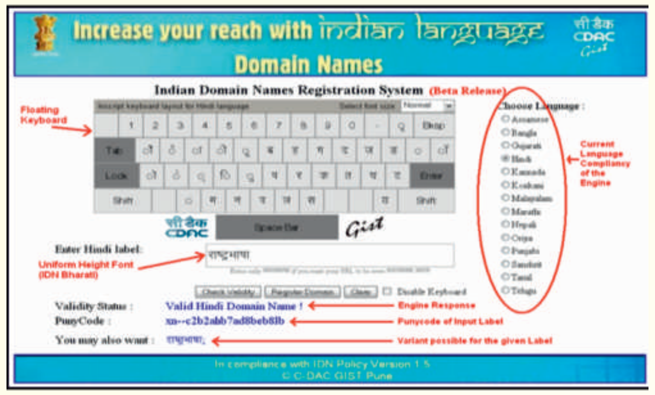
Standardization of International Phonetic Alphabet (IPA) for Indian Languages
C-DAC has developed the IPA for Bangla language, and is currently developing IPAs for Assamese, Manipuri, Bodo, Punjabi and Marathi.

International Domain Names (IDN)
C-DAC is assisting NIXI in supporting Indian language domain names.


Increase your reach with indian language Domain Names

सी डैक CDAC Gist

Indian Domain Names Registration System (Beta Release)

Floating Keyboard

Inscript keyboard layout for Hindi language — Select font size: Normal

Enter Hindi label: राष्ट्रभाषा

Uniform Height Font (IDN Bharati)

Check Validity | Register Domain | Clear | □ Disable Keyboard

Validity Status : Valid Hindi Domain Name ! — Engine Response

PunyCode : xn--c2b2abb7ad8beb8lb — Punycode of Input Label

You may also want : राष्ट्रभाषा; — Variant possible for the given Label

Choose Language :
○ Assamese
○ Bangla
○ Gujarati
● Hindi
○ Kannada
○ Konkani
○ Malayalam
○ Marathi
○ Nepali
○ Oriya
○ Punjabi
○ Sanskrit
○ Tamil
○ Telugu

Current Language Compliancy of the Engine

In compliance with IDN Policy Version 1.5
© C-DAC GIST Pune

Language Learning/Tutors
C-DAC has developed LILA [Learn Indian Languages through Artificial intelligence] series of language tutors, namely, LILA Prabodh, LILA Praveen, LILA Pragya, LILA Foreigner, etc. These tutors are available on CDs, the Internet and mobile MMCs. Other systems developed are Certificate course in Hindi (Online Praveshika), e-Mahashabdkosh – Domain based Bi-Lingual and Bi-Directional Hindi/English Dictionaries with Pronunciation as well as Online Examination System for Hindi Prabodh, Praveen and Pragya.

Marathi Tutor is an online system for learning Marathi from vocabulary to sentence formation. It follows a non-grammar based approach, modelled on the way children learn their first language.

Heritage Computing

Standardization of Heritage Scripts
C-DAC is involved in standardization and representation of heritage scripts such as Grantha, Vedic, Samavedic, Modi, etc. in modern standards such as UNICODE.

JATAN: Virtual Museum Builder
JATAN: Virtual Museum Builder has been standardized by the Ministry of Culture for all the museums under the ministry. It has been deployed in various national and state museums.

Standard for Preservation Information Documentation (eGOV-PID) of Electronic Records
This standard proposes to capture most of the preservation information (metadata) automatically after the final e-record is created by an e- governance system. Such preservation information documentation is necessary only for those e-records that need to be retained for long durations (e.g. 10 years, 25 years, 50 years and beyond) and the e-records that need to be preserved permanently as per the requirements specified in the ISO 14721 Open Archival Information Systems (OAIS) reference model.

Centre of Excellence for Digital Preservation
Centre of Excellence for Digital Preservation at C-DAC has developed digital preservation guidelines and standard for e-governance to ensure that the electronic records are produced in preservable manner. The digital preservation standard has been notified and adopted for all e-governance applications by the Ministry of Communications and Information Technology, Government of India.

Digital Preservation System for Disposed Cases
Digital Preservation System for disposed cases for courts was developed under the National Digital Preservation program.

Digital Library of India
C-DAC has established regional mega scanning centres for digitization of rare and copyright free books of various regions of India including the North-Eastern region. The aim is to create a portal of heritage books and manuscripts for the Digital Library of India, which will foster creativity and free access to all human knowledge.

Contact Details
support@cdac.in

* Bengaluru * Chennai * Hyderabad * Kolkata * Mohali * Mumbai
* New Delhi * Noida * Pune * Silchar * Thiruvananthapuram

www.cdac.in
https://www.facebook.com/CDACINDIA and @cdacindia

सी डैक
CDAC

Language Technology and Heritage Computing

Standardization
Localization
Machine Aided Translation
OCR and OHWR
Multilingual Semantic Search
Speech Technologies
Indian Languages on Embedded Devices
Indian Language Fonts, Corpora, Dictionaries and Tools
Indian Language for Media
Heritage Computing

सी डैक
CDAC
प्रगत संगणन विकास केंद्र
CENTRE FOR DEVELOPMENT OF ADVANCED COMPUTING

India is a unique country in the world in terms of language diversity, having 22 scheduled languages besides heritage languages and over one hundred widely used languages with different scripts. Despite an impressive growth of computerization and the Internet over the past few decades, most of the content on the Internet and most of the ICT based solutions in India are still available only in English. This is in stark contrast to the fact that hardly 10% of Indians can use English as a language for communication. C-DAC realized long ago that penetration of IT to masses is possible in India only through tools and technologies to overcome the language barrier. For the last 25 years, C-DAC has been pursuing pioneering work in Language Technology and Heritage Computing.

Standards for Indian Languages

C-DAC has helped in establishing standards such as ISCII, Unicode, ISFOC, etc. for Indian language applications on computers and electronic media. It is also working for standardization activities of W3C (languages on the Web), Internationalized Domain Names, Governance, linguistics formats, storage, input, display fonts, etc.

Application Localization

C-DAC has developed localization frameworks for applications in various domains like banking, insurance and finance, administration, travel, etc. C-DAC has provided support for localization of various kinds of applications such as web applications, desktop based applications, localized browser solutions, etc. Localization of various Open Source software such as Libre office, Firefox, e-Mail client, Multi Protocol Messenger, Content Management System and Operating System (Linux) is ongoing activity at C-DAC. Currently, localized versions of these software systems are available for all 22 official languages recognized by the constitution of India. Localized software can be downloaded from *http://ildc.in*

C-DAC has also developed localization frameworks for conversion of data. The data may be in the form of a database such as Oracle, MS SQL Server, MySQL, IBM DB2 or MS Access; or in the form of a MS Excel or MS Word document. The framework includes database translator, acronym handler, number-to-word conversion, date-time conversion and address field conversion routines.

Machine Aided Translation

C-DAC has a number of solutions for translation of English to Hindi and other Indian languages like Assamese, Bangla, Malayalam, Nepali, Punjabi, Telugu and Urdu. A number of software solutions have been developed in this area including MANTRA, AnglaBharati and MaTra. C-DAC has also developed translation systems for cognate languages like Hindi, Urdu and Punjabi. Some of these solutions have been developed as part of multi-institution consortium projects. These systems differ in their underlying approach to translation, translation language pairs and the applicable domains. Given the complexity of automated translation, a machine aided approach is followed in all the solutions. Translation systems are hosted on *http://translation.tdil-dc.in*. A mobile based translation application is available as well from *http://mgov.gov.in*.

Multilingual Semantic Search

Addressing the need for dissemination of digital information across languages, C-DAC has developed NLP based multilingual semantic search engines such as:

- Quester - An NLP based semantic search tool capable of processing various information modes (text, audio, video, etc.) as Enterprise Search with Database Management System (DMS)
- Government of India (GoI) Directory - A Scalable search platform for GoI websites in association with National Informatics Centre (NIC)

Optical Character Recognition (OCR)

As part of a consortia based initiative comprising of 11 institutes in India including C-DAC, an integrated OCR system is being developed for Bangla, Devanagari, Malayalam, Gujarati, Telugu, Tamil, Oriya, Tibetan/Nepal, Gurumukhi and Kannada with font and point-size independent recognition capability. The supported formats in OCR are .tiff, .png and .bmp.

Online Hand Written Recognition (OHWR)

Go-Write is an online handwriting recognition system developed by C-DAC, which can run on tablets as well as on smart phones and can input Hindi characters.

Speech Technologies

C-DAC is working in the area of speech recognition and synthesis. Some of the major technologies/solutions are:

- Text-to-Speech for Hindi, Malayalam, Bangla, Mizo and Nepali
- ShrutLekhan-Rajbhasha: Hindi Speech-to-Text (speaker-independent continuous speech recognition system)
- Pravachak-Rajbhasha: Hindi Text-to-Speech speech synthesis
- Shruti Drishti: An integrated Text-to-Speech and Text-to-Braille system
- ASR (Automatic Speech Recognition) system for Hindi, Bangla and Malayalam
- Speech-to-Speech translation system: An outcome of international consortium, Universal Speech Translation Advanced Research (U-STAR), catering to various languages of the world

Natural Language Processing (NLP)

C-DAC has developed core NLP strength as part of Artificial Intelligence expertise spanning over 20 years in products and solutions like Spell Checkers, Homophone Engines, Visual Thesaurus, Dictionaries, Grammar Checkers, Online Dictionaries and Kosh, Lemmatizer, Data De-duplication solutions (Namescape), etc.

Media Lab Asia
Shruti Drishti
श्रुति दृष्टि
including the excluded
Version 1.0
Based on Vachantar Technology
Browsing through Listening of Conference Websites

C-DAC has developed various technologies in NLP such as transliteration technologies, NLP based translation, speech and search technologies for Indian languages as well as Perso-Arabic script.

Indian Languages on Embedded Devices

Right from the development of the first commercial Application Specific Integrated Circuit (ASIC) in India, namely GIST 9000, for processing Indian languages in the 1980s, C-DAC has focused on all aspects of embedded devices to proliferate use of Indian languages. C-DAC has enabled Indian language scripting solutions on Pagers, Set-Top-Boxes, Dot Matrix Printers, Line Printers, Handheld devices, Digital cameras, etc. C-DAC has also developed solutions to enable use of Indian languages on mobile phones, allowing users to send SMS and e-mail in their own languages. The input mechanism is available both in predictive and non-predictive forms.

Android based keyboards for Aakaash tablet and Android devices have been developed and are available for download from *http://mgov.gov.in*

Fonts, Corpora and Dictionaries

C-DAC has developed several True Type Fonts (TTFs) and Open Font Format for various Indian languages. For UNICODE support in various applications, C-DAC has developed Open Type Fonts for all the available scripts in each of the 22 official Indian languages. Over 8000 fonts consisting of True Type, Open Type and Bitmap font-formats have been produced so far.

In language computing, corpora play a major role. Aligned corpora provide the basis for extraction of various linguistic resources, and are useful for building translation memory, cross-language information retrieval systems, terminology extraction, etc. C-DAC has also developed dictionaries in collaboration with the Language Boards and Academies of different linguistic regions in India.

C-DAC has developed speech corpora along with text for three Eastern Indian Languages, viz., Bangla, Assamese and Manipuri.

Indian Language Tools

In order to enable development of Indian language applications with greater ease, C-DAC has developed a plethora of tools, some amongst them are mentioned below:

- Intelligent Script Manager (ISM)
- Name translation tool from English to Indian Languages
- Indian language Software Development Kit (SDK)
- iPlugin (Web based development tool for Indian languages)

C-DAC has also developed award winning word processing systems such as iLEAP, LEAP Office and ISM, which have brought computing to Indian homes.

Indian Language for Media

C-DAC has pioneered development of multilingual technologies for media and broadcast. Porting of Indian languages onto the video medium has enabled millions of Indian viewers to watch films sub-titled in their language of choice. Tickers and banners on TV stations, news channels, cable operators, etc. use C-DAC's fonts and technologies. LIPS Live (Language Independent Program Subtitles on air) technology has enabled subtitling of movies and programs in Indian languages. Several products like MOVE CG 2001, simplify the subtitling of video programs with high resolution aesthetic fonts. MultiPrompter is a solution for teleprompting in Indian languages for TV channels that are thriving in the Indian subcontinent. MPEG-II based DVD authoring products are now available. Electronic Programming Guides (EPGs) on DTH networks are localized with the help of C-DAC technologies.