4. About TDIL Programme

Technology Development for Indian Languages Programme (TDIL) -An Overview

1. Need of the programme

The world is in the midst of a technological revolution nucleated around Information and Communication Technology (ICT). Advances in Human Language Technology will offer nearly universal access to information and services for more and more people in their own language. Today 80 % of the content on the Web is in English, which is spoken by only 8% of the World population and only 5% of Indian population. In a multilingual country like India, with 18 official languages & 10 scripts it is essential that information processing and translation software should be developed in local languages and available at low cost for wider proliferation of ICT to benefit the people at large and thus pave the way towards "Digital Unite and Knowledge for all" and arrest the sprawling Digital Divide.

2. Impediments

- Lack of industry involvement due to constrained demand;
- Sub-critical and Un-sustained demand in States;
- Negligible software tools and re-usable components in public domain
- Plurality of Internal Codes: ISCII-88, ISCII-91, UNICODE, and other propriety codes;
- Content Glyph-coded, not (ISCII) character-coded;
- No font Standard;
- Non-availability of Human resources in the domain of Language Technology in both Computational Linguistics (CL) and knowledge Engineering (KE)
- Reluctance of Academia to focus on productising the R&D.

3. TDIL Mission

In this context, a number of initiatives have been taken towards development of software, tools and human–machine Interface systems in Indian Languages under the TDIL programme.

Vision

Digital unite and knowledge for all.

Mission

Communicating without language barrier & moving up the knowledge chain.

Objectives

- To develop information processing tools to facilitate human machine interaction in Indian languages and to create and access multilingual knowledge resources/content.
- To promote the use of information processing tools for language studies and research.
- To consolidate technologies thus developed for Indian languages and integrate these to develop innovative user products and services.

3.1 Focus Areas

Translation Systems

- Machine aided Translation system (MAT)
- Parallel corpora, Lexware, Multilingual dictionaries, Wordnet, speech databases
- Speech-to-Speech Translation system

Human Machine Interface systems

- Optical Character Recognition
- Speech Recognition
- Text to Speech system

Language processing and Web tools

Word processors, spell checkers, converters and portals

Localization

Adapting IT tools and solutions in Indian languages.

Standardization

UNICODE, XML, ISCII (Indian Standard Code for Information Interchange), INSFOC (Indian Standard Font Code), INSROT (Indian

Script to Roman Transliteration), Standard for Lexware format, etc.

Evaluation and Benchmarking

Evaluation and testing of the prototype technologies against benchmarks, product compliance testing against specifications and validations of standards.

4. Technology Development Phases

Development of Language Technology in India may be categorized in three phases:

1976-1990 : A-Technology Phase
Focus was on Adaptation Technologies; abstraction of requisite technological designs and competence building in R&D institutions.

1991-2000 : B-Technology Phase
Focus was on developing Basic Technologies- generic information processing tools, interface technologies and cross-compatibility conversion utilities. TDIL (Technology Development for Indian Languages) programme was initiated.

2001-2010 : C-Technology Phase
Focus is on developing Creative Technologies in the context of convergence of computing, communication and content technologies. Collaborative technology development is being encouraged to realise.

5. Program Management Structure

All the above activities are being implemented through various Resource Centres and Localization (CoIL-Net) Centres. The programme is being technically monitored by TDIL Working Group consisting of experts drawn from academia, industry, R &D organizations and Government. The Resource centres are divided into regional clusters for peer review of the progress of the technologies and solutions developed by them.

Resource Centres

The Department has thirteen Resource Centres for Indian Language Technology Solutions at various educational institutes and R&D organizations who have developed several important tools and technologies for Indian Language support.

CoIL-Net (Content Creation & IT Localisation Network

CoIL-Net programme is formulated encompassing a vision for all pervasive socio-economic developments by proliferating use of language specific IT based content, solutions and applications for bringing the benefits of IT revolution to the common citizen; requisite technology development and consequently bridging the existing digital divide in the Hindi speaking states of India. The objective is to provide a boost to IT localization based socio-economic developments and help bridge the existing digital divide by appreciably improving IT penetration and awareness levels, using Hindi as medium of delivery, in the Hindi speaking states of MP, Chattisgarh, UP, Uttaranchal, Bihar, Jharkhand, and Rajasthan.

CoIL-Tech

The MAIT Consortium on Innovation & Language Technology (CoIL-Tech) since its inception in September 2001, has been actively coordinating various activities with the Industry and the TDIL (Technology Development in Indian Languages) Department of the Ministry of Communications & Information Technology. The consortium today has active participation from both Indian and MNC companies with a focus to promote industry participation in collaborative R&D in language technology, to coordinate Open Source Software supporting Indian languages, to evolve consensus on standards, benchmarks, and certification of LT products, to collectively interface with government and academia, to conduct market surveys, organize technology shows, promote technology transfers and expand market collectively.

6. Major Achievements of the TDIL Programme

Optical Character Recognition (OCR)

Optical Character Recognition is an indispensable tool for digitizing the content and is essential for development of Knowledge networks such as digital libraries. Optical Character Recognition (OCR) technology offers the facility to scan and store the printed text. There are three essential elements to OCR technology - scanning, recognition and then reading text. Initially, a camera scans a printed document. OCR software

then converts the images into recognized characters and words. The OCRs developed are being tested and benchmarked independently by STQC.

OCR with more than 97% accuracy has been developed for seven Indian Languages viz Hindi, Marathi, Bangla, Tamil, Telugu, Punjabi, Malayalam.

The OCR technologies for Assamese, Oriya, Malayalam and Gujarati scripts are in the advanced stages of development.

Spell Checkers

Spell checker is the language specific software component used in word processing, OCR post processing and text processing applications. Spell Checkers for Hindi, Marathi, Bangla, Tamil, Telugu, Punjabi, and Malayalam have been developed. The Spell checkers for other Indian Languages are in advanced stages of development.

Machine Aided Translation System (MAT)

MAT system provides rough translation from a source language (say English) into a target language (say Hindi). This output may be post edited with specially designed tools for enhancing the translation accuracy.

Anglabharati English to Hindi Translation

Support System has been developed by IIT Kanpur (Resource Centre for Indian language in Hindi and Nepali language) and the same is available in Public Domain at **http://anglahindi.iitk.ac.in**

The utility of text to speech synthesis has also been integrated with Anglabharati and works on the Linux platform.

The Machine Translation system for **Indian Languages to Hindi and Hindi to English** are in the process of development.

Support on Indian Language LINUX Operating System (INDIX)

Localization Linux operating system at x-windows level has been done. INDIX (Localized LINUX) operating system has been developed by NCST, Mumbai.

It is now possible on Linux to give file names, domain names in Devanagari. Indian language support on Linux in open domain will also ensure Indian Languages support on all Open Source Tools Kits. INDIX supports GNOME desktop environment (of LINUX) to enable applications to create, edit and display contents in Indian Languages. The R&D for supporting other Indian Languages have been initiated.

Standardization

Unicode Standards are widely being used by the Industry for the development of Multilingual Software. Unicode Standard 3.0 includes standard code sets for Indian scripts based on ISCII-1988 document. Some modifications are necessary to incorporate in the Unicode Standard for adequate representation of Indian Scripts.

Department of Information Technology, Ministry of Communications & IT, is the voting member of the Unicode Consortium. Proposed changes in the existing Unicode Standards has been finalized in consultation with respective State Government and Indian IT Industry. These have been published in TDIL Newsletter Vishwabharat@tdil.

The Indian Scripts Font Code (INSFOC) to overcome the problem of interoperability of various software products developed by different vendors has been completed for Hindi, Malayalam, Gurmukhi and Gujarati. The work is in progress for other Indian languages.

An "Indian Script to Romanization Table (INSROT)" has been worked out to facilitate no-Hindi users to work in Hindi. Efforts are also on to Standardise the Indian Lexware Format.

Parallel Corpora (Gyan Nidhi)

A parallel corpus is an important input for MAT system. **One Million pages Parallel Corpora** is presently in the process of development. The Graphic User Interface for viewing the multiple Language Corpora has also been developed. The aim is to develop parallel corpora of One million pages. So far 450, 000 pages parallel corpora has been developed (as on May 31, 2003).

Bi-lingual Dictionaries

Bi-lingual electronic dictionaries are the basic linguistic resource which is useful for any NLP related research development and testing progress. These are useful in the development of Word processors in Indian Languages. English – Hindi, English – Telugu, English- Tamil, English – Kannada, English- Bangla, Bangla-Bangla, English-Punjabi, English-Oriya, English-Malayalam bilingual dictionaries have been developed by RCs.

IT Localisation Clinic

Workshops were organised by DIT during November 2002, to accelerate the project deliverables such as Content development in Hindi, Creation of Repository of language technology products, Trainers Training programmes, IT Localisation Clinics, IT Curricula for Schools, Website Development in Hindi. Installation of test beds in the domains of e-governance (Land Records, Agri-Net, Query System), e-education (Schools, ITIs), e-business (Small Business), e-tourism, e-health.

TDIL Portal

The TDIL website is bi-lingual (English and Hindi) and contains information about the TDIL Programme, it's initiatives and achievements. It provides access to Indian scriptures, standards (Indian scripts, keyboard layout, font layout, etc), articles and reviews. The unique services provided by the website includes E-mail in Hindi and Online Machine Translation from English to Hindi and vice versa.

The website also provides downloadable softwares and tools in Indian Languages viz. Plug-in such as Akshar for Windows (Plug-in for MS-Word), Shrilipi Bharti (Plug-in with keyboard driver having Devanagari fonts), Indian Language Word Processors, NLP tools, NLP Resource for Windows/Linux.

Quarterly on Indian Language Technologies

The **VishwaBharat@tdil** is a quarterly Newsletter of Indian Language Technologies, which consolidates in one place information about products, tools, services, activities, developments, achievements in the area of Indian Language software. It serves as a means of sharing ideas among technology developers. This creates awareness in the society regarding the availability of language technology resources. It is issued quarterly is widely circulated

Eight issues viz. Jan 2001, May 2001, Sept. 2001, Jan 2002 , May 2002, July 2002, October 2002, Jan 2003 published. All the issues Accessible through TDIL Web site and all the issues are available in a single CD.

7. Targets for 2003-04

- Integration and Deployment of ***OCR, TTS*** and Document processing technologies
- Initiation of development of **Machine aided Translation System** between English and Indian Languages
- Linux Operating System enabled with Indian languages support
- Enforcing **UNICODE** revision and Vedic code standardization
- Initiation of ***Intelligent Cognitive system (KUNDALINI)*** project for bringing synergy in traditional Sanskrit Shastras and Information Technology.
- Cross Lingual Information Retrieval for Universal Access
- IT Localization Clinics for incubation of IT technologies
- 16 Bit Unicode compliant Hindi fonts for public domain.

8. New Projects Initiated

Intelligent Cognitive System

A new project in the domain of Intelligent Cognitive system named Knowledge

UNDerstanding & Acquisition of Languages, INferencing and Interpretation (KUNDALINI) is being initiated with the following objectives:

- To develop methodologies and tools for knowledge representation, extraction, mining, gisting, inferencing and interpretation
- To develop knowledge frameworks and access mechanisms based on Indian tradition
- To develop Sanskrit based Networking Language as Machine Translation Interlingua. (Concept based Networking Languages : *CNL*)

Gyanaudyog (ज्ञानोद्योग)

Encouraged by the success of the **Gyanaudyog** Workshop held on April 7 –9, 2003 a new project **Gyanaudyog** is being initiated to promote Small Office & Home Entrepreneurship in the area of Information Technology for catalyzing IT enabled services (ITES), specifically, Content Creation, Content localization & application software localization, Remote customer interaction Services, Computer Aided Design with support for ***Technology Mentoring, Financial Support guidance and Market information.***

Anusrijan (अनुसृजन)

There is addition of over 25 Million pages in R&D in the field of Science &Technology. It is planned to prepare books/monographs in local language on emerging areas of Information & Communications Technology (ICT). These will be available for translation into other Indian Languages so as to create averseness about recent development to promote innovation and entrepreneurial aptitude.

SOHE - Ganak Bharti (सोहे-गणक भारती)

A new project named as Ganak Bharti is being initiated to promote development and deployment of Low cost PC with Indian Language support and Open Source Software for small business men, women at home and children aiming at catalyzing the growth of Small Office, Home and Education (SOHE) environment.

Evaluation and Benchmarking of the Indian Language Technology Tools

Testing, Evaluation and Benchmarking of the Indian Language Technology tools is essential for wider acceptance of the Indian Language technology products. In view of this, Standardization Testing and Quality Certification (STQC) division of DIT has been designated as the third party for evaluation of the language technologies tools and products developed TDIL programme as per international standard.

9. Digital Library Initiative of India

Digital libraries are a form of Information Technology in which social impact matters as much as technological advancements. Future knowledge networks will rely on scalable semantics, on automatically indexing the community collections so those users can effectively search within the Interspace of a billion of repositories. Just as the transmission networks of the Internet are connected via switching machines that switch packets, the knowledge networks of the Interspace will be connected via switching machines that switch concepts. Connectivity and training continue to be the principal barriers to integrating the global network of libraries.

Among a number of major DL initiatives in USA is the Universal Digital Library (UDL) project that aims at creating a free-to-read, searchable collection of **one million books**, primarily in the English language, available to everyone over the Internet. This project involves participation from many countries including USA, Australia, India China, Egypt, Srilanka . The funding agency is National Science Foundation (NSF),USA, with a funding of US $ 4.0 Million

The contribution form various industries of USA is US $ 6.0 million. The overall coordination of the project is done by Carnegie Mellon University (CMU), USA.

India participates in the UDL project and makes efforts to put e-content of vast Indian knowledge base in the Indian language as much as possible. The Indian activity of Digital Library is coordinated by Indian Institute of Science, Bangalore

9.1 Objectives

- To digitize and index the heritage knowledge.
- To provide a test bed that will support other research domains such as scanning techniques, optical character recognition etc
- Inter-ministerial collaboration for e-Content.
- To promote life long learning in the society (A necessity of the Knowledge-based society).
- To promote collaborative creativity and building up knowledge teams across borders.
- Involvement of Resource Centers for Indian Language Technology Solutions and COILNet Centers to digitize and web-enable contents in Indian Languages.

9.2 Participating Institutions of UDL program

- Indian Institute of Science, Bangalore (Overall coordination)
- Anna University, Chennai, TamilNadu
- Arulmigu Kalasligam College of Engineering, Madurai,Tamil Nadu
- Goa University, Goa
- Indian Institute of Information Technology, Allahabad, Uttar Pradesh
- City and State Central Library, Andra Pradesh
- Indian Institute of Information Technology, Hyderabad, Andra Pradesh
- Shanmugha Art, Science Technology & Research Academy, Thanjavore, TamilNadu
- Sri Sri Sharda Peetam, Sringeri Mutt, Sringeri, Karnataka
- Tirumala Tirupati Devasthanam, Tirupati, AndraPradesh
- Maharasthra Industrial Development Corporation, Mumbai, Maharashtra
- University of Pune
- Centre for Development of Advanced Computing, Noida
- Kanchi University, Kanchi, Tamil Nadu
- Indian Institute of AstroPhysics, Karnataka
- Indira Gandhi National Centre for Arts, New Delhi
- Rashtrapathi Bhavan, New Delhi
- Punjab Technical University, Punjab

9.3 Targets for 2003-04

- Development and Deployment of on-line Multilingual Machine Translation system, OCRs in Indian Languages and Universal Dictionary Programme through TDIL programme.
- Establishment of Regional Mega Centres at the institutions with proven capability of scanning at least 5000 pages per day. The Mega-centres may be created with State government's initiative for meeting operational costs.
- Up gradation of the connectivity of the Mega-Centres of scanning with 2Mb/s connectivity.
- Creation of " Digital Library Act" – Provision for Tax deduction etc for the purpose of providing a Digital version on web for public benefit.
- Creation of 4C (Consortium for Compensation for Creative content)
- Evolution of Project Management mechanism and productivity measurement strategy of the Digital Library Centres and criteria for setting up of Mega Scanning Centres.
- Setting up of Inter-ministerial committee to integrate Digital Library efforts in India
- Evolution of National Mission on Digital Library in consultation with Prof. M.G.K.Menon.
- Participation with lead role in
  - WSIS (World Summit on Information Society) organized by ITU at Geneva in Dec 2003)
  - UN Decade of Literacy (2003-2013)
  - TWA (Third World Academy)
- Promotion of Spin-off technology from Digital Library programme.

- TDIL Programme Team