Informatics Studies. ISSN 2320 – 530X. Vol. 3, Issue 4,
Fourth Quarterly Issue. October – December, 2016. P 15-30

Language Technology and Digitization of Ancient Records in Indian Local Scripts

K. Rajendran and R. Raman Nair

Abstract

The paper presents the aims, Digital versions of recorded of knowledge are important cultural and economic resources. Stored in efficient and accessible digital systems, they can enable preservation and wider distribution of our knowledge heritage through World Wide Web. Digitized records can allow quick search for phrases, words, and combinations of words in any record if appropriate language technology is used. There are many projects currently active worldwide that attempts to put in electronic form ancient texts on different subject areas. Taking six digitization projects in Tamil and Malayalam as samples this study evaluates the status of digital libraries in Indian languages, and discusses their problems and possible solutions. The paper stresses the need for giving priority for the development of digital library packages that can process Indian local scripts to make digitization projects fruitful.

Keywords: Digital Archiving, Language Technology, Malayalam, Tamil, Indian Scripts

Recorded Knowledge

We have millions of records remaining in numerous media like stone, clay, parchments, leather, palm and other leaves, paper as well as in the new media like hard disks, CD ROM, audio and video tapes etc. Each individual document or record is a powerful force in itself. But its full potency can be achieved only if the content can be connected to the concerned in time and for application in the society.

Documents support the transfer of knowledge between generations. It makes it possible for sharing human experience down through time, and casts our vision of life forward into a future (Raman, 2008). Document collections forms the essential instrument for conserving and disseminating knowledge for the benefit of present and future generations. Thousands of clay tablets and writings in seals found in Indus Valley and Harappa indicate that people of India developed the art and science of recording knwoledge thousands of years before Christ and the system was even older than sumerian writing. The letters used the compostion of the tablets and seals and the perfection of the pictures they drew with unpolished materials they used as writing mehnisms are many times more than what we achieve with highly powerfull ICT based tools.

Before paper based printed books came each and every document recorded in traditional methods are unique and invaluable. For there will be no second copy and if the available copy is lost information contained in them is lost to the humanity forever. Printing

enabled production of numerous copies of the same document so that even if one or ten copies are lost hundreds of copies will remain at different locations. In the present day the major parts of recorded knowledge remain as printed text. Now digital documents also are increasing tremendously which can transcend the barriers of space and time.

Digital Media

Electronic versions of printed texts of ancient literatures and knowledge are important pedagogic and scholarly resources. Stored in easily accessible digital archives, they can permit preservation and wider distribution of knowledge around the globe through the means of World Wide Web (Raman and Jayapradeep, 2013). Digital versions of documents also allow quick search for phrases, words, and combinations of words denoting subjects, ideas or concepts in any document. There are many projects currently active worldwide that attempts to put in electronic form our ancient records.

Now digital technology has developed numerous formats and techniques, which can extend the life of the content of the document to infinity. If one format becomes obsolete it is now possible to transfer it to latest formats and media with ease. Documents in all formats can now be copied into digital media. The text, image, audio, video and animation in their different forms of traditional media can be transferred to digital media and digitized documents in different formats can be pooled together and if we could develop concerned language technology for languages covered; thousands of documents can be searched in a single stroke for a specific term or concept.

Efficient Management of Records

The new technologies help conservation of documents, organization of documents and easy, speedy and efficient retrieval of information from them. All these can be done in traditional formats using traditional systems. Even though Indian documents have some difficulties at present most of the digital archives of the foreign countries, have the power to search through the content of the documents pooled together and retrieve specific information. This is possible because each and every software used internationally have the power to process the text and languages used in most of the European countries, which we consider as international languages. But India has got more than 22 languages many times more script sets and any search mechanism to be established at the front-end of the digital collections of India will need a very efficient retrieval system that can process concerned Indian language or languages and scripts.

But so far no Indian language has developed reliable and efficient version of a package that can process all or any Indian language/s (Hussain, Et al, 2005). This paper deals with the relevant and specific questions related to the full text processing and retrieval in Tamil and Malayalam and development of digital library of Malayalam and Tamil Documents. It also covers packages available for developing digital libraries in Tamil and Malayalam with power to process the full text in Tamil and Malayalam languages and scripts.

Digitization of Traditional Collections

Digitization of document collections has different objectives like preservation, space saving easy organization and retrieval, speedy and efficient transfer of documents beyond time and space etc. Of this preservation is the most important purpose. Preservation is very important on numerous contexts especially for knowing about the developments in the past. In all sphere of human activity past is important. Hence history is important and these records preserve history. In writing history of a nation- political, social, cultural or economic if it is to be meaningful it needs to cover the history of each and every region. So writing

regional histories is a preliminary requirement for writing the country's history. Primary sources for regional history are documents that originated in the region, which will normally be in regional language (Sulochana, 1997). They can be recordings from temple or church, diaries of personalities, records of local administration, minutes of organizations etc. Each and every house used to preserve various types of documents. They are documents related to property, birth certificate of family members, ration cards, voters lists and testimonials of their education and employment photos and albums of the family members, diaries, writing of the people and correspondence between various members of the house as well as their communication to the people from out side. Public records have been created by social reformers, religious leaders, local political leaders military officials etc in the events in the region, their role in achieving reformers etc.

Local Records and Regional Studies

The earliest records such as those of the government offices like village office, Panchayat office, local offices of political parties etc. give details of the past administration of social, political and economic matters which are considered as a guide to be followed by present administrators. Based on these local history is built up: one best example is 'Vikasana Rekhakal' of Kerala which is historical documents covering the regions and all aspects under a Panchayath. This is a most important set of documents running to lakhs of pages and it is in regional language and if such documents are to be digitized and made useful, technology to process regional language is important.

Numerous search tools are available to locate appropriate sources and without these search tools, the chance of finding relevant information from the ocean of recorded information in digital collections or web is very slim. But for regional scripts none of

Figure 1: Writing in a Harappan Seal (200 BC)

the search tools are useful. Even with the help of search tools, users must be able to use sophisticated searching techniques and strategies of respective search tools in order to find relevant information. In the case for regional language generating inverted files and full text retrieval is still impossible and no reliable search mechanism are available for Indian regional language.

Context of the Present Evaluation

Libraries and information centers are attempting to transfer their collections from print to digital media. Information and Communication Technology is applied in most of the functions in the Libraries in order to perform house keeping operations as well as information storage and retrieval. India has an ocean of records in different local languages. The problems related to digitization in Indian languages require research and development of solutions to solve the problems. This can be done only in the regions itself for experts in the language and scripts and ICT professionals and programmers knowing the language and script will be available in concerned regions only. The ICT enabled services like digital libraries and archives provide efficient and maximum facilities to clientele. In this context the present study is intended to

evaluate digitization projects in Malayalam and Tamil and assess the present status and evaluate the aptness of technologies used.

Objectives of the Study

The main objective of the study is to Assess the usefulness of the digitization projects of India that consumes a substantial amount from public funds; which even after two decades fail to properly provide access to at least 5% of the digitized content. The objectives of this enquiry can be listed elaborately as follows:

- To examine the use of language technology for Tamil and Malayalam in digital libraries and archives.
- To evaluate the facilities available in digital libraries holding documents in local scripts.
- To assess the status of Informatics applications for documents in Indian languages and scripts.
- To suggest packages and standards for digital library development in Indian languages.

Local Script Related Projects

In Tamil and Malayalam there are numerous digital library projects for documents on their literature and culture as well as books on different subjects including science and technology in those languages. This paper covers six randomly selected digitization projects - three in Tamil and three in Malayalam. The Tamil Digital Libraries are; The Digital library of Tamil Virtual Academy, Virtual Library of Tamil Heritage Foundation and Project Madura. Malayalam Projects are Online Digital Library of Kerala Sahithya Akademi, Kerala Gazette Archives of State Central Library of Kerala and Mahatma Gandhi University Open Access PhD theses Digital Library. Of these only two projects have effectively applied Informatics for search and retrieval of knowledge through local scripts. Tamil and Malayalam Language are very closely connected and findings of most research on computational linguistics in regard to both these language can provide results applicable to both languages in specific and also to other Indian languages to some extent.

Tamil Language and Script

Tamil belongs to the southern branch of the Dravidian languages, a family of around 26 Indian languages. Tamil language family, which alongside Tamil proper includes the languages of about 35 ethno-linguistic groups such as the Irula and Yerukula languages. According to Chattampi Swami a scholar saint who lived in the second half of nineteenth century who has done some original research in linguistics it was in the continent that existed to the east of Ceylon where life originated and that the first language also evolved there which is Dravidian. In his theses named 'Adhibhasha' he established that all the languages originated from a mother language, which was Tamil. According to him it is the term Thaymozhi that later became Tamil.

Tamil is a of the classical languages of the world with a literary history of more than two millenniums spanning from the

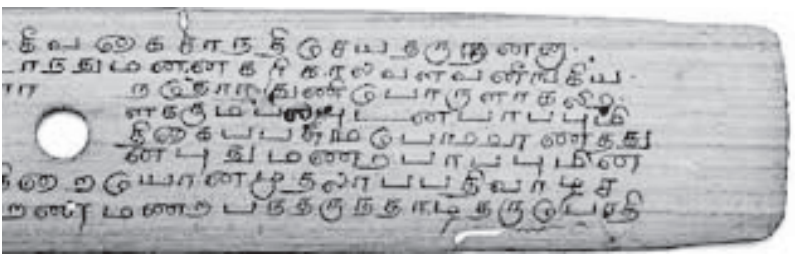


Figure- 2: An Ancient Tamil Work in Palm Leaf

Sangham age (300 BC – 200 AD). Traditionally, Tamil was written on palm leaves – one of the oldest medium of writing in ancient India. The precise origin and history of palm leaves writing are unknown but the practice is believed to have existed since the Sankam period. The use of palm leaf medium continued for several centuries until late twentieth century. A medium with such historicity has a negligible literature on its writing system, medium and the influence on script evolution. The existing literature on evolution of script only focuses on stone and metal inscriptions. Other traditional mediums such as palm leaf manuscripts have not been explored and researched (Udayakumar, 2010).

The closest major relative of Tamil is Malayalam; the two began diverging around the 9th century CE. Although many of the differences between Tamil and Malayalam demonstrate a pre-historic split of the western dialect, the process of separation into a distinct language, Malayalam, was not completed until sometime in the 13th or 14th century.

அகர முதல எழுத்தெல்லாம்
ஆதி பகவன் முதற்றே உலகு.

Figure- 2: Tamil Script in Print

In ancient days Tamil was written using a script called the vattezhuthu amongst others such as Grantha and Pallava script. The current Tamil script consists of 12 vowels, 18 consonants and one special character, the aytam. The vowels and consonants combine to form 216 compound characters, giving a total of 247 characters. All consonants have an inherent vowel a, as with other Indic scripts. This inherent vowel is removed by adding a title called a pulli, to the consonantal sign. Many Indic scripts have a similar sign, generically called virama, but the Tamil script is somewhat different in that it nearly always uses a visible pulli to indicate a dead consonant. In addition to the standard

Figure- 4: *A mythological Illustration from Tamil Magazine Cover*

characters, six characters taken from the Grantha script, which was used in the Tamil region to write Sanskrit, are sometimes used to represent sounds not native to Tamil, that is, words adopted from Sanskrit, Prakrit and other languages. All these sophistications make the standardization of fonts for information retrieval from digital content using Tamil difficult than in languages like English.

Introduction of Printing in India

In the late sixteenth century the introduction of printing technology gradually displaced the traditional writing system. In 1556, a printing press sent to Abyssinia for missionary works landed in India by an accident. Soon the Christian evangelists adapted the press for the conversion of natives. Initiatives of the missionaries led to the spread and establishment of printing. Progressively, the press became one of the major medium of communication and began to dominate the hand written manuscripts in the later centuries. The script on the other transformed itself with respect to the medium. The shift from handwritten

palm leaves to printing led to the transformation of letterforms. And also there has been an influence of the letterpress medium and western typography on the native script. This early transformation is critical for typographers and type designers to understand the script and its evolution from hand written letterforms to standardized letters seen today. This understanding can bridge the knowledge gap between the evolutional findings of archeologist, epigraphists, historians and the influence of other medium (Udayakumar, 2010).

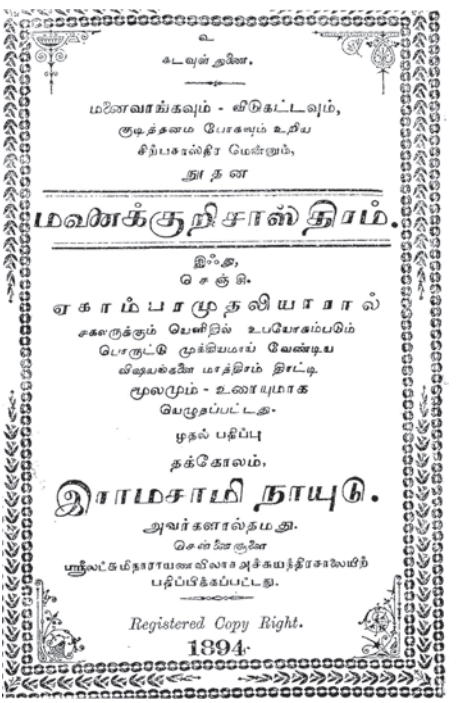
Tamil Virtual University (TVU) Digital Library

The Government of Tamil Nadu established Tamil Virtual University in 2001. The university provides internet-based educational resources and opportunities for the Tamil Diaspora as well as for others interested in learning the Tamil language and acquiring knowledge of the history, art, literature and culture of the Tamils. Tamil Virtual University offers different academic programmes from UG to PG as well as facility for research in Tamil. The digital library of TVU provides literature, glossaries and dictionaries. It accommodates literature starting from Sangam Era to the present day, with the features like: classified sections of books, ancient and medieval literature, with their commentaries, romanized versions of 'Tolkappiyam', 'Patthuppaattu', 'Ettutthogai' and similar classic texts etc. The library has subject-indexing and search facilities.

Tamil Heritage Foundation Digital Library

Tamil Heritage Foundation (TMHDL) is formed as a non-profit, non-political, non-governmental organization to serve Tamils around the world. It engages in the digital preservation of South Indian Heritage Materials. By web enabling the ancient South Indian scripts that were once read and understood only by Tamil experts and academic community in the higher levels of society, the archive now make resources available to all sections of Tamil society, International students of Tamil language and culture and to scholars in general. TMHDL has digitized palm leaf manuscripts that still remain unpublished from various Governmental and private libraries, museums and family collections. This includes manuscripts dealing with indigenous medicine especially, mathematics, astronomy, chemistry, engineering (ship building etc.,) architecture, philosophy, religion with stress Saivasiddhantha, literature, music and arts.

The major projects it plans to undertake are: data basing of various manuscript catalogues that are currently available in print form and bring-in new entries as well, acquisition of the manuscripts for digitization, digital preservation, critical reading and editing; publishing (electronic and print versions), developing methods of electronic


கடவுள் துணை.

மனைவாங்கவும் - விடுகட்டவும்,
குடித்தனம் போகவும் உரிய
சிற்பசாஸ்திர மென்றும்,
நூ தன

மனைக்குறிசாஸ்திரம்.

இஃது,
செஞ்சி.
ஏகாம்பரமுதலியாரால்
சகலருக்கும் பெரிதில் உபயோகம்படும்
பொருட்டு முக்கியமாய் வேண்டிய
விஷயங்களை மாத்திரம் திரட்டி
மூலமும் - உரையுமாக
வெழுதப்பட்டது.

முதல் பதிப்பு

தக்கோலம்,
ராமசாமி நாயுடு.
அவர்களால்தமது.
சென்னைரூல
ஸ்ரீலட்சுமிநாராயணவிலாசஅச்சுயந்திரசாலையிற்
பதிப்பிக்கப்பட்டது.

Registered Copy Right.
1894.

Figure- 5: *A Rare Book from Tamil Heritage Collections*

Figure- 6: An Old Image in Madurai Project

discussions across borders and generating worldwide interest that stimulate new discoveries, inventions, patents, audio-video documentation of traditional practices in the preparation and treatment of diseases, pharmaceutical evaluation of Indian herbs for drugs, developing research and development support for traditional healers, abstraction of scientific concepts from ancient manuscripts for experimentation and evaluation and developing indigenous curriculum. Tamil Heritage Foundation Library is an important initiative to preserve and understand Tamil heritage in a proper scientific way so that Tamil history, science, and technology is understood correctly, conserved and propagated.

Project Madurai

Project Madurai is an open and voluntary initiative to collect and publish free electronic editions of ancient Tamil literary classics. This means either typing-in or scanning old books and archiving the text in one of the most readily accessible formats for use on all popular computer platforms.

These e-texts are distributed in both web/html and PDF formats through the World Wide Web servers. Anyone located anywhere may download a copy for personal use or read what the project publish in the Internet, free of charge. Since its launch in 1998, Project Madurai e-texts are released in Tamil script form as per TSCII (Tamil Script Code for

Information Interchange) encoding. Since 2004 it is releasing e-texts in Tamil Unicode as well. The project has also collected numerous images (Figure- 6)

Malayalam Language and Script

Chattampi Swami and most linguists opine that Malayalam originated from Tamil (Chentamil) in the 6th century. Another view proposes a split in even more ancient times. Malayalam incorporated many elements from Sanskrit through the ages and today major portion of the vocabulary of Malayalam in scholarly usage is from Sanskrit. Before Malayalam came into being, Old Tamil was used in literature and courts of a region called Tamilakam, which included the present day Kerala also. An example is Tamil Classic 'Chilappatikaram' which was written by a Kerala prince Ilango Adigal from Cochin region.

Modern Malayalam still preserves many words from the ancient Tamil vocabulary of Sangam literature of the ancient Tamils. The earliest script used to write Malayalam was the Vattezhuttu script, and later the Kolezhuttu, which were derived from Tamil.

Figure- 8: *The First Printed Malayalam Book in Digital Collections*

അആഇഉഋഎഏഒകഖഗഘങ
ചഛജഝഞടഠഡഢണതഥദ

Figure- 7: Malayalam Script in Print

As Malayalam began to freely borrow words as well as the rules of grammar from Sanskrit. Grantha script was later adopted for writing and it came to be known as Arya Ezhuttu. This developed into the modern Malayalam script. The oldest literary work in Malayalam, distinct from the Tamil tradition, is dated between the 9th and 11th centuries. Due to its lineage deriving from both Tamil and Sanskrit, the Malayalam alphabet has the largest number of letters among the Indian language orthographies. The Malayalam script includes letters capable of representing almost all the sounds of Indo-Aryan and Dravidian languages. Historically, several scripts were used to write Malayalam. Among these scripts were Vattezhuthu, Kolezhuthu and Malayalam scripts. But it was the Grantha script, another Southern Brahmi variation, which gave rise to the modern Malayalam script. It is syllabic in the sense that the sequence of graphic elements means that syllables have to be read as units, though in this system the elements representing individual vowels and consonants are for the most part readily identifiable. In the 1960s Malayalam dispensed with many special letters representing less frequent conjunct consonants and combinations of the vowel /u/ with different consonants.

Malayalam script consists of a total of 578 characters. The script contains 52 letters including 16 vowels and 36 consonants, which forms 576 syllabic characters, and contains two additional diacritic characters named anusvara and visarga. The earlier style

of writing has been superseded by a new style in 1981. This new script reduces the different letters for typesetting from 900 to fewer than 90. This was mainly done to include Malayalam in the keyboards of typewriters but later it complicated things for computer manipulation. All these make processing of Malayalam for digital libraries and search and retrieval of full text content difficult. To solve the problems, in 2000 Rachana Akshara Vedi produced a set of free fonts containing the entire character repertoire of more than 900 glyphs, which was used in most of the successful digital library projects for Malayalam.

Kerala Sahithya Akademi Online Library

Kerala Sahitya Akademi was established for the promotion and development of Malayalam language and literature. The major activities of the Akademi consist of collection, organization, conservation, preservation and dissemination of documents including manuscripts, printed books, journals, audio and video records, digitized databases and other information sources. Akademi libraries hold a major portion of published works in Malayalam and such a reliable collection is not available elsewhere on the region as well as reordered in that language.

Akademi collection contains one lakh and thirty thousand books 1200 audio files and 500 video files as well as microfilm reels of 1500 old and rare Malayalam books. The library has also a unique collection of manuscripts pictures and paintings. Kerala Sahitya Akademi has developed a large digital library of which selected items are made available for access through the web. The Digital Library and information system of the Akademi already completed scanning and processing more than 12 lakh pages. At present Akademi has 6000 digitized books as well as the digital copies of Malayalam journals published from the last quarter of 19th century and the early 20th century. Examples of some of the journals in the

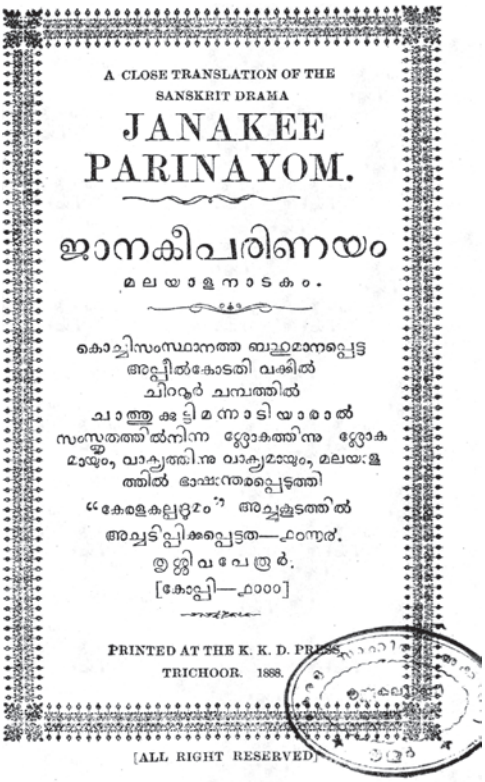

A CLOSE TRANSLATION OF THE
SANSKRIT DRAMA
JANAKEE
PARINAYOM.
ജാനകീപരിണയം
മലയാളനാടകം.
കൊച്ചിസംസ്ഥാനത്ത ബഹുമനപ്പെട്ട
അപ്പീൽകോടതി വക്കിൽ
ചിറയ്ക്കൽ ചമ്പത്തിൽ
ചാത്തുക്കുട്ടിമനാടിയാരാൽ
സംസ്കൃതത്തിൽനിന്ന ശ്ലോകത്തിനെ ശ്ലോക
മായും, വാക്യത്തിനെ വാക്യമായും, മലയാള
ത്തിൽ ഭാഷാന്തരപ്പെടുത്തി
"കേരളകല്പദ്രുമം" അച്ചുകൂടത്തിൽ
അച്ചടിപ്പിക്കപ്പെട്ടത—ഫാന്നൂർ.
ഇംഗ്ലിഷ പേത്രർ.
[കോപ്പി—ഫാനാ]
PRINTED AT THE K. K. D. PRESS,
TRICHOOR. 1888.
[ALL RIGHT RESERVED]

Figure- 9: An old book from Kerala Sahithya Akademi Collection

digital collections are Mangalodayam, volumes published from the year 1877, Vidya Vinodhini volumes published from the year 1889 and Bashaposhini volumes published from the year 1893. The collection includes almost full set of most of the old Malayalam journals published for short periods in the last quarter of 19th century and the first half of 20th century.

The oldest books in the Digital Library of the Akademi include 'Samkshepa-vethartham', the first Malayalam Book published in 1772 (Figure-8), 'Cheru Paithangalkku Upakarartham: Englishil Ninnum Paribhashappeduthiya Kathakal' the first Malayalam text printed in Kerala in 1824, 'Seethankal Thullal' by Kunjan Nambiar published in 1853, 'Ambareeshacharitham' published in 1897 etc. The collection includes most of the

SELECT DECISIONS

OF THE

SADR COURT OF TRAVANCORE.

Figure- 11: An Old Travancore Government Document

books published in 18th and 19th centuries. Manuscripts of eminent Malayalam writers, speeches of eminent personalities, photos, pictures etc are also included in the digital collections of the Akademi. These digital documents are available to the researchers from the line digital library at the Akademi.

Akademi provide online access to more than 1000 copyright free books, which are already uploaded in the web. Users through the Digital Library of India (DLI) can also down load it. The meta-data was also prepared and published in the web. At present Akademi is the only contributor of Malayalam books to Digital Library of India and its name is listed in the home page of DLI. From the links available with retrieved metadata any one can easily download the print version of the document they want from the Akademi site. Akademi is trying to collect the digital versions of almost all known titles in Malayalam published between 1772 to 1900

Digital Archive of the Writers of the By gone Era

Kerala Sahitya Akademi has also developed a digital archive of 'Malayalam Sahitya Parambara', an interactive Multimedia on writers of the by gone era, which is distributed as a CD Rom publication. It is an archive of Malayalam Authors and can be considered a companion project of Akademis' Online Library. Most of the prominent Malayalam authors of by gone ea are covered in this database. It contains files related 255 eminent Malayalam writers of the last two centuries. It contains photos, speeches, biographical sketches, handwritings, list of books written by the writer, cover page of the major works and other data. This database in the digital format has the capability to search the writers through alphabetized lists in Malayalam and English as well as their year of birth or period of life. This database is also available to the public through Kerala Sahitya Akademi website.

Travancore, Travancore-Cochin, Kerala Gazettes

It is a project of the Kerala State Central Library. The State Central Library is one of the oldest Libraries in India. It was established in the year 1829 A.D during the reign of His Highness Sree Swathi Thirunal Maharaja of Travancore with technical support from Col. Edward Cadogan, the then British Resident who was the grand son of Sir Hans Sloan, the founder of the British Museum.

Travancore, Travancore-Cochin, Kerala Gazettes online developed by the Library is a collection of major government documents in Malayalam, Tamil, English and Kannada created since 1840 consisting of Administrative reports, Government Notifications, Parliamentary debates, Bills, Acts, Statistical Reports, Budget documents, Committee and Commission reports etc and covering areas like educational development,

Figure- 10: *An illustration by Nambuthri from a Journal Digitized at Kerala Sahithya Akademi*

populations control measures, health management, land distribution, river distribution etc. These important Government Documents were in a state of neglected condition. They are in Indian languages and consisted mainly of documents in Tamil, Malayalam, Kannada and English. Gazette Digital Library contains more than five lakhs scanned pages of gazette with index correlated to the content with details such as person, institution, house, place, year, order number etc. In this archive proper query formulation can catch relevant pages of full text, which can be viewed directly on the computer screen to locate and ascertain the required information. This digitization project is an inspiring example of how people's need for a valuable piece of information connected to their personal and social life can be accomplished in minutes, which in manual search may extend to days and weeks. Gazette Archive can be a productivity tool to government departments also with immense possibilities. The Kerala Gazette Archive developed using and indigenous package compliant Unicode and Malayalam Language Technology is known to have faced many language related issues in its development, which were not earlier encountered by any information systems in Malayalam for no earlier project attempted search mechanisms that can use regional script. Successful creation and web launching of digital library of gazettes contributes significantly to the methodology of making information systems in Malayalam. It is a model and forerunner for all information systems that will come up in Malayalam and Tamil in future.

Digital Archives of MGU Dissertations

Open Access Digital Archives of Doctoral Dissertations launched by Mahatma Gandhi University (MGU) in 2008 consisting of more than 2000 dissertation in Malayalam and other languages is a notable venture using Indian language technology. MGU was established in 1983. It has to cater to the higher education of Central Kerala and has

seven faculties for Science, Fine Arts, Commerce, Engineering and Technology, Technology and Applied Science, Aired, and Homeopathy. MGU Online Digital Archives of Dissertations applies a special archiving package for hosting dissertations in the web. The package has multilingual search facility and satisfies UNICODE standards. The search mechanism is based on Nitya Digital Archive (dArch). The specificity in search and retrieval offered is commendable. MGU Online Dissertations Archives is the only bibliographic and full text information system presently existing in Kerala having multilingual search capability using English, Malayalam, Tamil and Hindi scripts. A visual keyboard for Malayalam, Tamil etc is provided to construct queries for those unfamiliar with Inscript keying. A sample search interface for local scripts in mgutheses.org is shown in Figure- 13 and a Malayalam dissertation retrieved from the archive is shown in Figure-14.

Images that Can Conserve Scenes of the Social Life

One of the interesting aspects that came to the notice of the author during the examination of digitized collections of Tamil and Malayalam documents in the sample projects is the very valuable periodical collections from the second half of nineteenth century to present in those projects. These periodical collections as well as many books contain illustrations of importance to researchers on this regions people and culture. The documents mainly the local journals from the region contain millions of illustrations made by renowned artists for the stories novels and other articles as well as for cover pages of the magazines. The cover and story illustrations in Tamil magazine Kalki (Figure –3) will be of lasting interest to general public, art students as well as researchers on culture. Some artists who have done illustration during the last hundred years like Nambuthiri, Karunakaran, AS etc can stand with similar artists found any where in the world or who have done such work to magazines and journal of international repute.

These illustrations can help to trace out the evolution of the peoples physical features, dress, customs, architecture, culture and many other aspects. No special care has been given for their digitization. Example is the illustration by the famous artist Namputhiri in one of the magazines digitized during the process is shown above (Figure -9). At the same time author has tried to digitize one illustration from a magazine by the same artist with special care which resulted in a unique archival image sample that depicts dress features and expressions of the people of a specific time and region in Kerala which will be of great interest to future researchers. The illustrations like the one given here (Figure -11) requires a little special attention as part of the book or as separate entity while digitizing such documents. People involved need to have devotion and love for such work if we are to preserve such heritage items for posterity.

Language Technology for Tamil and Malayalam

The initial study of the six digitization project of Tamil and Malayalam as well as a casual scanning of similar resources available in the web revealed that, of these only two projects have provision to search and retrieve information using Indian scripts like Tamil and Malayalam. All other projects use only hyper linking to document files from a front-end database or list. Metadata in front end is also in English Transliteration. Even in the projects in which search through Tamil and Malayalam script is possible the indexing, and navigation through full text are done manually which will be an impossible task when we consider the huge quantum of records remaining to be digitized in India.

The status of the present Digital Library initiatives from India and the study of the sample projects from Tamil and Malayalam

Figure- 12: An Illustration by Nambuthiri for a Malayalam Story

reveal that Government and ICT organizations have to give priority to development of packages and solutions that can solve problems specific to India. India needs digital library solutions that can process Indian languages and scripts. This is to be a priority item if conservation and use of our knowledge heritage is to become possible.

India also needs Optical Character Recognition (OCR) for Indian languages including Tamil and Malayalam that can convert document images into text so that they become searchable through networks or web using search engines. Unless we achieve, the required language processing technologies, capabilities all digitization programmes will end as huge unusable collections of digital images. Even though character recognition is possible with existing packages for all the International languages the recognition of characters of Indian and other oriental languages is not possible (Sharada, 2005). The diversity and complexity of documents archived in Indian digital libraries complicate the problem of document analysis and understanding. This necessitates urgent attention to developing programmes that can make the valuable archived records searchable and usable in

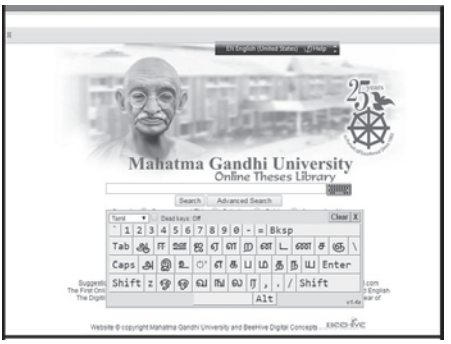


Figure-13: Search Interface for Tamil in Nitya D Arch Software

order to achieve the objectives of costly digitization. Currently research is going on in similar direction to develop language technology for the computers which can make access to digital documents imminent through automatic indexing, inverted file generation and retrieval by content specification. Of the projects Evaluated for this study only Kerala Gazettes Archive and Mahatma Gandhi University Digital Archives of Doctoral Dissertations only have used a digital library packages, which can process Tamil and Malayalam and other South Indian Scripts. The study has made sample searches for Tamil and Malayalam in these sites and found that the packages used in these two projects can customized for all projects in these languages.

The packages used in these two projects being Unicode compliant it is assumed that they can be used for any Indian language. Beehive Digital Concepts, Cochin, developed the base package Nitya Archive on which packages of these two projects are based, in close association with Center for Informatics Research and Development (CIRD). This special archiving package can be customized for hosting any Indian language document repository in the web. The package has multilingual search facility and satisfies UNICODE standards.

The search mechanism of Nitya Digital Archive (dArch) is unique in Indian context when its capability to process Indian scripts is considered. The specificity in search and retrieval offered is commendable and its metadata can be made OAI-PMH (Open Access initiative – Protocol for Metadata Harvest) compliant. The important features of the package are that it can be used to host any number of dissertations in English, Malayalam, Hindi, Sanskrit, Tamil and Kannada, has facility for multi keyword search using advanced Boolean search for accurate and relevant search, facilitates multilingual data input and Boolean Search, and is compatible with any browser on any operating system. A customized version of the same has been used in Gazette Archives of State Central Library. A search through the net reveals that it is the only package with search mechanism compliant to Tamil and Malayalam presently used for digital library development in India.

Character Recognition of Indian Languages

Most OCR systems perform with high accuracy for English in presence of printing variations and document degradation. For Indian and many other oriental languages, OCR systems are not yet able to successfully recognize printed document images of varying scripts, quality, size, style and font. Compared to European languages, Indian languages pose many additional challenges. Some of them have Large number of vowels,

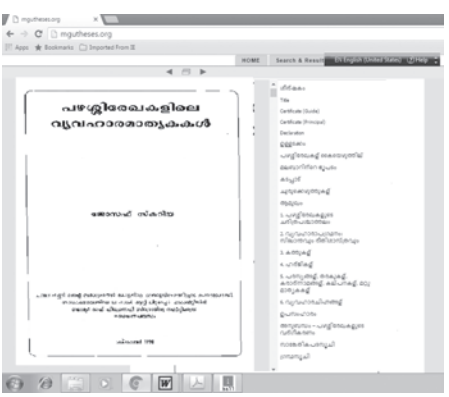


Figure- 14: A Malayalam Thesis Retrieved from mgutheses.org

consonants, and conjuncts (Sesh Kumar, 2006). Most scripts spread over several zones, inflectional in nature and having complex character grapheme, lack of statistical analysis of most popular fonts and/or databases, lack of standard test databases of the Indian languages, lack of standard representation for the fonts and encoding, lack of support from operating systems, browsers and keyboard, and lack of language processing routines, add to the complexity of the design and implementation of a document image retrieval system (Sachin, 2006).

General Observations

Most of the Digital Libraries in the web are not designed with even minimum aesthetic sense. When compared to their foreign equivalents these sites look like trails and attempts made by novices in website designing as part of their training. Home pages and facilities of MGU theses archive and Travancore Gazette collections are the only exceptions. The functionality of archives available in Tamil are very poor, but have rich content.

Digitized documents in the sample projects are from diverse languages and hence vary in scripts. They are extremely poor in image quality. They are also varying in scripts, fonts, sizes and styles. The pdf pages or images are not even properly trimmed. More than that inclusion of watermarks without any aesthetic sense distracts the use in most of the samples checked for this study. Accessing from these complex collections of document images is a challenging task. Most of the sites are very slow and it takes hours to get a document open or get downloaded.

Even though much fund has been availed and claims have been put forward on completion of digitization of large collections; even 5 % of the claimed completed documents are available for public access in the web. Even though it was not one of the samples of this study; a quick evaluation conducted at the National Library Digital Collection available on the web showed that against the substantial amount of expenditure for digitization there even 1% percent of result for expenditure in terms of digital documents accessible to public has been achieved. The site compared to MGU Theses Archive, Kerala Gazette Archive or Tamil Virtual University Archive presents a very poor face in regard to technology used and content. Hence all these samples from local projects deserves appreciation for they are achieved with minimum cost and technical support.

Priorities for Digitization Initiatives

There is large-scale digitization of documents. Most of the digitization projects generate oceans of digital documents from which retrieving a specific items will be more difficult than in their traditional counterparts. The search through the sample digital libraries and other Indian projects as mentioned above show that the quality of the full text files are highly inferior and this can lead to a dangerous situation if we require perfect files in future or scaling up becomes necessary.

The Tamil and Malayalam digitization projects can use facilities like Digitize India Platform initiated by the Government of India to provide digitization services for scanned document images or physical documents which can enable ensuring quality of digitized files as well as development of language technology and their application.

The Digitization projects covered here require OCR packages in Tamil, Malayalam etc to automatically generate indexes from the digitized pages for easy information retrieval. Already a few OCR programs have been developed by organizations like C-Dac under Government of India and private groups like ind.senz. It is seen that none of the packages were used in Tamil or Malayalam digitization programmes. These existing OCR packages need to be tested in real life situations. Except in MGU theses Archive

and Travancore Gazetteer Project there is no search interface for Tamil or Malayalam Scripts. It is a shame to note that Even though crores of rupees is spent annually for digitization projects in organizations from national Library to regional level institutions we have failed to develop even simple information retrieval mechanisms for the Indian language documents which makes the resources spent not useful to society. In addition even 0.1% of the documents digitized in the Indian institutions including those in Tamil and Malayalam in total are accessible to public. The quality of what is accessible is very poor. So we can imagine what will be the condition of those, which are not accessible but claimed to have been digitized. As the digitization projects of last two decades have failed to give good results and as required technologies are yet to be developed an urgent technical audit or a combined financial and technical audit needs to be undertaken at regional and national levels before further supporting of any major digitization project.

References

Balakrishnan, N (2006). Digital Library of India: A Test bed for Indian Language Research. TCDL Bulletin, 3 (1).

Chattampi Swami (2005). Adhibhasha. Ed by K Maheswaran Nair. Calicut, Mathrubhumi.

Chattampi Swami (2011). Tamizhakavum Dravida Mahatmyavum. Compiled with a Study by Suresh Madav. Quilon, Panmana Ashram.

Govi, K M (1970). Malayalam Granthasooji. Thrissur, Kerala Sahithya Akademi, 1970.

Kerala Sahithya Akademi (2015). http://www.keralasahityaakademi.org/online_library/index.html. Accessed on 25.12.2015.

Kerala, State Central Library (2015). Shttp://statelibrary.kerala.gov.in/gazette/index.php. Accessed on 21.12.2015.

Project Madhurai (2015). http://www.projectmadurai.org/. Accessed on 08.01.2015.

Raman Nair, R (1998). Development of an Information System for Regional Literature and Culture with INFLIBNET Support. *CALIBER*. Nagpur University, 1998.

Raman Nair, R (2002). Digital Archive on Kerala Literature and Culture: Report on digitization of the rare and antique collections at the Kerala Sahitya Akademy.Thrissur, 2002.

Raman Nair, R (2008). Application of Information and Communication Technology for Conservation of Cultural Heritage of Thalasseri. National Seminar on Heritage Tourism in Thalassery. Thalassery, 20th March 2008.

Raman Nair, R (2011). Information Systems in Malayalam. SRELS, Journal of Information Management Vol 48, No. 4.

Sachin Rawat, K. S. Sesh Kumar, MM Et al (2006). Adaptive OCR for Digital Libraries. In: Proc. of 7th IAPR Workshop on Document Analysis Systems (DAS), Nelson, NewZealand, (LNCS 3872), 2006.

Sesh Kumar, K.S, Anoop M Nambudiri and C V Jawahar (2006). Learning Segmentation of Documents with Complex Scripts. In: 5th Indian Conference on Computer Vision, Graphics and Image Processing, Madurai, India LNCS, 2006.

Sharada, BA (2005). Digitizing Documents. In: Linguistics and Indian Languages: Experiences At The Central Institute Of Indian Languages, Mysore. Information Studies, 11 (2).

Tamil Heritage Foundation (2015). http://www.tamilheritage.org/. Accessed on 31.12.2015.

Tamil Virtual University (2015). http://www.tamilvu.org/. Accessed on 05.01.2015.

Udayakumar, D(2010). Transformation of Tamil letterforms from palm leaf manuscripts to early letterpress printing. Theses. Bombay, IIT.