

Draft 3: Revised on 15 August 01

## **Preserving Telugu Language Heritage via Digitization**

### **A White paper**

NOTE: Please send your input, via E-mail, with copies to both [vemuri@attbi.com](mailto:vemuri@attbi.com) and [srao@shanthabiotech.com](mailto:srao@shanthabiotech.com) for prompt action.

#### **GOAL**

The vision of this project is to collect, compile, digitize, catalog, preserve and publish via digital media all the available Telugu language works. Wherever suitable, the scope of the effort shall also include oral histories and folklore – memories, thoughts and motivations behind people’s actions, the details of day-to-day life and perspectives often marginalized from the mainstream historical record using audio and video media as well.

The scope of initiatives like this can range from the digitization of specific collections of materials to elaborate state, national or international organizations devoted to developing innovative technologies for providing a range of information services. Both the modest and the large-scale versions aim at high-quality digital content – and its selection, design, management, preservation, and user-friendly tools to ensure its widespread availability and usage.

A modest short-term goal of this effort is to focus on a narrower aspect of the vision as a confidence-building pilot project to demonstrate feasibility and flush out implementation problems.

#### **PREAMBLE**

Of the more than 6,000 languages currently being spoken, fewer than half are likely to survive the next century. When a language is gone, it is gone forever ([sapir.ling.yale.edu/~elf/](http://sapir.ling.yale.edu/~elf/)) – and along with it a culture. A language may fall out of favor and may even become extinct, but there is no reason to lose that literary and cultural heritage forever. In this age, we have an unprecedented opportunity to digitize the available literature and preserve it suitably for posterity. This is a challenging and formidable task. However, the successful competition of such a project for Chinese literature under the leadership of Mr. Gabriel Yu is a most inspiring one ([www.skqs.com/](http://www.skqs.com/)). Among other efforts of this kind, the one to preserve Maori culture in New Zealand (*Communications of the ACM*, May 2001) and Project Gutenberg

( <http://sailor.gutenberg.org>) quickly comes to mind. We wish to take inspiration from these efforts and accomplish the same for our Telugu literary heritage.

## **TELUGU LANGUAGE**

Telugu is one of the southern Indian languages ([www.sil.org/ethnologue/countries/](http://www.sil.org/ethnologue/countries/)) and is the official language of the state of Andhra Pradesh (AP), India. It is spoken by approximately 76 million people ([www.censusindia.net/profiles/apd.html](http://www.censusindia.net/profiles/apd.html)), giving it a rank of 15<sup>th</sup> or higher among the world's languages ([www.sil.org/ethnologue/top100.html](http://www.sil.org/ethnologue/top100.html)). Its written records date back to the 6<sup>th</sup> century AD and has a rich literature since 1120 AD ([members.iquest.net/~vepachedu/Nannayya.htm](http://members.iquest.net/~vepachedu/Nannayya.htm)). However, with pressures from the use of Hindi as a national language and English as a global language, chinks are beginning to appear on the Telugu fabric and concerns are being expressed about the functional survival of Telugu.

## **EARLY EFFORTS AT PRESERVING TELUGU LITERARY HERITAGE**

For centuries, much of the Indian literature, including Telugu literature, had been preserved and transmitted through the oral tradition. On rare occasions some literary works were preserved as inscriptions on copper plates, but by and large the bulk of the early Telugu literary output was preserved as palm-leaf manuscripts. Many of these palm-leaf manuscripts did not even appear in print, however, there are several Telugu literary works available in this form at a site that is currently under construction ([xlweb.com/heritage/asian/manuscri.htm#ind](http://xlweb.com/heritage/asian/manuscri.htm#ind)). Charles P. Brown (1798-1884), a British civil service officer, took the job of collecting palm-leaf manuscripts and editing them. He left behind a vast treasure of Telugu literary/historical documents. Innumerable works are in the Government Oriental Manuscripts Library, Madras, India Office Library, London, and at Sri Venkateswara University, Tirupati, A. P., India. It is in the interest of Telugu people to have these materials well preserved and researched not only because of their literary value, but also because they provide a valuable window into Telugu history and social life (<http://www.engr.mun.ca/~adluri/telugu/modern/people/cpbrown1.html>).

Subsequent to Brown, who began collecting Telugu literary works and printing them on paper using the available technology of the 1800's, several standard works in Telugu have been collected and printed by different groups. Several rare copies of such printed books are still available in various libraries in India and abroad; several more are falling victim to the ravages of time and disasters like fires and floods. At the Government Archives in Tarnaka, Hyderabad, there are several administrative records of historical significance, as well as Gurazada Apparao's diaries as well as old Telugu (Late 1800s, early 1900s)

magazines that were transferred to Andhra Pradesh from Tamilnadu. Lists of several collections of Telugu books are available on the web and in print in Andhra Pradesh. A brief historical account, with milestones and personalities of Telugu literature, is available at [www.infovani.com/english/samskruthi/as\\_sahityacharitra.asp](http://www.infovani.com/english/samskruthi/as_sahityacharitra.asp).

## **BENEFITS ACCRUED BY THE PROPOSED PROJECT**

First, the collection, cataloging, and indexing of literary material that is available in all forms (palm-leaf manuscripts, printed books, oral versions, and so on) will, in itself, be a great service for the preservation of Telugu heritage and literature. Standardization and cataloging will generate further interest and research activity. Digitization, with the capability to search and the ability to convert into different existing and upcoming formats, will act as an incubator for several new software projects. On the dream frontier, digitization can bring together capabilities to capture oral renditions, a perhaps-yet-to-be-developed Telugu sign language, display material in various calligraphies, hypermedia links, and so on. This project, in short, not only preserves our literary and linguistic heritage but also stimulates and provides a special place for Telugu literature in the world arena. A product such as this, and services that grow out of the availability of such a product, will open up new commercial opportunities.

Second, this project will help us to identify, articulate and pursue technological, textual, literary and historical issues uncovered during the digitization process, and will serve serves as a liaison with other organizations carrying out such research. In short, the thrust of this effort is not just to create a museum for a dying language, but also to create a living resource for the growth and thriving of a culture. Finally, this project may serve as a model for others to follow.

## **RELATED EFFORTS**

1. The Digital South Asia Library Project (DSAL), at the University of Chicago, which is currently in its first stages, will produce the online digitization of South Asian documents and materials. The DSAL Project is funded by a Title VI Grant awarded to CRL by the Department of Education for \$540,000 (Rebecca Moore: [rebecca@thyme.uchicago.edu](mailto:rebecca@thyme.uchicago.edu)), (<http://www.crl.uchicago.edu/info/focus/mayjune00.htm>) and (<http://www.lib.uchicago.edu/e/su/southasia/dsal2.html>).

2. A project to preserve palm-leaf manuscripts in Asia is actively pursuing the digitization of literature for different languages (<http://xlweb.com/heritage/asian/palmleaf.htm>).
3. There are several university-based efforts in the U. S. directed toward creating digital dictionaries for South Asian languages ([www.lib.uchicago.edu/e/su/southasia/digital-dict.pdf](http://www.lib.uchicago.edu/e/su/southasia/digital-dict.pdf)).
4. Eco Foundation, a non-profit organization based in California, is in the process of creating a web-based Telugu-English and English-Telugu digital dictionary with built in capabilities for word search and spell checking (<http://www.ecofoundation.org>).
5. Andhra Pradesh government, under the guidance of Potturi Venkateswara Rao, is sponsoring a project to digitize (via image scanning) Telugu newspapers.
6. National Digital Libraries Project: According to a National Informatics Policy Submission Paper, prepared by Mr. Randeep Sudan, “there is an immediate need for giving an impetus to this area. Digitization of educational content should be taken up on a large scale in a co-ordinated manner. A national level task force should coordinate such digitization among states in order to avoid duplication and in order to ensure that relevant literature can be digitized in the fastest possible manner.”

## **PUTTING BOUNDS ON THE SIZE OF THE EFFORT**

Admittedly, the vision being articulated here is an ambitious one. To bring a semblance of feasibility to this vision, it becomes necessary to define some boundaries and develop an action plan for its implementation. There are several ways to bound this project.

(1) For a comparison, Gabriel Yu’s digitization project handled 4.7 million pages or 800 million Chinese characters. In 1772, at the behest of Emperor Qianlong, an army of scholars, scribes, and clerks catalogued 3,460 works under four categories. This work came to be known as Siku Quanshu – a complete library in four branches of literature. The works themselves were not digitized, but only their summaries and author biographies. Everything was input via keyboard using Unicode character set amounting to 110 Giga bytes of computer storage. An equivalent task for the proposed project would probably correspond to the digitization of Kasinathuni Nageswara

Rao's "Amdhra vAj~maya sUcika" (Prachee publications). This probably can be done at a fraction of Gabriel Yu's project cost.

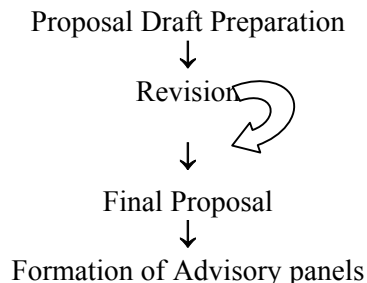
(2) Another way to confine the scope of this work is to focus, first, on preserving the content of the palm-leaf manuscripts that have not been printed. As the task of reading a palm-leaf manuscript is rapidly becoming a vanishing art, the process of inputting the characters via the keyboard will be laborious, error prone and expensive. Optical character recognition (OCR) technology is not robust and mature enough to recognize handwritten Telugu characters on palm-leaves. As the maximum life-span of a well maintained palm-leaf manuscript is no more than 400 years, and many manuscripts have reached or surpassed this limit, there is some merit in preserving the images of the pages themselves, rather than digitization. Such a project can easily be undertaken with the help of (a) a PC attached to a digital camera or (b) a PC attached to a scanner and distribute the end product on CD ROM's with a reasonable probability of recovering the costs.

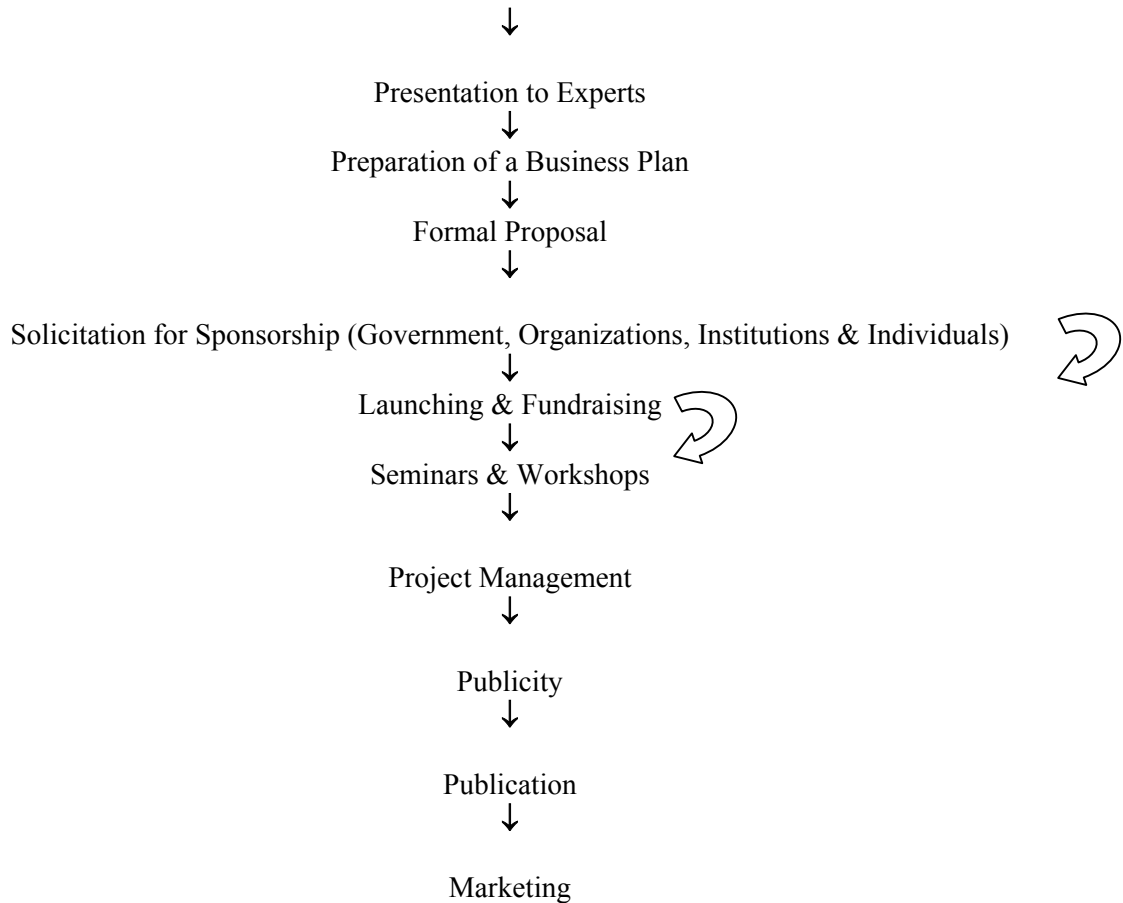
## **PROJECT PLAN**

Implementation of this project – in small or large-scale - requires a well-qualified team of experts from various fields. Since future generations will depend on this resource, accuracy and forward compatibility with future technologies will be the foremost objectives. Four major resource categories have been identified for a successful implementation of this project:

1. Human Resources
2. Literary resources
3. Technology resources
4. Financial Resources

A flow chart and several aspects of the project are presented below. This is followed a brief description of the resources required.





## 1. Human Resources

### 1.1. Key Players

A team of key players, drawn from all over the world, will serve on an advisory panel to provide guidance on management, financial, marketing, technology and literary issues. The government of India's Culture and Language Departments, the Andhra Pradesh government, all universities in and outside of India with an interest in Telugu (numbering, approximately 20), all institutions interested in Telugu language and people (numbering, approximately 50), voluntary organizations and individuals who are already making token contributions, Telugu associations around the world (estimated at about 150), and philanthropists are being contacted for necessary inputs and to identify the key players of this project.

### 1.2. Organizational Structure

A full-time staff will coordinate several teams (Management, Financial, Legal, Documentation and Archiving, Literary, Information Technology, Library Science, Quality Control, Liaison and Publicity).

One main coordinator will be responsible for all information flow and scheduling meetings and communications. There will be scope to integrate the efforts of various voluntary activities.

### 1.3. Legal Counsel

A team of legal experts will be requested to advise from time to time on all legal matters, including copyright and intellectual property issues.

### 1.4. Publicity

The publicity will be primarily through electronic media including a well designed website.

### 1.5. Collaborations

All universities ([www.andhraachuki.com/universities.htm](http://www.andhraachuki.com/universities.htm)) with Telugu departments and institutions with technical expertise in areas in which the project needs assistance will be approached for participation. The faculty and students at these institutions will be recruited for the project work at various stages. Several graduate theses and publications will be expected as spin-offs of this activity.

Adequate scope will be provided to recruit the help of amateur and non-professionals in some aspects of the effort, somewhat analogous to the way the SETI project is harnessing the idle time of thousands of home computers in its hunt for extra-terrestrial life. As another example, Dr. Kalyana Sundaram and others have built an organization of volunteers worldwide to input Tamil works. They have first standardized on a 8-bit mixed (Tamil & English ) character set. They post the progress of input, proofing, and publishing of Tamil works on a website.

## **2. Literary (Content) Resources**

### 2.1. Material Identification and Collection

A viable material collection plan will be developed via workshops organized for this purpose, both in India, and in the U. S. Necessary procedures to acquire material will be identified. All information will be made available to the other teams for proper coordination. Collection of all works, cataloguing and storage will be taken up according to a generally approved plan.

The scope of the collection (of written matter) is expected to be from the 11<sup>th</sup> century (Nannayya, the first person to produce a literary piece in Telugu lived in the 11<sup>th</sup> century) to the end of the 19<sup>th</sup> century. The

bulk of the material from the 20<sup>th</sup> and 21<sup>st</sup> centuries will be taken up during a later phase of the project. Inclusion of sounds and images will become relevant for this period.

## 2.2. Cataloging

With the help of teams of experts a master catalogue for all the available matter will be assembled. Established Libraries like Gouthami Grandhalayam in Rajahmundry, Vatapalem Library, Krishnadevaraya Library in Hyderabad and other government wings will be consulted.

## 2.3. Storage

All volumes collected and procured physically will be organized by the documentation team and housed in an appropriate condition in a library with necessary precautions. All information about the volumes collected will be maintained on computers and the most economical way to obtain the volumes will be planned. All the information available will be added to the master database.

## 2.4. Evaluation and Selection

A team of experts will be given charge to evaluate and prioritize the material to be included and make recommendations. Periodic meetings with experts from different specialties will be convened to make necessary decisions regarding the matter to be digitized.

## 2.5. Proofing and Quality Control

All digitized material will be proof read before making a final version. The volumes that are to be digitized will be checked for copy variations, printing and written mistakes in manuscripts. Experts will add necessary corrections with notes.

# **3. Technology Resources**

## 3.1. Hardware and Software Resources

A team of experts will be convened, and a workshop will be conducted for discussing and making necessary recommendations to carry out this project. Relevance of technologies such as optical character recognition, voice recognition, data mining and data base management will be evaluated to the fullest extent possible.

## 3.1. Standardization and Coping with Change

Keeping pace with rapidly changing technology is essential if the project is not to become obsolete before it is completed (The Paradox of Digital Preservation, *IEEE Computer*, March 2001). An approach that anticipates and plans for change is essential. This requirement makes it essential to adhere to international standards and open systems, and to avoid proprietary solutions for hardware or software. It is also essential that many of the standard practices in software development be practiced to avoid costly mistakes of retroactively making changes. This can only be achieved by involving the appropriate persons from the start: librarians, catalogers, indexers, and archivists should work hand in hand with engineers and language experts who can provide value judgments on the semantics of the content.

For this project to succeed, the design should keep users in mind. It is necessary to ensure open access to content. An important will be the possible threat to open access arising from intellectual property (IP) rights. The issue here is not just copyright, but handling of the new IP rights emanating from the almost universal availability of the information in the new format.

### 3.2. Digitization

Based on the technology available as the project progresses and also from the advice of experts in this area, the actual process of digitization will take place in India. Some portable/mobile units will be established to digitize special collections in libraries that do not lend books (for example, circa 1600 AD Telugu-French dictionary in Bibliothèque Nationale, Paris, France).

## **4. Financial Resources and Administrative Issues**

Realization of the vision articulated in this document requires a long-term effort, with a heavy commitment of financial and human resources. No formal estimate of the cost of this effort is attempted at this time. It is known that Mr. Gabriel Yu spent over US\$6 million. As Mr. Yu was able to spend his own money, he probably enjoyed better control on the project's management, costs and time table. Content selection, it is believed, turned out to be easier for him, however, as culling of the material had been done for him by an earlier Chinese emperor.

Since a project like this will take a long time to take shape, a skeletal structure of the entire project will be developed initially, laying down the basic frame work and making provision to add details as time goes by. By doing so "something" can be made available to the public even at an early stage of the project. Launching a simple pilot project, that takes advantage of the ongoing efforts in various places under the

rubric of digital libraries, may help flush out the details. If individual strands of grass can be made into a rope to tie an elephant down, so can we – as a group, if only we have the will.

Although massive volunteer effort self-help is anticipated, a project like this cannot be implemented without a regular infusion of money and management. The Government of Andhra Pradesh has a special program called Janmabhoomi (<http://www.andhrapradesh.com/>) with an annual budget allocation of Rs. 7500 lakh (~ US\$16 million ) (<http://www.ap.gov.in/apbudget/APLAN01.htm>). A project like this may well find a niche in this program. Although information technology is a key for the success of this project, employment opportunities generated by this effort will reach out and touch other sectors where highly talented people are heavily under employed. Because of this leveraging feature foundations may find this project attractive for funding.

### **Signatories of This Draft Document**

The people/organizations listed hereunder contributed ideas, provided critiques, endorsed this document or helped in some way during the drafting of this document.

Eco Foundation, Pleasanton, CA USA

Kalasapudi Srinivasa Rao, Ph. D. (Biochemistry)  
Long Island Jewish Medical Center, New Hyde Park, NY USA

Vemuri Venkateswararao, Ph. D. (Engineering)  
University of California, Davis, CA USA

Sriram Vemuri, Ph. D (Pharmacy)  
SciClone, Milpitas, CA USA

K. V. Baparao, Ph. D. (Computers)  
Computer Consultatnt, Los Angeles, CA USA

Parigi Madan Mohan (Computers)  
Computer Consultant, Fremont, CA USA

Kodavatiganti Rohiniprasad, Ph. D. (Physics)  
BARC, Mumbai, India

Behara Bhaskar Rao, Ph. D. (Economics)

University of NSW, Sydney, Australia

Akkiraju Ramapathi Rao, Ph. D (Telugu), India

See next for the ad hoc Honorary Advisory Board

### **Honorary Advisory Board**

(Alphabetical listing of all the members. Several of them were contacted and consent and support for the proposed project was obtained. A few of them could not be reached in Hyderabad and are being contacted. A list of the same with their specialization will be mailed later. )

Sri Bhadriraju Krishna Murthy

Sri Chandrashekar Rao RVR

Sri Eluripati Anataramayya

Sri Gopi N

Sri Narayana Murthy K

Sri Narayana Reddy C

Smt Nayani Krishna Kumari

Sri Perala Bharata Sarma

Sri Pervaram Jagannatham

Sri Potturi Venkateswara Rao

Sri Prasad ABK

Sri Prasadaraya Kulapati

Sri Rajendra Prasad H

Sri Rama Raju B

Sri Ramachandra Murthy K

Sri Ramalinga Raju, Satyam Computers

Sri Ramoji Rao, Eenadu

Sri Sastry BN

Sri Sastry MVR, Editor, Andhra Bhoomi

Sri Sundaram

Smt Viajay Bharati

