

Content Development in Regional Languages: Experiences from Vikaspedia Portal

C. Kathiresan
Principal Technical Officer
C-DAC, Hyderabad

About C-DAC

- Centre for Development of Advanced Computing
- Scientific Society under Ministry of Communications & Information Technology, Government of India
- Primarily an R&D Institution on advanced computing; Started in 1988
- India's first 'Super Computer' – PARAM in 1991
- PARAM Yuva-II supercomputer being ranked 1st in India, 9th in the Asia Pacific Region and 44th in the world among the most power efficient computer systems

C-DAC's core Research Areas

- High Performance Computing & Grid Computing
 - Multilingual Computing
 - Professional Electronics & Ubiquitous Computing*
- Software Technologies, including FOSS*
 - Cyber Security* & Cyber Forensics
 - Health Informatics
- Education & Training*

Local Language content usage in the web

- India - 418 languages; 22 official languages
- Population over 1.2 billion
- 910 million mobile users (TRAI Report, May 2014)
- 243 million internet users in India
- 64% of rural users search content in Local Languages
- 25% of urban users search content in Local Languages

India, being a Non-English speaking country, has more than 1.5 million websites in English, an insignificant 20,000+ in various Indian languages.

Source: Internet & Mobile Association of India (IAMAI) & Indian Market Research Bureau (IMRB)

Analysis of websites - Key Findings

- ✓ Most of the websites (80%) are available in English only
- ✓ 20% of the websites host bilingual content (Hindi/regional language)
- ✓ User Interactive features such as discussion forum, feedback forms, content search facility, etc) are minimal in government websites

- ✓ The scope of content is limited to the particular department / institution with little cross-linking / provision of useful links.

(Source: Study conducted by InDG team, C-DAC, Hyderabad in July 2012)

Analysis of websites of line Ministries (Central & State Government)

	Category	No of websites studied
1	Ministry level websites	55
2	Department level websites	401
3	Knowledge Portals	61
4	Others (websites of institutes, commissions, cooperatives, Universities etc..)	235
Total		752

Source: Study conducted by InDG team, C-DAC, Hyderabad in July 2012

Search for regional language content increasing...internet usage in Mobiles

1. In rural India, entertainment, social networking and email remain the primary purpose of using online local language content. Among these users, 27% of the users use Hindi to access online content
2. Among the current users of the Internet, the local language usage penetration among active Internet users is around 42%.
3. Internet users in India could increase by 24% if local language content is provided on the Internet
4. Mobile phone penetration, social media applications and local language content available in the internet are the key drivers for reaching the unreached

Source: Survey by Internet and Mobile Association of India (IAMAI), February 2014

Web based KM Services in Agriculture hosting content in regional languages

- AGMARKNET
- AGRISNET
- Vikaspedia
- Rice Knowledge Management Portal
- Agri University Portals
- Farmers' Portal
- Agropedia
- Digital Green- video portal

Mobile based Services implemented in India :

Mobile based services attempted in regional languages

- IKSL (voice)
- Reuters Market Light
- Fisher Friend
- Annapurna
- mKrishi
- Nokia Life Tools
- MOTHER
- Apps from Vikaspedia

Major issues in local language content development

- Complexity in Indian languages
- Issue of understanding fonts
- Understanding the concept of 'Unicode'
- Issue of font conversion
- Awareness on language tools & technologies (Translation, TTS, OCR etc...)
- Issue in digitization of existing print material
- Issue of dependency on proprietary softwares

Character encoding standards

- To display content correctly, the browser must know what character-set to use.
 - **ASCII** (American Standard Code for Information Interchange)
 - **ACII** (Alphabetic Code for Information Interchange
 - **ISCI** (Indian Script Code for Information Interchange)
 - **UNICODE** (Universal Character Encoding Standard)
- **Unicode**
- Unicode is an industry standard character set encoding developed and maintained by the Unicode Consortium.
- Unicode provides a unique number for every character
 - no matter what the platform
 - no matter what the program
 - no matter what the language.
- Unicode is not a font. Fonts are built on Unicode Standard
- **Unicode**
- The Unicode Standard has been adopted by all industry leaders including Apple, HP, IBM, Microsoft, Oracle, SAP, Sun, Google etc..
- It is supported in many operating systems, all modern browsers, and many other ICT products

	090	091	092	093	094	095	096	097
0	ी	ऐ	ठ	र	ी	ॐ	ॠ	ॡ
1	ँ	ऑ	ड	र	ऌ	ॠ	ॡ	ँ
2	ं	ओ	ढ	ल	ॡ	ॢ	ॣ	।
3	ः	ओ	ण	ळ	॥	॥	॥	॥
4	ॐ	औ	त	ळ	॥	॥	॥	॥
5	अ	क	थ	व	ॠ	ॡ	ॢ	ॣ
6	आ	ख	द	श	॥	॥	॥	॥
7	इ	ग	ध	प	॥	॥	॥	॥
8	ई	घ	न	स	॥	॥	॥	॥

- Each character is assigned with a unique Numeric value & Name
- Provides the capacity to encode all of the characters used for the written languages of the world

Rendering of scripts by computer

Script	Example of rendering		Mac OS X inbuilt support				Linux or BSD inbuilt support		Windows inbuilt support						
	Correct rendering	Your computer	OS 10.3	OS 10.4	OS 10.5/6	OS 10.7/8	KDE with Qt	GNOME with Pango	9x/ME or NT	2000	XP or 2003	XP SP2	Vista	7	8/8.1
Devanagari	क + ि → कि	क + ि → कि	Yes	Yes	Yes	Yes	Yes	Yes	no	Yes	Yes	Yes	Yes	Yes	Yes
Tamil	க + ி → கி	க + ி → கி	needs font	Yes	Yes	Yes	Yes	Yes	no	Yes	Yes	Yes	Yes	Yes	Yes
Gujarati	ક + િ → કિ	ક + િ → કિ	Yes	Yes	Yes	Yes	Yes	Yes	no	no	Yes	Yes	Yes	Yes	Yes
Gurmukhi	ਕ + ਿ → ਕਿ	ਕ + ਿ → ਕਿ	Yes	Yes	Yes	Yes	Yes	Yes	no	no	Yes	Yes	Yes	Yes	Yes
Kannada	ಕ + ಿ → ಕಿ	ಕ + ಿ → ಕಿ	needs font	needs font	needs font	Yes	Yes	Yes	no	Yes	Yes	Yes	Yes	Yes	Yes
Telugu	క + ి → కి	క + ి → కి	needs font	needs font	needs font	Yes	Yes	Yes	no	no	Yes	Yes	Yes	Yes	Yes
Bengali, Assamese	ক + ি → কি	ক + ি → কি	needs font	needs font	needs font	Yes	Yes	Yes	no	no	no	Yes	Yes	Yes	Yes
Malayalam	ക + ി → കി	ക + ി → കി	needs font	needs font	needs font	Yes	Yes	Yes	no	no	no	Yes	Yes	Yes	Yes
Tibetan	ཀ + ཨ + ཨ + ཨ → ཀྨྨྨ	ཀ + ཨ + ཨ + ཨ → ཀྨྨྨ	needs font	needs font	Yes	Yes	needs font	needs font	no	no	no	needs patch	Yes	Yes	Yes
Sinhala	ඵ + ඵ → ඵඵ	ඵ + ඵ → ඵඵ	needs font	needs font	needs font	Yes	Yes	Yes	no	no	no	needs font	Yes	Yes	Yes
Oriya	କ + ି → କି	କ + ି → କି	needs font	needs font	needs font	Yes	Yes	Yes	no	no	no	needs font	Yes	Yes	Yes
Thai	ก + ิ → กิ	ก + ิ → กิ	Yes	Yes	Yes	Yes	Yes	Yes	no	Yes	Yes	Yes	Yes	Yes	Yes
Lao	ກ + ື + ື → ືກ	ກ + ື + ື → ືກ	needs font	needs font	needs font	Yes	Yes	Yes	no	no	no	needs font	Yes	Yes	Yes
Khmer	ក + វ + ្រ + ្រ → ក្រវ្រ	ក + វ + ្រ + ្រ → ក្រវ្រ	needs font	needs font	needs font	Yes	Yes	Yes	no	no	no	needs font	needs font	Yes	Yes
Burmese	ယ + ိ → ယ	ယ + ိ → ယ	needs font	needs font	needs font	Yes	no	needs font	no	no	no	needs font	needs font	needs font	Yes

Speech Recognition Technologies from C-DAC

- Text-to-Speech for Hindi, Malayalam, Bangla, Mizo and Nepali
- Shruti Drishti : An Integrated Text-to-Speech and Text-to-Braille System
- ASR (Automatic Speech Recognition) System for Hindi, Bangla and Malayalam

C-DAC, Hyderabad

Language tools for mobiles

- [GIST, C-DAC](#)
 - GIST Android basic editor
 - GIST Android Soft Keyboards (Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Oriya, Punjabi, Tamil, Telugu and Urdu)

Domain names in Indian Languages

- Internationalized Domain Names (IDNs) are domain names represented in local language characters. Such domain names could contain letters or characters from non-ASCII scripts.
- In India, the IN Registry is also responsible for registering IDN in Indian Languages.
- Registration at - <https://www.registry.in/node/883>

Technologies developed for Indian Languages - Online resources

- [C-DAC Language Tools](#)
- [Technology Development for Indian Languages \(TDIL\)](#)
- [Microsoft Indic Language support](#)

Vikaspedia Portal

- www.vikaspedia.in
- Aimed at creating a collective knowledge repository
- Focus on social development sectors
- At present 10 languages (Assamese, Hindi, Marathi, Gujarati, Bengali, Kannada, Tamil, Telugu, Malayalam and English);
- To be made in 22 languages
- At present 6 domains (Agriculture, Health, Education, Energy, Social Welfare and e-Governance)

Vikaspedia Platform

- Content Management System - Plone
- Database- Zope
- Content uploading/Reviewing/Publishing
- Multilingual

Interactive features (Discussion forum, Polls, Comments, Feedbacks etc..)

www.plone.org

Collaborative Content Creation

- **‘Crowdsourcing Model’** inviting community for active contribution of content in their own language
- Any individual/volunteer can contribute content, edit / comment on the existing content. Similar to Wikipedia, but, with more reliable and authentic content in local languages specific to the region
- Contributed content is validated by identified experts and moderated by State Nodal Agency in the respective states.

Vikaspedia Portal

User of this portal can

- ❑ **Contribute** new content / topic
- ❑ **Edit** the existing content / topic
- ❑ **Comment** on the available information
- ❑ **Participate** in the discussions
- ❑ **Share** the information in the social media

Information Services

Information / Knowledge in local languages

(Success stories, technologies, best practices, Government schemes etc..)

- Multimedia products
- e-Learning Courses
- Value Added Services (e-Vyapar, Ask an Expert etc.)
- Mobile Apps related to key livelihood sectors

Vikaspedia currently hosts

- **Web pages** – 7000 (~ 70,000 A 4 pages)
- **Multimedia content**
 - Audio- 2330 mins, Video -13,923 mins, Flash -12.5 GB
- **Online services**
 - Ask an Expert – an online advisory application
 - MOTHER, Mobile based maternal health awareness
 - GK quiz – a self assessment app for students
 - e-learning courses – multilingual, web based
 - Recaller – web based alerts on financial products

Mobile apps - 8

State Nodal Agencies (SNAs)

- Managing the regional portal
- Coordinating the collaborative content creation
- Organises awareness programmes and trainings
- Support the volunteers on technical issues



Outreach

- Orientation and training programmes for SNAs
- State level content mapping workshops
- State expert group meetings
- District level capacity building workshops
- Community level events with CSCs
- Vikaspedia Awareness programmes for development stakeholders
- Coverage of events in print and mass media



Dept of Agriculture, Andhra Pradesh - adopted this service for AGRISNET

Ask An Expert – Multilingual

- ✓ Piloted in 6 domains with 14 experts
- ✓ Queries posted – 2038; Answered – 2029
- ✓ Dept of Agriculture, AP- adopted this service for AGRISNET

Mobile Applications Download

Mobile Applications (Android Based)



Ask An Expert
Online expert solution
Downloads:846
[Know more.....](#)



CSC Finder
Common Services Centers
Information
Downloads:561
[Know more.....](#)



KVK Khoj
Agricultural Science Centre
finder
Downloads:499
[Know more.....](#)



MOTHER
MOBILE based maTernal
HEalth awaReness
Downloads:340
[Know more.....](#)



SHELTOR
Student HostEL locaTOR
Downloads:220
[Know more.....](#)



Diabetes Tool
Diabetes Risk Assessment
Tool
Downloads:36
[Know more.....](#)



Vikaspedia Browser
Browser app for vikaspedia
portal
Downloads:78
[Know more.....](#)



Right Weight Checker
Body Mass Index Calculator
Downloads:44
[Know more.....](#)

Opportunities

- Support from government for ICT based Knowledge Management – Mobile Seva service from Ministry of Communications & IT (free bulk SMS for Govt organisations) – www.mgov.gov.in
- Scope for partnership with other stakeholders including private
- Growing broadband connectivity at rural level
- Affordability of the service (decreasing cost due to competition among the service providers)