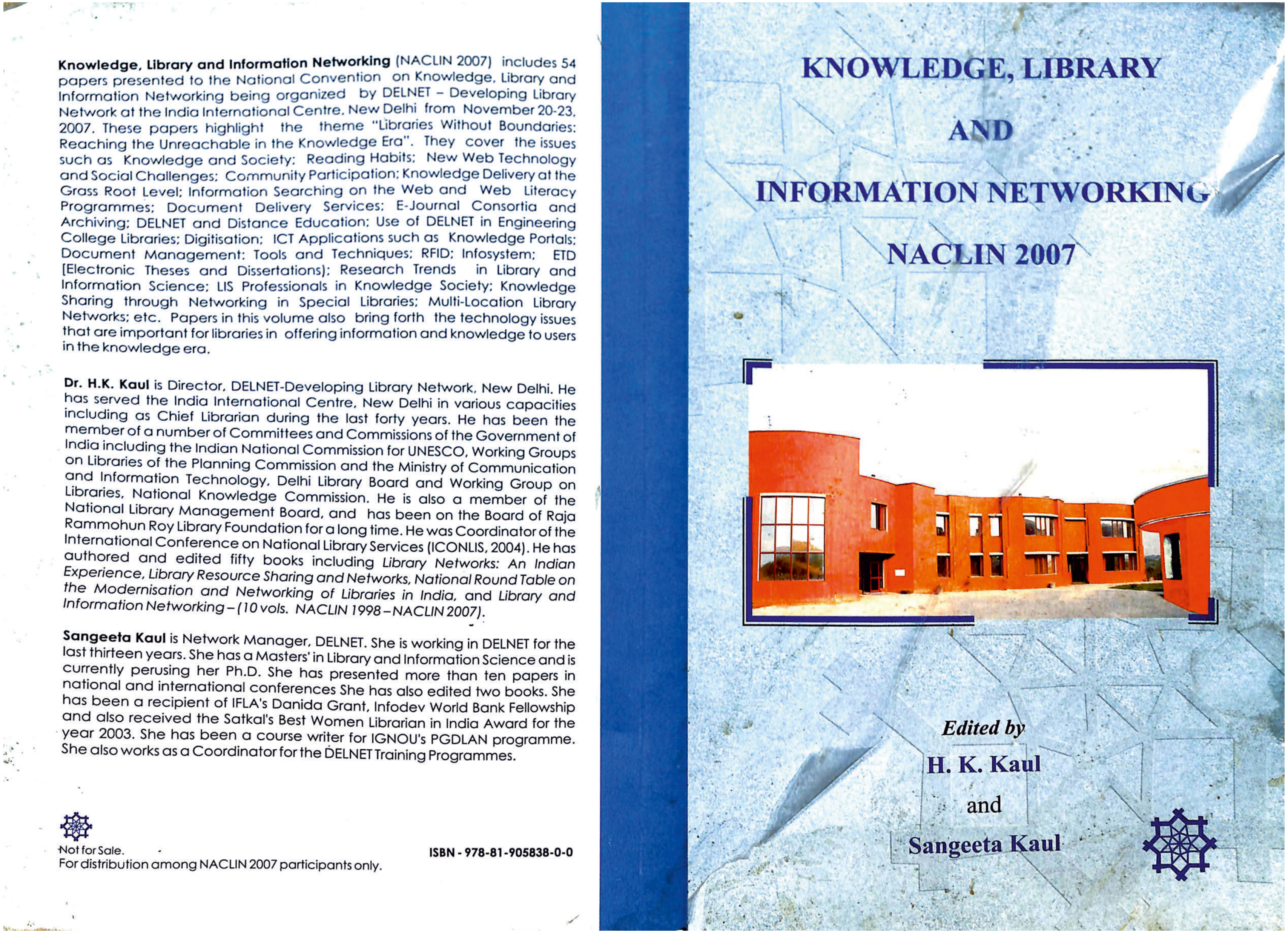


Knowledge, Library and Information Networking (NACLIN 2007) includes 54 papers presented to the National Convention on Knowledge, Library and Information Networking being organized by DELNET – Developing Library Network at the India International Centre, New Delhi from November 20-23, 2007. These papers highlight the theme "Libraries Without Boundaries: Reaching the Unreachable in the Knowledge Era". They cover the issues such as Knowledge and Society; Reading Habits; New Web Technology and Social Challenges; Community Participation; Knowledge Delivery at the Grass Root Level; Information Searching on the Web and Web Literacy Programmes; Document Delivery Services; E-Journal Consortia and Archiving; DELNET and Distance Education; Use of DELNET in Engineering College Libraries; Digitisation; ICT Applications such as Knowledge Portals; Document Management: Tools and Techniques; RFID; Infosystem; ETD [Electronic Theses and Dissertations); Research Trends in Library and Information Science; LIS Professionals in Knowledge Society; Knowledge Sharing through Networking in Special Libraries; Multi-Location Library Networks; etc. Papers in this volume also bring forth the technology issues that are important for libraries in offering information and knowledge to users in the knowledge era.

Dr. H.K. Kaul is Director, DELNET-Developing Library Network, New Delhi. He has served the India International Centre, New Delhi in various capacities including as Chief Librarian during the last forty years. He has been the member of a number of Committees and Commissions of the Government of India including the Indian National Commission for UNESCO, Working Groups on Libraries of the Planning Commission and the Ministry of Communication and Information Technology, Delhi Library Board and Working Group on Libraries, National Knowledge Commission. He is also a member of the National Library Management Board, and has been on the Board of Raja Rammohun Roy Library Foundation for a long time. He was Coordinator of the International Conference on National Library Services (ICONLIS, 2004). He has authored and edited fifty books including *Library Networks: An Indian Experience, Library Resource Sharing and Networks, National Round Table on the Modernisation and Networking of Libraries in India,* and *Library and Information Networking – (10 vols. NACLIN 1998 – NACLIN 2007).*

Sangeeta Kaul is Network Manager, DELNET. She is working in DELNET for the last thirteen years. She has a Masters' in Library and Information Science and is currently perusing her Ph.D. She has presented more than ten papers in national and international conferences She has also edited two books. She has been a recipient of IFLA's Danida Grant, Infodev World Bank Fellowship and also received the Satkal's Best Women Librarian in India Award for the year 2003. She has been a course writer for IGNOU's PGDLAN programme. She also works as a Coordinator for the DELNET Training Programmes.

Not for Sale.
For distribution among NACLIN 2007 participants only.

ISBN - 978-81-905838-0-0

KNOWLEDGE, LIBRARY AND INFORMATION NETWORKING

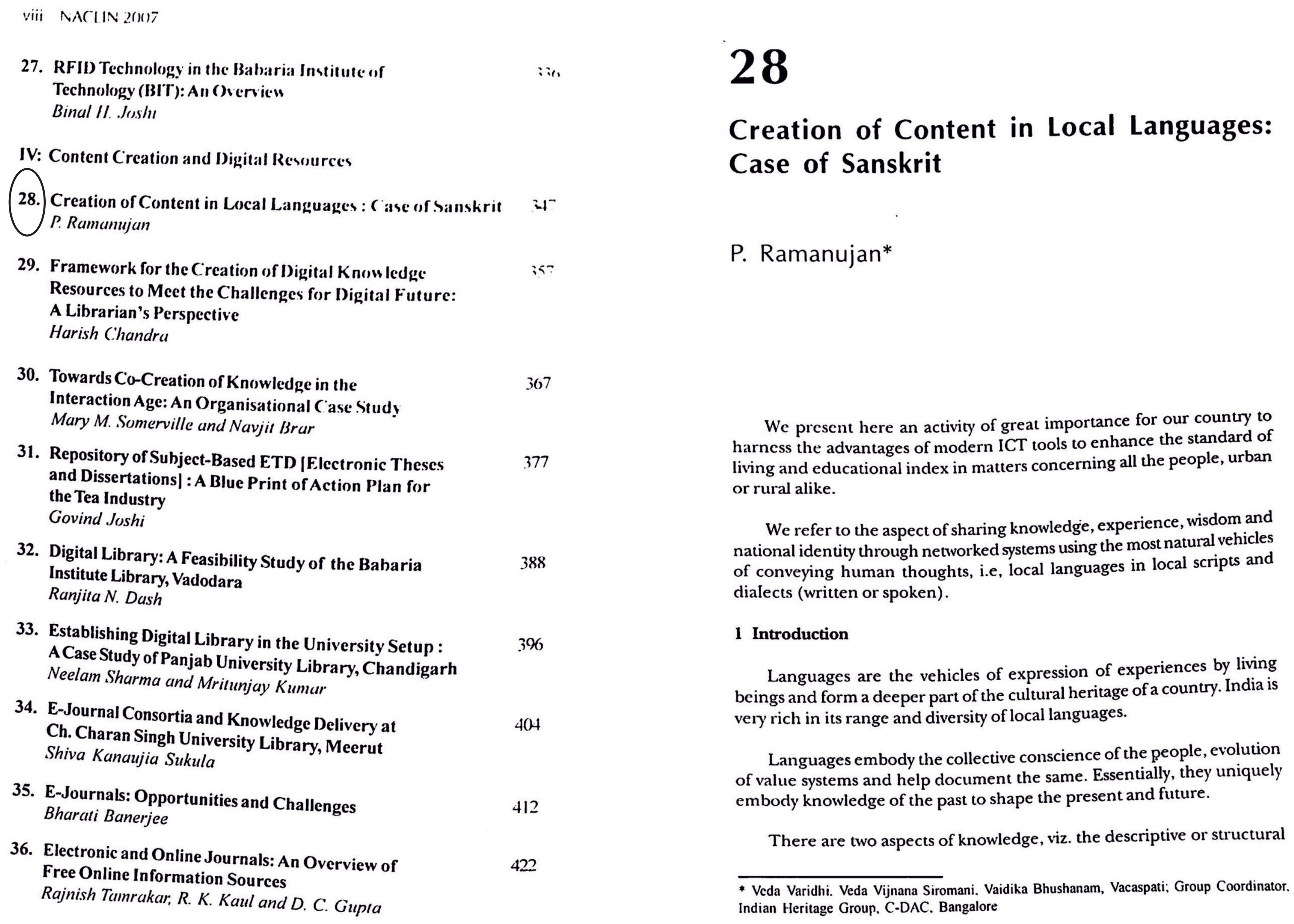
NACLIN 2007

Edited by

H. K. Kaul

and

Sangeeta Kaul



27. RFID Technology in the Babaria Institute of Technology (BIT): An Overview 336
Binal H. Joshi

IV: Content Creation and Digital Resources

28. Creation of Content in Local Languages : Case of Sanskrit 347
P. Ramanujan

29. Framework for the Creation of Digital Knowledge Resources to Meet the Challenges for Digital Future: A Librarian's Perspective 357
Harish Chandra

30. Towards Co-Creation of Knowledge in the Interaction Age: An Organisational Case Study 367
Mary M. Somerville and Navjit Brar

31. Repository of Subject-Based ETD [Electronic Theses and Dissertations] : A Blue Print of Action Plan for the Tea Industry 377
Govind Joshi

32. Digital Library: A Feasibility Study of the Babaria Institute Library, Vadodara 388
Ranjita N. Dash

33. Establishing Digital Library in the University Setup : A Case Study of Panjab University Library, Chandigarh 396
Neelam Sharma and Mritunjay Kumar

34. E-Journal Consortia and Knowledge Delivery at Ch. Charan Singh University Library, Meerut 404
Shiva Kanaujia Sukula

35. E-Journals: Opportunities and Challenges 412
Bharati Banerjee

36. Electronic and Online Journals: An Overview of Free Online Information Sources 422
Rajnish Tamrakar, R. K. Kaul and D. C. Gupta

28

Creation of Content in Local Languages: Case of Sanskrit

P. Ramanujan*

We present here an activity of great importance for our country to harness the advantages of modern ICT tools to enhance the standard of living and educational index in matters concerning all the people, urban or rural alike.

We refer to the aspect of sharing knowledge, experience, wisdom and national identity through networked systems using the most natural vehicles of conveying human thoughts, i.e, local languages in local scripts and dialects (written or spoken).

1 Introduction

Languages are the vehicles of expression of experiences by living beings and form a deeper part of the cultural heritage of a country. India is very rich in its range and diversity of local languages.

Languages embody the collective conscience of the people, evolution of value systems and help document the same. Essentially, they uniquely embody knowledge of the past to shape the present and future.

There are two aspects of knowledge, viz. the descriptive or structural

* Veda Varidhi. Veda Vijnana Siromani. Vaidika Bhushanam, Vacaspati; Group Coordinator. Indian Heritage Group, C-DAC, Bangalore

Creation of Content in Local Languages – Case of Sanskrit

Veda Varidhi, Veda Vijnana Siromani, Vaidika Bhushanam, Vacaspati
Dr. P. Ramanujan
Group Co-ordinator, Indian Heritage Group, C-DAC, Bangalore

Abstract

We present here an activity of great importance for our country to harness the advantages of modern ICT tools to enhance the standard of living and educational index in matters concerning all the people, urban or rural alike.

We refer to the aspect of sharing knowledge, experience, wisdom and national identity through networked systems using the most natural vehicles of conveying human thoughts, i.e., local languages in local scripts and dialects (written or spoken).

Introduction

Languages are the vehicles of expression of experiences by living beings and form a deeper part of cultural heritage of a country. India is very rich in its range and diversity of local languages.

Languages embody the collective conscience of the people, evolution of value systems and help document the same. Essentially, they uniquely embody knowledge of the past to shape the present and future.

There are two aspects of knowledge, viz. the descriptive or structural part for representation and modeling; and the cognitive part, which inheres in the animate (sentient) user. Let us explore the aspects in some practical system-building perspective.

Requirements

While aiming to employ advanced ICT technologies for human endeavours, we are required to look at the issues of proper human-machine interfaces, inputting/ outputting tools, processing utilities, technology for content creation, sharing, enhancing, accessing, processing etc.

Features

Multiple modes of human-machine interactions are possible in orthographic, photographic or spoken forms. Languages in handwritten/printed text, manuscript/inscription, recorded or synthetic speech modes and media forms like images (scanned and movie), text, sound waves etc. need to be handled.

Attempts

There have been many efforts in this direction at national and international levels, by both Govt., as well as private teams and many success stories are well-known. They have evolved representation standards, NLP tools and utilities, editors, fonts, converters, DTP systems, web-enabling plug-ins and SDKs, Indian language support to various popular word-processors, data processing systems, programming environments across platforms and OS's. Optical Character Recognition, Text-To-Speech, Speech-To-Text or speech recognition, learning systems, Machine Translation systems, portals, corpora and databases are developed.

Our Contribution

At IHG, C-DAC, starting from the evolution of PC-ISCII standard for Vedic character set representation, Computational rendering of Panini's grammar, developing *DESIKA* – NLU System for Sanskrit, RgVeda database, Mahabharata Database, *C-VYASA,* Sanskrit Authoring System with knowledgebase, *Pandu-lipi Samshodhaka* for assisting study and collation of Manuscripts for critical editions etc., Sanskrit Self-learning multi-media CD-ROM, Editorial assistance for reference works like dictionary, encyclopedia, thesauri etc. have been attempted in this direction.

A Case-study of a project related to Sanskrit and Vedas

The recently undertaken project titled **'Development of Analytical Tools for Large, Scientific Knowledge bases in Grid Computing environment'**, involves creation of web-based Vedic and Sanskrit Knowledgebase, developing analytical and search capabilities and deploying in *Garuda* Grid Environment.

IHG, C-DAC's *Sakala-Shastra-Sutra-Kosha* (A repository of all knowledge-related original treatises) contains the fourteen *Vidya-sthana-s* or disciplines of study in Sanskrit, covering the four *Veda-s* (RgVeda, YajurVeda (Sukla & Krsna), SamaVeda, AtharvaVeda), six *Vedanga-s* (Shiksha, Vyakarana, Chandas, Nirukta, Jyautisha, Kalpa) and four *Upanga-s* (Mimamsa (includes Vedanta), Nyaya, Puranas, Dharma Shastra) embodying the main tenets of these systems. Here, digital web contents are created for all the fourteen vidyasthana-s.

In Vedanga-s, hyperlinking to Veda-s is being provided dynamically. Application programs for each vidya-sthana and an integrated, hyper-linked 'concordance' program are being developed. All the contents can be seen in Indian scripts with transliteration facility and Vedic texts will be with proper accent-markers in the corresponding scripts.

This is the proposed *'Digital Library of Indian Heritage - A Reference Compendium of Original Treatises'* with features for browse, search, index etc.

Major outputs of this project are expected to be:

Tools to study, collate and critically edit works of interest.

Scanning/digitizing palm-leaf and other Vedic manuscripts with accents in Indian scripts.

Digital web contents along with possible commentaries and translations.

Application programs for hyperlinked concordance.

Heritage portal with the digital library mentioned above.

A Computer-based Manuscript (Palm-leaf) Editor

IHG, C-DAC has developed a comprehensive Manuscript Processing Software **Pāndu-Lipi Samśodhaka** for the purpose of the critical edition of Sanskrit texts. We have now undertaken to extend computing help to the deciphering of manuscripts in scripts like Grantha, Nandināgarī, Telugu, Malayālam etc., of Sanskrit/Vedic texts, many of which are not yet published. We also help collation of various version/variant forms of texts from different sources for critical editions of such rare, unpublished works in these domains.

The **Pāndu-lipi Samśodhaka** is useful in the collation, search, view, print etc. We describe the salient features of this further.

Features

The functional modules of the system cover acquisition, formatting, inputting, indexing, creating database, searching, locating, printing, collation and publishing. The range of texts covered include Śāstric texts, ṚgVeda, Kṛṣṇa Yajurveda, Sāmaveda, Lakṣaṇa Granthas, texts in Tamil, a combination of Tamil and Sanskrit, called Maṇipravāla, etc., in a variety of scripts. The sample includes about fifty works, one hundred and twenty manuscripts, six scripts and many domains. There are about 3500 leaves (pages) as images to accompany the PC-ISCII texts. Two of these texts, viz. Ṣaḍvimśati Sūtra and Yohi Bhāṣya, are chosen, for illustration and possible publication of a critical edition with a Sanskrit commentary.

Description of the modules:

Acquisition

a) Start with consulting catalogues, indices, lists, reports, etc., of Manuscript collection of desired texts through a number of sources Bibliographic survey.

b) Shortlist the ones feasible to obtain from these. Give balanced representation for various regions, scripts and versions (i.e., with commentaries, with accents etc.)

c) Acquire copies xeroxed/scanned /microfilmed

d) Convert/export to a single (uniform) format. Factors like clarity, condition of original, resolution of scanning and size of the image files, all influence the choice of common format used.

Formatting

The inputs come in various forms, when raw, i.e., from the institutions/library collections. We may have 3, 5 or even 10 folios per scanned image. Here the two sides of the folios are in different files and the job of sequencing the image as per text and separation of folios is involved. Numbering them serially according to the text is done. An important task here, in the case of manuscript bundles containing different texts, is separation of the texts and folios belonging to multiple texts. They must be present in all the works concerned. Usually, libraries offer separation, if catalogued already.

Inputting

We strongly recommend the entry of the data contained in the manuscripts for the purpose of study, word-split, index, search (phrases), editing and collation. This, of course, requires domain experts who are difficult to get. However, IHG can offer expertise in this endeavour. We also have another possible source for data entry, which is loading text, if the work in the manuscript is one of available digital texts from our repository. (A list of about 250 texts from all Vidyāsthānas is available. C-DAC Indian Heritage Portal would make this available on the web soon).

Adding commentaries, translations, hyperlinks, annotations for collation etc., are the factors necessitating data-entry. Also, transliteration, training in rare scripts etc., is enabled. However, efforts may be launched to develop efficient OCR or speech recognition systems of high quality simultaneously and when these mature, we can minimize data entry needed.

Editing

This step involves aligning the data entered, with the original manuscript, line by line and page by page. This also can be done in an edit box/window below (or adjacent to) the image of the manuscript or entered through Vedic Editor and inserted into database. The pages and line boundaries are as before. Adding information for retrieval, hyperlinks etc., can also be done. Multilingual texts, currently require LEAP-like software for data entry and use in RTF format in the system for further processing. Here ISCII-ISFOC conversions are employed. Currently Vedic texts of Sāmaveda Gāna require use of only Grantha script and transliteration is not available. Śrautam and Guruparamparā Prabhāva etc. are multilingual samples. These are typed in LEAP and processed through rtf controls.

Creating database

The PC-ISCII text files (*.pci) created by data entry or loading data are to be converted into database format. This is either Microsoft Access or Microsoft FoxPro format covering various fields for facilitating information retrieval. Databases of works, institutions, manuscripts, books etc., are also created and linked in the application list of abbreviations. Scheme of data for reference in these texts etc. are also created as tables.

Searching

This is the crux of the system and helps in providing word or phrase level search (with and without accent-markers) across the database, text-wise, and lists the manuscript reference numbers where the search string occurs. In future, this can be extended across texts if need be (this feature is there in our Vedic Editor, wherein a string occurring in any of the 250+ texts are listed as a concordance). Choice of script, facility to transliterate, and seeing the results in the same manner of alignment as the manuscript are the useful aspects.

Locating

This refers to locating the search string in the image of the particular page of the manuscript where it occurs, including the line number and location in it. We see the string 'highlighted' in the text window by choosing 'find' in the page and physically looking in the corresponding line and 'location' in it on the image above by selecting view in 'search' mode. The text window is provided with line numbers to facilitate this manual locating in the image.

Printing

Provision to print the texts in database, search results etc., in any script of choice or script of the original etc. so that further reference or insertion into documents can be enabled. Report generation kind of printing needs can also be addressed. List of texts, institutions, reference details etc. can be printed.

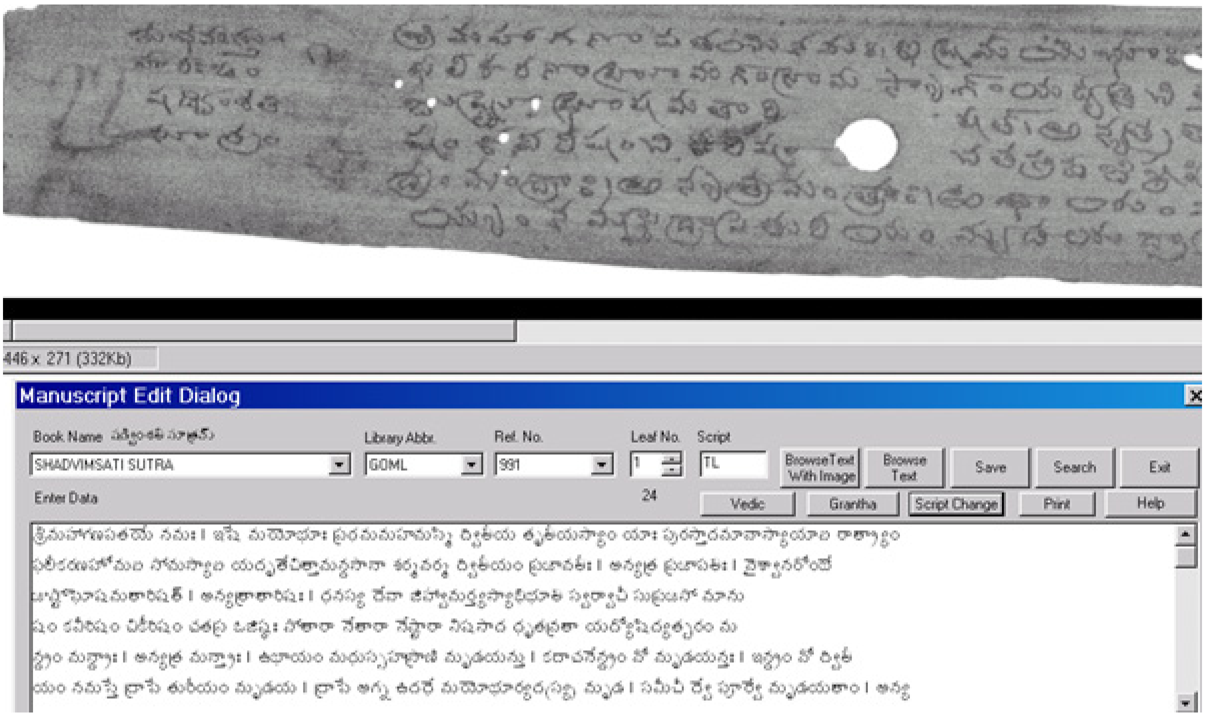
Collation

From the search function, we can organise the readings of different texts (like 'file compare') across the manuscripts combined with report generators. A scheme for annotating can be devised to assist here. Work will follow to enrich features here.

Publishing

Publication through Desk-top-publishing can be done by exporting to some DTP software and adding embellishments as desired.

Screen shots of some of the modules



Book Name షడ్వింశతి సూత్రమ్

Book Name: SHADVIMSATI SUTRA

Library Abbr.: GOML

Ref. No.: 991

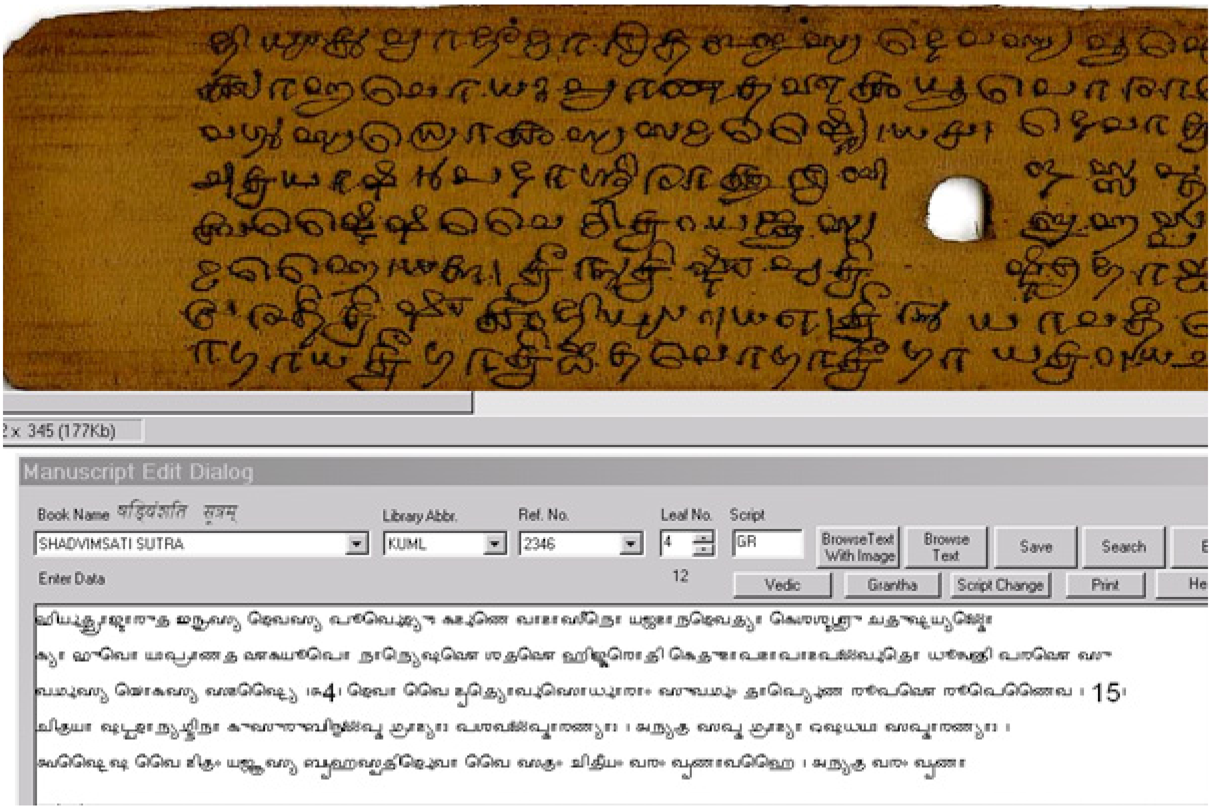
Leaf No.: 1

Script: TL

24

Enter Data

శ్రీమహాగణపతయే నమః । ఇషే మహాభూతాః ప్రథమమహమస్మి ద్వితీయ తృతీయసాక్షం యాః పురస్తాదమావాస్యాయా రాత్ర్యం
ఫలీకరణహోమం సోమస్యేం యద్వృతేచిత్తామన్త్రసానా కర్మవర్మ ద్వితీయం ప్రజాపతిః । అన్యత్ర ప్రజాపతిః । వైశ్వానరోంతే
బర్హిర్యోనాసమతారిషత్ । అన్యత్రతారిషుః । ధనస్య దేవా జిహ్వామద్యసాక్షీభూతి స్వరాట్పి సుప్రజసో మాను
షం కపీరసం వీరసం పత్ర ఒజస్యః సోతారా నేతారా నేష్టారా నిషసార ధృతవ్రతా యద్యోక్షిద్యశ్వరం మ
న్త్రం మన్త్రాః । అన్యత్ర మన్త్రాః । ఉపాయం మధుస్సహస్రాణి మృడయన్తు । కదాచనేష్ఠం వో మృడయన్తుః । ఇష్టం వో ద్వితీ
యం నమస్తే బ్రాహే తురీయం మృడయ । బ్రాహే అగ్న ఉదరే మహాభూతార్థరస్య మృడ । సమీపే ద్వే పూర్వే మృడయతాం । అన్య



2 x 345 (177Kb)

Manuscript Edit Dialog

Book Name षड्विंशति सूत्रम्

SHADVIMSATI SUTRA

Library Abbr. KUML

Ref. No. 2346

Leaf No. 4

Script GR

Enter Data 12

Salient Features of the Mahābhārata Database

The database can be browsed for any desired parvan (and sarga) as text (optionally as word-split and marked-up or *tagged* form as well) with choice of scripts from among Assamese, Bengali, Devanagari, Gujarati, Kannada, Malayalam, Oriya, Punjabi, Roman (with diacritics), Tamil and Telugu [Hindi, Konkani, Marathi, Nepali and Sanskrit share Devanagari Script]. Additional details provided are sarga name and Antar-parvan name. Prose form of the text are wrapped around to a new line after about 50 characters. On-line help is provided in all screens.

In the retrieval mode, multiple ways of choosing/searching the desired information are provided like parvan, sarga, śloka number, word (split form), part of a śloka (phrase search includes blanks and multiple words also), by sarga name, speaker name, topic name (prakaraṇa/viṣaya), *prātipadika* or nominal stem search, be it the initial, middle or final member of a compound word and Boolean search to cover these with logical operators like AND, OR, NOT etc.

Details like parvan name, sarga name, speaker name, antar-parvan name are available. Script change, marked-up form view, help and back (exit) are standard features. In the search by number option, on selecting the parvan, admissible limits of sarga numbers and thereupon, those of śloka numbers are displayed for valid values to be entered.

In the 'search word' option, on entering few initial characters (even one), all admissible words beginning with the typed characters are listed and choosing anyone thereupon, the śloka numbers are displayed with the ref. no. scheme for desired values to be selected. The selected śloka with all other details is displayed as before.

In the 'phrase search' option, any particular parvan(s) or all can be chosen and the phrase can be typed. All occurrences, with statistics and details of information are displayed. On choosing any desired number, the particular śloka is shown. Sarga, speaker and topic names are also similarly selected. In prātipadika search, the desired stem as beginning, middle or end are additionally singly or severally selectable, and with statistics, detailed display of the ślokas containing those compound words are shown.

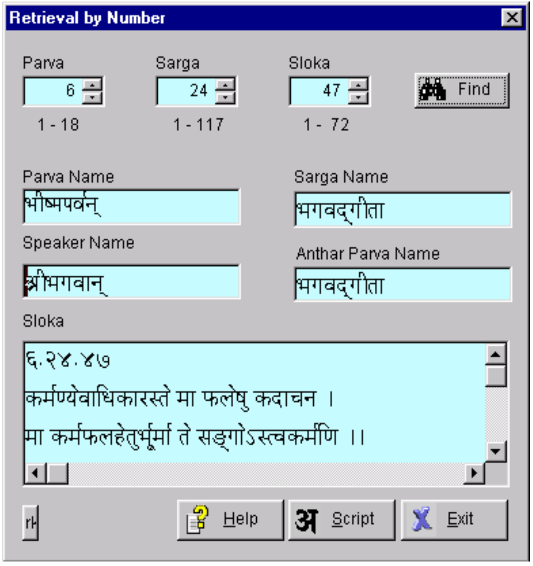
In the 'Index mode', word, name, sarga, śloka, samāsa and speaker indices are provided. Śloka index covers the entire Mahābhārata and samāsa index has two-tier selection for first-level and subsequent detailed types. Help files are also accessible from any screen by pressing F1 key or Help button. There is also a demo of the program as a guided tour included in the CD-ROM.

The Samāsa Mark-up Scheme Used for the Mahābhārata Database Project

The scheme for mark-up has the name of the compound and its notation (tag) to be used for mark-up. Multiple word compounds are also marked distinctly. The bracketing in the tags also has distinction between words where compounding proceeds sequentially (i.e, from left to right) and where there is a change in its direction. These are also suitably illustrated. Nested brackets could be used later with our programs.

Multiple mark-ups for words denoting different possibilities/interpretations or allegory etc. are also encouraged to be indicated by using? and placing alternate tag(s) in curly brackets. e.g, **word>**_*Tag1*/{?*Tag2*...}.

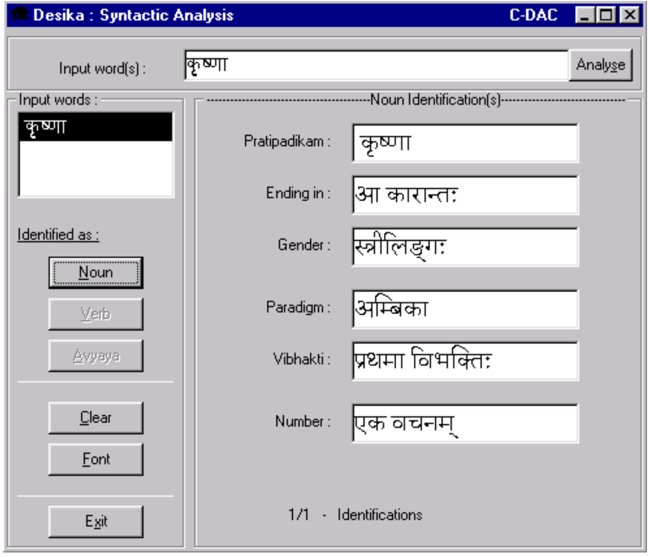
We have suggested certain specific features like not marking *samāsānta taddhita* suffixes etc. in the end where possible (with exceptions as in +t option in Dvigu, for example) and also included *lyabanta avyayas* as samāsa etc. In all cases, a compound word has at least a pair of *angle brackets* (< & >). The *underscore* character (_) will be the delimiter for tags. *Hyphens* separate stems. **The marked-up files are processed for tags and statistically**


Retrieval by Number

Parva: 6 (1 - 18)
Sarga: 24 (1 - 117)
Sloka: 47 (1 - 72)

Parva Name: भीष्मपर्वन्
Sarga Name: भगवद्गीता
Speaker Name: श्रीभगवान्
Anthar Parva Name: भगवद्गीता

Sloka:
६.२४.४७
कर्मण्येवाधिकारस्ते मा फलेषु कदाचन ।
मा कर्मफलहेतुर्भूर्मा ते सङ्गोऽस्त्वकर्मणि ।।


Desika : Syntactic Analysis — C-DAC

Input word(s) : कृष्णा

Input words : कृष्णा

Identified as : Noun

Noun Identification(s)

Pratipadikam : कृष्णा
Ending in : आ कारान्तः
Gender : स्त्रीलिङ्गः
Paradigm : अम्बिका
Vibhakti : प्रथमा विभक्तिः
Number : एक वचनम्

1/1 - Identifications

Road Ahead

By a strategic approach, we can ensure creation, sharing, accessing, research and propagation of the treasures lying hidden in the contents in local languages systematically. We could prepare the registry and catalogs of manuscripts and published works collected under the National Mission for Manuscripts (NMM), New Catalogus Catalogorum (NCC), Veda Lakshana Bibliography etc., employing experts, students, home makers etc., in freelance mode by suitable training and development of computational tools/utilities as mentioned.

Conclusions

Beginning with inclusion of Computational study of regional languages and introducing the concepts of ‘language lab’ and ‘project work’, web content creation in Indian languages, creation of reference works like Dictionary, Encyclopedia, Thesaurus etc, in regional language study at Universities at undergraduate through doctoral research levels, we can ensure trained and competent resource persons for such advanced projects. Dissemination of information on Government sponsored projects in these areas to academic community in Arts, Humanities etc., would help in better synergy and unearthing treasures in the cultural, linguistic and heritage domains, resulting in India once again leading the world through humane use of technology and even IT R&D.

We would also be bridging the digital divide by integrating and involving by positive contribution to society, the maximum number of citizens in all regions equitably and in an even distribution across spheres in intellectual endeavours.