International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-5, March 2019

Evaluation of Automatic Metadata Schema for Indian Palm Leaf Manuscripts

Nagendra Panini Challa, R.Vasanth Kumar Mehta

Abstract: ***India is home for various treasures of knowledge which are inscribed, stored and passed on from one generation to the next through a unique medium – palm leaves. In the present scenario, the need to preserve our rich heritage and provide an easy interface is the major challenge, using modern technology and techniques, thereby enhancing accessibility, applicability and appreciation for the repository of knowledge. A well-built catalogue is a primary requirement to facilitate effective and efficient information retrieval. The main aim of this research is to provide users with a standard means for intellectual access to digitized materials. Hence the outcome of this research can be useful in two ways firstly to prioritize the least/high damaged manuscript to perform restoration and secondly to obtain accurate search results from two methods proposed using TF-IDF and crowdsourcing approach. These can be widely utilized in various digital libraries across the globe. This metadata schema can be incorporated into an enhanced search engine for obtaining better precision and recall results.***

Index Terms**:** ***Palm Leaf Manuscripts, Digitization, Information Retrieval, TF-IDF, Crowdsourcing, Libraries, Precision, Recall.***

I. INTRODUCTION

The palm leaf manuscripts are identified as the original source and key to the knowledge in various fields that have been passed on from our ancestors. The valuables found in these manuscripts are considered as treasures to the future development of India, and plays a major role in nation building. This is done through the research, collection, categorization, and preservation of manuscripts as historical evidence. The study of manuscripts is still facing a lot of issues because they are not systematically collected and managed. Public access is limited to these manuscripts, because some of the manuscripts being lost or damaged even after proper catalogues have already been made. Besides, there is a huge lack of experts who can read and translate palm leaf manuscripts which are sometimes fragile and prone to damages. The benefits and the importance of the contents of manuscripts were recognized and many have tried to find the way to access and bring out this knowledge without destroying the original copy through the management via information technology through digitization. This allows users to access any part of the manuscript while allowing the experts to make changes and publish the work on the computer network to serve users' background. Automatic metadata generation is known to be more efficient, less costly, and more consistent than human-oriented processes. Researchers have advised and concluded globally to all the peers that the most effective means of metadata creation is combine both human and machine oriented methods. In order to access the digitized document effectively, a structured form of metadata needs to be created. Due to the lack of quality issues like mismatches, irrelevancy (between the actual and catalog data) in various manual extraction techniques, an automatic metadata extraction schema is proposed to increase efficiency in search, access, management of these manuscripts, which increases the quality of the retrieval process.

The main objective of this research is

1) To facilitate well defined automatic metadata extraction schema:
   Metadata has become one of the most important aspects of modern ICT based system since it directed many people between structured and unstructured data. This research work describes a schema for automatic metadata extraction from Indian palm leaf manuscripts to ease metadata creation process. Extracting metadata from images automatically has many problems. There are many problems which need a special mention out which the most important one is the quality of extracted metadata [11].
2) Access mechanism to extract embedded metadata from the image: In this research work along with predefined standards, additional metadata are also proposed in this work, which facilitates the accuracy for the digital restoration process.
3) To implement search mechanism based on enhanced metadata schema:

The metadata which are embedded in the form of exif tags are extracted with the help of keyword extraction procedures. Out of many automatic keyword extraction approaches like Alchemy Keyword Extractor, Rapid Automatic Keyword Extraction (RAKE), crowdsourcing where the simplest approach is the Term frequency – Inverse document frequency (TF-TDF).

Manuscript published on 30 March 2019.
***Correspondence Author(s)**
Nagendra Panini Challa, Department of Computer Science & Engineering, SCSVMV, Kanchipuram, Tamil Nadu, India.
R.Vasanth Kumar Mehta, Department of Computer Science & Engineering, SCSVMV, Kanchipuram, Tamil Nadu, India.

This approach work efficiently in text/image documents as it finds the terms efficiently within that document/image.

II. Literature Survey

The main objective of this works is to develop a metadata scheme for the management of digitized Palm Leaf Manuscripts (PLMs) to increase efficiency in various mechanisms like search, access, management and use. Chamchong et.al [12] proposed a research framework which was based on functional requirements, while the metadata development schema was developed on the idea of Metadata Life Cycle Model (MLM). They analyzed their framework into 3 main parts like:

(1) Analyzing the user needs, expectations, requirements and content with respect to the manuscripts,

(2) Development of a metadata schema based on the results of the above analysis; and

(3) Implementation and evaluating the final metadata schema.

The evaluation results in this research give us a perception that the manuscripts management usefulness is relatively significant with user's experience. Finally, the schema for manuscripts was developed which consisted of 76 properties based on several standards that describes all formats scripts.

Many aspects of manuscripts were thoroughly reviewed [13] which focusses on highlighting their importance, the process of inscribing on the leaves, physical and biological factors of deterioration, the classification and cataloguing process of manuscripts, and different prospects of digital preservation of manuscripts and the attempts taken by various manuscript libraries for digitization.

Manuscripts in the earlier stage were used in Southeast Asia to store early written knowledge about subjects such as medicine, Buddhist culture and astrology [14]. Therefore, these historical manuscripts play an important role in many individual's lives who like to learn about historical documents. Some researchers proposed an image segmentation approach of these historical handwritten manuscripts. This approach is composed of three steps:

1) background elimination to separate text

2) line segmentation and

3) character segmentation.

These results were applied to optical character recognition (OCR) engine method for Indian languages.

Digitized manuscripts should enable resource discovery, retrieval, collation, analysis of the most valuable content in palm leaf manuscripts. Generally, they are merely digital objects which are represented by complex data format, which is known as a metadata schema [15]. To facilitate the metadata; repositories adopt metadata schemas that are proposed through many standard organizations. Some of the metadata schemas used for digital documents are MARC, Dublin Core, EAD, MODS [16], etc. Certain factors such as type of documents, characteristic of collection, etc. are considered for selecting appropriate metadata schemata.

Some researchers have a different approach for management of manuscripts such as interpretive approach because as it is based on subjectivist and relativist assumptions. Weber [17] explains that this is a scientific process which reveals the cause and effect of social actions by using interpretive understanding. His study understands the context of palm leaf manuscripts management in the community to reveal the tacit and explicit ideas of community members.

Metadata based search [18] can play an important role while retrieving, although this would be dependent on the search goals and strategies of the users, according to Besser [19]. Another advantage of metadata is using "existing text search techniques" [20] to retrieve music information. The advantages of metadata also extend to retrieval of many annotated and reference records. As per the findings of Milosavljevic [21], using a metadata component for the retrieval of reference records enhances flexibility of the indexing process and efficiency of the retrieval system.

At sometimes, in a mixed environment where metadata from other systems are created by library staff can bring challenges to creation, management and access of information [22]. So there is need of customization to make information retrieval an easy process for the user. Generally, metadata is viewed as a system's data dictionary, capturing definitions of data entities and the relationship among them [23] which maybe complex but, decision makers need to determine the contextual relevance and improve retrieval efficiency.

1) Dominika Tkaczyk et.al [1] addressed these research problems by proposing an automatic, accurate and flexible algorithm for extracting wide range of metadata directly from scientific articles in born digital form. It includes basic document metadata, structured full text and bibliography section. It is implemented by deploying various machine-learning algorithms trained on large & diverse datasets. They evaluated the performance of proposed metadata extraction algorithm which showcased good results which is reliable and accurate solution to the problem of metadata extraction from documents.
2) Alan Pinto Souza et.al [2] proposed a metadata extraction method called 'Artic' which is based on a probabilistic framework based on some conditional random fields. It aims at identifying the main sections and for each section of the metadata information.
3) Christopher Clark et.al [3] proposed a model which identifies and extracts figures, tables along with their captions from articles (academic) is essential for article summarizations that seek to gain deeper, semantic understanding of the articles.

His approach analyzes the structure of individual pages of a document by detecting parts of body text, and locates the figures or tables within that text [4]. The evaluation of this model achieved 96% precision at 92% recall when tested against many scholarly articles, surpassing previous state of the art.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

4) Mario Lipinski, Kevin Yao et.al [5] evaluated the performance of various metadata extraction tools from scientific articles. He made a comparative study between different extraction tools and made them available for developers to analyze accurate and effective metadata. The GROBID, arXiv collection delivered the best results, followed by Mendeley Desktop depending on the metadata type to be extracted.
5) Francesco Ronzano et.al [6] proposed a schema to automatically extract, enrich and characterize several structural and semantic aspects of scientific publications. Based on the scientific Text Mining Framework the papers were analyzed and developed. On-line access for their work like Web visualizations are generated which are obtained by mining the scientific papers dataset.
6) Alexander G. et.al [7] provided a convenient access to web-based data and intelligent systems which are developed to construct a knowledge base from several huge chunks of unstructured information. The rich metadata extracted has been used for many data mining projects which provides free access to many full-text academic documents. He gave a brief architectural overview which highlights several research driven developments for processing academic information.
7) Nitin Kumar et.al [8] mainly focused on extracting metadata information from web pages and HTML pages with some good accuracy. It supported different types of documents like News Articles, Patents, in PDF format, Blogs, Websites and more [9]. This algorithm comprises of different machine learning phases which is obtained by natural language processing techniques.

III. PROPOSED WORK

Metadata has become one of the most important aspect of modern ICT based system since it directed many people between structured and unstructured data. However, describing annotating metadata through manual process is time-consuming, labor-extensive, and expensive. This research work describes a schema for automatic metadata extraction from Indian palm leaf manuscripts to ease metadata creation process. Extracting metadata from images automatically has many problems. There are many problems which need a special mention out which the most important one is the quality of extracted metadata [11].

A. Design and Development of Metadata Schema

Many researchers basically divided metadata into different categories [24] like descriptive metadata which consists of creator, title, subject, location; Structural metadata which describes how a metadata item is structured; Administrative metadata which includes the different descriptions like ownership and other related data; Rights management metadata which deals with intellectual property rights; Preservation metadata which deals with archival information; Technical metadata which includes document characteristics to a highly detailed hierarchy level. Many different metadata schemes are being developed in a variety of user environments and disciplines [4]. The metadata are general with the help of high level language, which are encoded into digital form to machine – readable. Different languages are used for encoding metadata standard like HTML, XML, SGML, XRENT, ODRL and SMIL [5].

Metadata Standard	HTML	XML	XRML	ODRL	SMIL
DC	Y	Y	N	Y	Y
AACR	Y	Y	Y	N	Y
EAD	N	Y	Y	N	Y
TEI	N	Y	Y	N	N
METS	N	Y	N	Y	Y
MODS	N	Y	N	Y	Y
MARC	N	Y	N	Y	N

Table: 1- Table showing XML is the most widely used language [8] for Encoding Metadata

B. Development and Identification of accurate Data Elements

Dublin Core standard is one of the most popular which proposed 15 core metadata elements like Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage and Rights. Their main idea is to define a set of elements that could be used by the researchers and experts to describe their own web sources. On each metadata element the researcher and expert's idea may vary according to their perception. The initial metadata consists of Title, Author, Language, Subject, Bundle number, accession number, Total folios, Condition, Catalog source, Date, Original Title (in Indian Language), Script, Material, Catalog Type, Missing Folios. All the metadata mentioned here are given to the system through some human intervention which leads to a mismatch between actual scripts to existing scripts at some instances. So automatic extraction of metadata from these manuscripts is implemented which minimizes the mismatching between manuscripts which eventually increases the quality of metadata.

S.No	Metadata Element	Description	Element Available In Dublin Core
1	Title	It defines the main theme of the manuscript image.	Y
2	Author	It refers to the creator of the document image.	Y

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

3	Subject	It refers to the related are that the author has worked.	Y
4	Description	It refers to the quick abstract of the manuscript image.	Y
5	Publisher	It refers to the one who owns the manuscript image.	Y
6	Language	It specifies the language of the manuscript.	Y
7	Date	It refers to the actual date when the document was prepared.	Y
8	Type	It refers to the file type of the manuscript image.	Y
9	Edge Damage Percentage	It refers to the total edge damage occurred to the manuscript image.	N
10	Source	It refers to the place from where the manuscript is collected.	Y
11	Bundle Number	It refers to the number in the university library where the script is stored.	N
12	Accession Number	It refers to the actual identity of the manuscript in the bundle.	N
13	Coverage	It refers whether all the subjects is covered in the manuscripts.	Y
14	Rights	It refers to the University copyrights property.	Y
15	Total Folios	It refers to the total number of leaves in a single bundle.	N
16	Condition	It specifies the present condition of the manuscript.	Y
17	Word Count	It refers to the total number of words in a manuscript image.	N
18	Script	It refers to the type of script present in the manuscript image.	N
19	In Boundary Damage Percentage	It refers to the total damage inside the manuscript image.	N
20	Material	It refers to the type of palm leaf used to write the information.	N
21	Missing Folios	It specifies the number of missing leaves in a single bundle.	N
22	Contributor	It specifies the actual donor of the manuscript.	Y
23	Identifier	It refers to the unique number which is system generated automatically.	Y
24	Keywords	It refers the accurate search information specified by the user.	Y

Table:2 Metadata Elements proposed in this research

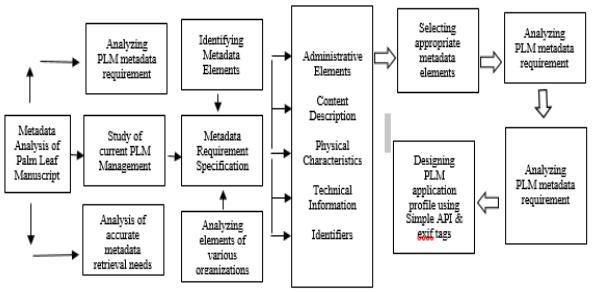


Figure:1 Proposed Metadata Schema

These core properties of this schema have been divided into various categories such as

1) Content Description – which consists of Title, Subject and other aspects.
2) Physical Characteristics – which consists of Word Count, In-boundary Damage and other aspects.
3) Administrative Elements: which consists of Contributor, Rights and other aspects.
4) Technical Information: which consists of format, date and other aspects.
5) Identifiers: Relation, Source and other aspects.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

There 3 elements of metadata that are proposed in this research which are as follows:

1) Edge Damage: It estimates the total damage percentage on the boundary of the palm leaf manuscript.
2) In-boundary Damage: It estimates the total damage percentage inside the manuscript body.
3) No: of words in an Image: It detects and counts the total number of words that exist in a manuscript.

C. Development of Well-Defined Access Mechanism

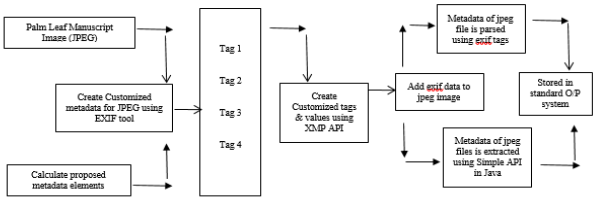


Figure: 2 Access Mechanism proposed in this research

The metadata which are embedded in the form of exif tags are extracted with the help of keyword extraction procedures. Out of many automatic keyword extraction approaches like Alchemy Keyword Extractor, Rapid Automatic Keyword Extraction (RAKE), crowdsourcing where the simplest approach is the Term frequency – Inverse document frequency (TF-TDF). This approach work efficiently in text/image documents as it finds the terms efficiently within that document/image.

Keyword Search using TF-IDF

This algorithm efficiently categorizes relevant metadata which are further utilized for efficient query retrieval. It calculates values for each metadata element in a manuscript image through an inverse proportion of the frequency of the word in a particular manuscript image to the total number of metadata elements. Each words have its respective TF and IDF scores. The product of TF and IDF scores of a term is called TF and IDF weight of that term. Generally, the elements with high TF-IDF values are denoted as strong bond with the manuscript image, which implies efficient search result. The formal approach for implementing TF-IDF works as follows:

$$W_d = f_{w.f} * \log (|F| / f_{w,F})$$

Where F is the file collection of manuscripts and an individual file € F.

$f_{w.f}$ equals the number of times w appeared in file f,

|F| is the size of manuscripts corpus

$f_{w,F}$ equals the number of files in which w appears in F.

Clearly, TF-IDF can find the words with the related metadata elements that are frequent and determine whether they are relevant to that document/image.

In short the TF-IDF algorithm is as follows:

Step:1 – The term count of the metadata element is the number of times the word appears in the collection which is denoted as follows:

$$Tf_{(i,j)} = \frac{n(i,j)}{\sum_k n(k,j)}$$

where n(i,j) is the number of occurrences of the element in the collection, denominator is the collection of the number of occurrences of all elements in the collection.

Step:2 – The inverse document frequency is measured as

$$Idf_i = \log \frac{|D|}{|\{j : t(i) \in d(j)\}|}$$

Step:3 – The tf-idf weight of the term t_i in a particular collection is denoted as the product of tf and idf which is shown as

$$(tf - idf)_{i,j} = tf_{i,j} * idf_i$$

Keyword Search using Crowd sourcing

Many traditional web-based systems provide metadata search and retrieval services through a user friendly search GUIs. This enables the users to find the relevant data via keywords and some other metadata information such as title, author etc. proposed in this research. The most important aspect is to correctly identify and extract the relevant metadata elements as well as some common relationships among them. The first challenge is to identify and extract each metadata element correctly, and secondly it is to identify the contents and other key phrases, thirdly is to extract the various semantic relationships between the metadata elements.

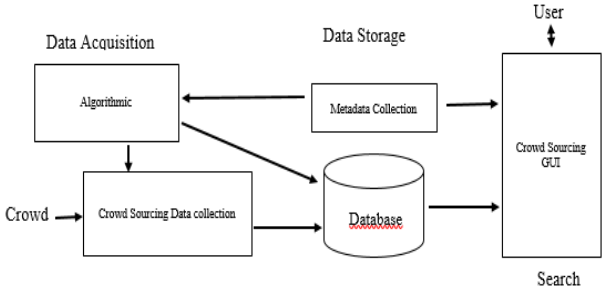


Figure:3 Crowdsourcing Architecture

The algorithm for the crowd sourcing proposed in this research is as follows:

Step:1 - The Manuscript Collection is denoted by M and its metadata elements as E.

Step:2 - All answered and unanswered pairs are listed with respect to manuscript collection.

Step:3 - When the Metadata elements (E>0) is greater than 0, then

- The similar elements are paired,
- Some queries are posted to the users to convey their opinion.
- Later the answers are collected from the crowd and the inferences are estimated.

Step:4 - The Resultant R is calculated against the manuscripts collection and its elements.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

IV. EVALUATION OF METADATA

The results of the metadata extracted using the proposed schema are evaluated according to the Precision, Recall & F-score in this research.

Precision: It is the amount of property values received that were relevant relative to the total number of property values retrieved overall. It is represented as

$$\text{Precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved doduments}\}|}{|\{\text{retrieved documents}\}|}$$

Recall: It is the proportion of relevant property values received to the total amount of relevant property values possible. It is represented as

$$\text{Recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved doduments}\}|}{|\{\text{relevant documents}\}|}$$

F-Measure: Generally, Precision and recall tend to be inversely related, can be used to combine them into a single measure. It is defined as

$$\text{F-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

S. No	Metadata Element	Precision (%)	Recall (%)	F-Score (%)
1	Title	97.19	98.00	97.59
2	Author	94.40	94.38	94.39
3	Subject	96.13	96.41	96.27
4	Description	98.11	98.68	98.40
5	Publisher	91.38	88.32	89.82
6	Language	94.75	93.20	93.97
7	Date	95.67	97.48	96.57
8	Type	92.39	79.12	85.24
9	Edge Damage Percentage	99.51	98.14	98.33
10	Source	89.42	87.14	88.27
11	Bundle Number	99.41	95.15	95.28
12	Accession Number	98.67	99.76	99.21
13	Coverage	97.19	96.35	96.77
14	Rights	94.30	95.32	94.81
15	Total Folios	96.13	97.14	96.63
16	Condition	98.11	97.68	97.89
17	Word Count	94.38	96.47	95.42
18	Script	94.75	93.37	94.06
19	In Boundary Damage Percentage	98.51	97.43	97.97
20	Material	89.42	96.87	93.14
21	Missing Folios	98.51	96.56	97.83
22	Contributor	89.56	92.45	91.00
23	Identifier	98.63	99.12	98.87
24	Keywords	94.75	93.56	94.15
Average		95.46	94.92	95.07

Table:3 Evaluation results of metadata elements extracted proposed schema

Method	Precision (%)	Recall (%)	F-Score (%)
Manual Metadata Extraction	82.22	76.96	79.34
Automatic Metadata Extraction	95.46	94.92	95.07

Table:4 Overall Performance of Metadata elements extracted using proposed schema

3.5.1 Extraction Evaluation of proposed metadata schema

The evaluation results of the metadata extracted are showcased and hence compared with the results of other existing metadata extraction tools.

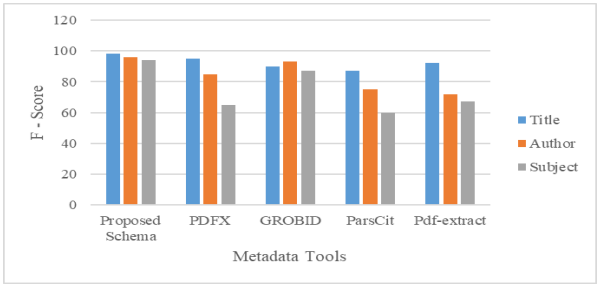


Figure:3 Evaluation Results of different metadata extraction tools

The proposed metadata schema and their respective resultant results are compared with the different existing metadata tools in order to assess the evaluate the performance of proposed metadata schema.

The evaluation consists of 24 metadata elements compared with 4 different metadata tools like PDFX, GROBID, ParsCit and Pdf-extract. It consists of different combination of extraction tasks like: extracting title of the manuscript image, keywords, author and many others.

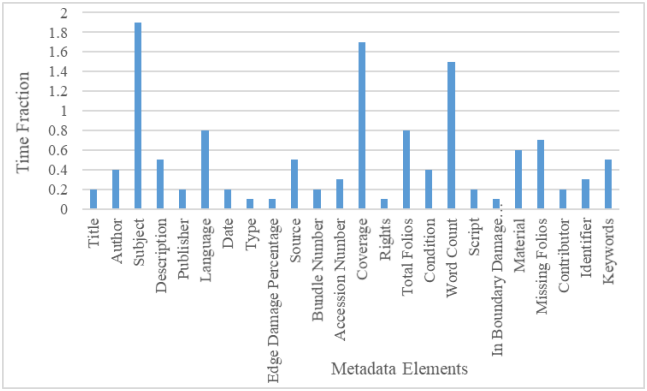


Figure:5 Percentage of time spent on each Metadata Element

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

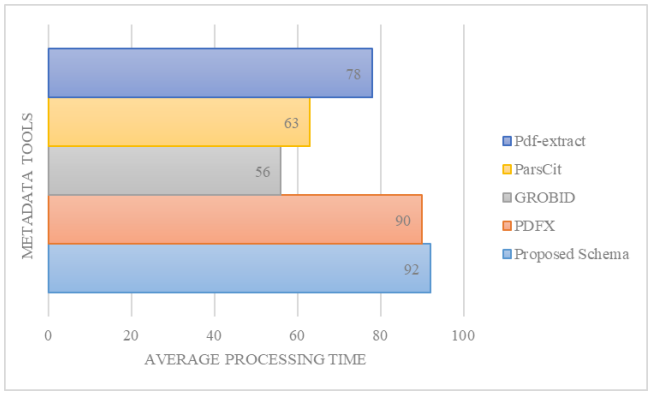


Figure:6 Average processing time of all metadata extraction Tools

V. CONCLUSION

The objective of this research work is to highlight the importance of automatic metadata extraction and its impact on the digitized palm leaf manuscripts and to provide an accurate schema using various metadata standards. The current metadata consists of 76 metadata elements making them complex in the practical environment.

The analysis of current work showed that scope exists to apply the proposed metadata extraction schema in the following ways:

1) The text retrieved after OCR implementation can be compared with the original manuscript which ensures accurate text content retrieval.
2) The proposed metadata elements in this research are used as an evaluation metric to verify the quality of original manuscript from further degradation which constitutes effective prioritization of the Digital Restoration process.

The main contributions of this research work include the following

- A detailed study on the existing manual metadata extraction techniques are presented.
- The current metadata elements are compared, and new metadata elements has been proposed with respect to Indian palm leaf manuscripts.
- Based on the proposed elements, a metadata schema consisting of 24 essential properties are designed in this research.
- The proposed elements can be used as an evaluation metric to verify the quality of the original manuscript from further degradation.

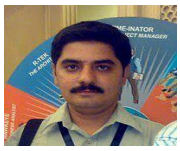
REFERENCES

1. G A Survey on the Application of Image Processing Techniques on Palm Leaf Manuscripts by Tulasi Krishna, VK Mehta, Prashant in International Journal of Advanced Engineering Research and Science (IJAERS), 2016.
2. Rapeeporn Chamchong, Chun Che Fung, Text Line Extraction Using Adaptive Partial Projection for Palm Leaf Manuscripts from Thailand, International Conference on Frontiers in Handwriting Recognition, DOI 10.1109/ICHFR.2012.280, 2012.
3. R Vasanth Kumar Mehta and Nagendra Panini Challa, "Facilitating Enhanced User Access Through Palm-Leaf Manuscript Digitization – Challenges and Solutions", Proceedings of IEEE International Conference on Electrical, Computer and Communication Technologies, February 2017.
4. A survey on Script and Language identification for Handwritten document images by Prasanthkumar P V , Midhun T P , Archana Kurian in IOSR Journal of Computer Engineering (IOSR-JCE), Volume 17, Issue 2, Mar – Apr. 2015.
5. Cross-Linking Between Journal Publications and Data Repositories by Sarah Callaghan, Jonathan Tedds, Rebecca Lawrence in International Journal of Digital Curation, 2014.
6. Metadata Development for Palm Leaf Manuscripts in Thailand by Lampang Manmart, Vilas Wuwongse in Proc. Int'l Conf. on Dublin Core and Metadata Applications 2012.
7. Digital Repository Best Practices for USCultural Heritage Organizatio ns by Katherine Kott on February 3, 2012.
8. Nagendra Panini Challa and R Vasanth Kumar Mehta, "Automatic Data Acquisition- A Major Challenge", Proceedings of IEEE International Conference on Electrical, Computer and Communication Technologies, February 2017.
9. Ontology Based Search Mechanism in Bilingual Database Resource by Norasykin Mohd Zaid, and Sim Kim Lau in The 11th International DSI and the 16th APDSI Joint Meeting, Taipei, Taiwan, July 12 – 16, 2011.
10. Rafael C. Gonzales, Richard E. Woods, "Digital Image Processing", Second Edition, Tata McGrawHill Education, 2010.
11. Nagendra Panini Challa and R.Vasanth Kumar Mehta, "Applications of Image Processing Techniques on Palm Leaf Manuscripts- A Survey", Proceedings of International Conference on - "Cognitive Science and Artificial Intelligence, Sree Vidya Niketan Engineering College, Tirupathi, July 2017
12. Image Segmentation of Historical Handwriting from Palm Leaf Manuscripts by Olarik Surinta and Rapeeporn Chamchong from Department of Management Information Systems and Computer Science Faculty of Informatics, Mahasarakham University Mahasarakham, Thailand, 2008.
13. Marcia L.Zeng and Jian Qin, "Metadata", 1st Edition, New York, 2008.
14. An Evaluative Study of Some Selected Libraries in India Undergoing the Process of Digitization by Anup Kumar Das, Jadavpur University, 2008.
15. Metadata Creation System for Mobile Images by Rista Sarvas, Errik Herratte in MobiSys'04, June 6-9, Boston, Massachusetts, USA, 2004.
16. R. Liu, W. Huang, and C. L. Tan. Extraction of vectorized graphical information from scientific chart images. In Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, volume 1, pages 521–525. IEEE, 2007
17. L. Lopez, J. Yu, C. Arighi, H. Huang, H. Shatkay, and C. Wu. An automatic system for extracting figures and captions in biomedical pdf documents. In Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on, pages 578–581. IEEE, 2011.
18. X. Lu, S. Kataria, W. J. Brouwer, J. Z. Wang, P. Mitra, and C. L. Giles. Automated analysis of images in documents for intelligent document search. IJDAR, 12(2):65–81, 2009.
19. M. Savva, N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, and J. Heer. Revision: Automated classification, analysis and redesign of chart images. In Proceedings of the 24th annual ACM symposium on User interface software and technology, pages 393–402. ACM, 2011.
20. V. Prasad, B. Siddiquie, J. Golbeck, and L. Davis. Classifying computer generated charts. In Content-Based Multimedia Indexing, 2007. CBMI '07. International Workshop on, pages 85 –92, June 2007.
21. Han, H., Giles, C.L., Manavoglu, E., Zha, H., and Zhang, Z. "Edward A. Fox: Automatic Document Metadata Extraction Using Support Vector Machines". JCDL 2003, p. 37-48, 2003.

AUTHORS PROFILE

Nagendra Panini Challa is an Assistant Professor from Department of CSE, SCSVMV. He has done his both M. Tech in Department of CSE from JNTU Kakinada and B.E from SCSVMV. His area of research includes digital image and video processing, computer vision related approaches.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)



Dr. R.Vasanth Kumar Mehta is currently working as an Associate Professor and Head, Department of CSE at SCSVMV. He obtained his Ph.D. in CSE from SCSVMV and M.Sc (Tech) and B.Sc in CSE from BITS Pilani His research interests are data mining, machine intelligence and image processing.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)