State Level Awareness Workshop for the state of Kerala on e-Governance Standards & Guidelines, 20th -21st Jan 2023

C-DAC, TechnoPark, Thiruvananthapuram

Localization and Language Technology Standards

Chandrakant Dhutadmal
Joint Director, C-DAC, Pune
chandrakantd@cdac.in, Mobile - 9823050212

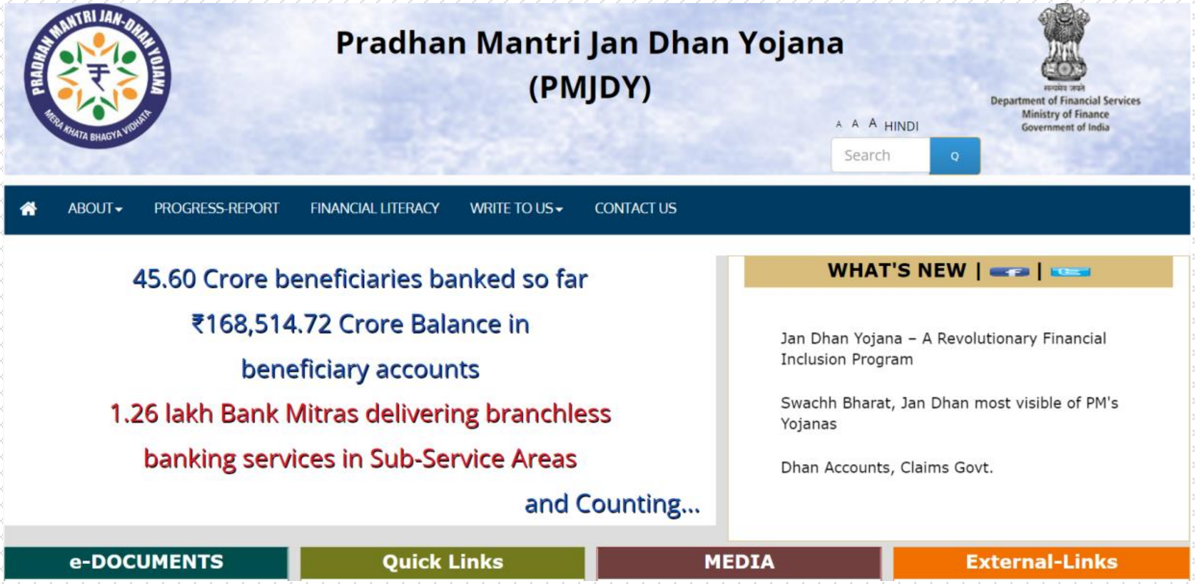
Standards ensure seamless sharing of information & interoperability of e-Governance applications

Outline of the Presentation

- Background and Importance of Localization
- Basics of Indian Language Computing
- Linguistic scenario in India
- Working with English vis-à-vis Indian Languages
- Standards & Guideline for Language Computing
- Availability of tools for Indian languages
- Database and Website search
- Localization Standards and Best Practices
- Localization Guidelines Minimal Set - Recommendations

Background – Importance of Localization

- Most computer systems, software solutions and devices even today are designed and developed keeping English as primary language.
- Digital India aims to transform India into a digitally empowered society and knowledge economy.
- Also, aims to ensure that the Government's services are made available to citizens electronically.
- More than 90% of the population of the country are Non-English speakers.
- Attempt to reach out to these 90% population through localisation of applications, data, reports, code, services, devices...... for Indian Languages



Pradhan Mantri Jan Dhan Yojana (PMJDY)

A A A HINDI

Search

Department of Financial Services
Ministry of Finance
Government of India

ABOUT ▾ PROGRESS-REPORT FINANCIAL LITERACY WRITE TO US ▾ CONTACT US

45.60 Crore beneficiaries banked so far

₹168,514.72 Crore Balance in beneficiary accounts

1.26 lakh Bank Mitras delivering branchless banking services in Sub-Service Areas

and Counting...

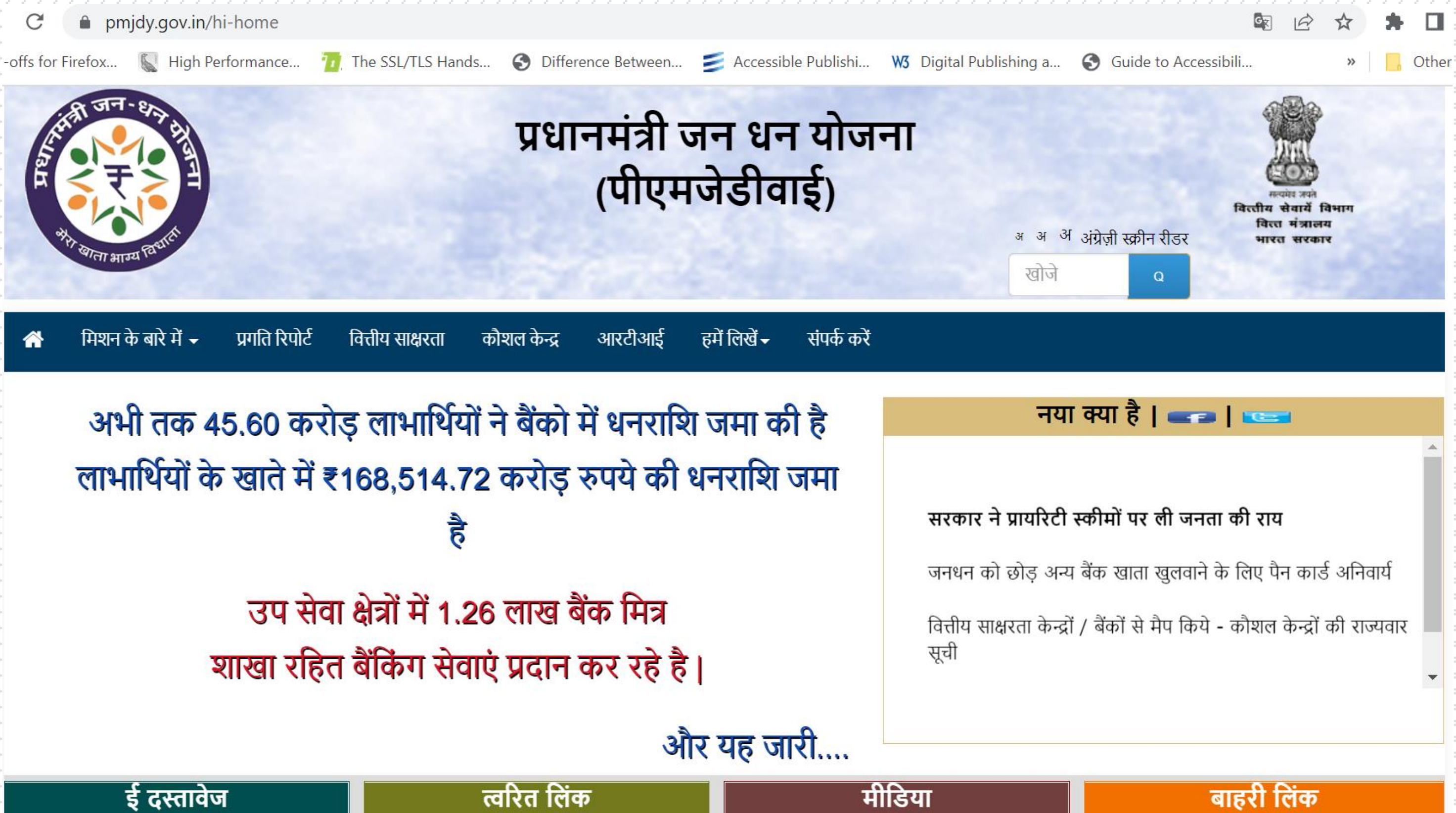
WHAT'S NEW |

Jan Dhan Yojana – A Revolutionary Financial Inclusion Program

Swachh Bharat, Jan Dhan most visible of PM's Yojanas

Dhan Accounts, Claims Govt.

e-DOCUMENTS | Quick Links | MEDIA | External-Links



pmjdy.gov.in/hi-home

-offs for Firefox... High Performance... The SSL/TLS Hands... Difference Between... Accessible Publishi... Digital Publishing a... Guide to Accessibili... » Other

प्रधानमंत्री जन धन योजना (पीएमजेडीवाई)

अ अ अ अंग्रेज़ी स्क्रीन रीडर

खोजे

वित्तीय सेवायें विभाग
वित्त मंत्रालय
भारत सरकार

मिशन के बारे में ▾ प्रगति रिपोर्ट वित्तीय साक्षरता कौशल केन्द्र आरटीआई हमें लिखें ▾ संपर्क करें

अभी तक 45.60 करोड़ लाभार्थियों ने बैंको में धनराशि जमा की है

लाभार्थियों के खाते में ₹168,514.72 करोड़ रुपये की धनराशि जमा है

उप सेवा क्षेत्रों में 1.26 लाख बैंक मित्र शाखा रहित बैंकिंग सेवाएं प्रदान कर रहे है |

और यह जारी....

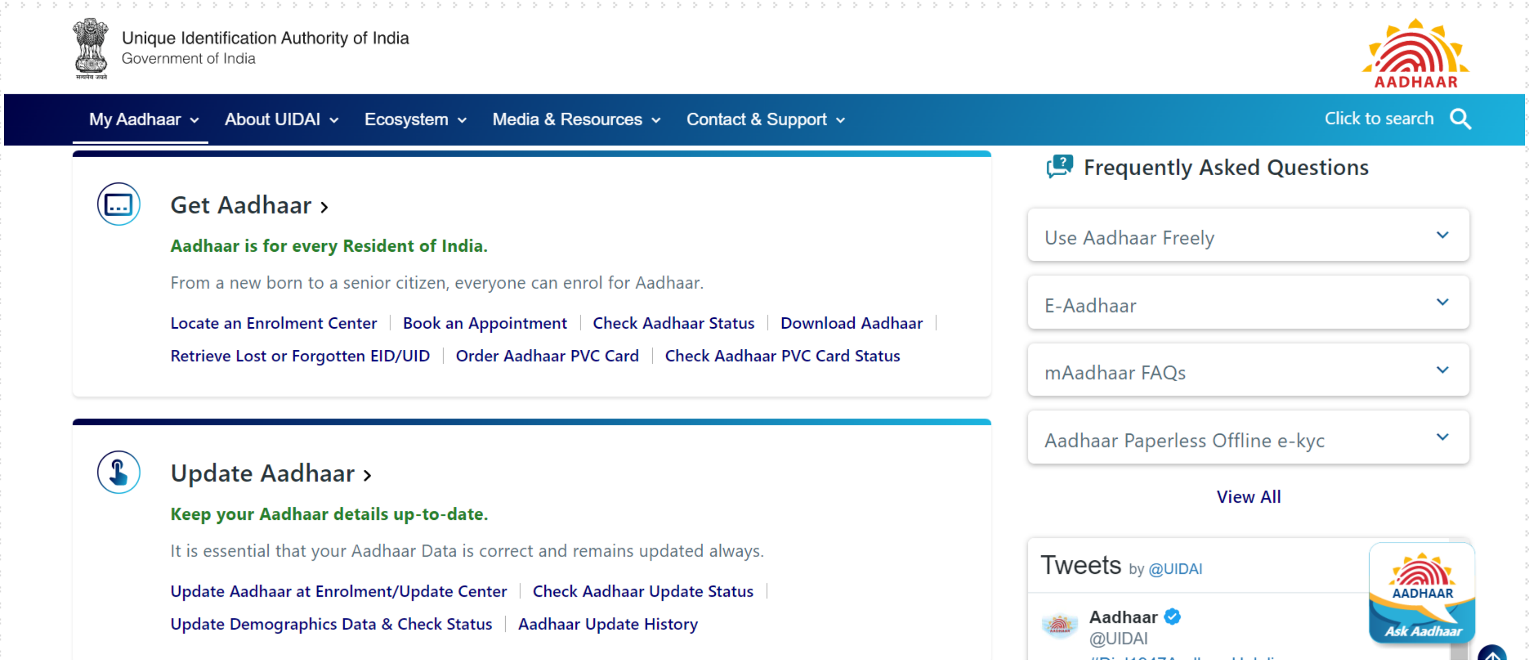
नया क्या है |

सरकार ने प्रायरिटी स्कीमों पर ली जनता की राय

जनधन को छोड़ अन्य बैंक खाता खुलवाने के लिए पैन कार्ड अनिवार्य

वित्तीय साक्षरता केन्द्रों / बैंकों से मैप किये - कौशल केन्द्रों की राज्यवार सूची

ई दस्तावेज | त्वरित लिंक | मीडिया | बाहरी लिंक



Get Aadhaar ›

Aadhaar is for every Resident of India.

From a new born to a senior citizen, everyone can enrol for Aadhaar.

Locate an Enrolment Center | Book an Appointment | Check Aadhaar Status | Download Aadhaar | Retrieve Lost or Forgotten EID/UID | Order Aadhaar PVC Card | Check Aadhaar PVC Card Status

Update Aadhaar ›

Keep your Aadhaar details up-to-date.

It is essential that your Aadhaar Data is correct and remains updated always.

Update Aadhaar at Enrolment/Update Center | Check Aadhaar Update Status | Update Demographics Data & Check Status | Aadhaar Update History

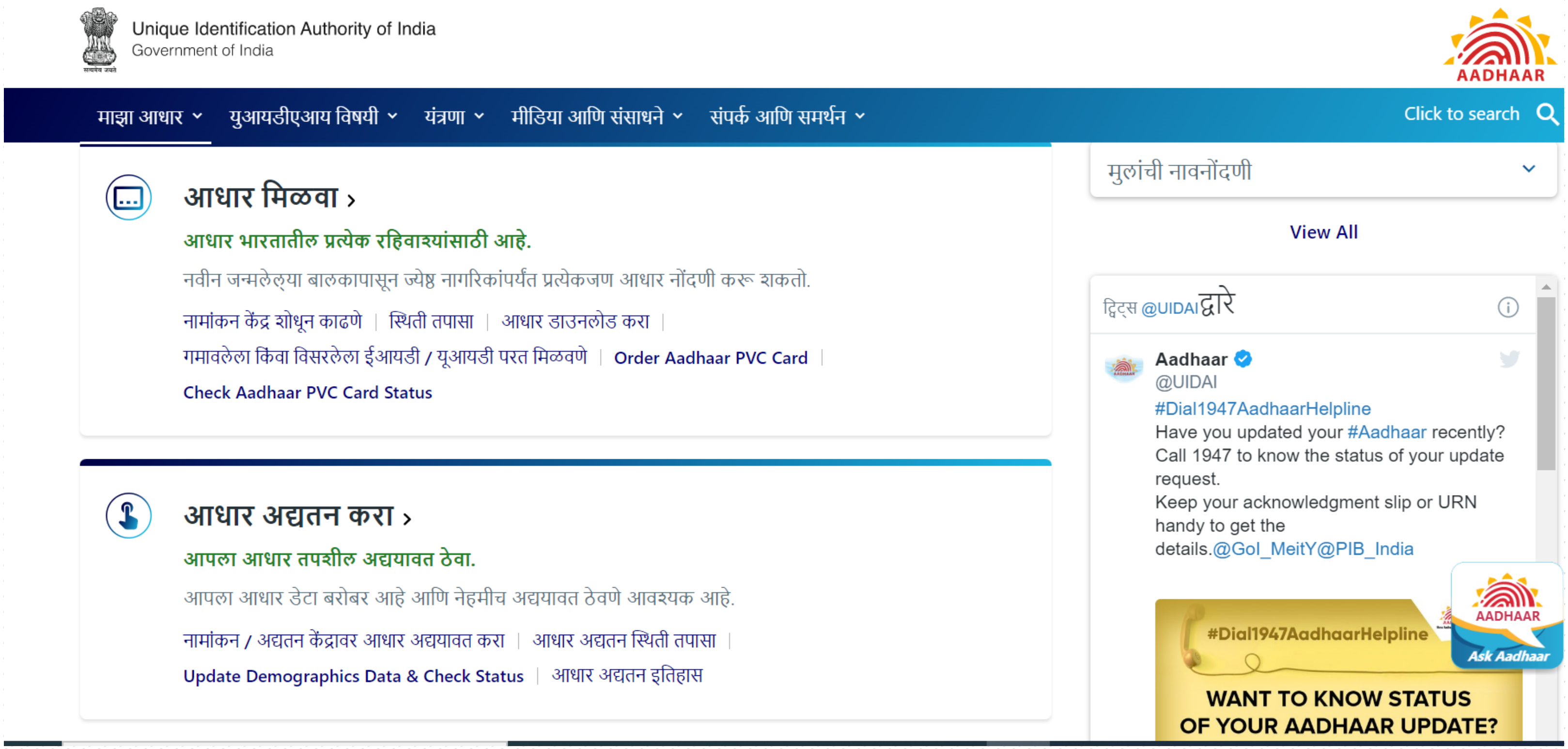
Frequently Asked Questions

- Use Aadhaar Freely
- E-Aadhaar
- mAadhaar FAQs
- Aadhaar Paperless Offline e-kyc

View All

Tweets by @UIDAI

Aadhaar
@UIDAI



आधार मिळवा ›

आधार भारतातील प्रत्येक रहिवाश्यांसाठी आहे.

नवीन जन्मलेल्या बालकापासून ज्येष्ठ नागरिकांपर्यंत प्रत्येकजण आधार नोंदणी करू शकतो.

नामांकन केंद्र शोधून काढणे | स्थिती तपासा | आधार डाउनलोड करा | गमावलेला किंवा विसरलेला ईआयडी / यूआयडी परत मिळवणे | Order Aadhaar PVC Card | Check Aadhaar PVC Card Status

आधार अद्यतन करा ›

आपला आधार तपशील अद्ययावत ठेवा.

आपला आधार डेटा बरोबर आहे आणि नेहमीच अद्ययावत ठेवणे आवश्यक आहे.

नामांकन / अद्यतन केंद्रावर आधार अद्ययावत करा | आधार अद्यतन स्थिती तपासा | Update Demographics Data & Check Status | आधार अद्यतन इतिहास

- मुलांची नावनोंदणी

View All

ट्विट्स @UIDAI द्वारे

Aadhaar
@UIDAI
#Dial1947AadhaarHelpline
Have you updated your #Aadhaar recently?
Call 1947 to know the status of your update request.
Keep your acknowledgment slip or URN handy to get the details.@GoI_MeitY@PIB_India

Basics of Indian Language Computing

All systems can be broken down into three parts

- Inputting
  - First INSCRIPT keyboard standard was published by BIS in 1991 (IS 13194: 1991)
  - Enhanced INSCRIPT keyboard layouts (IS 16350: 2016).
  - Indian Language Support for Mobile Phones as per IS 16333 (Part 3) : 2016
  - Phonetic / Transliteration, Typewriter
- Storage / Processing
  - UNICODE de-facto standard for storage of Multilingual text. Current version is 15.0 releases on 13th Sept 2022.
- Output
  - UNICODE compliant Open Font format Fonts

Applications for Indian Languages should have support throughout the lifecycle of the system – rather than being an after thought.

Enhanced INSCRIPT keyboard layouts (IS 16350: 2016)

- Bureau of Indian Standards (BIS), the national standards body of India has published a standard on ‘Enhanced Inscript keyboard layouts’ (IS 16350: 2016) on 07th June 2016.

- The standard covers code charts, characters and character names for 11 Indian scripts which cover all 22 official languages of India which are identical to ISO/IEC 10646 (Universal Coded Character Set).

- Additionally, this standard gives enhanced versions of the Inscript keyboard layouts and mapping of the code points with the respective keyboard layouts.

- Further, this standard is also referred in published standards IS 16333: 2016 Indian Language Support for Mobile Phone Handsets

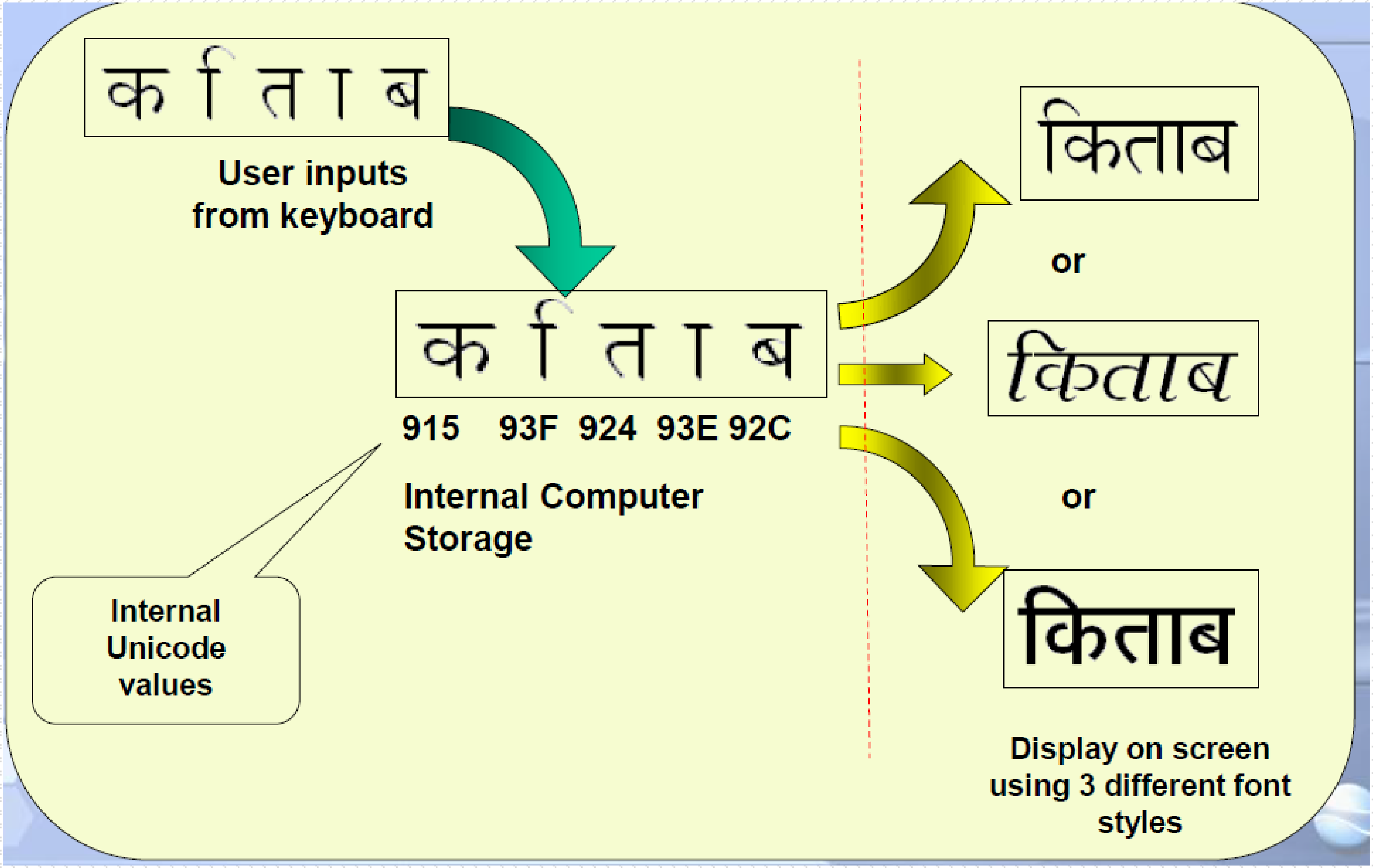
Indian Language Support for Mobile Phones as per IS 16333 (Part 3) : 2016

- Indian Language Support for mobile phones as per IS 16333 (Part 3): 2016 was mandated by MeitY with effect from 1 July 2017 vide order dated 24 October 2016 bringing it under the provisions of "Electronics and Information Technology Goods (Requirements of Compulsory Registration) Order, 2012".

- As per the new mandate, all mobile phones in the country will have to provide text reading support in all Indian languages on their devices.

- The standard mandates mobile phone companies to provide message typing facility in English, Hindi and a regional language of their users choice.

Akshara (Syllable) Formation



क ि त ा ब
User inputs from keyboard
क ि त ा ब
915 93F 924 93E 92C
Internal Computer Storage
Internal Unicode values
किताब
or
किताब
or
किताब
Display on screen using 3 different font styles

Indian Language Nuances and complexities

Alternate Spellings

हिंदी : ह + ◌ि + ◌ं + द + ◌ी

हिन्दी : ह + ◌ि + न + ◌् + द + ◌ी

तसवीर : त + स + व + ◌ी + र

तस्वीर : त + स + ◌्+ व + ◌ी + र

Alternate forms

अन्न : अ + न + ◌् + न

अन्‌न : अ + न + ◌् + ◌् + न

अन्‍न : अ + न + ◌् + INV (ZWJ) + न

बिट्टु : ब + ◌ि + ट + ◌् + ◌् + ◌् + ट + ◌ु

बिट्‌टु : ब + ◌ि + ट + ◌् + ट + ◌ु

ट्‌ठ : ट + ◌् + ठ : ट्ठ : ट + ◌् + ◌् + ◌् + ठ

द्‌ध : द + ◌् + INV + ध : द्ध : द + ◌् + ध

ड्‌ड : ड + ◌् + ड : ड्ड : ड + ◌् + ◌् + ◌् + ड

Reordering

क ि त ा ब

क ि त ा ब

(179+219+194+218+202)

किताब

(ि+क+त+ा+ब)

Linguistic Scenario in India

One script: many languages

- **Devanagari –** Hindi, Marathi, Konkani, Rajasthani, Sindhi, Nepali, Dogri, Santhali, etc.
- Thus the data in Devanagari (code page) can support all languages using that particular script.
- However tools like synonym Dictionaries, spellcheckers, and search engine crawlers and indexers, etc. are language dependent and require language information along with the data.
- Though the contents would reveal the language used, it would be ideal if a special attribute code to indicate the language is inserted. <p lang="mr-IN">सोलापुरातील डॉक्टरांना मिळणार एक हजार खोल्यांचा नवीन वसतिगृह.</p>

One language: many scripts

- Konkani is written in **Roman, Devanagari, Malayalam and Kannada**.
- Sindhi is written in Gurmukhi (Punjabi), Arabi (Perso-Arabic), Devanagari, Gujarati and also Roman.
- Sindhi has adopted the Perso-Arabic script for representing their language. In case of Konkani, Devanagari is used as official script.

Language	ISO 639 code	Official Language	Family	Script
Assamese	asm	Assam	Indo-Aryan	Assamese
Bengali	ben	Tripura and West Bengal	Indo-Aryan	Bangla
Manipuri	mni	Meitei	Tibeto-Burman	Bangla Meitei-Meyek
Boro	brx	Assam	Tibeto-Burman	Devanāgarī (modified)
Dogri	dgo	Jammu and Kashmir	Indo-Aryan	Devanāgarī (modified)
Hindi	hin	Andaman and Nicobar Islands, Bihar, Chandigarh, Chhattisgarh, Delhi, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttar Pradesh and Uttaranchal	Indo-Aryan	Devanāgarī
Konkani	kok	Goa	Indo-Aryan	Devanāgarī Roman (Latin)
Maithili	mai	Bihar	Indo-Aryan	Devanāgarī
Marathi	mar	Maharashtra	Indo-Aryan	Devanāgarī
Nepali	nep	Sikkim	Indo-Aryan	Devanāgarī

Language	ISO 639 code	Official Language	Family	Script
Sanskrit	san	Pan-Indian	Indo-Aryan	Devanāgarī
Gujarati	guj	Dadra and Nagar Haveli, Daman and Diu, and Gujarat	Indo-Aryan	Gujarati
Punjabi	pan	Punjab	Indo-Aryan	Gurmukhi
Kannada	kan	Karnataka	Dravidian	Kannada
Malayalam	mal	Kerala and Lakshadweep	Dravidian	Malayalam
Santali	sat	Jharkhand	Munda	Ol Ciki
Oriya	ori	Orissa	Indo-Aryan	Oriya
Kashmiri	kas		Indo-Aryan	Perso-Arabic Devanāgarī
Sindhi	snd	Pan-Indian	Indo-Aryan	Perso-Arabic Devanāgarī Gujarati Roman (Latin)
Urdu	urd	Jammu and Kashmir	Indo-Aryan	Perso-Arabic
Tamil	tam	Tamil Nadu and Pondicherry	Dravidian	Tamil
Telugu	tel	Andhra Pradesh	Dravidian	Telugu

Working with English

Inputting

- The keys on the keyboard are mapped to ASCII characters.
- One to One mapping between keys and the English characters

Display

- The glyph representing the character pressed is displayed.
- The English font contains the glyphs at the position specified by the ASCII character set.
- One to One mapping between the characters and the glyphs

Example: Hi = H + I

Storage

- The ASCII value of the characters is stored.

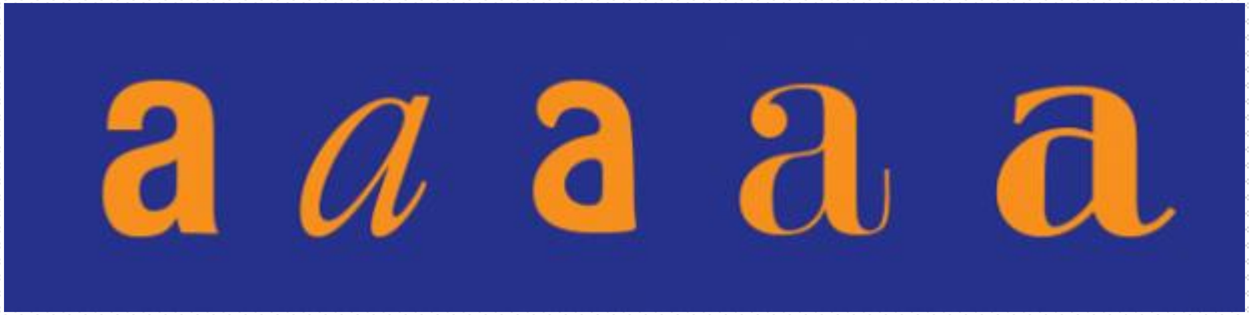
Printing

- The glyphs representing the ASCII characters are printed.

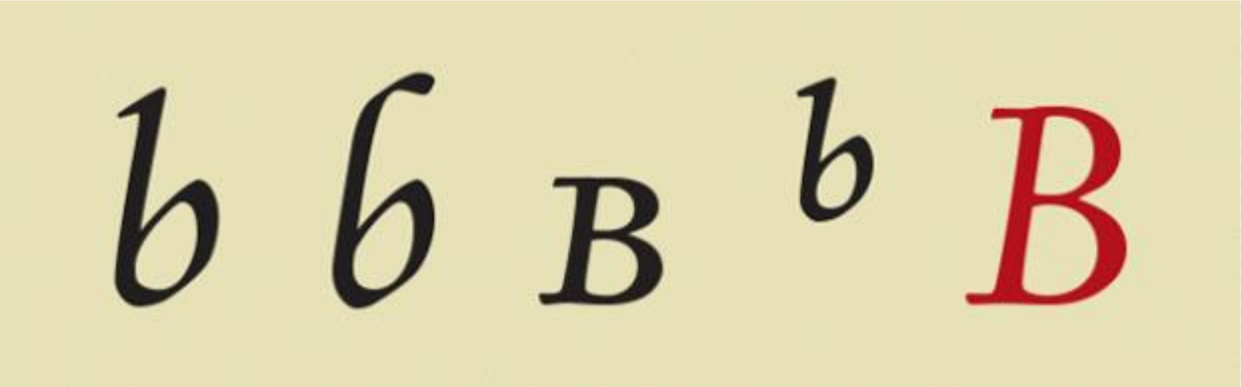
Character vs. Glyph

A character is the symbol representing a letter.

A glyph is the specific shape, design, or representation of a character.



Collection of glyphs representing the character "a".



(left) The lowercase "b" character is represented by four glyphs in Jenson italic: the standard b, an alternate, a small cap, and the superscript. (right) The capital "B" is a different character.

Complexity of Indian Languages

- Character Set
  - Consonants (k, kh, g, gh)
  - Vowels (Ae, e, E)
  - Vowel Sign (Matras)
  - Vowel Modifiers (Chandrabindu, anuswar, visarg)
  - Others (Halant, Nukta)

- Shape of a character does not remain constant

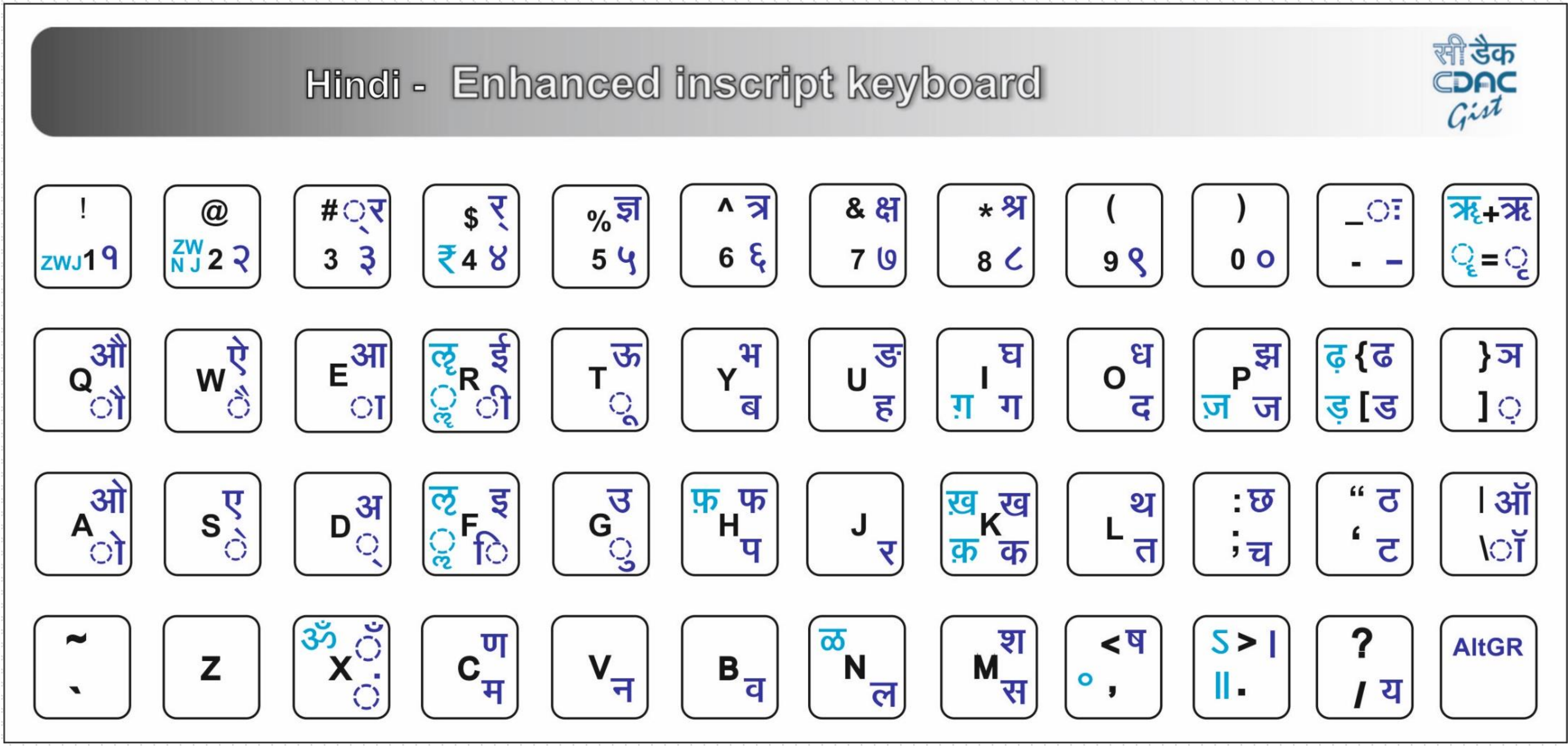
Example: क्या

According to Devnagari Script Rules क + ् + य + ा = क्या

- Here the shape of the क gets modified.
- Hence no one to one mapping between character and its shape.

Working with Indian Languages

Enhanced INSCRIPT keyboard layouts (IS 16350: 2016)



Working with Indian Languages

Inputting

- All the combinations of consonants and vowels cannot be mapped to limited set of keyboard keys.
- So a standard set of characters representing all the basic shapes is defined by Bureau of Indian Standards called as Indian Script Code for Information Interchange (ISCII).
- Each character in the set is assigned a unique value.
- One to one mapping between the keys and the ISCII characters possible.

Working with Indian Languages

Display

- The characters that are inputted through keyboard are mapped to glyphs in the font.
- There is many to many mapping between characters entered and the glyphs displayed due to complexity of Indian Languages.

  - e.g. क्या will be entered as क + ् + य + ा

  क + ् will be represented by single glyph क्

- There might be repositioning of characters before the actual display

  - e.g. किताब will be entered as क + ि + त + ा + ब

  but ि will be displayed before क

Working with Indian Languages

Storage

- The data can be stored in various formats like ISCII, ISFOC, Unicode etc.

Printing

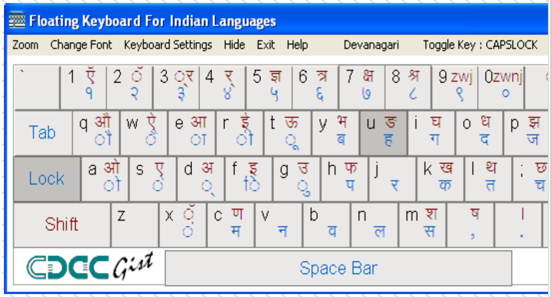
- Printing is also based on storage.
- If the storage is in font code, the font information should be there with the data.
- If the storage is ISCII based, the printer should be enabled to print ISCII else the data should be converted to font code before sending to printer.

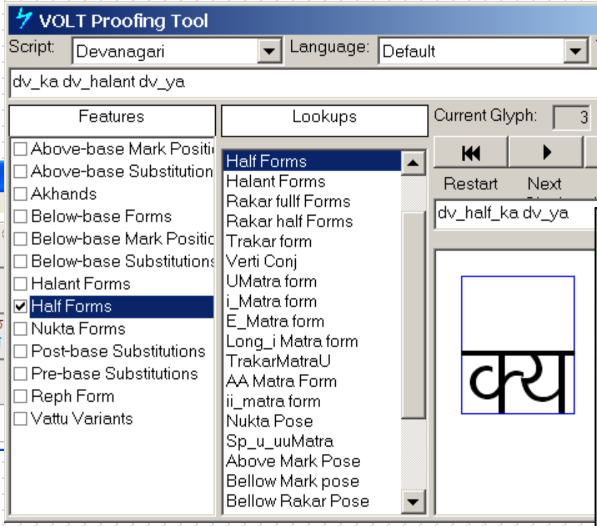
Standard

"a document, established by consensus and approved by a recognized body, that provides, for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context".

Adherence to standards ensures compatibility, safeguarded data, avoids vendor locking, proper exchange of data between various systems, applications, databases, devices, etc.

Indian Language Standards





Script: Devanagari
Language: Default
Text Flow: LTR

dv_ka dv_halant dv_ya

Current Glyph: 3
Current Lookup:

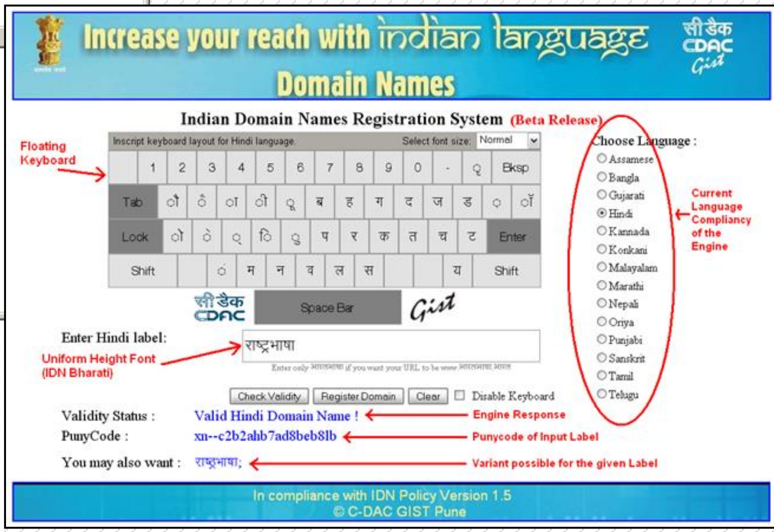
dv_half_ka dv_ya

Features:

- □ Above-base Mark Positi
- □ Above-base Substitution
- □ Akhands
- □ Below-base Forms
- □ Below-base Mark Positic
- □ Below-base Substitutions
- □ Halant Forms
- ☑ Half Forms
- □ Nukta Forms
- □ Post-base Substitutions
- □ Pre-base Substitutions
- □ Reph Form
- □ Vattu Variants

Lookups:

- Half Forms
- Halant Forms
- Rakar fullf Forms
- Rakar half Forms
- Trakar form
- Verti Conj
- UMatra form
- i_Matra form
- E_Matra form
- Long_i Matra form
- TrakarMatraU
- AA Matra Form
- ii_matra form
- Nukta Pose
- Sp_u_uuMatra
- Above Mark Pose
- Bellow Mark pose
- Bellow Rakar Pose


Increase your reach with indian language
Domain Names
CDAC Gist
Floating Keyboard
Uniform Height Font (IDN Bharati)
Current Language Compliancy of the Engine
Engine Response
Punycode of Input Label
Variant possible for the given Label
Bksp
Tab
Lock
Enter
Shift
Space Bar
Check Validity
Register Domain
Clear

Indian Domain Names Registration System (Beta Release)

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

Choose Language :

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

Enter Hindi label: राष्ट्रभाषा

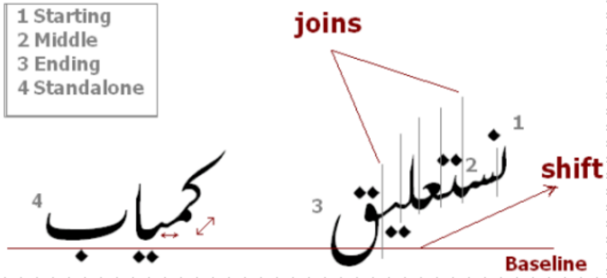
□ Disable Keyboard

Validity Status : Valid Hindi Domain Name !

PunyCode : xn--c2b2ahb7ad8beb8lb

You may also want : राष्ट्रभाषा;

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



UNICODE
IDN – ICANN, IANA
Enhanced INSCRIPT

ISCII ,
INSCRIPT,
PASCII

W3C for IL on Browsers
and Mobile devices

Indian language standards

Storage (UNICODE)

Inputting (INSCRIPT)

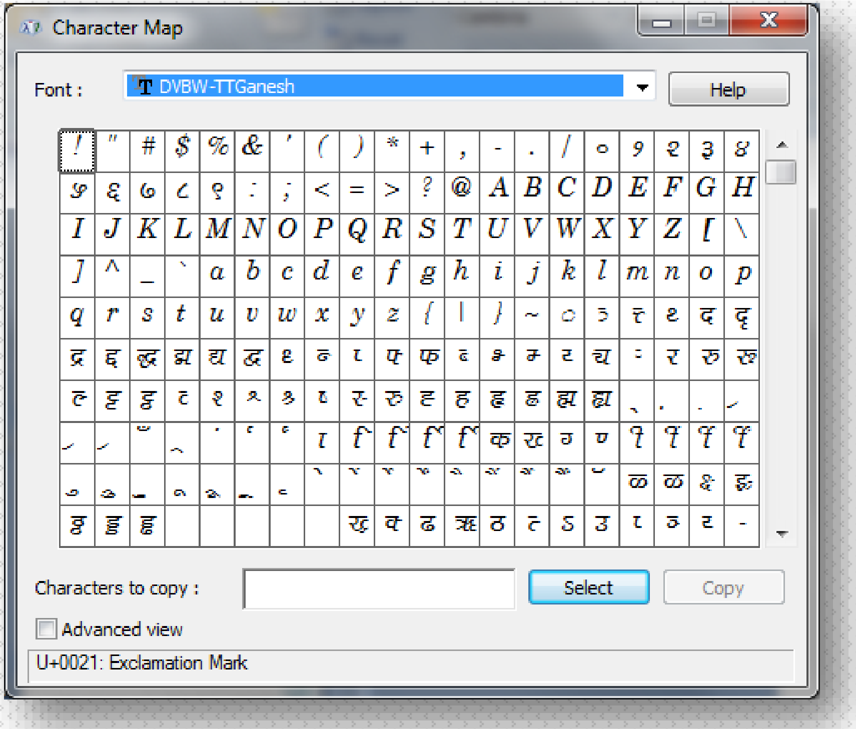
Display (Open Font Format)

Pre-Unicode Era

- Displaying multilingual data would require fonts.

- Font is a set of well defined shapes to display symbols (letters, punctuation marks, special characters of the language).

- An 8-bit font can represent upto 256 glyphs by giving unique index (called glyph index) and name to each glyph/shape.

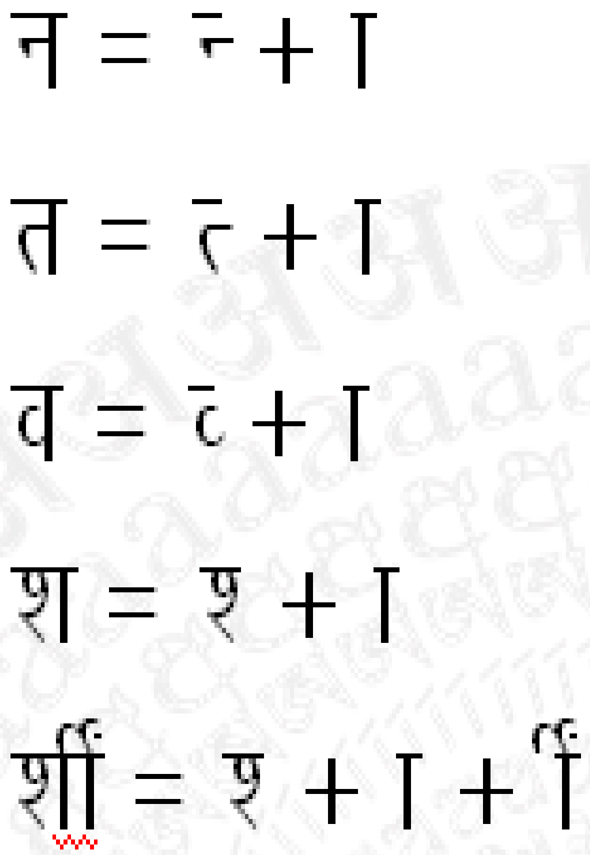


Font : DVBW-TTGanesh

Characters to copy :

U+0021: Exclamation Mark

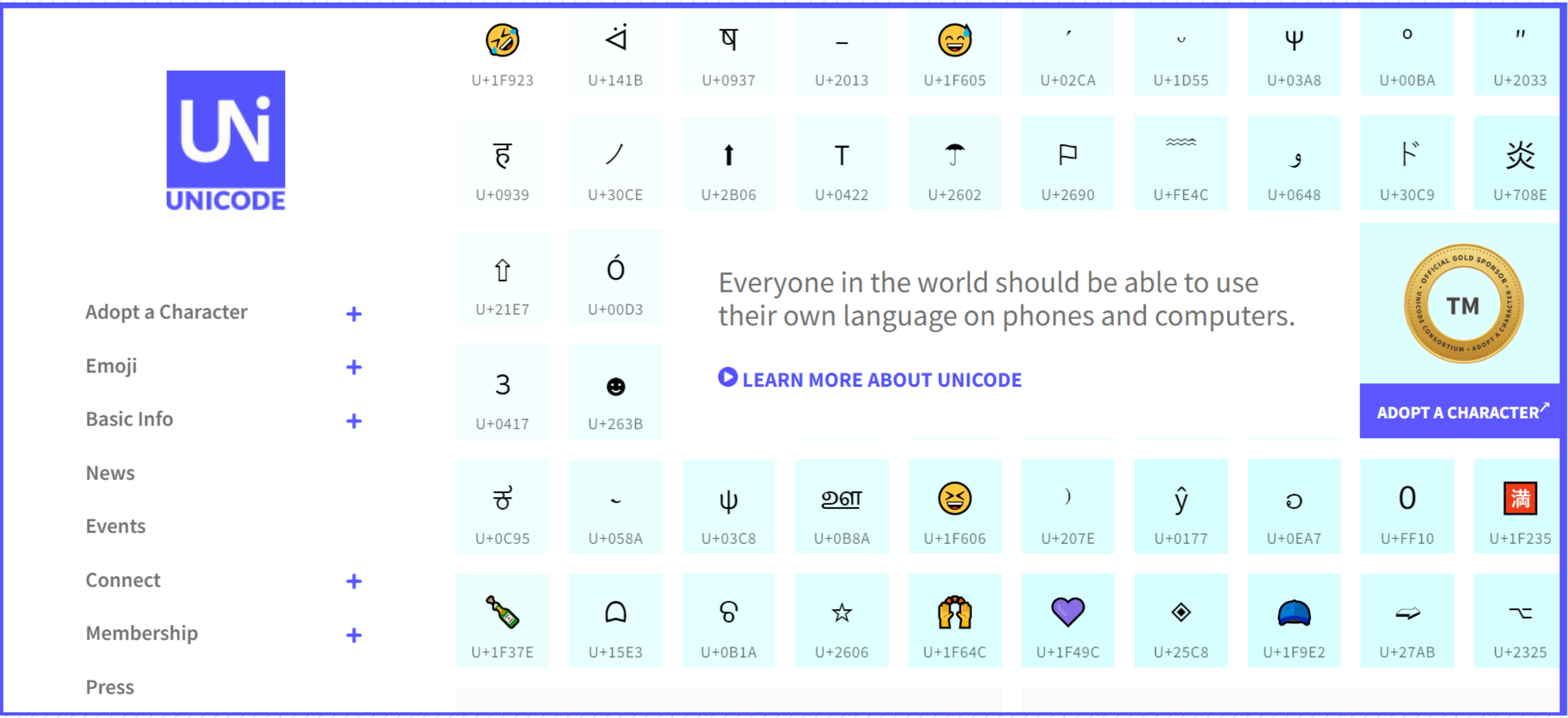
Pre-Unicode Era



UNICODE

- Storage standard
- What ASCII is for English, Unicode is for other languages of the world
- Enables seamless exchange of data – desktops, printers, databases, browsers, devices.

UNICODE



🤣 U+1F923	ᐛ U+141B	ष U+0937	– U+2013	😅 U+1F605	ˊ U+02CA	ᵕ U+1D55	Ψ U+03A8	º U+00BA	‴ U+2033
ह U+0939	ノ U+30CE	⬆ U+2B06	Т U+0422	☂ U+2602	⚐ U+2690	﹌ U+FE4C	و U+0648	ド U+30C9	炎 U+708E
⇧ U+21E7	Ó U+00D3								
З U+0417	☻ U+263B								
ಕ U+0C95	֊ U+058A	ψ U+03C8	ஊ U+0B8A	😆 U+1F606	⁾ U+207E	ŷ U+0177	ວ U+0EA7	０ U+FF10	🈵 U+1F235
🍾 U+1F37E	ᗣ U+15E3	ଚ U+0B1A	☆ U+2606	🙌 U+1F64C	💜 U+1F49C	◈ U+25C8	🧢 U+1F9E2	➫ U+27AB	⌥ U+2325

Everyone in the world should be able to use their own language on phones and computers.

Latest stable version – 15.0

UNICODE

- Unicode consortium defines Unicode as :

- “Unicode is the universal character encoding, maintained by the Unicode consortium. This encoding standard provides the basis for processing, storage and interchange of text data in any language in all modern software and information technology protocols.”

- It is the superset of all the languages in the world which also includes punctuation, special characters (shapes), currency symbols, mathematical symbols etc. Using Unicode, more than 65000 different characters can be represented. Unicode comprises of many code charts.

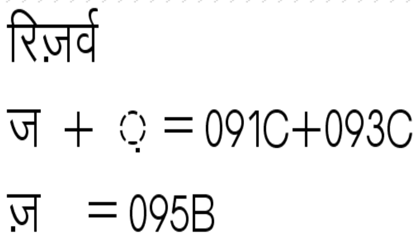
- The Unicode code charts can be referred at: http://www.unicode.org/charts

UNICODE

- Various editors / applications / development environments / databases / browsers need to understand how to read in the given Unicode data and interpret the same.
- Various encoding schemes to represent Unicode are UTF-8, UTF-16, UTF-32 with a combination of endian-ness. (Little & Big Endian -two ways of storing multi-byte data-types)
- There are normalization rules which are required to be followed for data compatibility between various applications / underlying environment.
- Non adherence to some of these may lead to wrong interpretation of data and will also pose problems in searches as well.

Normalization in Unicode

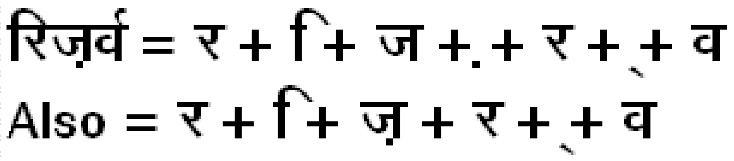
- The Unicode data requires normalization.
- There are many cases where a character can be entered in more than one ways.
- If application or database does not normalize, searching becomes difficult.



रिज़र्व

ज + ़ = 091C+093C

ज़ = 095B



रिज़र्व = र + ि + ज + ़ + र + ् + व

Also = र + ि + ज़ + र + ् + व

Character semantics

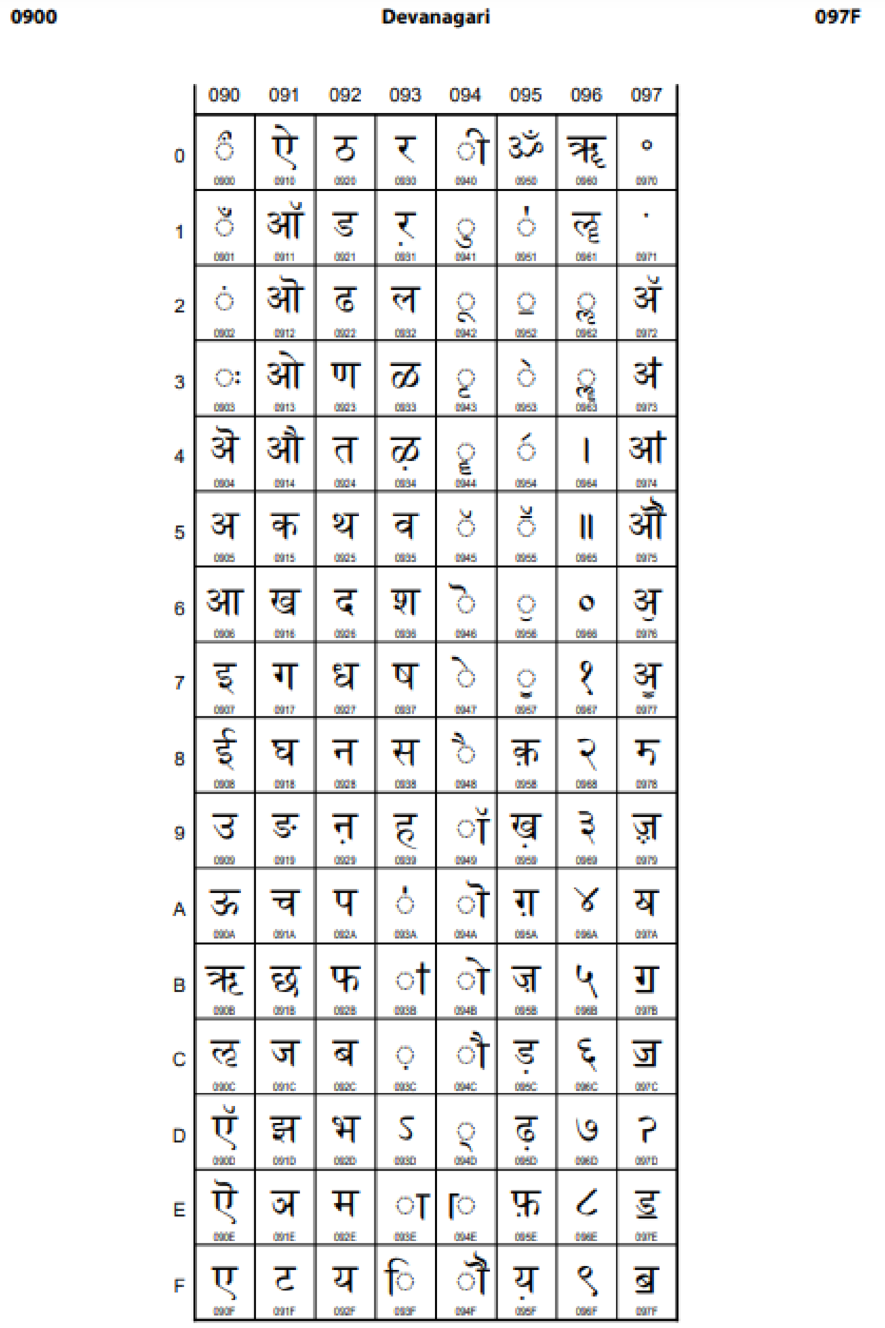
- The Unicode standard includes an extensive database that specifies a large number of *character properties*, including:
  - Name
  - Type (e.g., letter, digit, punctuation mark)
  - Decomposition (Å can be formed with U+00C5 or U+0041 & U+030A)
  - Case and case mappings (for cased letters)
  - Numeric value (for digits and numerals)
  - Combining class (for combining characters)
  - Directionality (Left to Right or Right to Left)
  - Line-breaking behavior
  - Cursive joining behavior (in Arabic, a character used at the beginning of a word looks completely different when it is used as the last character of a word)

https://util.unicode.org/UnicodeJsps/character.jsp

Advantages of UNICODE

- Character based encoding.
- Can be ported on any platform and any OS.
- Can be ported on hand held and mobile devices
- Different scripts have different code page.
- All Indian languages are supported along with all other languages.
- Allows multiple languages in the same data.

UNICODE Devanagari Code Page



0900 Devanagari 097F

	090	091	092	093	094	095	096	097
0	◌ऀ 0900	ऐ 0910	ठ 0920	र 0930	◌ी 0940	ॐ 0950	ॠ 0960	॰ 0970
1	◌ँ 0901	ऑ 0911	ड 0921	ऱ 0931	◌ु 0941	◌॑ 0951	ॡ 0961	ॱ 0971
2	◌ं 0902	ऒ 0912	ढ 0922	ल 0932	◌ू 0942	◌॒ 0952	◌ॢ 0962	ॲ 0972
3	◌ः 0903	ओ 0913	ण 0923	ळ 0933	◌ृ 0943	◌॓ 0953	◌ॣ 0963	ॳ 0973
4	ऄ 0904	औ 0914	त 0924	ऴ 0934	◌ॄ 0944	◌॔ 0954	। 0964	ॴ 0974
5	अ 0905	क 0915	थ 0925	व 0935	◌ॅ 0945	◌ॕ 0955	॥ 0965	ॵ 0975
6	आ 0906	ख 0916	द 0926	श 0936	◌ॆ 0946	◌ॖ 0956	० 0966	ॶ 0976
7	इ 0907	ग 0917	ध 0927	ष 0937	◌े 0947	◌ॗ 0957	१ 0967	ॷ 0977
8	ई 0908	घ 0918	न 0928	स 0938	◌ै 0948	क़ 0958	२ 0968	ॸ 0978
9	उ 0909	ङ 0919	ऩ 0929	ह 0939	◌ॉ 0949	ख़ 0959	३ 0969	ॹ 0979
A	ऊ 090A	च 091A	प 092A	◌ऺ 093A	◌ॊ 094A	ग़ 095A	४ 096A	ॺ 097A
B	ऋ 090B	छ 091B	फ 092B	◌ऻ 093B	◌ो 094B	ज़ 095B	५ 096B	ॻ 097B
C	ऌ 090C	ज 091C	ब 092C	◌़ 093C	◌ौ 094C	ड़ 095C	६ 096C	ॼ 097C
D	ऍ 090D	झ 091D	भ 092D	ऽ 093D	◌् 094D	ढ़ 095D	७ 096D	ॽ 097D
E	ऎ 090E	ञ 091E	म 092E	◌ा 093E	◌ॎ 094E	फ़ 095E	८ 096E	ॾ 097E
F	ए 090F	ट 091F	य 092F	◌ि 093F	◌ॏ 094F	य़ 095F	९ 096F	ॿ 097F

Availability of tools

- UNICODE is not vendor specific
- It is Backward compatible
- Major database, OS, browser players support some form UNICODE encoding

- Data Migration services can be provided for e-governance developers
- Office documents such as .doc/.docx, .xls/xlsx, .txt can be converted to UNICODE
- Database migration tools can also be made available.

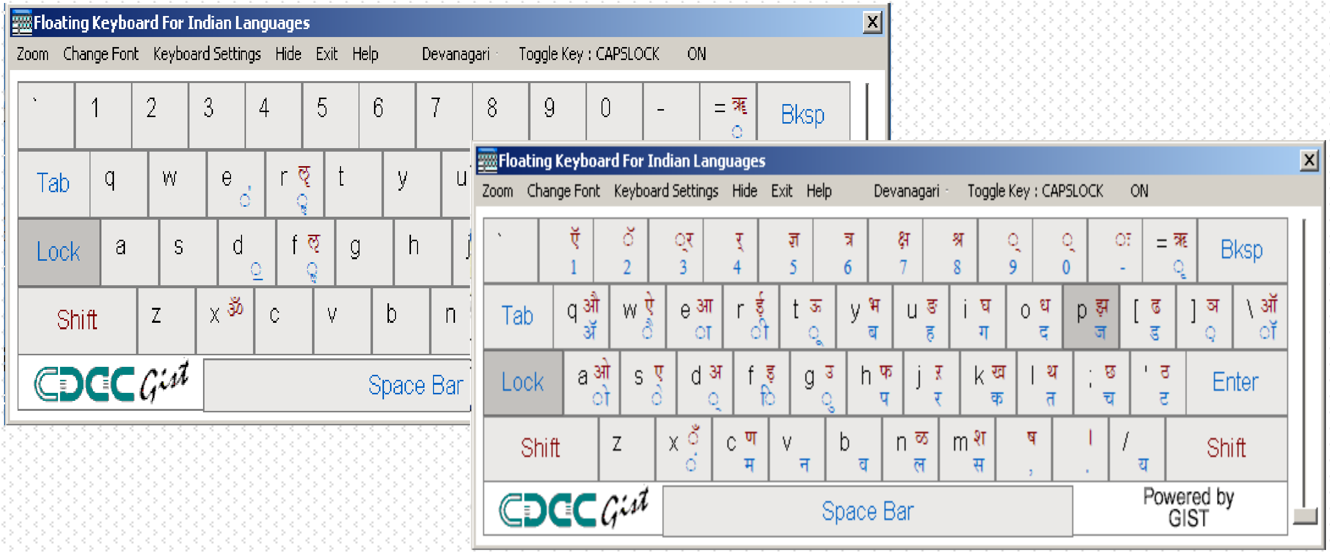
https://localisation.gov.in/

Enhanced INSCRIPT (2.0)

- INSCRIPT is part of BIS ISCII document (IS 13194:1991)
- Enhanced INSCRIPT allows user to type latest UNICODE characters such as Rupee symbol.
- Unlike the phonetic or transliteration mechanism, it does not expect the user to know English to type Indian language and so caters to rural audiences as well.
- Fast typing is possible as consonants are typed by one hand while vowels are typed by left hand

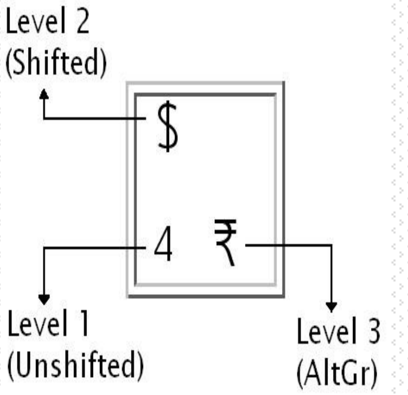
Enhanced INSCRIPT Standardization for Latest Unicode Version

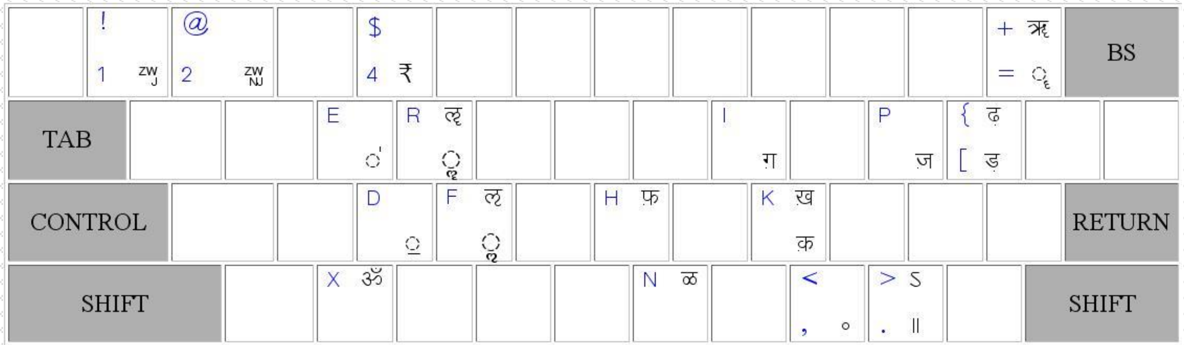
- Study and Research for Keyboards of various languages Normal layer and Extended layer
- Along with teams from – Microsoft, Redhat and IBM

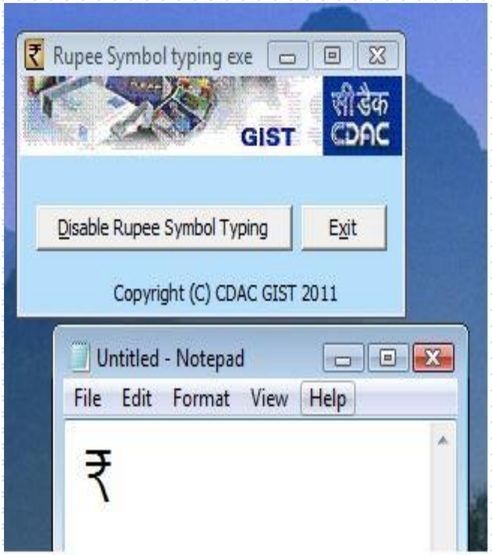


- The Enhanced INSCRIPT keyboard layout provides three layers and this to accommodate all the extra characters and yet make the keyboard as ergonomic and efficient as possible

Standardization of Rupee Symbol Inputting







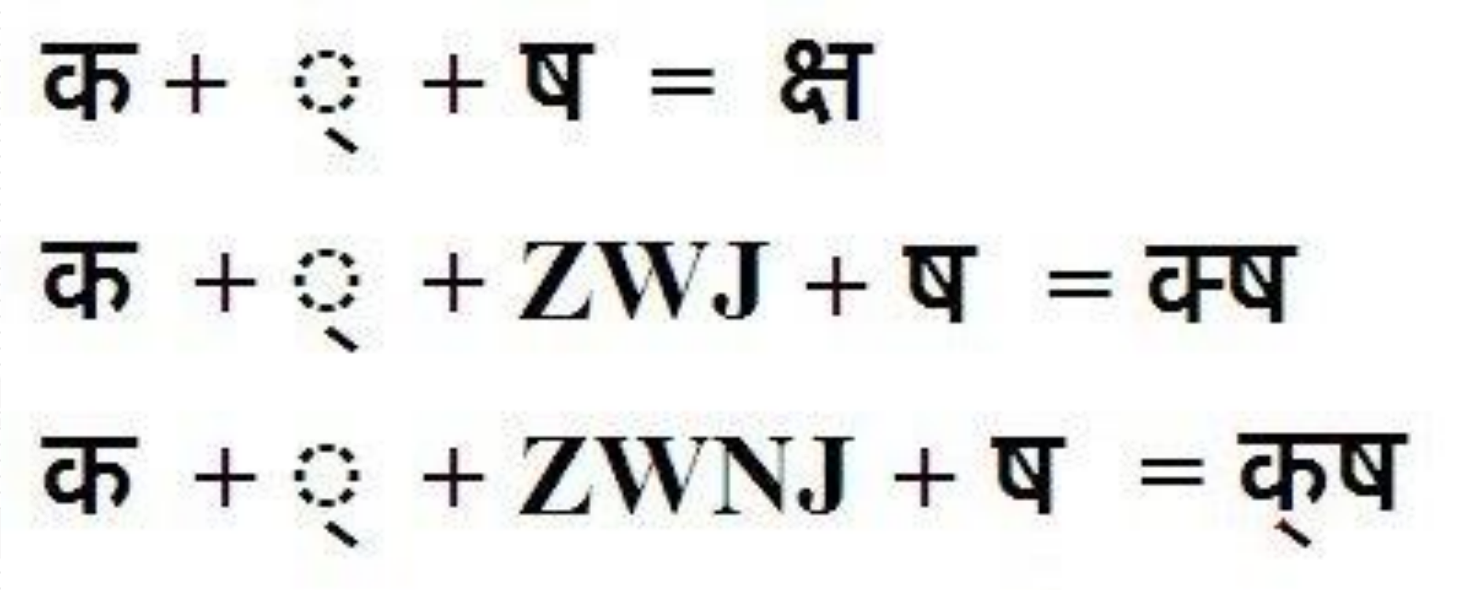
Copyright (C) CDAC GIST 2011

Made available for free download on

http://tdil-dc.in

ZWJ and ZWNJ

- Two special characters in Unicode
- ZWJ - 200D, ZWNJ - 200C



क + ◌् + ष = क्ष

क + ◌् + ZWJ + ष = क्‍ष

क + ◌् + ZWNJ + ष = क्‌ष

Tools Availability

- UNICODE typing Tool is available for free download from http://ildc.in and http://www.cdac.in

- It has all 22 languages and supports enhanced INSCRIPT layout including the ₹ Symbol

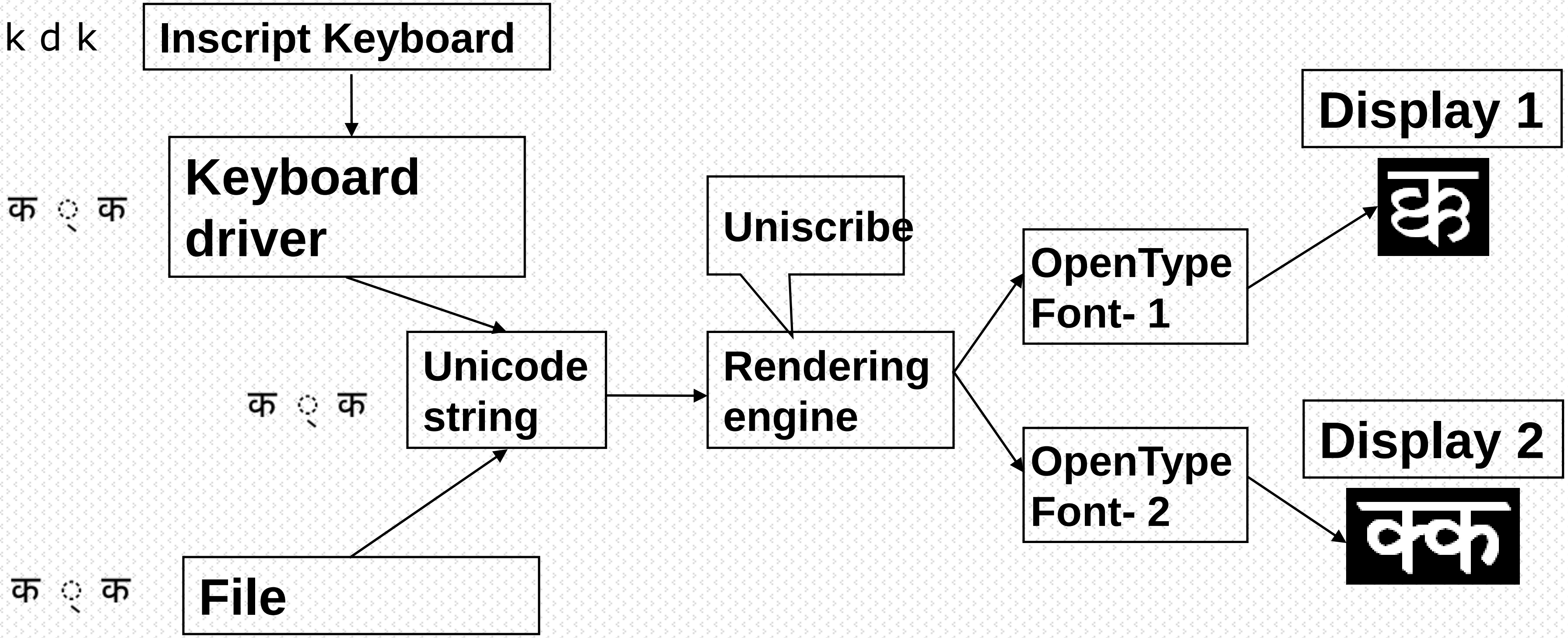
- The keyboard sticker layouts are also available for download from http://ildc.in

- Onscreen Javascript code for websites is made available free of cost to all e-governance developers

Display : Open Font format -Fonts

- Joint effort by Adobe and Microsoft.
- 16-bit Unicode compliant, more glyphs possible.
- Glyph substitution & positioning logic built into the font
- Storage-to-display conversion is done by the rendering engine
- Data is not stored in glyph codes rather in Unicode
- No issue of data portability
- No need to have a font glyph standard

Unicode and OpenType Fonts



k d k

Inscript Keyboard

क ् क

Keyboard driver

क ् क

Unicode string

Uniscribe

Rendering engine

OpenType Font- 1

OpenType Font- 2

Display 1

Display 2

क ् क

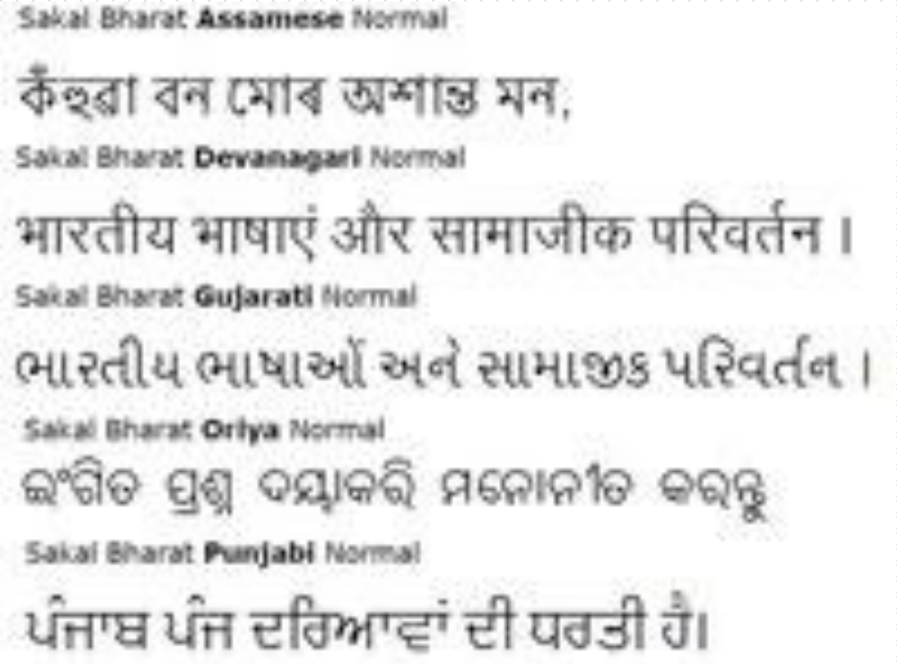
File

Windows Rendering Mechanism

Sakal Bharati font

- A single font which contains all the Indic scripts has been developed by CDAC Pune. This font has got consistent look and feel across various Indian Scripts including English language.
- This font can be downloaded from the url: http://www.tdil-dc.in

کنپیوٹر کی دوڑ میں ہم کہاں ہیں ؟



Sakal Bharat **Assamese** Normal

কঁহুৱা বন মোৰ অশান্ত মন,

Sakal Bharat **Devanagari** Normal

भारतीय भाषाएं और सामाजीक परिवर्तन ।

Sakal Bharat **Gujarati** Normal

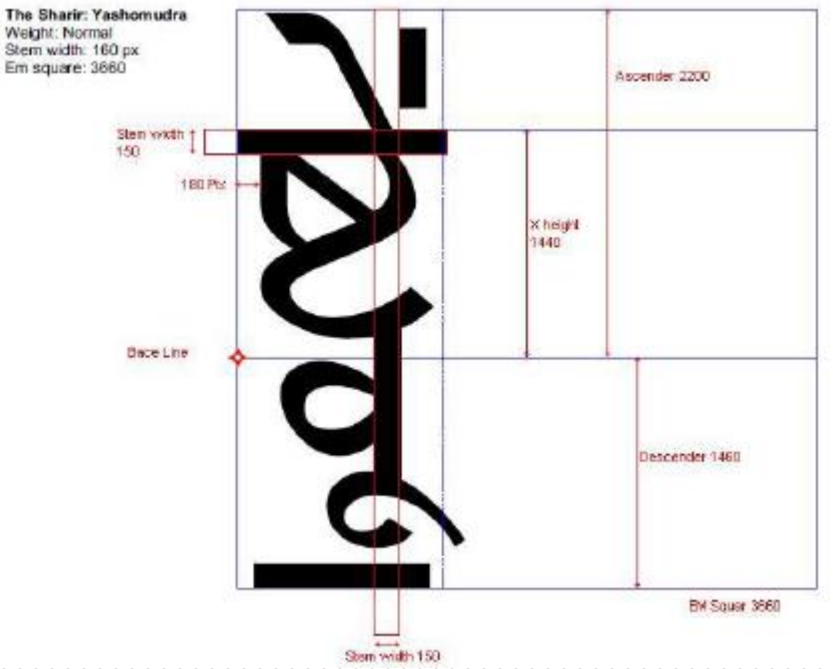
ભારતીય ભાષાઓં અને સામાજીક પરિવર્તન ।

Sakal Bharat **Oriya** Normal

ଇଂଗିତ ପୃଷ୍ଠ ଦୟାକରି ମନୋନୀତ କରନ୍ତୁ

Sakal Bharat **Punjabi** Normal

ਪੰਜਾਬ ਪੰਜ ਦਰਿਆਵਾਂ ਦੀ ਧਰਤੀ ਹੈ।



https://github.com/RajyaMarathiVikasSanstha/Yashomudra

Storing data in Databases

You need to set UTF-8 Collation for the Table to store the various languages text characters

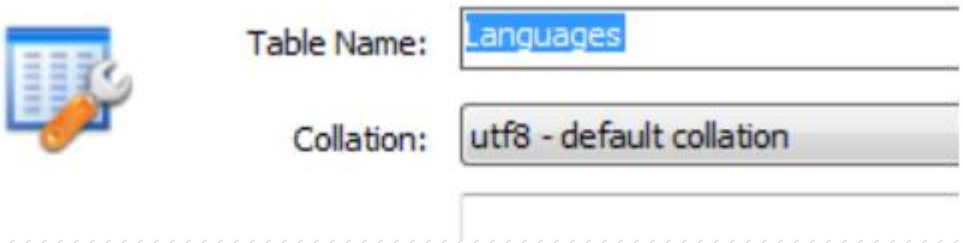


Table Name: Languages

Collation: utf8 - default collation

INSERT INTO Languages VALUES('English', 'World is beautiful.');

INSERT INTO Languages VALUES('Hindi', 'दुनिया सुंदर है।');

INSERT INTO Languages VALUES('Marathi', 'जग सुंदर आहे.');

INSERT INTO Languages VALUES('Gujarati', 'વિશ્વ સુંદર છે.');

Searching in Databases

- Several words have **multiple correct spellings** and **Alternate representation forms**

- E.g.: the word Hindi may be written with a bindi on top of the first syllable or with a half na.

हिंदी हिन्दी

- What should happen in case of using database queries

- So also with the representations of the word vitthal विट्ठल विठ्ठल

- Indian Language numerals are not mapped to English numerals.

Searching in Indian language Websites

- Provide Relevant titles for WebPages
- Main / summarized information of activity to appear in index page
- Images, pdfs, video, audio files should have proper naming convention and names should be appropriate to the contents of the images / pdf.
- Use of Search Engine optimization techniques.

Terminology in Translations

Localisation of strings

- Translation v/s Transliteration
- Technical Term v/s common man’s Term
- Physical-size of localized equivalent strings
- 3 out of 22 languages are right to left oriented
- Location / Layout
  - Positioning of back-next buttons, scroll bar positions for applications supporting right to left scripts.
  - coexisting along with English (Bi-directional support)

Localization of strings

- Context and Domain specific meanings
  - Example
    - the word ‘Bank’ (Financial Entity, River bank, to trust on someone/thing, etc.)
    - the word ‘Fire’ (may very in meaning depending on context) – If it is as a verb (such as fire an event) then it may suggest some action to be undertaken, If noun then the meaning changes completely

  - Multi-Domain expertise as well as context may be required apart from linguistic know-how

Localization of strings

- Technical terminology
  - Differentiating between similar meaning such as cancel, abort, terminate
  - Translation v/s Transliteration (IPR and registered copyrights and trademarks)

  - What should be Localized string for :
    - Windows
    - Mouse
    - FireFox
    - Internet Explorer
    - Double click
    - Dock Windows

Common Locale Data Repository (CLDR)

- The locale is an explicit model and definition of a native-language environment.
- The notion of a locale is explicitly defined and included in the POSIX standard which can be accessed through http://opengroup.org.
- A program's locale defines its code sets, date and time formatting conventions, monetary conventions, decimal formatting conventions, and collation (sort) order.

http://cldr.unicode.org/

Common Locale Data Repository (CLDR)

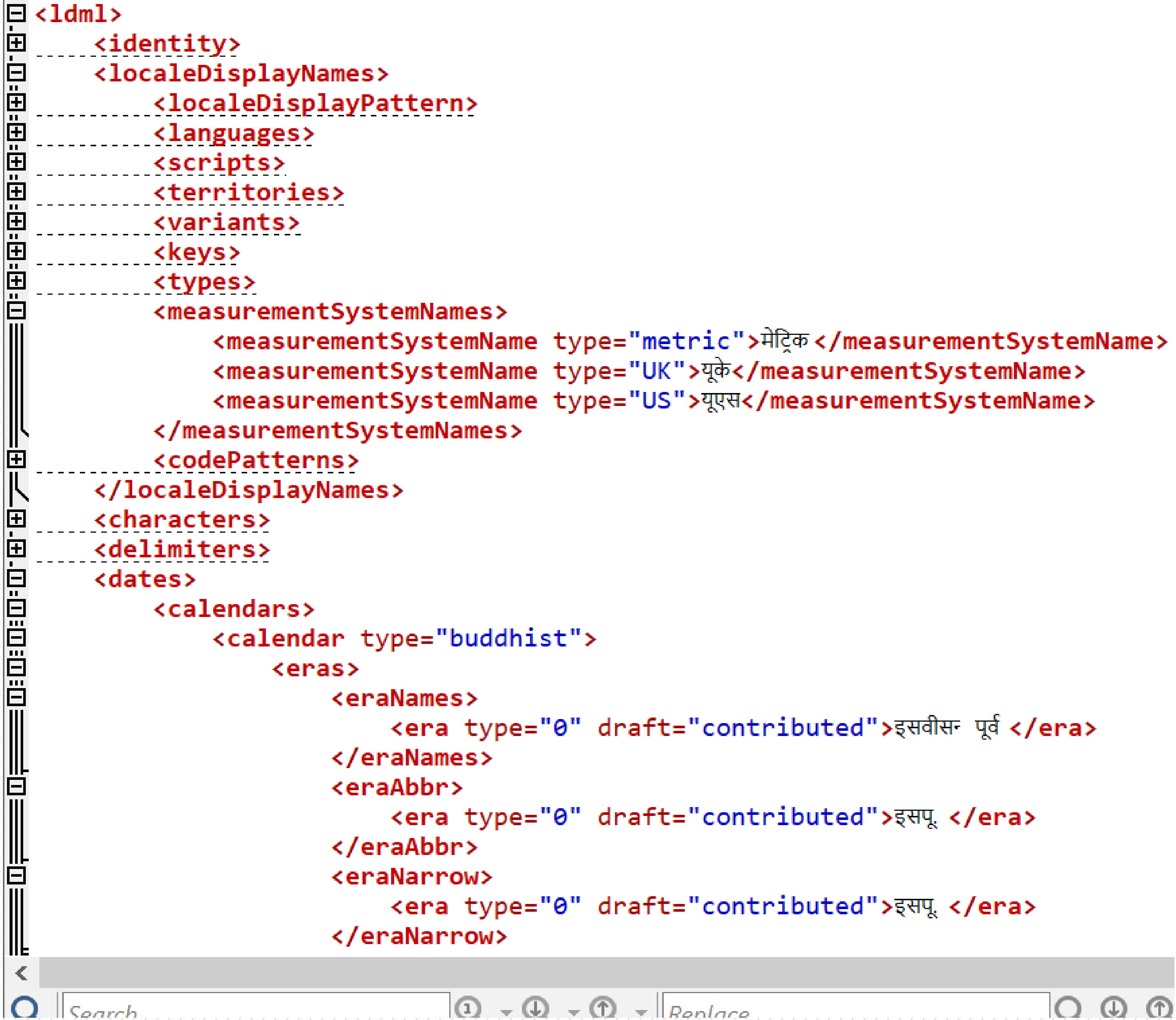
- The CLDR provides key building blocks for software to support the world's languages.
- **Locale-specific patterns for formatting and parsing:** dates, times, timezones, numbers and currency values, measurement units.
- **Translations of names:** languages, scripts, countries and regions, currencies, eras, months, weekdays, day periods, time zones, cities, and time units, emoji characters and sequences (and search keywords)
- **Language & script information:** characters used; plural cases; gender of lists; capitalization; rules for sorting & searching; writing direction; transliteration rules; rules for spelling out numbers; rules for segmenting text into graphemes, words, and sentences; keyboard layouts
- **Country information:** language usage, currency information, calendar preference, week conventions
- **Validity:** Definitions, aliases, and validity information for Unicode locales, languages, scripts, regions, and extensions,

Common Locale Data Repository (CLDR)

Who uses using CLDR?

- Apple (macOS, iOS, watchOS, tvOS, and several applications; Apple Mobile Device Support and iTunes for Windows; ...)
- Google (Web Search, Chrome, Android, Adwords, Google+, Google Maps, Blogger, Google Analytics, ...)
- IBM (DB2, Lotus, Websphere, Tivoli, Rational, AIX, i/OS, z/OS,...)
- Microsoft (Windows, Office, Visual Studio, ...)

*and many others, including:*ABAS Software, Adobe, Amazon (Kindle), Amdocs, Apache, Appian, Argonne National Laboratory, Avaya, Babel (Pocoo library), BAE Systems Geospatial eXploitation Products, BEA, BluePhoenix Solutions, BMC Software, Boost, BroadJump, Business Objects, caris, CERN, CLDR Engine, Debian Linux, Dell, Eclipse, eBay, elixir-cldr, EMC Corporation, ESRI, Firebird RDBMS, FreeBSD, Gentoo Linux, GroundWork Open Source, GTK+, Harman/Becker Automotive Systems GmbH, HP, Hyperion, Inktomi, Innodata Isogen, Informatica, Intel, Interlogics, IONA, IXOS, Jikes, jQuery, Library of Congress, Mathworks, Mozilla, Netezza, OpenOffice, Oracle (Solaris, Java), Lawson Software, Leica Geosystems GIS & Mapping LLC, Mandrake Linux, OCLC, Perl, Progress Software, Python, Qt, QNX, Rogue Wave, SAP, Shutterstock, SIL, SPSS, Software AG, SuSE, Symantec, Teradata (NCR), ToolAware, Trend Micro, Twitter, Virage, webMethods, Wikimedia Foundation (Wikipedia), Wine, WMS Gaming, XyEnterprise, Yahoo!, Yelp



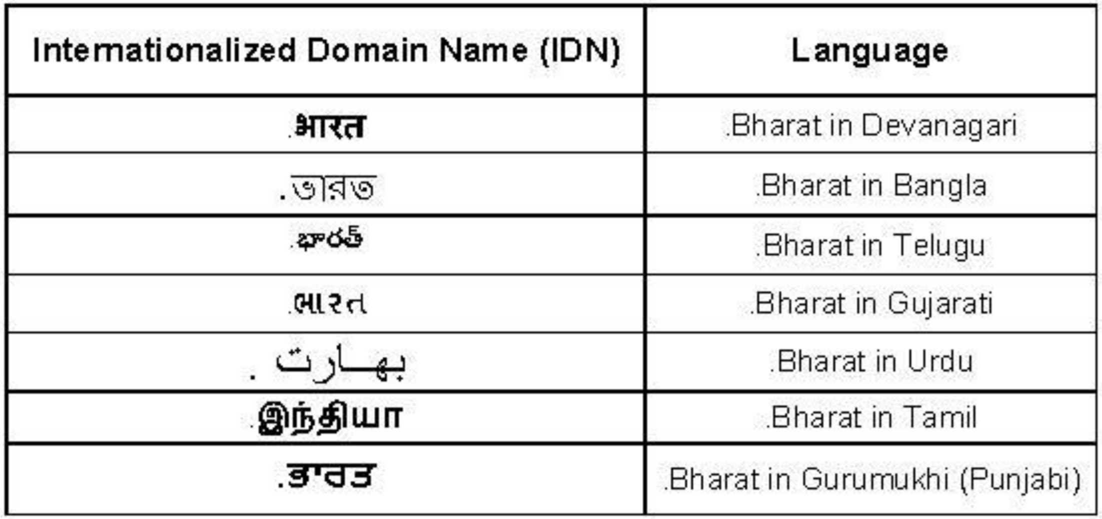
```
<ldml>
    <identity>
    <localeDisplayNames>
        <localeDisplayPattern>
        <languages>
        <scripts>
        <territories>
        <variants>
        <keys>
        <types>
        <measurementSystemNames>
            <measurementSystemName type="metric">मेट्रिक</measurementSystemName>
            <measurementSystemName type="UK">यूके</measurementSystemName>
            <measurementSystemName type="US">यूएस</measurementSystemName>
        </measurementSystemNames>
        <codePatterns>
    </localeDisplayNames>
    <characters>
    <delimiters>
    <dates>
        <calendars>
            <calendar type="buddhist">
                <eras>
                    <eraNames>
                        <era type="0" draft="contributed">इसवीसन् पूर्व </era>
                    </eraNames>
                    <eraAbbr>
                        <era type="0" draft="contributed">इसपू. </era>
                    </eraAbbr>
                    <eraNarrow>
                        <era type="0" draft="contributed">इसपू. </era>
                    </eraNarrow>
```

Internationalized Domain Names (IDNs)

- Internationalized Domain Names (IDNs) enable people around the world to use domain names in local languages and scripts.

- IDNs are formed using characters from different scripts, such as Arabic, Chinese, Cyrillic or Devanagari.

- National Internet Exchange of India (NIXI) is a not for profit organization who is responsible for .IN Registry, managing and operation of .IN country code domain and .भारत IDN domain for India



Internationalized Domain Name (IDN)	Language
.भारत	.Bharat in Devanagari
.ভারত	.Bharat in Bangla
.భారత్	.Bharat in Telugu
.ભારત	.Bharat in Gujarati
.بھارت	.Bharat in Urdu
.இந்தியா	.Bharat in Tamil
.ਭਾਰਤ	.Bharat in Gurumukhi (Punjabi)

Localization Guidelines Minimal Set

Localization Guidelines Minimal Set

- Default Homepage in Local Language
- All Subsequent Web-Pages in Local Language. The subsequent linked pages should be in Local Language. Many of the pages, pdf files linked are in English.
- All Menu titles of the web pages in Local Language.
- All Web-pages should be developed using UTF-8 encoding.
- Lang attributes lang="mr" should be specified (replace “mr” with local language).
- Meta tags should be defined in Local Language. <meta name="keywords" content="HTML,CSS,XML,JavaScript">
- Image ALT/Captions, titles and text should be in Local Language.
- Font for the website should be provided through latest HTML5 Font SRC provisions.
- All Downloads (PDF, DOC, Excel, etc.) should be in Local Language with Unicode compliance.
- All page titles in Local Language.

Localization Guidelines Minimal Set

- Numbers should be in Local Language on all pages, Documents, PDFs, Images, etc.
- Contact us information given in Local Language.
- All user defined alert/error/pop-up messages are in Local Language.
- Feedback form should be available in Local Language.
- Typing facility in Local Language should be given for Interactive website.
- Typing should support Enhanced / INSCRIPT layout.
- Onscreen Floating keyboard should be made available.
- Provision for increasing font size is available.
- In-site search support is available for Local Language.
- Website works on Hand held devices.
- Site map of website is in Local Language.

धन्यवाद

Nurturing living languages

जीवित भाषाओं का पोषण