pre-processing (or quality enhancement) are to be merged into a tool that can be made available to all the digitization centers for automatic evaluation and enhancement of quality of the palm leaf manuscripts. Post the pre-processing, there is a need for tools to segment the data, restore the data lost due to degradation of the physical leaf and for handwriting recognition etc. which are generic and applicable for the variety of scripts in our country.

Volume 3 Issue 2 August 2022